Modeling/Computation

31

Global Organization of Metabolic Fluxes in the Bacterium, *Escherichia coli*

E. Almaas[1], B. Kovács[1,2], T. Vicsek[2], Z. N. Oltvai[3] and **A.-L. Barabási**[1] (alb@nd.edu)

[1]Department of Physics, University of Notre Dame, Notre Dame, IN; [2]Biological Physics Department and Research Group of HAS, Eötvös University, Budapest, Hungary; and [3]Department of Pathology, Northwestern University, Chicago, IL

Cellular metabolism, the integrated interconversion of thousands of metabolic substrates through enzyme-catalyzed biochemical reactions, is the most investigated complex intercellular web of molecular interactions. While the topological organization of individual reactions into metabolic networks is increasingly well understood, the principles governing their global functional utilization under different growth conditions pose many open questions. We have implemented a flux balance analysis (FBA) of the *E. coli* MG1655 metabolism, finding that the network utilization is highly uneven: while most metabolic reactions have small fluxes, the metabolism's activity is dominated by several reactions with very high fluxes[1]. *E. coli* responds to changes in growth conditions by reorganizing the rates of selected fluxes predominantly within this high flux backbone. The identified behavior likely represents a universal feature of metabolic activity in all cells, with potential implications to metabolic engineering.

To identify the interplay between the underlying topology[2,3] of the *E. coli* K12 MG1655 metabolic network and its functional organization, we focused on the global features of potentially achievable flux states in this model organism with a fully sequenced and annotated genome. In accordance with FBA[4-7], we first identified the solution space (i.e., all possible flux states under a given condition) using constraints imposed by the conservation of mass and the stoichiometry of the reaction system for the reconstructed *E. coli* metabolic network. Assuming that cellular metabolism to be in a steady state and optimized for the maximal growth rate, FBA allows us to calculate the flux for each reaction using linear optimization, providing a measure of each reaction's relative activity. A striking feature of the obtained flux distribution[1] is its overall inhomogeneity: reactions with fluxes spanning several orders of magnitude coexist under the same conditions. To characterize the coexistence of such widely different flux values, we plot the flux distribution for active (non-zero flux) reactions of *E. coli* grown in a glutamate- or succinate-rich substrate. The distribution is best fitted with a power law with a small flux constant, indicating that the probability that a reaction has flux v follows $P(v) \sim (v + v_0)^{-\alpha}$, where the constant is $v_0 = 0.0003$ and the flux exponent has the value $\alpha = 1.5$. The observed power-law is consistent with published experimental data as well[1,8].

We further examined whether these observed flux distributions are independent of the exocellular conditions by mimicking the influence of various growth conditions by randomly choosing 10%, 50% or 80% of the 96 potential substrates that *E. coli*

can consume in this *in silico* model. Optimizing the growth rate, we find that the power law distribution of metabolic fluxes is in fact independent of the external conditions. Moreover, the implementation of a "hit-and-run" method, which samples the solution space in 50,000 non-optimal states, confirms that the power law flux distribution also is independent of the assumption of optimality[1].

The observation and theoretical prediction of a power-law load distribution in simple models, as well as the presence of a power law in both the optimal and non-optimal flux states, suggests that the metabolic flux organization is a direct consequence of the network's scale-free topology. As all organisms examined to date are characterized by a scale-free metabolic network topology, the observed scaling in the flux distribution is likely not limited to *E. coli*, but characterizes all organisms from eukaryotes to archaea. As FBA is available for an increasing number of prokaryotic- and eukaryotic organisms, this prediction could be verified both experimentally and theoretically in the near future. Hence, the observed uneven local and global flux distribution appears to be rooted in the subtle, yet generic, interplay of the network's directed topology and flux balance, channeling the numerous small fluxes into high flux pathways. The dependence of the scaling exponents characterizing the flux distributions on the nature of the optimization process, as well as the experimentally observed exponent, may serve as a benchmark for future structural and evolutionary models aiming to explain the origin, the organization and the modular structure of cellular metabolism.

References

1. E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai and A.-L. Barabási, *Nature*, in press.
2. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai and A.-L. Barabási, *Nature* **407**, 651-4 (2000).
3. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai and A.-L. Barabási, *Science* **297**, 1551-5 (2002).
4. J.S. Edwards and B.O. Palsson, *Proc Natl Acad Sci U S A* **97**, 5528-33 (2000).
5. J.S. Edwards, R.U. Ibarra, and B.O. Palsson, *Nat Biotechnol* **19**, 125-30 (2001).
6. R.U. Ibarra, J.S. Edwards and B.O. Palsson, Nature 420, 186-9 (2002).
7. D. Segre, D. Vitkup and G.M. Church, *Proc Natl Acad Sci U S A* **99**, 15112-7 (2002).
8. M. Emmerling et al., *J Bacteriol* **184**, 152-64 (2002).

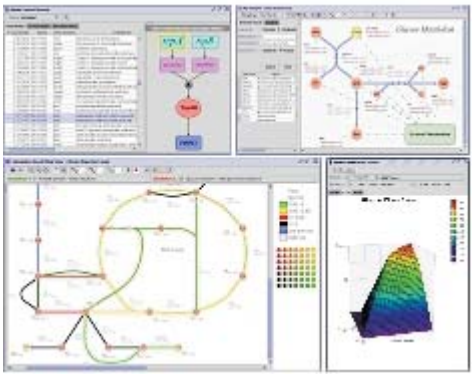
32

SimPheny™: Establishing a Computational Infrastructure for Systems Biology

Christophe H. Schilling (cschilling@genomatica.com), Sean Kane, Martin Roth, Jin Ruan, Kurt Stadsklev, Rajendra Thakar, Evelyn Travnik, and Sharon Wiback

Genomatica, Inc., San Diego, CA

The Genomics:GTL (GTL) program has clearly stated a number of overall goals that will only be achieved if we develop "a computational infrastructure for systems biology that enables the development of computational models for complex biological systems that can predict the behavior of these complex systems and their responses to the environment." At Genomatica we have developed the SimPheny™ (for Simulating Phenotypes) platform as the computational infrastructure to support a model-driven systems biology research paradigm. SimPheny™ enables the efficient development of genome-scale metabolic models of microbial organisms and their simulation using a constraint-based modeling approach.



We are currently utilizing this platform for a number of DOE-related projects that are discussed in accompanying posters and abstracts including:

1. Analysis and Design of Genome-scale Metabolic Networks (co P.I. Christophe Schilling, Costas Maranas, Penn State University)

2. *In Silico* Modeling to Improve Uranium Bioremediation and Energy Harvesting by *Geobacter* species (P.I. Derek Lovley, University of Massachusets, Amherst)

3. Development and Application of a Genome-scale Metabolic Model for *Pseudomonas fluorescens* (P.I. Sung Park, Genomatica, Inc.)

Recently we have launched another SBIR research program to address the problem of how to effectively deploy and deliver a system such as SimPheny to the academic research community to effectively promote collaborative research around systems biology. Ultimately we seek to establish an academic/institutional access program for the distribution, support, and training of our systems biology software platform to enable a broader usage of model-driven research for enhanced biological discovery. To accomplish this we are systematically addressing key issues related to deployment

strategies, collaborative requirements, experimental data integration needs, as well as modeling and simulation requirements. This research will be accomplished by working with a number of existing collaborators and groups involved with the DOE Genomics:GTL program that represent different types of user groups. Collectively, success with this program will facilitate the research activities of laboratories involved in various microbial genome programs and provide a much-needed solution to their data integration needs through the introduction of model centric databases. The results of these research activities will also provide valuable information on the collaborative needs and system requirements for the development of complementary software platforms that may be under parallel development by other groups. Perhaps most importantly, success with this program will further one of the core aims of the Genomics:GTL program, namely the development and distribution of a computational infrastructure for systems biology research. Establishing such an institutional/academic technology access program will also enable Genomatica to distribute and license non-energy related microbial models to the general scientific community for applications related to both medical and industrial biotechnology.

33

Analysis and Design of Genome-Scale Metabolic Networks

Costas D. Maranas[1] (costas@psu.edu), Anthony P. Burgard[1], Evgeni V. Nikolaev[1], Priti Pharkya[1], and Christophe H. Schilling[2]

[1]Department of Chemical Engineering, Pennsylvania State University, University Park, PA and [2]Genomatica, Inc., San Diego, CA

An overarching attribute of metabolic networks is their inherent robustness and ability to cope with ever changing environmental conditions. Despite this flexibility, network stoichiometry and connectivity do establish limits/barriers to the coordination and accessibility of reactions. The recent abundance of complete genome sequences has enabled the generation of genome-scale metabolic reconstructions for various microorganisms(1,2). Here we introduce the Flux Coupling Finder (FCF) framework for elucidating the topological and flux connectivity features of genome-scale metabolic networks(3). The framework is demonstrated on genome-scale metabolic reconstructions of Helicobacter pylori, Escherichia coli, and Saccharomyces cerevisiae(4-6). The analysis allows one to determine if any two metabolic fluxes, v_1 and v_2, are (i) directionally coupled, if a non-zero flux for v_1 implies a non-zero flux for v_2 but not necessarily the reverse; (ii) partially coupled, if a non-zero flux for v_1 implies a non-zero, though variable, flux for v_2 and vice-versa; or (iii) fully coupled, if a non-zero flux for v_1 implies not only a non-zero but also a fixed flux for v_2 and vice-versa. Flux coupling analysis also enables the global identification of blocked reactions, which are all reactions incapable of carrying flux under a certain condition, equivalent knockouts, defined as the set of all possible reactions whose deletion forces the flux through a particular reaction to zero, and sets of affected reactions denoting all reactions whose fluxes are forced to zero if a particular reaction is deleted. The FCF approach thus provides a novel and versatile tool for aiding metabolic reconstructions and guiding genetic manipulations.

The advent of genome-scale metabolic models has also laid the foundation for the development of computational procedures for suggesting genetic manipulations that lead to overproduction. Here the computational OptKnock framework is introduced for suggesting gene deletions strategies leading to the overproduction of

chemicals or biochemicals in *E. coli*(7,8). This is accomplished by ensuring that a drain towards growth resources (i.e., carbon, redox potential, and energy) must be accompanied, due to stoichiometry, by the production of a desired product. Computational results for gene deletions for succinate, lactate, and 1,3-propanediol (PDO) production are in good agreement with mutant strains published in the literature. While some of the suggested deletion strategies are straightforward and involve eliminating competing reaction pathways, many others suggest complex and non-intuitive mechanisms of compensating for the removed functionalities. The OptKnock procedure, by coupling biomass formation with chemical production, hints at a growth selection/adaptation system for indirectly evolving overproducing mutants.

