Milestone 3

GTL Computational Biology Environment

Section 1

Computing Infrastructure and Education

121

Research Highlights from the BACTER Program

Julie C. Mitchell* (mitchell@math.wisc.edu), Timothy J. Donohue, George N. Phillips Jr., Nicole Perna, Qiang Cui, David Schwartz, Mark Craven, Paul Milewski, and Stephen Wright

University of Wisconsin, Madison, WI

The BACTER Institute for Computational Biology (http://www.bacter.wisc.edu) has completed its first full year of graduate training activities. We will present vignettes of the research being done by BACTER students and postdocs, in collaboration with the above-mentioned faculty at the University of Wisconsin. Our training and research efforts are directed in three areas: Comparative Genomics, Cellular and Molecular Models, and Biological Pathways. To help focus and interrelate the research of our students, *Rhodobacter sphaeroides* and *Shewanella oneidensis* are the model organisms to which they apply their modeling and informatics toolkits. In comparison with frequent model organisms like yeast or *E. coli*, there is less experimental data available for our model systems. Thus, all of our students, whether in genomics or structural biology, must actively seek out the most recently acquired experimental data and deal with incomplete information. In some cases, our students have initiated their own collaborations with experimental groups to generate the data needed for their research.

Projects in Biological Pathways

We have several students and postdocs working on biological pathways. One student has created a model for the Calvin Cycle in *Rhodobacter*, an organism that has two distinct forms of Rubisco. A BACTER postdoc has performed a highly accurate fitting of experimental data for metal reduction in *Shewanella*, within the context of high-level mathematical models that can be used to predict reaction rates in the presence of multiple metals and metabolites. Two additional students are working on different aspects of microarray data analysis for *Rhodobacter* and *Shewanella*, with the goal of mapping the interactomes of these organisms via machine learning techniques and automated model generation.

Projects in Cellular and Molecular Models

Molecular dynamics and quantum mechanical calculations are being used to predict rate constants across entire biological pathways in *Rhodobacter*. In cases where data is missing, or available only for other organisms, feasible parameter values can fill in the gaps. This undertaking will greatly help advance the research of the Biological Pathways group, and even drive experimental research to help refine and adapt their predictive models. In addition, BACTER is developing accurate protein docking methods able to model large conformational changes with significant dimension reduction in the free variables. Finally, chemotaxis is being studied from the perspective of coupling spatial and

signaling phenomena into a single mathematical model, which might be neither a partial differential equation nor a system of nonlinear equations, but clearly must have some properties of each.

Projects in Comparative Genomics

We are applying optical mapping technology to *Rhodobacter* genomes, to discover which essential genes are conserved across multiple isolates of its species. The ability to accurately characterize the photosynthetic machinery and carbon sequestration abilities of simple organisms cannot be underestimated. In addition, large-scale comparative genomics tools are being applied to recently sequenced *Shewanella* genomes from JGI, and research into the genetic basis of chemotaxis in *Rhodobacter* will provide nice ties to the research of BACTER's Cellular and Molecular Models group.

122

UC Merced Center for Computational Biology

Michael Colvin[1]* (mcolvin@ucmerced.edu), Arnold Kim[1], Masa Watanabe[1], and Felice Lightstone[2]

[1]University of California, Merced, CA and [2]Lawrence Livermore National Laboratory, Livermore, CA

We have established a Center for Computational Biology (UCM-CCB) at the newest campus of the University of California. The UCM-CCB is sponsoring multidisciplinary scientific projects in which biological understanding is guided by mathematical and computational modeling. The center is also facilitating the development and dissemination of undergraduate and graduate course materials based on the latest research in computational biology. This project is a multi-institutional collaboration including the new University of California campus at Merced, Rice University, Rensselaer Polytechnic Institute, and Lawrence Livermore National Laboratory, as well as individual collaborators at other sites.

