GTL Milestone 3

Develop the Knowledgebase, Computational Methods, and Capabilities to Advance Understanding and Prediction of Complex Biological Systems

Section 1
Computing Infrastructure, Bioinformatics, and Data Management

Center for Computational Biology at the University of California, Merced

Michael Colvin1* (mcolvin@ucmerced.edu), Arnold Kim,1 Masa Watanabe,1* and Felice Lightstone2

1School of Natural Sciences, University of California, Merced, California and 2Biosciences Directorate, Lawrence Livermore National Laboratory, Livermore, California

Project Goals: The goals of the UCM-CCB are 1) 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, 2) and to facilitate the development of multidisciplinary research programs in computational and mathematical biology.

The Center for Computational Biology (UCM-CCB) was established at the newest campus of the University of California in fall, 2004. 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 has brought many prominent computational biologists to speak at UC Merced over the past year.

Additionally, 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. Over the past year, members of the UCM-CCB have visited several other universities to describe this new major. All electronic, modular course materials produced by the UCM-CCB being released under an open public license and have been facilitating linkages to feeder schools at the state university, community college, and high school levels. Last summer the UCM-CCB ran a six-week compu-
tional biology internship program, that culminated in a one-week visit to the National Center for Supercomputing Applications at the University of Illinois.

The long-term goal of the UCM-CCB is 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 Genomics:GTL 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.

113

Projects from the DOE-BACTER Institute at the University of Wisconsin, Madison

Julie C. Mitchell* (mitchell@math.wisc.edu), Julie Simons, Paul Milewski, Peter Koenig, and Qiang Cui

University of Wisconsin, Madison, Wisconsin

Project Goals: BACTER was initiated through a DOE call for proposals to address the critical need for computational modelers in the area of bioenergy systems modeling. BACTER faculty, students and postdocs work on DOE-relevant modeling projects, ranging from molecules, to pathways and cells, to entire populations.

Selected here are two of many projects ongoing at the DOE-BACTER Institute. BACTER was initiated through a DOE call for proposals to address the critical need for computational modelers in the area of bioenergy systems modeling. BACTER faculty, students and postdocs work on DOE-relevant modeling projects, ranging from molecules, to pathways and cells, to entire populations.

*R. sphaeroides and E. coli: a comparison of population-level behavior

Understanding the behavior of bacteria in populations is of importance in biofilm formation, our potential to harness bacteria for use in such contexts as bioremediation, and understanding how simple organisms display complex behavior. We are interested in modeling the chemotactic behavior of *Rhodobacter sphaeroides* in an effort to elucidate why this species behaves differently from the better understood *Escherichia coli* in different environments. Specifically we performed “swarm-plate” experiments, inoculating *R. sphaeroides* and *E. coli* in agar plates with varying concentrations of the chemotactant L-aspartate. Separately we also measured growth rates for the bacteria in liquid cultures.

From the data obtained, we examine the behavior seen via a set of partial differential equations known as the Keller-Segel equations. These equations allow us to differentiate growth effects from chemotaxis and point to sensitivities of *R. sphaeroides* as compared to *E. coli* in terms of their abilities to create and respond to gradients of chemotactant and move within the agar media.

Perturbative analysis of potential of mean force simulations

Enzymes reach their high efficiency in catalyzing reactions. Bioenergetic enzymes interconvert energy with unsurpassed efficiency. Gaining insight into how enzymes modulate the energetics of a reaction is a key step to their understanding.
Molecular dynamics simulations of chemical reactivity have become routine. The simulation of the free energy along a reaction coordinate, not only describes the chemical reaction per se, but is also able to capture the dynamic response of the reacting species and their environment. Despite the advantages of these simulations, the methods for their analysis lag behind. For instance, there is no established methodology for probing the quality of the potential functional or the importance of interactions and contributions.

We suggest a perturbative approach for examining the importance of interactions for the energetics of a reaction. Beyond simple \textit{in silico} mutations, this technique allows the exploration of components not easily accessible to experimental study, e.g. the electrostatic effect of a whole part of a protein or solvation.

We present the application of this method for studying reactions in a number of enzymes. We have also applied this methodology to study the mechanism of proton blockage in aquaporin, which is an ongoing subject of discussion in the literature. This membrane channel efficiently conducts water molecules but prevents the decoupling of bioenergetic processes due to proton leakage. With the perturbative approach introduced here, the different contributions to the energetics could be dissected.