References

1. J. L. Reed and B. O. Palsson (2003). "Thirteen years of building constraint-based in silico models of Escherichia coli." *J Bacteriol* **185**(9): 2692-9.

2. M. W. Covert, C. H. Schilling, I. Famili, J. S. Edwards, I. I. Goryanin, E. Selkov and B. O. Palsson (2001). "Metabolic modeling of microbial strains in silico." *Trends Biochem Sci* **26**: 179-186.

3. A. P. Burgard, E. V. Nikolaev, C. H. Schilling and C. D. Maranas (2004). "Flux coupling analysis of genome-scale metabolic network reconstructions." *Genome Res*, in press.

4. C. H. Schilling, M. W. Covert, I. Famili, G. M. Church, J. S. Edwards and B. O. Palsson (2002). "Genome-scale metabolic model of Helicobacter pylori 26695." *J Bacteriol* **184**(16): 4582-93.

5. J. S. Edwards and B. O. Palsson (2000). "The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities." *Proc Natl Acad Sci U S A* **97**(10): 5528-33.

6. J. Forster, I. Famili, P. Fu, B. O. Palsson and J. Nielsen (2003). "Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network." *Genome Res* **13**(2): 244-53.

7. P. Pharkya, A. P. Burgard and C. D. Maranas (2003). "Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock." *Biotechnol Bioeng* **84**: 887-899.

8. A. P. Burgard, P. Pharkya and C. D. Maranas (2003). "Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization." *Biotechnol Bioeng* **84**(6): 647-57.

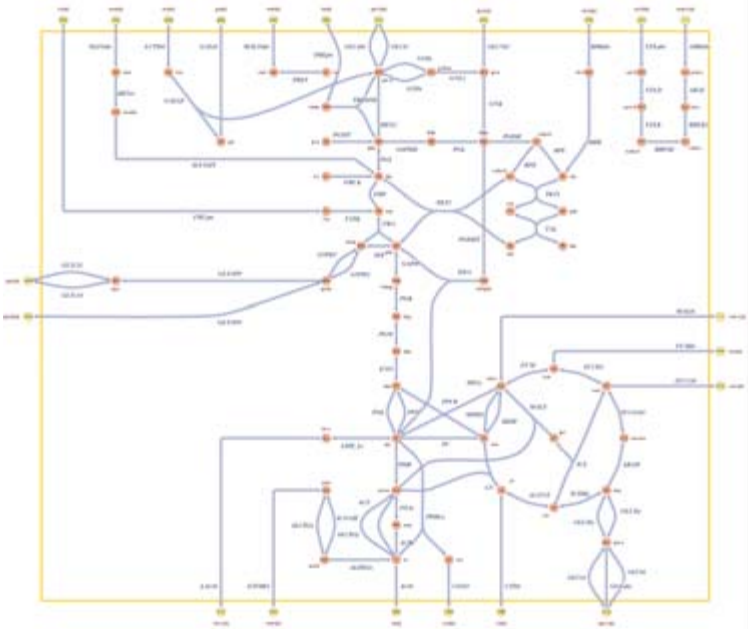
34

Development and Industrial Bioprocessing Application of a Genome-Scale Metabolic Model for *Pseudomonas fluorescens*

Sung M. Park[1] (spark@genomatica.com), Christophe H. Schilling[1], Tom Ramseier[2], and Charles Squires[2]

[1]Genomatica, Inc., San Diego, CA and [2]Dow Chemical Company

Innovative approaches are needed to utilize the information generated from genome research in an integrated fashion to analyze, interpret, and predict the function of biological systems and assist the advancement of biotechnology on the whole. This work addresses these needs with novel engineering approaches for studying the systemic capabilities of metabolism in completely sequenced bacterial genomes. The overall goal of this entire SBIR research program is to demonstrate the utility of constraints-based modeling to drive metabolic engineering and the design of bioprocesses utilizing *Pseudomonads*. There are two main commercial applications for altering the metabolism of these organisms, which include their use as a catalyst for the fermentative production of various biologics (e.g. industrial enzymes, and chemicals) as well as their use in bioremediation treatment strategies. In collaboration with the Dow Chemical Company, we have been addressing the commercial needs to fully implement the model for metabolic engineering objectives. The plan represents an integrated effort including computational and experimental components along with the necessary software development required to support these efforts.



This poster focuses exclusively on the development and implementation of a genome-scale metabolic model of *Pseudomonas fluorescens* that we have accomplished through our SBIR Phase I effort. This model is now the subject of further enhancement and utilization in a Phase II program currently underway. A comprehensive *in*

silico metabolic model of *P. fluorescens* will be shown within SimPheny. Metabolic model reconstruction of *P. fluorescens* was primarily based on the genome sequence with additional information obtained from the literature. The model includes all of the major metabolic pathways in this organism and contains 928 balanced chemical reactions accounting for 1244 genes (~20% of the total genes in *P. fluorescens*). Simulations with the reconstructed model show a range of metabolites that can be taken up and be degraded.

Modeling and simulation strategies for systems biology can now be used to guide experimental design, facilitate biological discovery, and produce the next generation of enhancements to metabolism-dependent bioprocesses. This model of *P. fluorescens* provides another platform to demonstrate power of genome-enabled science and the potential for using modeling technology to drive biological research.

35

Parallel Scaling in Amber Molecular Dynamics Simulations

Michael Crowley, Scott Brozell, and **David A. Case** (case@scripps.edu)

Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA

Large-scale biomolecular simulations form an increasingly important part of research in structural genomics, proteomics, and drug design. Popular modeling tools such as Amber and CHARMM are limited by both state-of-the-art hardware capabilities and by software algorithm limitations. Current macro-molecular systems of interest range in size to several hundred thousand atoms, and current simulations generally simulate one to tens of nanoseconds. With a 2 fs timestep, and each force evaluation involving millions of interactions to be calculated, a simulation requires many gigaflops to finish in a reasonable period of time. A parallel implementation of the calculation can provide the required performance by using the power of many processors simultaneously. However, communication speed between nodes has not progressed as rapidly as CPU processing power in recent years. Here, we address some weakness of the current parallel molecular dynamics implementation in Amber (and in a comparable program such as CHARMM). The work is aimed at making affordable a new generation of increasingly sophisticated biomolecular simulations.

Atom-Based Decomposition in Amber

Of the many ways to distribute the work of a force calculation in parallel [1,2], the method of replicated data (or "atom decomposition") has traditionally been used in Amber and CHARMM. This sort of parallel implementation is based on dividing each portion of the force calculation evenly among the processors, while keeping a full set of coordinates on all processors. This is very flexible, and relatively straightforward to program. Each processor is assigned an equal number of bonds, angles, dihedrals, and nonbond interactions. In this way, the work is balanced in each part of the force calculation, and the computation time scales well as the number of processors increases. However, in each part of the force calculation a node computes forces for different subsets of atoms. For this reason, each processor requires a complete set of up-to-date coordinates and is assumed to have components of forces for all atoms. At each step, the forces computed for all atoms on each node must be summed and distributed, and updated coordinates must be collected from each

node and sent complete to all nodes. There are hence two all-to-all communications at each step. Even with binary tree algorithms for distributed sums and redistribution, the communication time becomes a significant fraction of the total time by 32 processors, even on the most sophisticated parallel machines. This limitation eliminates the possibility of efficient parallel runs at large numbers of processors, and puts a restriction on the size and length of simulations that a researcher can attempt even when large parallel computational resources are available. Still, for systems up to about 32 processors, these codes are more efficient for typical solvated simulations than are popular alternatives such as CHARMM or NAMD.

Spatial Decomposition in Amber

The second-generation parallel Amber, now under development, implements a "spatial decomposition" method [1,2] in which the molecular system is divided into regions of space where approximately equal amount of force computation is required. The method works when contributions to the force on an atom come primarily from interactions with other atoms that are relatively close and are neglected for atoms that are beyond a fixed cutoff. (This condition is valid in modern MD simulations except for long-range electrostatics, which use Ewald-based methods discussed below.) In this approach, a processor is assigned the atoms located in a slice of space and it is responsible for the coordinates, forces, velocities, and energetic contributions of those atoms. In order to compute the forces for its *owned* atoms, the processor must be able to compute the contributions from interactions with atoms that are within the cutoff, including any that are assigned to other processors. A processor keeps a copy of all such *needed* atom coordinates and forces as well as its *owned* atom coordinates and forces. At each step, a processor determines the force contributions due to all interactions in its *owned* and *needed* atoms. It sends all force contributions on needed atoms to the processors that own those atoms and receives any force contributions for its *owned* atoms that were calculated by other processors. When the force communications are complete, the coordinate integration is performed on the owned atoms. Each message in all the above communications is at most the size of the *owned* atom partition and will often be considerably smaller.

This conversion of the Amber codes is complex, since there are complications inherent in spatial decomposition that do not arise in the replicated data method; these are mainly in the treatment of bonded interactions, constraints, long-range electrostatics, and bookkeeping. The first two complications arise when molecules (chemical bonds) or distance constraints span the spatial boundaries. Most bonds, angles, dihedrals, restraints, and constraints can be assigned according to ownership of atoms. When the atoms involved are owned by distinct processors, an algorithm must be implemented to insure that the interactions are considered but only once, and that the coordinates necessary are current and correct. Bond-length constraints (using the so-called "SHAKE" approach) are more complicated, since they redefine the positions of atoms after the computed forces have been applied to owned atoms. In this case, the updated positions of all atoms involved in a constraint must be known in order to adjust positions of owned atoms regardless of whether they are owned or not. Besides these complications lie the bookkeeping needed to keep track of which forces and coordinates are being sent and received. Finally, we must optimize scaling of the Ewald method of treating long-range electrostatics in periodic system, and in particular, the PME implementation of Ewald sums. We are exploring several methods of reducing the communications costs of PME in highly parallel systems.

Current Code Status

Two separate implementations of spatial decomposition are in the final testing stage, and will be released in March, 2004, as a part of Amber 8. The first, called *psander*, builds upon "classic" Amber code, and promises to minimize communication times, particularly on systems such as clusters of relatively low-end machines. The second, called *pmemd*, is in some ways a more ambitious effort: it involves an extensive re-write of major portions of the code in a controlled F90 environment, carefully moving subsets of features in as they can be validated. *Pmemd* is best suited for large numbers of processors that have good communications; for example, it scales well to 128 or 256 processors on systems such as IBM SPx architectures or the Lemiuex supercomputer at the Pittsburgh Supercomputer Center. Timings, capabilities, and prospects for future development will be presented in our poster.