The UCM-CCB is sponsoring a number of research projects that emphasize the role of predictive simulations in guiding biological understanding. This research is being performed by post-docs, graduate and undergraduate students and includes mathematical models of cell fate decisions, molecular models of multiprotein machines such as the nuclear pore complex, new mathematical methods for simulating biological processes with incomplete information, and mathematical approaches for simulating the interaction of light with biological materials. The UCM-CCB has run workshops to facilitate computational collaborations with many of the experimental biology programs at UC Merced and is hosting an ongoing seminar series that will bring five prominent computational biologists to speak at UC Merced in Winter and Spring 2006.

Additionally, the UCM-CCB is working to translate this research into educational materials. The UCM-CCB is having a central role in enabling the highly mathematical and computationally intensive Biological Science major, which is currently the largest major at UC Merced and has attracted a very large number of applications for next year. In Fall 2005, the first semester of undergraduate instruction at UC Merced, UCM-CCB materials and expertise were used in computational biology laboratories for two large undergraduate biology courses, and such materials will be used in several more courses in Spring 2006. All course materials are being released under an open public license, and we are in the process of translating these materials into modules in the Connexions courseware system developed at Rice University. The electronic, modular course materials produced by the UCM-CCB are also facilitating linkages to feeder schools at the state university, community college, and high school levels.

* Presenting author

The long-term impact of the CCB will be to help train a new generation of biologists who bridge the gap between the computational and life sciences and to implement a new biology curriculum that can both influence and be adopted by other universities. Such scientists will be critical to the success of new approaches to biology, exemplified by the DOE Genomes to Life program in which comprehensive datasets will be assembled with the goal of enabling predictive modeling of the behavior of microbes and microbial communities, as well as the biochemical components of life, such as multiprotein machines.

123

The BioWarehouse System for Integration of Bioinformatics Databases

Tom Lee, Valerie Wagner, Yannick Pouliot, and **Peter D. Karp*** (pkarp@ai.sri.com)

SRI International, Menlo Park, CA

BioWarehouse[1] is an open-source toolkit for constructing bioinformatics database (DB) warehouses. It allows different users to integrate collections of DBs relevant to the problem at hand. BioWarehouse can integrate multiple public bioinformatics DBs into a common relational DB management system, facilitating a variety of DB integration tasks including comparative analysis and data mining. All data are loaded into a common schema to permit querying within a unified representation.

BioWarehouse currently supports the integration of UniProt, ENZYME, KEGG, BioCyc, NCBI Taxonomy, CMR, Gene Ontology, and the microbial subset of Genbank. Loader tools implemented in the C and Java languages parse and load the preceding DBs into Oracle or MySQL instances of BioWarehouse.

The BioWarehouse schema supports the following bioinformatics datatypes: chemical compounds, biochemical reactions, metabolic pathways, proteins, genes, nucleic acid sequences, features on protein and nucleic-acid sequences, organism taxonomies, and controlled vocabularies.

BioWarehouse is in use by several bioinformatics projects. An SRI project is developing algorithms for predicting which genes within a sequenced genome code for missing enzymes within metabolic pathways predicted for that genome[2]. BioWarehouse fills several roles within that project: it is used to construct a complete and nonredundant dataset of sequenced enzymes by combining protein sequences from the UniProt and PIR DBs, and by removing from the resulting dataset those sequences that share a specified level of sequence similarity. Our current research involves extending the pathway hole filling algorithm with information from genome-context methods such as phylogenetic signatures, which are obtained from BioWarehouse thanks to the large all-against-all BLAST results stored within CMR. Another SRI project is comparing the data content of the EcoCyc and KEGG DBs using BioWarehouse to access the KEGG data in a computable form.

BioWarehouse is supported by the Department of Energy and by DARPA through the DARPA BioSPICE program for biological simulation.

References

1. BioWarehouse Home Page http://bioinformatics.ai.sri.com/biowarehouse/
2. Green, M.L. and Karp, P.D., "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases," *BMC Bioinformatics* 5(1):76 2004 http://www.biomedcentral.com/1471-2105/5/76.