114 GTL

VIMSS Computational Core

Paramvir S. Dehal\textsuperscript{1,2*} (PSDehal@lbl.gov), Eric J. Alm\textsuperscript{1,3}, Dylan Chivian\textsuperscript{1,2}, Katherine H. Huang\textsuperscript{1,2}, Wayne Huang\textsuperscript{1,2}, Janet Jacobsen\textsuperscript{1,4}, Marcin P. Joachimiak\textsuperscript{1,2}, Keith Keller\textsuperscript{1,4}, Morgan N. Price\textsuperscript{1,2} and Adam P. Arkin\textsuperscript{1,2,4,5,6} (aparkin@lbl.gov)

\textsuperscript{1}Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov/; \textsuperscript{2}Lawrence Berkeley National Laboratory, Berkeley, California; \textsuperscript{3}Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; \textsuperscript{4}University of California, Berkeley, California; \textsuperscript{5}Howard Hughes Medical Institute, Chevy Chase, Maryland; and \textsuperscript{6}Department of Bioengineering, University of California, Berkeley, California

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics: GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

Background: The VIMSS Computational Core group is responsible for data management, data integration, data analysis, and comparative and evolutionary genomic analysis of the data for the VIMSS project. We have expanded and extended our existing tools sets for comparative and evolutionary genomics and microarray analysis as well as creating new tools for our proteomic and metabolomic data sets. Our analysis has been incorporated into our comparative genomics website MicrobesOnline (http://www.microbesonline.org) and made available to the wider research community. By taking advantage of the diverse functional and comparative datasets, we have been able to pursue large evolutionary studies.

* Presenting author
Data Analysis: During the course of analysis of various stress responses of DvH the computational core has continued to develop new statistical analyses of data that take advantage of the predicted regulatory structures (operons, regulons, etc.) from our comparative analyses. This year we have used these analyses to investigate the response of DvH to oxygen stress and chromium stress. Our analysis has focused on the combined results from both transcriptomic and proteomic datasets to interpret oxygen stress. Additionally, we have begun preliminary work to examine metabolomic datasets within the framework of predicted metabolic activities.

Data Management: All data generated by ESPP continues to be stored in our Experimental Information and Data Repository (http://vimss.lbl.gov/EIDR/). Researchers have access to datasets from biomass production, growth curves, image data, mass spec data, phenotype microarray data and transcriptomic, proteomic and metabolomic data. New functionality has been added for storage of information relating to mutants and protein complex data, in addition to new visualization for assessing existing data sets such as the phenotype microarrays.

The MicrobesOnline Database: The MicrobesOnline database (http://www.microbesonline.org) currently holds 330 microbial genomes and will soon be expanded to over 600 genomes, providing an important comparative genomics resource to the community. New functionality added this year includes the addition of a phylogenetic tree based genome browser that allows users to view their genes and genomes of interest within an evolutionary framework, tools allowing users to search for novel regulatory binding site motifs or matches to existing regulatory binding motifs across a user selected set of genes using our gene carts, tools to compare multiple microarray expression data across genes and genomes and more integration with the RegTransDb of experimentally verified regulatory binding sites.

MicrobesOnline continues to provide an interface for genome annotation, which like all the tools reported here, is freely available to the scientific community. To keep up with the rapidly expanding set of sequenced genomes, we have begun to investigate methods for accelerating our annotation pipeline. In particular we are researching methods to speed up the most time consuming process, homology searching through HMM alignments and all against all BLAST. Over the next year we will be releasing methods that will allow us to deal with the many millions of gene sequences generated from metagenomics.

Over the next year, several new features will be added to the MicrobesOnline resource. Microarray expression data will be added from the NCBI GEO database, in addition to datasets generated from the VIMSS team. To supplement the analysis tools we already have, enrichment of functional genes and operon-wise analysis, we will provide tools for comparing multiple experiments across multiple genomes. We will also expand our regulatory binding motif search to incorporate co-expression data to support predictions.