References

1. S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys*. **117**, 1-19 (1995).

2. L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K Schulten. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.* **151,** 283-312 (1999).

36

Bioinformatics Methods for Tandem Mass Spectrometry

Andrey Gorin[1] (agor@ornl.gov), Tema Fridman[1], Robert M. Day[1], Jane Razumovskaya[2], Andrei Borziak[1], and Edward Uberbacher[2]

Computational Biology Institute, [1]Computer Science and Mathematics Division, [2]Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

The importance of computational tools and algorithms for mass spectrometry (MS) is hard to overstate. Efficient and reliable software for database search identification of proteins is the foundation of today high-throughput protein identification based on mass spectrometry data. At the same time unrelenting pace of scientific research creates a constant demand for a higher precision, efficiency and novel capabilities of the bioinformatics and computational algorithms used to analyze MS data.

We are developing a set of novel computational algorithms for reliable and comprehensive protein identification through detailed analysis of the tandem MS (MS/MS) data. The principal idea could be described as "micro analysis" of the spectra: analysis of the patterns typical for individual peak categories and relationships between individual peaks. Our approach is to design a probabilistic classification algorithm, aimed to establish identities of individual peaks in terms of belonging to the specific peak categories. A large set of positively identified peptide spectra have been used to determine "neighborhood" patterns for b- and y-ions as well as conditional probabilities of other important observable attributes for peak categories of interest. The established patterns have been applied to determine peak identities in other tandem MS spectra. The identification is done in a probabilistic manner, so the results have the form of probabilistic statements (e.g., "peak number 123 is a b-ion with a 0.8 probability"). The robustness of the method (named Probability Profile Method (PPM)) was investigated on a large set (>5000) of positively verified peptide spectra. Preliminary results indicate that a large majority of the useful peaks in MS/MS spectra could be identified with a surprising level of confidence, providing founda-

tion for a range of new algorithmic capabilities. An incomplete of the possible directions includes: (1) spectra can be edited sorting out desirable peak categories; (2) overall characteristics of MS/MS spectra, such as parent ion charge or total number of the present useful b- and y-ions, can be very rapidly estimated with a high precision; (3) labeled peaks of the same category, e.g. b-ion peaks, can be connected into *de novo* peptide tags providing a way for protein identification without strong reliance on the sequence database.

Two specific applications of the PPM algorithm will be discussed in details.

First, we report a novel tool for differentiation of parent ion charge states. For each spectrum we predict number of the fragments ions with charges 1 and 2. Spectra of the parent charge 3 have those fragments roughly equally 1:1, as one would intuitively expect with a splitting of +3 charge. At the same time the 2++ parent ion has 7-fold more single charged fragments compare to double charged. We demonstrate that the total number for each type is very accurately computed without any prior assumptions about what parent charge state is. As a result the PPM-based tool is fast and has 99% accuracy while being applicable to a wide range of peptide spectra. Importantly the parent charge differentiation capability not only to 2-times acceleration of the identification process, but also may eliminate some hard-to-catch misidentification originating from the wrong parent mass estimate.

Second, we demonstrate a PPM-based approach to construction of de novo peptide tags. Efficient separation of "noble" b- and y-ions dramatically simplifies algorithmic challenges, as we can easily generate and score all connectable paths for de novo tags, without "prefixing" or elaborated optimization techniques. Our method is capable of finding peptide tags 3 to 10 amino acid long for ~80% of MS/MS spectra from our testing set. When only a single top scoring tag was considered for the answer, more than a half of the constructed tags were correct ones. While additional tests and development are needed before de novo sequencing could be declared a solved problem, the approach holds a strong promise to substantially improve performance of several bioinformatics tools depending on de novo methodology for MS/MS data analysis.

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).

37

The Use of Microarray Technology and Data Mining Techniques to Predict Gene Regulation and Function in *Geobacter sulfurreducens*

Barbara Methé[1]* (bmethe@tigr.org), Kelly Nevin[2], Jennifer Webster[1], and **Derek Lovley**[2]

*Presenting author

[1]The Institute for Genomic Research, Rockville, MD and [2]University of Massachusetts, Amherst, MA

Geobacter sulfurreducens is a member of a family of prokaryotes which possess the ability to oxidize organic compounds to carbon dioxide with Fe(III) or other metals serving as the electron acceptor. As such they play critical roles in the global cycles of these metals and carbon. Additional interest in *Geobacter* spp. stems from their potential as agents of bioremediation via an ability to precipitate soluble metals including uranium and a capacity to create electricity that can be captured via energy-harvesting electrodes. Completion of the entire *G. sulfurreducens* genome sequence has provided the critical foundation for the creation of a whole genome microarray to examine global gene expression patterns.

Two categories of experiments for querying whole genome PCR-based arrays are currently being pursued: 1) test wild type *G. sulfurreducens* gene expression profiles under relevant physiological conditions and 2) test mutants in which a selected gene has been knocked out versus their wild type counterpart. cDNA probes for querying the array were derived from mRNA of cells grown in most instances in chemostats. For each condition tested competitive hybridizations were performed with Cy-Dye labeled cDNA probes. Post-hybridization the intensity of the two dyes for each gene (treatment vs. control) was measured by scanning the slide with a laser. The TM4 package (www.tigr.org/software/tm4) which consists of a suite of open-source programs developed at The Institute for Genomic Research was employed for microarray data analysis. Background corrected intensity values were normalized prior to examination for significant changes indicating up or down regulation of a gene.

An initial test of the array queried cells grown under nitrogen fixing conditions to a wild type control. The expression of at least ninety genes was determined to have changed significantly using the Significance Analysis of Microarrays (SAM) algorithm which incorporates gene-specific t-tests and false discovery rates to determine significant changes in expression. These included genes known to be vital to the nitrogen fixation process for example the up regulation of the genes responsible for nitrogenase, the key enzyme in nitrogen fixation. In addition, others genes not immediately predicted to be expressed under nitrogen limiting conditions were determined including the up regulation of several hypothetical genes and a sensor histidine kinase and response regulator.

Analysis of microarray data is also proving beneficial in corroborating predictions from annotation and analysis of whole genome data. Genome analysis predicted that *G. sulfurreducens* is not a strict anaerobe and is capable of using oxygen. A microarray analysis of *G. sulfurreducens* growth on 5% oxygen versus standard anaerobic conditions revealed up regulation of genes predicted to be involved in an oxidative metab-

olism including the high oxygen affinity oxidase, cytochrome d ubiquinol oxidase as well as ruberythrin which is capable of scavenging oxygen radicals.

In addition to examining gene expression profiles from individual physiological conditions and mutants, another facet of this project is to apply further data mining methods to the resulting array data across multiple experiments with the goal of elucidating functional roles and regulatory patterns in this organism. A variety of statistical and clustering techniques are currently being utilized. For instance, the software application, Expression Analysis Systematic Explorer (EASE), is being evaluated as a tool for determining biological themes present in significantly up and down regulated genes across multiple experiments. In one example, conditions tested included: growth with a chelated iron source as electron acceptor, attached growth on energy-harvesting electrodes and growth in the presence of 5% oxygen all versus the same standard control (growth in suspension under anaerobic conditions). EASE was used to elucidate statistically significant over representation of biological categories based on Gene Ontology (GO) assignments from the up regulated and down regulated gene lists. A modified Fisher exact probability test (EASE score) and correction via bootstrapping was used to determine significance ($p < 0.05$). The GO biological process of electron transport and the GO molecular function of transport activity were among the over represented assignments in the up regulated gene lists. These results confirm the importance of electron transport across diverse physiological conditions while suggesting the importance of other metabolic processes such as transporter activity in this organism.

An Analysis of Variance (ANOVA) of this same data set was used to look for genes with statistically significant changes in their gene expression profiles between the three experiments versus those that did not. The two resulting categories of gene expression data were then examined using clustering techniques. This analysis revealed a cluster of genes that based on their physical location and coordinate regulation describe a putative operon with genes for a c-type cytochrome, a C_4-dicarboxylate transporter and several genes of unknown function. The operon is down regulated under growth with a chelated iron source, up regulated when grown as a biofilm on graphite electrodes and expression does not vary greatly when grown in the presence of 5% oxygen suggesting that it may in part be repressed by the presence of iron and important in biofilm growth. Conversely, another cluster includes a group of transporters putatively related to heavy metal efflux and a transcriptional regulator from the mercuric reductase family of regulators all of which are up regulated in a similar manner across each of the three experiments.

Additional data mining techniques being evaluated include the use of template matching algorithms in which the expression of one or more genes can be used as a template and genes with expression profiles similar to the template can be identified between the template and genes in the data set. These matches may indicate genes related by function and/or regulation. In the three experiment data set the mean expression values of a family of heat shock protein genes were used as a pattern to look for genes regulated in a similar fashion. This technique was successful in matching other heat shock and chaperone genes with similar expression profiles as well as two periplasmic c-type cytochromes and several genes whose functions may be related to membrane structure suggesting that in addition to coordinate regulation these genes may collectively participate in the assembly or degradation of the functional c-type cytochromes. These findings reveal the power of microarray technology coupled with a variety of data mining techniques to suggest new functional roles and regulatory patterns in this organism.

38

In Silico Elucidation of Transcription Regulons and Prediction of Transcription Factor Binding Sites in *Geobacter* Species Using Comparative Genomics and Microarray Clustering

Julia Krushkal[1]* (jkrushka@utmem.edu), Bin Yan[1], Daniel Bond[2], Maddalena Coppi[2], Kelly Nevin[2], Cinthia Nunez[2], Regina O'Neil[2], Barbara Methé[3], and **Derek Lovley**[2]

*Presenting author

[1]University of Tennessee Health Science Center, Memphis, TN; [2]University of Massachusetts, Amherst, MA; and [3]The Institute for Genomic Research, Rockville, MD

Geobacter species are important for bioremediation of a variety of environments contaminated with metal, metalloid, and organic waste compounds, and their ability to harvest electricity also suggests them as a possible source of alternative fuel for the future. Therefore, we are developing a model of the physiological responses of *Geobacteraeae* to different environmental conditions in order to more rationally optimize bioremediation and energy-harvesting strategies. As part of this effort, we are using a computational approach that utilizes genome sequence information and whole genome expression data to elucidate the transcription regulatory circuitry of the *Geobacteraceae*.

We are employing a combination of complementary computational strategies to most efficiently predict operons, regulons, and transcription factor binding sites in *Geobacteraceae*. These computational methods can be divided into three categories, i.e. those that (1) are based on individual genome sequences of *Geobacter* species; (2) compare genome sequences from several closely related species of *Geobacteraceae,* and (3) use microarray clustering of *G. sulfurreducens* genes.

Single-genome analyses of the completed genome sequence of *G. sulfurreducens* and draft contig assemblies of *G. metallireducens* and *Desulfuromonas acetoxidans* provided whole genome predictions of operon organization. For example, we identified 1418 putative operons and transcription units in the *G. sulfurreducens* genome. As a first step in interpreting genome information obtained from the *Geobacteraceae* sequencing projects, we currently perform routine operon structure predictions with each new round of contig assembly of each genome (Table 1).

Table 1. An example of an NADH-quinone oxidoreductase operon predicted in the *D. acetoxidans* genome

Gene No. in the operon	Location (bp) in contig 548	Putative gene function
1	18892 –19248	NADH:ubiquinone oxidoreductase subunit 3 (chain A)
2	19239 –19748	NADH:ubiquinone oxidoreductase 20 kD subunit and related Fe-S oxidoreductases
3	19790 -20272	NADH:ubiquinone oxidoreductase 27 kD subunit
4	20306 –21496	NADH:ubiquinone oxidoreductase 49 kD subunit 7
5	21525 -22022	NADH:ubiquinone oxidoreductase 24 kD subunit
6	22067 -23848	NADH:ubiquinone oxidoreductase, NADH-binding (51 kD)
7	23882 -26362	Uncharacterized anaerobic dehydrogenase
8	26388 -27350	TPR-repeat-containing protein
9	27412 -28449	NADH:ubiquinone oxidoreductase subunit 1 (chain H)
10	28477 -28872	NADH:ubiquinone oxidoreductase 23 kD subunit (chain I)
11	28893 -29396	NADH:ubiquinone oxidoreductase subunit 6 (chain J)
12	29422 -29724	NADH:ubiquinone oxidoreductase subunit 11 or 4L (chain K)
13	29768 -31750	NADH:ubiquinone oxidoreductase subunit 5 (chain L)
14	31797 -33353	NADH:ubiquinone oxidoreductase subunit 4 (chain M)
15	33402 -34862	NADH:ubiquinone oxidoreductase subunit 2 (chain N)

We further analysed genome sequences of *G. sulfurreducens* and *G. metallireducens* by providing predictions of potential transcription regulatory elements using similarity searches to over 60 position-specific matrices of established transcription factor binding sites from other prokaryotes and by the neural network approach. As a result of these searches, we have developed databases of predicted transcription regulatory elements in each genome, along with software tools for querying these database in user-specified locations (Table 2).

Table 2. An example summary of the most significant positive predictions of transcription regulatory elements in the *G. sulfurreducens* genome based on whole genome similarity searches:

Transcription factor binding site	Number of highly significant hits in the *G. sulfurreducens* genome		Transcription factor binding site	Number of highly significant hits in the *G. sulfurreducens* genome
ArgR	2		metJ	4
CpxR	3		metR	12
Crp	58		narL	2
CytR	12		ompR	39
dnaA	57		rpoD15	547
FarR	62		rpoD16	446
Fis	123		rpoD17	1590
Fnr	2		rpoD18	244
FruR	2		rpoD19	422
Fur	4		rpoS17	227
GlpR	33		rpoS18	3
Hns	1396		soxS	41
Ihf	240		torR	1
Lrp	1117		tyrR	9
MalT	514			

Using **comparative genome analyses**, we verified our operon predictions in *Geobacteraceae* genomes by identifying clusters of genes conserved across three species from that family*: G. sulfurreducens, G. metallireducens*, and *D. acetoxidans.* Many of these genes participate in essential cell functions, e.g., DNA replication and protein biosynthesis. Interestingly, one of these conserved gene clusters involved genes related to flagellar proteins that are likely related to cell motility. We have developed and are maintaining a database of putative orthologs in these genomes. To date, the across-genome comparisons of gene clusters have allowed us to identify both conserved operons and operons unique to individual species of *Geobacter.* Using information from multiple genomes, we also searched for potential transcription regulatory elements by using the phylogenetic footprinting approach that identified conserved regions of noncoding DNA in different species of *Geobacteraceae*.

Further effective validation of operon and regulon predictions came from whole genome **microarray analyses** that involved hierarchical clustering of *G. sulfurreducens* genes based on their change in expression levels in *G. sulfurreducens* mutants as compared to the wild type. This approach allowed us to identify two groups of operons positively controlled by the fur regulator and one group negatively affected by this protein. Similarly, several groups of operons positively and negatively controlled by the RpoS regulator were identified. Microarray expression data are being used to predict transcription factor binding sites by identifying DNA elements conserved upstream of clusters of putative operons with similar expression patterns. A number of intriguing observations were made from these analyses. For example, both groups of operons positively controlled by fur contain fur binding sites and several other conserved motifs in their upstream noncoding regions, while the operons negatively controlled by fur do not seem to contain a fur box; instead they contain a motif highly similar to the lrp binding site. A group of operons under a strong positive control of RpoS contain in their upstream regions a -35/-10 box

along with other conserved motifs, suggesting that cooperative binding of transcription regulators affects the transcription of these operons.

The information obtained from the three computational strategies outlined here is being regularly compared and reconciled. This approach allows us to most efficiently identify putative transcription regulatory interactions among genes of *Geobacteraceae* and to identify groups of co-regulated genes and their putative DNA regulatory elements, as our first step toward the understanding of the complex network of regulatory interactions of *Geobacter.*

39

In Silico Modeling to Improve Uranium Bioremediation and Energy Harvesting by *Geobacter* species

R. Mahadevan[1], B. O. Palsson[1], C. H. Schilling[1], D. R. Bond[2], J. E. Butler[2], M. V. Coppi[2], A. Esteve-Nunez[2], and **D. R. Lovley**[2] (dlovley@microbio.umass.edu)

[1]Genomatica, Inc., San Diego, CA and [2]University of Massachusetts, Amherst, MA

Geobacter species are important organisms in the bioremediation of uranium-contaminated subsurface environments and for harvesting electricity from waste organic matter, but their metabolism is poorly understood. In order to better predict the response of *Geobacter* species under different environmental conditions and to further optimize bioremediation and energy harvesting applications, a genome-scale metabolic model of *Geobacter sulfurreducens* was developed using the constraints-based modeling approach. The metabolic model currently contains 523 reactions and 540 metabolites accounting for 583 genes (29% of the annotated genome).

The model has provided a number of new insights into the physiology of *G. sulfurreducens*. For example, one of the unsolved mysteries of anaerobic respiration in this organism is why growth yields with fumarate as the electron acceptor are 3-fold higher than during growth on Fe(III), despite the fact that Fe(III) has a higher mid-point potential than fumarate. The model has revealed the previously unsuspected importance of proton balance in the energetics of this organism and that more cytosolic protons are likely to be formed when Fe(III) is the electron acceptor. The energetic cost associated with the pumping of these extra protons to maintain the transmembrane gradient in cells growing on Fe(III) leads to a much lower energy yield than when fumarate serves as the electron acceptor. Thus, the model-based analysis has provided a likely explanation for the difference in biomass yields for growth with different electron acceptors. The current version of the model not only predicts these differences in growth yields, but also accurately predicts growth rates with Fe(III) or fumarate.

One of the most useful applications of the model has been to predict the phenotype of mutations generated for functional genomics studies. For example, a knock-out mutant that no longer produced succinate dehydrogenase could not grow with acetate as the electron donor and Fe(III) as the electron acceptor. However, the model made the non-intuitive prediction that this mutant would be able to grow better than the wild type on acetate and Fe(III) if fumarate was also provided. This prediction was experimentally verified. These types of predictions have the potential to

greatly accelerate functional analysis of genes and the understanding of the central metabolism of *G. sulfurreducens*.

Furthermore, *in silico* deletion studies can make laboratory mutational studies more efficient. For example, *in silico* deletion analysis revealed that deletion of genes associated with central metabolism led, in most cases, to either a lethal or a silent phenotype. Thus, this result has suggested that investing labor and time in making mutations in many of these genes may not be a fruitful line of investigation.

The model has also been helpful in elucidating the function of genes of unknown function and interpreting the results of microarray analysis of gene expression. For example, a knock-out mutation in a gene previously annotated as an Fe(III) reductase did not have the expected specific effect on Fe(III) reduction, but rather appeared to have a more general effect on metabolism. When the results of a microarray study comparing gene expression of this mutant with the wild type, were analyzed with the model, the gene expression changes were found to map closely with predicted changes in metabolism in an *in silico* mutation in NADPH dehydrogenase. This prediction of function from the model and other evidence has indicated that this gene encodes for a NADPH dehydrogenase, rather than an Fe(III) reductase, as previously proposed.

The model has also provided further insight into why *Geobacter* species predominate over other Fe(III)-reducing microorganisms, such as *Shewanella* and *Geothrix* species, in a diversity of subsurface environments. Previous studies have suggested that *Shewanella* and *Geothrix* species release extracellular electron-shuttling compounds in order to reduce Fe(III) whereas *Geobacter* species do not. Analysis of the energetic cost of producing an electron shuttle under conditions typically found in subsurface environments demonstrated that a microorganism, like *Geobacter* species, that did not need to produce a shuttle would grow 20-50% faster than an organism that produced an electron shuttle. This is a substantial difference that would provide *Geobacter* species with a significant competitive advantage.

The model is also helpful for predicting environmental manipulations that might stimulate the growth of *Geobacter* species, possibly accelerating bioremediation. For example, simulation studies demonstrated that there are a few amino acids, which if provided to *Geobacter*, would enhance its growth. These results are now being evaluated experimentally to determine if they represent a potential strategy to increase biomass yields.

These results demonstrate that this iterative modeling and experimentation approach to microbial physiology can rapidly accelerate discovery of gene function and provide important physiological and ecological insights. It is clear from these results that the *in silico* model of *G. sulfurreducens* has the potential to help guide the development of better strategies for the bioremediation of uranium and other contaminants as well as aid in the design of improved *Geobacter*-based fuel cells.

40

Continued Studies on Improved Methods of Visualizing Large Sequence Data Sets

George M. Garrity[1] (garrity@msu.edu), Timothy G. Lilburn[2] (Tlilburn@atcc.org), and Yuan Zheng[1] (zhangyu6@msu.edu)

[1]Michigan State University, East Lansing, MI and [2]American Type Culture Collection, Manassas, VA

We have continued our investigations into the use of graphical and analytical techniques drawn from the field of Exploratory Data Analysis to gain insight into the taxonomic relationships among prokaryotes, as currently defined by the 16S SSU rRNA. In our initial studies, we found that the dimensionality of extremely large sequence datasets ($n > 10^5$) could be reduced by methods such as Principal Components Analysis (PCA), allowing accurate projection of the data into 2D maps of the taxonomic space. In addition to revealing the overall topology of the taxonomic space, each strain could be accurately located within that space. While such plots were useful in delineating major groups, they proved less useful in resolving precise placement of individuals within some families and genera. Subsequently, we reported on the use of heatmaps, a form of colorized matrix, to directly visualize sequence similarity data. These plots readily revealed that many of the discrepancies detected by PCA could be attribute to a variety of taxonomic and sequence annotation errors. We also reported on the development of a self-organizing self-correcting classifier (SOSCC) that allowed for automatic detection and resolution of such errors; the SOSCC automatically optimizes the classification of the sequences, correctly positioning misplaced sequences based on their relationship to a set of validated nearest neighbors. Here we report on: (1) recent refinements in the SOSCC algorithm and porting of the algorithm to StatServer for deployment as an interactive web application, (2) the production a web-based taxonomic atlas of prokaryotes based on PCA plots and interactive heatmaps, (3) how this methodology has been applied to resolve a number of outstanding taxonomic anomalies, and (4) the development of a set of vetted sequences that can be used to improve the accuracy of identification of prokaryotes by 16S rRNA sequence analysis. Two spin-off projects applying this technology will also discussed: (1) integration of the application with the RDP to pipeline sequence data and (2) the use of interactive heatmaps as a graphical interface to access networked data resources.

41

PQuad for the Visualization of Mass Spectrometry Peptide Data

Bobbie-Jo Webb-Robertson[1] (Bobbie-Jo.Webb-Robertson@pnl.gov), Susan L. Havre[2], Deborah A. Payne[3], and Mudita Singhal[2]

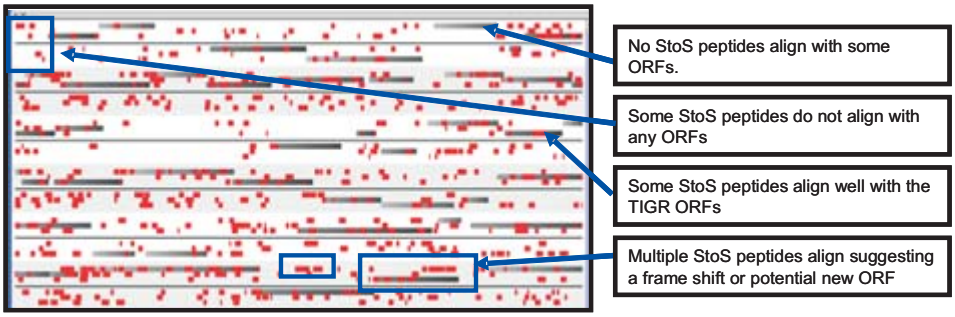
[1]Statistics & Quantitative Sciences, [2]Scientific Computing Environments, and [3]Information Analytics, Pacific Northwest National Laboratory, Richland, WA

Mass spectrometry has come forward as one of the most promising technologies for high-throughput proteomics. Thus, the development of supporting techniques and software tools for analyzing MS/MS data has been a high priority in recent years. The basic process clips a protein into peptides via proteolyic digestion prior to subjecting to the MS process. The resulting spectra are then subjected to a peptide identification tool. When the number of identified peptides becomes large, navigating and analyzing the dataset becomes a time-consuming and challenging task. The problem is exacerbated when the scientist attempts to integrate the results with other biological data or compare two or more sets of identified peptides, for instance, peptide sets collected under different experimental conditions. Information visualization and statistical methods are emerging as the technologies of choice in integrating and analyzing very large complex data. No tools comprising both these technologies currently exist to help scientists analyze MS/MS results. We have developed a proof-of-concept interactive visualization prototype, PQuad (Peptide Permutation and Protein Prediction), for the analysis of identified peptides in the context of an annotated DNA sequence. PQuad shows great promise in assisting in the evaluation and validation of both gene annotation and peptide identification software.

Currently, PQuad visualizes the identified peptides, the associated open reading frames (ORFs), and the actual nucleotide sequence of the DNA. The prototype provides three basic levels of resolution or views: summary, intermediate, and detail. The summary view is a miniaturized visualization of the ORFs and/or peptides overlaid on the complete DNA sequence; it provides a bird's eye view. The detail view shows a readable stream of both DNA nucleotide strands and the six possible amino acid reading frames; the location of the ORFs and peptides are indicated by highlighting the appropriate sections of the DNA strands and reading frames. The intermediate view is more flexible in both the resolution and the amount of information. Sequence letters are not provided in the intermediate view, but the user can view a section of the DNA that shows individual ORFs in the context of neighboring ORFs. Each view continuously reports the sequence indices, ORF names, and peptides under the cursor for easy exploration. Selecting an ORF in either the summary or intermediate view is propagated to all views so that the selected ORF is featured in the detail view; centered, highlighted, and surrounded by neighboring ORFs in the intermediate view; and highlighted in the summary view.

More recently, we have been working with a *Deinococcus radiodurans* spectral dataset that has been run through SEQUEST (a popular commercial peptide identification tool) with both TIGR and Stop-to-Stop (StoS) annotations. The difference in the number of identified peptides and ORFs is striking. The StoS identifications overlaid on the StoS annotation appears noisy, most likely due to a large number of false positives; many ORFs are observed with a single peptide hit. However, there are many places where groups of peptides cluster in apparent confirmation of the underlying ORF.

As a next step, we relaxed our rule that PQuad show the peptides against the ORF annotation that was used by SEQUEST to identify the peptides. Now PQuad can show the peptides (red) identified using the StoS annotation against the TIGR ORF annotation (gray boxes). The results of this combination are startling. It is easy to see where groups of StoS peptides align with TIGR ORFs. A few ORFs do not have associated peptides, but interestingly groups of StoS peptides cluster where there is no TIGR ORF. We believe that these peptides may suggest missed start positions, potential new genes, or gene modifications not included in the TIGR annotation.



42

Computational Framework for Microbial Cell Simulations

Haluk Resat[1] (haluk.resat@pnl.gov), Linyong Mao[1], Heidi Sofia[1], Harold Trease[1], Samuel Kaplan[2], and Christopher Mackenzie[2]

[1]Pacific Northwest National Laboratory, Richland, WA and [2]University of Texas Medical School, Houston, TX

Development of integrated sets of computational tools is needed to achieve the level of sophistication necessary to bridge experimental and computational biology studies. Because of the complexity of the biological data associated with the cellular processes, use of mathematical and computational methods are needed to decipher the information hidden in the experimental results and to design new experiments. As part of this project, we have been developing a wide range of prototype computational biology and bioinformatics analysis tools, and new algorithms and methods. New tools are employed to investigate the flux and regulation of fundamental energy and material pathways in *Rhodobacter sphaeroides*. The prototype components are designed in such a way that, when combined later, they will form the backbone of a comprehensive microbial cell simulation environment.

Our recent efforts to develop a computational framework have concentrated on the following research areas:

Gene regulatory networks: We have developed a new probabilistic algorithm to model the stationary properties of the gene regulatory networks and our new algorithm was implemented in the object oriented stochastic simulation software NWGene. We applied the new algorithm to simulate the expression patterns in a library of synthetically engineered gene regulatory networks. The agreement between the model predictions and the experimental data was very good.

Stochastic kinetic simulations: We have further improved the computational efficiency of the NWKsim program, a kinetic simulation package that uses stochastic Gillespie algorithm and its variants. We used the NWKsim program to investigate the cell receptor signaling networks using kinetic models.

Imaging of bacterial cells and image recontruction: We have obtained electron tomography images of *R. sphaeroides* using the TEM imaging facility at UCSD. Utilizing our image reconstructed software NWGrid, we have computationally reconstructed a 3-D geometry of *R. sphaeroides'* surface features from the obtained series of tilted digital images.

Mesh grid based simulation framework: We are using the VMCS (Virtual Microbial Cell Simulator), which is based on the biological version of NWGrid/NWPhys (http://www.emsl.pnl.gov:2080/nwgrid), to simulate spatial and temporal growth of microbial cell communities. The model includes explicit representations of individual microbial cells, derived from TEM image data or computational geometry. The evolution of the model allows for the generation of communities that can take the form of biofilms or free floating flocs. The VMCS model imports reaction/diffusion models and can include environmental conditions such as flow velocity, shear flows, and structures. As a test application of the software, we are simulating the biofilm growth in a two-organism syntrophic bacterial system.

Genome comparison data mining: Genome comparison data mining detects larger patterns useful in understanding complex biological processes from large quantities of sequence data across many species. Our Similarity Box software provides a sensitive and accurate method for extracting several important types of genome comparison results, including conserved gene neighborhood relationships, which are informative for protein function and binding partners. We have now incorporated a high-throughput version of this approach to gene neighbor analysis in a HERBE database implementation designed to support *R. sphaeroides* genome annotation. A user can enter a single *R. sphaeroides* protein identifier and receive a useful view of all relevant conserved relationships. Using the new implementation, for example, we found that the *R. sphaeroides* RpoH1 protein belongs to a cluster of heat shock factors with a conserved association with pseudouridine synthases, in contrast to the *E. coli* RpoH which is linked to the FtsYEX proteins. This strategy will be made available to the *R. sphaeroides* annotation community.

Determination of regulatory mechanisms for the photosynthesis genes of *R. sphaeroides*: Although most of the regulators of the photosynthesis genes of *R. sphaeroides* have been determined using biochemical methods, detailed understanding of the regulatory mechanisms is still lacking. We are using the recent genome (DOE sponsored JGI) and microarray (UT-Houston) data to investigate the regulators of the photosynthesis genes of *R. sphaeroides*. Using a combination of clustering analysis and DNA motif finding methods, we have investigated the DNA recognition motifs of the known regulators. We were able to confirm the binding motifs of the regulators FNR and PpsR, and we have derived a statistical distribution of the binding motif of another regulator PrrA. We have also developed a new approach to search for motifs in DNA sequences. Our approach, which is computationally intensive, combines techniques from combinatorial and numerical optimization, and was implemented with a parallel genetic algorithm.

43

Optimization Modules for SBW and BioSPICE

Vijay Chickarmane, **Herbert M. Sauro** (hsauro@kgi.edu), and Cameron Wellock
Keck Graduate Institute, Claremont, CA

Parameter estimation and model validation are essential components to model building. As part of the DARPA BioSPICE project, we have developed a series of optimization modules which enable experimentalists to fit time series data to ordinary differential equation (ODE) based models. Given a model and a set of experimental data, the optimization modules compute estimates for the model parameters, the sums of squares of the final fit and standard errors on the certainly of the fitted parameter values.

The modules are written in Matlab and employ a new Systems Biology Workbench (SBW)/Matlab interface which makes integration of Matlab scripts into SBW extremely easy and enables us to leverage existing SBW tools such as model designers and simulators. The modules themselves employ a number of novel approaches to optimization, some of which we believe are suitable for optimizing large systems. The SBW integration permits the modules to be automatically used by the BioSPICE Dashboard interface and thus to appear as building blocks in a Dashboard workflow diagram. The use of SBW also permits developers to write additional modules and have them automatically integrated into the BioSPICE/SBW with very little effort. Note that the operation of the modules does not require Matlab to be installed on the client machine.

In addition to the modules themselves, we also provide an Optimization Controller GUI which enables users to easily employ these modules in their research. The controller permits different optimization methods to be applied either individually or in succession. During the optimization a real-time graphical display is generated that enables one to judge the effectiveness and progress of the optimization. Optimizations may be stopped and started, and different methods applied during the optimization. Users can also graphically compare the fits with the experimental data.

Due to integration into SBW, models can be developed under a variety of tools (eg JDesigner, CytoScape, CellDesiginer) via SBML. Simulation engines such as Jarnac or Dizzy can be exploited by the optimization modules to compute the solutions to the ODEs which leads to significantly improved performance. Due to the integration into SBW/BioSPICE additional control of the optimization procedures is available via Python and Perl scripts. This option provides great flexibility, for example a user can implement Monte Carlo fitting for those systems where the non-linearities in the model do not permit accurate estimates for the standard errors. In the first release the following optimization modules will be made available:

- Levenberg-Marquardt
- Nelder & Mead Simplex
- Simulated Annealing/Simplex Hybrid
- Genetic Algorithm
- Genetic Algorithm/Simplex Hybrid

The software will be released on our web site (www.sys-bio.org) by the time of the 2004 GTL meeting.

44

The Docking Mesh Evaluator

Roummel Marcia[1] (marcia@math.wisc.edu), Susan D. Lindsey[2] (lindsey@sdsc.edu), J. Ben Rosen[3] (jbrosen@ucsd.edu), and **Julie C. Mitchell**[1] (mitchell@math.wisc.edu)

[1]Departments of Mathematics and Biochemistry, University of Wisconsin, Madison, WI; [2]San Diego Supercomputer Center, University of California, San Diego, CA; and [3]Department of Computer Science and Engineering, University of California, San Diego, CA

Introduction

The Docking Mesh Evaluator (DoME) is a software for predicting a bound protein-ligand docking configuration by determining the global minimum of a potential energy function. Our present energy model is based on solvent effects defined implicitly using the Poisson-Boltzmann equation, as well as a pairwise Lennard-Jones term.

Description

Our approach consists of two phases. The first involves scanning the energy landscape for favorable configurations. This phase can be done once as a preprocessing step and need not be done again. The second phase involves the iterative underestimation of successive collections of local minima with convex quadratic functions, using the configurations from the first phase as initial seed points for optimization. The minima of the underestimators are then used as predicted values for the global minima. Both serial and parallel versions of this "coupled" optimization have been successfully implemented. Preliminary results are reported in [2].

Currently, our research is focused on optimizing parameters in the energy function, in order to obtain the best accuracy in predicting known docking configurations. In particular, we consider the benchmarking set of Chen et al. [1] for testing protein-protein docking algorithms. Of the 59 test cases it contains, 22 are enzyme-inhibitor complexes, 19 are antibody-antigen complexes, 11 are various diverse complexes, and 7 are difficult test cases whose solutions have significant conformational changes. These optimized parameters are expected to yield realistic results for biological problems whose solutions are unknown.

Flexibility in the protein-ligand model is being implemented using a hybrid of global optimization and rotamer search. Near the surface interface, subtle side-chain rearrangements are often necessary to model induced fit between the receptor and the ligand. These rearrangements can be modeled using candidate residue conformations, called rotamers. Using this approach, the protein backbone is held fixed while residues are allowed to take on various configurations. Such pseudo-flexibility is a more viable alternative to full backbone and side-chain flexibility, which requires inordinately many free variables, thus making the computational cost prohibitively expensive. Local shape complementarity analysis performed using the Fast Atomic Density Evaluator [3] will provide added efficiency by highlighting regions in which shape mismatches occur.

References

1.

R. Chen, J. Mintseris, J. Janin, and Z. Weng, "A protein-protein docking benchmark," *Prot. Struct. Fun. Gen.*, **52**, pp. 88–91, 2003.

2. R. F. Marcia, J. C. Mitchell, and J. B. Rosen, "Iterative convex quadratic approximation for global optimization in protein docking," *Comput. Optim. Appl.*, Submitted, 2003.

3. J. C. Mitchell, R. Kerr, and L. F. Ten Eyck, "Rapid atomic density measures for molecular shape characterization," *J. Mol. Graph. Model.*, **19**(3), pp. 324–329, 2001.

45

Functional Analysis and Discovery of Microbial Genes Transforming Metallic and Organic Pollutants: Database and Experimental Tools

Lawrence P. Wackett (wackett@biosci.cbs.umn.edu) and **Lynda B. M. Ellis** (lynda@mail.ahc.umn.edu)

Center for Microbial and Plant Genomics, University of Minnesota, St Paul, MN

Microbial metabolism is vast and much remains to be catalogued and characterized. Characterizing this metabolism is a major task of microbial functional genomics. Over time, these data will impart much greater predictive power onto microbial science. The research conducted on this project seeks to better assemble existing metabolic data, discover new microbial metabolism, and predict microbial metabolic pathways for compounds not yet in the databases.

One goal of the project, compilation of information relevant to the metallic and metalloid elements that comprise half of the periodic table, has been completed. The web-based University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) has been expanded to include information on microbiological interactions with 77 chemical elements (1). This is part of a comprehensive study of how microbes interact with all of the chemical elements (2). For each element, a webpage has been created with annotation on the major microbial interactions with that element, links to Medline, and access to further UM-BBD information. The project has added hundreds of new linkages to UM-BBD compound pages (3). For example, the mercury element page has 4 links, arsenic has 9 links, and chlorine has 145 links to UM-BBD compounds, respectively.

Another important goal of the current project is to discover new metabolism and functionally analyze the novel microbial enzymes and genes involved (4). A review of the natural product literature has revealed that on the order of one hundred chemical functional groups are produced by biological systems (5). Yet, only about fifty functional groups have been studied with respect to metabolism. Metabolism must exist for the remaining 50 chemical groups. This lack of information represents a knowledge gap contributing to the problem of incomplete microbial genome sequence annotation. In the current project, we are uncovering metabolism of chemical functional groups that have previously not been studied. To date, we have discovered new metabolism relevant to bismuth compounds, boronic acids (6), azetidine ring compounds, and novel organonitrogen compounds.

A third goal of the project has been to develop a computer software that predicts microbial metabolism using the UM-BBD as a knowledge base (7). The user of the software gets to see one or more plausible biodegradation pathways for the com-

pound they have entered into the system. The metabolism prediction software is based on rules that broadly describe microbial reactions such they can be applied to new compounds. At present, there are over 250 rules in the biotransformation rule database, each specifying the atoms and their positions in a functional group and the biotransformation reaction that they undergo. The software has been validated by comparison against: (i) known biodegradation reactions, (ii) expert predictions, and (iii) microbial growth studies. The system is freely available on the web (8). The system will be expanded with input from our Scientific Advisory Board and the broader scientific community.

References

1. UM-BBD Biochemical Periodic Tables: http://umbbd.ahc.umn.edu/periodic/

2. Wackett, L.P. A.G. Dodge, and L.B.M. Ellis. (2004) Microbial genomics and the periodic table. *Appl. Environ. Microbiol.* (in press).

3. Ellis, L.B.M., B.K. Hou, W. Kang and L.P. Wackett (2003) The University of Minnesota Biocatalysis/Biodegradation Database: Post genomic data mining. *Nucl. Acids Res.* **31**:262-265.

4. Wackett, L.P. (2002) Expanding the map of microbial metabolism. *Environ. Microbiol.* **4**: 12-13.

5. Wackett, L.P. and C. Douglas Hershberger (2001) *Biocatalysis and Biodegradation: Microbial Transformation of Organic Compounds*. American Society for Microbiology Press.

6. Negrete-Raymond, A.C., B. Weder, and L.P. Wackett (2003) Catabolism of arylboronic acids by *Arthrobacter nicotinovorans* strain PBA. *Appl. Environ. Microbiol.* **69**:4263-4267.

7. Hou, B.K., L.P. Wackett, and L.B.M. Ellis (2003) Microbial pathway prediction: A functional group approach. *J. Chem. Inf. Comp. Sci.* **43:**1051-1057.

8. UM-BBD Pathway Prediction System: http://umbbd.ahc.umn.edu/predict/

46

Comparative Genomics Approaches to Elucidate Transcription Regulatory Networks

Lee Ann McCue* (mccue@wadsworth.org), Thomas M. Smith, William Thompson, C. Steven Carmack, and **Charles E. Lawrence**

*Presenting author

The Wadsworth Center, New York State Department of Health, Albany, NY

The ultimate goal of this research is to delineate the core transcription regulatory network of a prokaryote. Toward that end, we are developing comparative genomics approaches that are designed to identify complete sets of transcription factor (TF) binding sites and infer regulons without evidence of co-expression. This approach has two components: motif identification via phylogenetic footprinting, and regulon identification via the clustering of motifs. The phylogenetic footprinting step requires the genome sequences of several closely related species, and employs an extended Gibbs sampling algorithm to analyze orthologous promoter data to identify individual transcription factor binding sites and the associated motif model of

common binding patterns. The accuracy of these predictions has been evaluated by comparison with sets of sites reported for 166 genes in *Escherichia coli*, revealing that 75% of predicted sites overlap the experimentally verified sites by 10 bp or more. We have also developed a novel Bayesian clustering algorithm to predict regulons via clustering of the motifs identified in the footprinting step. Again, we validated this technique by comparison with reported regulons in *E. coli*. This inference of regulons utilizes only genome sequence information and is thus complimentary to and confirmative of gene expression data generated by microarray experiments. Here we describe preliminary results of our applications of these technologies to the *Synechocystis* PCC6803 genome.

Project ID: DE-FG02-01ER63204

47

Elucidating and Evaluating Patterns of Lateral Gene Transfer in Prokaryotic Genomes: Phylogenomic Analyses using GeneMarkS Gene Predictions

John Besemer[1], **Mark Borodovsky**[1] (mark.borodovsky@biology.gatech.edu), and John M. Logsdon, Jr.[2]

[1]Schools of Biology & Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA and [2]Department of Biological Sciences, University of Iowa, Iowa City, IA

Is the extent of lateral gene transfer (LGT) in prokaryotic genomes so large as to preclude using phylogenetic methods to elucidate evolutionary relationships among major bacterial lineages? The evolutionary mode of many ribosomal components has apparently been vertical since phylogenetic trees of these genes are in considerable agreement. Nonetheless, phylogenetic trees derived from many genes are either non-resolvable or incongruent (*e.g.* with a presumed species tree). With a general goal of determining the utility of phylogenetic methods to understand prokaryotic evolution, and a specific goal of understanding the utility of predicted 'Atypical' genes as surrogates for laterally transferred genes, we have begun a phylogenetic comparative study of GeneMarkS-predicted genes in 121 selected prokaryotic genomes, including many from the DOE Joint Genome Institute. Each of the GeneMarkS-predicted genes (both typical and atypical classes) from six cyanobacterial genomes was subjected to an iterative BLASTP procedure designed to find complete homolog sets from a collection of 121 complete genomes. While 296 gene families that contained at least one atypical cyanobacterial member resulted, only 161 contained enough taxa to construct rooted phylogenetic trees. The trees were classified into three major groups: i) trees where the cyanobacteria formed a single clade (including the trivial subset which contained cyanobacterial genes exclusively), ii) trees where the cyanobacteria were split, and iii) trees with a single cyanobacterial taxon. In (i) cases, the focus of the search for potential LGT was among the cyanobacteria themselves. In the other (ii & iii) cases, potential transfers into and out of the cyanobacterial lineage were investigated. To extend our analysis to the larger question of LGT among all prokaryotes, we needed to derive a more reliable reference topology. Incongruities in trees built for particular genes compared to this reference tree are potentially indicative of LGT. We have been developing a reference tree from concatenated protein alignments built from groups of genes empirically selected based on their presence or absence in our set of 121 complete genomes (*e.g.* their phylogenetic distributions). We are experimenting

with several different criteria for selecting the genes and trimming the alignments, as well as testing two methods for phylogenetic tree construction: maximum likelihood distance (from TREE-PUZZLE) and Bayesian likelihood (from MrBayes). Our current reference topology has been generated from 37 different genes and totals more than 5000 amino-acid residues in length. With our ultimate goal to estimate the extent and pattern of LGT among all prokaryotes, we have randomly selected 3000 'typical' predicted genes and 5000 'atypical' predicted genes from a representative set of 26 genomes. Analysis of these sets is being used to determine if atypical codon usage is a reliable predictor of LGT as detected by the extent of phylogenetic incongruities with respect to the class of typical genes. The methods being developed herein are easily scalable to allow the inclusion of newly sequenced genomes into the analysis and allow the adjustment of important parameters (such as BLAST E-values). In addition to testing the validity of using atypical genes as surrogates for laterally transferred genes, the results of these analyses will provide solid, phylogenetically-based estimates for the rates of LGT in prokaryotic genomes.

48

Cell Modeling and the Biogeochemical Challenge

P. J. Ortoleva (ortoleva@indiana.edu)

Center for Cell and Virus Theory, Indiana University, Bloomington, IN

As the Genomics:GTL project matures, at CCVT we are completing our cell models Karyote® and CellX® and are planning the multiple space-time extensions needed to use them in environmental analysis. Our two cell models, a virus-intracellular feature model, and future perspective are as follows.

Karyote® is a cell model that accounts for transcription and translation (by step-by-step polymerization), metabolics and molecular exchange between organelles and cytoplasm or with the surroundings. The reaction-transport equations are solved using multiple timescale mathematical and computational techniques. It has interfaces that allow for the building of a compartmented eukaryotic cell or a simple or composite bacterium. Modules also have been developed for creating multi-cellular systems from models of single cell types, or for viewing results or the network of processes accounted for. The Karyote® system includes a database of cell properties, pathways and kinetic parameters, and procedures for model calibration using raw (e.g. NMR, mass spectral, microarray, microscopy) data.

CellX® is a finite element simulator that simultaneously solves reaction-transport equations on fibriles (1-D), membranes (2-D) and bulk medium (3-D), and the exchange among them through boundary conditions. CellX® has all the features of Karyote® and in addition accounts for gradients within each compartment (notably of proteins and other slowly migrating species)

VirusX® is designed to model the structure or function of a virus or other subcellular feature. It accounts for the supra-million atom structural detail of the virus or other object. VirusX® solves the molecular mechanics problem using space-warping and tree-code methods. We are developing mixed mesoscopic models wherein the capsid or other viral features can be described by continuous variables or wherein the dynamics of focus macromolecules is simulated using efficient computational techniques. The electrolyte host medium is treated using our continuum

position-orientation density description, a nonlinear, nonlocal dielectric model and associated free energy functional.

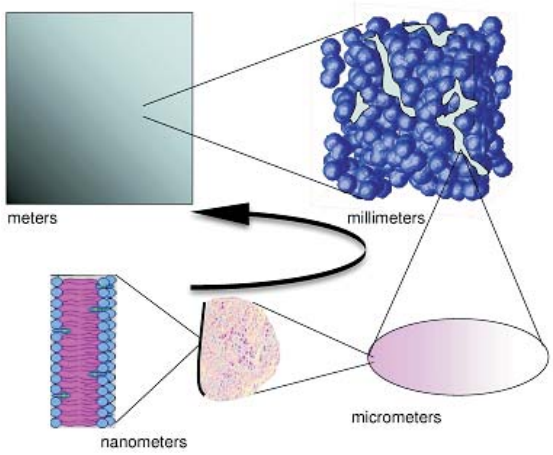
For computational biology to make a major contribution to microbial research, several grand challenges must be addressed. Greater predictive capability will only follow when reaction-transport models account for metabolic, proteomic, genomic and structure-forming (i.e. biomic) processes. Models accounting for this biomic process network will allow biologists to understand all the couplings among these processes and to simulate the life cycle of a microbe. This challenge is being addressed by our Karyote® model. The architecture of a microbe interior must be understood as a multi-dimensional system; this is being addressed via our CellX® simulator as described above. The impressive intracellular architecture and related functioning is strongly influenced by molecular-scale physics (e.g. fibrils along which molecules are trafficked between specific origins and destinations). Thus a fuller understanding of microbial behavior must involve the integration of molecular- and mesoscopic-scale physics and chemistry. In short, microbial models must be cast in the language of nanotechnology and mesoscopic theory, integrating atomistic to macroscopic variables (e.g. molecular structure to metabolite concentrations). In this way, microbial models will be cast in terms of the three scales (molecular, mesoscopic and macroscopic) at which biologists conceptualize and integrate their thoughts. This challenge is being addressed by our all-atom/mesoscopic simulator, VirusX®.

It must be admitted that for at least the next decade models will be incomplete and thus, even for well-understood processes, the rate and thermodynamic parameters are known only with great uncertainty or not at all. Even rate parameters determined from cell extracts may be far from those in the intracellular environment. Thus, procedures are needed that address calibration even for incomplete models. The data available (e.g. NMR, mass spectra) will only be indirectly related to the model parameters. The two endeavors (i.e. comprehensive biomic cell modeling and experimental data acquisition/interpretation) should be unified into one procedure to automate the building of models and the design of experiments to minimize (and assess) the uncertainties in both. Approaches must be developed that integrate the more complete biomic models into multi-cellular/sediment composite media systems. Effective models must account for the network of biomic intracellular processes that cannot be accounted for in available lumped models. Fundamental approaches starting from pore-scale detailed descriptions that are then rigorously upscaled to the field are needed to attain predictability as suggested in the figure. Mathematical/computational approaches involving homogenization theory should be applied to account for flow, diffusion, aqueous and mineral reactions, and the exchange with the evolving microbial colonies. Research should not be based on straightforward extension of simple lumped models, as they cannot easily be extrapolated to other sites using a calibration at a given site.

In summary, we must automate the integration of experimental and modeling technologies, develop a multi-scale platform to facilitate creative life sciences thinking, cast the approach in a manner that addresses the inherent uncertainties in experimental and computational life sciences, and simultaneously utilize the rapidly growing databases of genomic, proteomic, metabolic and structural information. For example, the genomic and proteomic information should be integrated into a meta-

bolic model via a detailed kinetic model of the polymerization of mRNA and proteins constituting transcription, translation and post-translational processing.

Schematic depiction of the multiple levels at which understanding of microbial systems is needed to attain quantitative predictability. Advances in the accuracy of the models at each level and methods for upscaling them are needed.



49

Rapid Reverse-Engineering of Genetic Networks via Systematic Transcriptional Perturbations

J. J. Collins (jcollins@bu.edu), T. S. Gardner, and C. R. Cantor

Department of Biomedical Engineering, Boston University, Boston, MA

The collection and assembly of large-scale genetic data into comprehensive databases is often regarded as the necessary first step in the elucidation of genetic network structure and function. However, it is not obvious if or how the disparate and partial data populating such databases can be assembled into unambiguous and predictive models of genetic networks. To address this problem, we have developed a reverse-engineering method that enables rapid construction of a first-order quantitative model of a gene regulatory network using no prior information on the network structure or function. The method, called Network Identification by multiple Regression (NIR), uses a series of steady-state transcriptional perturbations, coupled with RNA, protein, or metabolite activity measurements and multiple linear regression, to construct the model. In a pilot study, we successfully applied the NIR method to reverse engineer a 9-gene subnetwork in *E. coli* (Gardner et al., Science 301: 102, 2003).

Computational testing and our pilot *E. coli* study suggest that the NIR method can be applied on a large scale. In this project, we plan to extend the method to larger prokaryotic networks. Specifically, we plan to apply the method to *Shewanella oneidensis* electron-transport networks relevant to bioremediation. We will use *Shewanella* microarrays developed by DOE researchers to perform large-scale RNA profiling for our experiments. Our efforts on *Shewanella* will be designed to build

on and to support the existing experimental and computational efforts of Genomics:GTL (GTL) researchers.

The mapping and modeling of genetic networks in prokaryotes, a central objective of the GTL project, will provide the foundations for a variety of applications that advance the DOE mission needs in energy and the environment. Our method could significantly improve the efficiency of existing efforts of GTL researchers to map and model genetic networks in microbes.

50

Computational Hypothesis Testing: Integrating Heterogeneous Data and Large-Scale Simulation to Generate Pathway Hypotheses

Mike Shuler (info@gnsbiotech.com)

Gene Network Sciences, Ithaca, NY

Most prokaryotes of interest to DOE are poorly understood. Even when full genomic sequences are available, the function of only a small number of gene products are clear. The critical question is how to best infer the most probable network architectures in cells that are poorly characterized. The project goal is to create a computational hypothesis testing (CHT) framework that combines large-scale dynamical simulation, a database of bioinformatics-derived probable interactions, and numerical parallel architecture data-fitting routines to explore many "what if?" hypotheses about the functions of genes and proteins within pathways and their downstream effects on molecular concentration profiles and corresponding phenotypes. From this framework we expect to infer signal transduction pathways and gene expression networks in prokaryotes. The focus of this proposal is the:

1. Extension of accurate dynamical simulation methods to genome-size scales (i.e., 1000s).

2. Normalization of confidence levels for a wide variety of bioinformatics algorithms for extraction of a database of probable interactions.

3. Extension of numerical data-fitting techniques to large multi-scale cell simulations and exploitation of biological network properties for more efficient use of computational resources.

4. Integration of 1., 2., and 3. to derive an ensemble of hypothetical network structures and their corresponding molecular concentration profiles and phenotypic outcomes.

In order to create, refine and validate such a method for application to organisms of DOE interest where little functional data is available, our project will address a number of issues.

- Given the enormous cost in both time and money to collect genome wide data sets, it is important to determine what types of data and what quantities and qualities of data are necessary to lead to an inference of pathway circuitry at a given confidence level. Our proposed CHT platform would accomplish this determination through the integration of varying amounts of heterogeneous data for a

non-DOE model bacterial organism where much of the functional biochemical circuitry is known.

- Methods currently exist to extract static biochemical circuitry through the application of bioinformatics algorithms to whole genome sequences. This static circuitry does not directly link a particular molecular circuitry to the corresponding molecular concentration profiles and corresponding phenotypes. Our dynamic CHT is the tool that would quickly and inexpensively (in)validate the enumerable number of predicted network architectures that arise from the application of bioinformatics algorithms to whole genomes.
- As a more detailed functional understanding of DOE microbes is attempted, experimental efforts could greatly be accelerated if it were possible to investigate the multitude of hypotheses faster and cheaper than can be accomplished by experiment alone. We propose CHT as a method to investigate enumerable hypotheses on the computer with the goal of eliminating many improbable hypotheses and suggesting a more focused set of experiments.

CHT will be tested, validated, and applied to three different systems in decreasing order of completeness and transparency and ability to validate. The first is a synthetic network system that is created within our computer simulation framework (where by definition 100% of the circuitry and molecular functions are known). The second system is *E. coli K-12* (where 60–70% of the molecular functions are known). The third system and the one that is ultimately of direct interest to the DOE's objective is *Shewanella oneidensis* (where less than 10% of the circuitry and molecular functions are known).

51

Bacterial Annotation Tools

Owen White (owhite@tigr.org)

The Institute for Genomic Research, Rockville, MD

Manatee (MANual Annotation Tool, Etc Etc) is a graphical user interface designed to manage data in a common database that allows multiple users to simultaneously operate on that information. Production annotation teams at TIGR, as well as outside collaborators, have been using this system for the past year to obtain essential annotation information in a user-friendly way. The system supports making functional assignments using search results, paralogous families, and annotation suggestions generated from automated analysis. It also produces summaries that report the progress of each annotation project. Manatee runs using a web browser interface and easily allows installation of additional web scripts to facilitate frequent and rapid improvements. The Manatee suite contains documentation for installation and general use by scientists, and also contains complete documentation of the code libraries that can be modified by other software developers. The complete code base is available as open source and may be installed locally on Unix computer systems. Classes are now offered quarterly at TIGR giving instruction on the use of this software and on the "best practices" that have been formalized to generate uniform annotation.

Because of the relative ease of adding to the Manatee system, we are using this architecture as the basis of the Sybil software, a system that will allow scientists to discover, evaluate, and summarize intra-species variation. Sybil data storage is an

open source, modular relational schema called Chado that has been populated with data from closely related species. The Sybil API which makes calls to the database is operational, and many prototypic interfaces that display complex comparative data have been implemented. The system will be used for comparative analysis in two broad areas: 1) the interface will be incorporated into a web resource and used on-line by scientists interrogating data from TIGR resident in our database, and 2) users will be able to install the Sybil system on their local computers to perform custom analyses of their data. We anticipate a release of Sybil sometime before the end of year.

We have also developed a workflow system that performs routine operations required for most annotation pipelines. This system is based on simple configuration files and workflow templates. The configuration file is human-readable and contains the definitions used to generate a particular executable instance of a workflow, such as the Blast or HMMer programs. The workflow template is an XML document and describes the graph defining the overall pipeline. To begin the pipeline, the workflow execution engine executes an instance of the template; this "workflow instance" contains the complete description of the pipeline that has been invoked, and will be continuously updated with the status of the pipeline. Placing the status of completion in the workflow instance document allows for monitoring the process, and supports resuming fail or aborted workflows from the appropriate starting point. This engine is capable of handling a variety of complex workflows that may contain a combination of parallel and sequential processes and executes jobs in a distributed or grid computing environments.

The above projects have been developed in conjunction with DOE support for the Comprehensive Microbial Resource and NSF funding.

52

RELIC - A Bioinformatics Server for Combinatorial Peptide Analysis and Identification of Protein-Ligand Interaction Sites

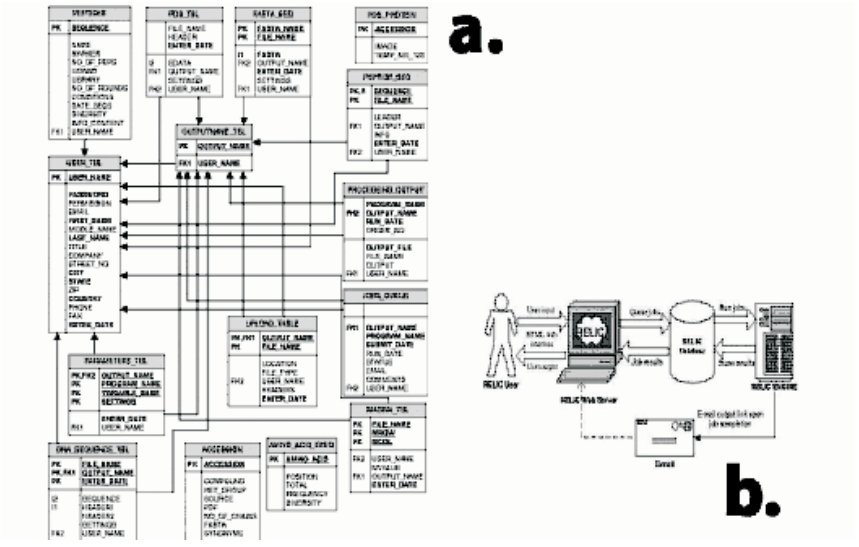
Suneeta Mandava*, Lee Makowski, Satish Devarapalli, Joseph Uzubell, and **Diane J. Rodi** (drodi@anl.gov)

*Presenting author

Biosciences Division, Argonne National Laboratory, Argonne, IL

The need for high throughput bioinformatic methods to characterize gene function is being driven by the generation of sequences at a rate far beyond our ability to carry out experimental functional analyses. In spite of the large number of analytical tools currently available, typically about 40% of predicted open reading frames remain functionally uncharacterized. An important clue to open reading frame function is the identification of binding partners. Phage display technology is a widely used tool for identifying either protein or small molecule binding partners. This project seeks to apply a novel approach to genome-wide identification of small molecule binding proteins. Preliminary results from our group has demonstrated that the similarity between the sequence of a protein and the sequences of affinity-selected, phage-displayed peptides can be predictive for protein binding to a small molecule ligand. Affinity-selected peptides provide information analogous to that of a consensus-binding sequence, and can be used in an analogous fashion to identify ligand binding sites.

In this project, libraries of phage-displayed peptides have been screened for affinity to the metabolites ATP and glucose, as well as other small molecule ligands. The sequences of affinity-selected peptides were determined and used as the basis of genome-wide analyses to identify proteins that have a high probability of binding to the screened ligands. The best set of affinity-selected peptides as validated through comparison with well-characterized proteins was used for genome-wide annotation of the *E. coli* genome as an initial test genome with a high percentage of functional annotation. During the course of this work, we have developed a suite of computational tools for the analysis of peptide populations and made them accessible by integrating fifteen software programs for the analysis of combinatorial peptide sequences into the REceptor LIgand Contacts (RELIC) relational database and web-server. These programs have been developed for the analysis of statistical properties of peptide populations; identification of weak consensus sequences within these populations; and the comparison of these peptide sequences to those of naturally occurring proteins. RELIC is particularly suited to the analysis of peptide populations affinity selected with a small molecule ligand such as a drug or metabolite. The order of the programs and their specific functionalities is specifically designed to aid a researcher in the combinatorial peptide field from the early stages of raw data acquisition to the final stage of protein epitope mapping. The flow of data-processing software starts with sequence translation programs, followed by physicochemical property mapping, sequence bias identification algorithms, and finally peptide/protein similarity mapping both within and in the absence of three-dimensional coordinates. In order to seamlessly integrate that biological data, RELIC is based on an object-oriented design using a relational database management system. For this particular project, the ORACLE 9i (Release 9.2) database system was chosen to store experimental data and the relevant genomic/structure information as it provides a wide array of database drivers for various programming languages (both for thin and thick clients). The figure at the lower left below (a.) is a diagram depicting the logical and relational model of the database by displaying all tables and intra-table relationships. The figure at the right (b.) shows a schematic of how users interact with the RELIC hardware. A RELIC user submits data for pro-



cessing via a web interface. The user input and job information is stored in a RELIC database. A job processing service periodically checks for pending jobs and processes them using the scientific algorithms developed in FORTRAN, using

COM+ interfaces. The user is sent an email upon completion of the job with a link to the output. Within this functional context, the ability to identify potential small molecule binding proteins using combinatorial peptide screening will accelerate as more ligands are screened and more genome sequences become available. The broader impact of this work is the addition of a novel means of analyzing peptide populations to the phage display community.

53

On Truth, Pathways and Interactions

Andrey Rzhetsky (ar345@columbia.edu)

Department of Biomedical Informatics and Columbia Genome Center, Columbia University, New York, NY

I will give an overview of our effort to automatically extract pathway information from a large number of full-text research articles (GeneWays system), automatically curate the extracted information, and to combine the literature-derived information with sequence and experimental (such as yeast two-hybrid) data using a probabilistic approach.