Evolutionary Analysis: The computational core continues work on understanding the evolution of regulatory networks. Transcription factors form large paralogous families and have complex evolutionary histories. Our analysis shows that putative orthology derived from bidirectional best hits across distantly related bacteria are usually not true evolutionary orthologs. Additionally, these false orthologs usually respond to different signals and regulate distinct pathways. Even in more closely related genomes, such as E. coli and Shewanella oneidensis, bidirectional best hits have a high error rate. By studying transcription factors with phylogenetic trees, we show that through the use of gene-regulon correlations, together with sequence analysis of promoter regions for confirmation, bacterial regulatory networks may evolve more rapidly than previously thought.
RegTransBase – A Resource for Studying Regulatory Interactions and Regulon Predictions in Bacteria

Michael J. Cipriano,1,5* Alexei E. Kazakov,2 Dmitry Ravcheev,2 Adam P. Arkin1,3,4,5 (aparkin@lbl.gov), Mikhail S. Gelfand,2 and Inna Dubchak1,5,6

1Lawrence Berkeley National Laboratory, Berkeley, California; 2Institute for Information Transmission Problems, Moscow, Russia; 3Howard Hughes Medical Institute, Chevy Chase, Maryland; 4Department of Bioengineering, University of California, Berkeley, California; 5Virtual Institute of Microbial Stress and Survival, http://vimss.lbl.gov; and 6Department of Energy Joint Genome Institute, Walnut Creek, California

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics: GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

RegTransBase, a database describing regulatory interactions in prokaryotes, is manually curated and based on published scientific literature. RegTransBase was created to supplement the data available in MicrobesOnline by including published experimental information on regulation. It describes a large number of regulatory interactions and contains experimental data from ~3000 articles from a wide range of taxa (including environmentally interesting organisms) which investigates regulation with known elements. RegTransBase additionally provides an expertly curated library of alignments of known transcription factor binding sites, and includes the exact location of the binding site on a published genome, the transcription factor, and links to published articles. RegTransBase builds upon these alignments by containing a set of computational modules for the comparative analysis of regulons among related organisms which guide a user through the appropriate steps of transferring known or high confidence regulatory binding site results to many other microbial organisms. An intuitive, interactive user-friendly interface makes this knowledge freely accessible to the larger microbiological research community.

RegTransBase creates access to the highest quality regulatory information about sequenced microbes and, with MicrobesOnline, is a critical tool to the inference of regulatory and stress response networks that are central goals of the VIMSS::ESPP project. RegTransBase is available at http://regtransbase.lbl.gov.

* Presenting author
MicrobesOnline: An Integrated Portal for Comparative Functional Genomics

Marcin P. Joachimiak,1,2* Katherine H. Huang,1,2 Eric J. Alm,1,3 Dylan Chivian,1,2 Paramvir S. Dehal,1,2 Y. Wayne Huang,1,2 Janet Jacobsen,1,4 Keith Keller,1,4 Morgan N. Price,1,2 and Adam P. Arkin1,2,4,5,6 (aparkin@lbl.gov)

1 Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov/; 2 Lawrence Berkeley National Laboratory, Berkeley, California; 3 Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts; 4 University of California, Berkeley, California; 5 Howard Hughes Medical Institute, Chevy Chase, Maryland; and 6 Department of Bioengineering, University of California, Berkeley, California

Project Goals: Environmental Stress Pathway Project (ESPP) is developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to the environmental conditions found in DOE waste sites. The research takes place within the Virtual Institute for Microbial Stress and Survival (VIMSS). Based at Lawrence Berkeley National Laboratory (LBNL), VIMSS supports an integrated and multi-institutional program to understand the ability of bacteria and other microorganisms to respond to and survive external stresses. VIMSS was established in 2002 with funding from the U.S. Department of Energy Genomics: GTL Program for Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria. LBNL is operated by the University of California for the U.S. Department of Energy.

The Virtual Institute for Microbial Stress and Survival (VIMSS, http://vimss.lbl.gov) funded by the Dept. of Energy's Genomics:GTL Program, is dedicated to using integrated environmental, functional genomic, and comparative sequence and phylogeny data to understand mechanisms by which microbes survive in uncertain environments while carrying out processes of interest for bioremediation and energy generation. To support this work, VIMSS has developed a Web portal with an underlying database and analyses for comparative functional genomics of bacteria and archaea. Since 2003 MicrobesOnline (http://www.microbesonline.org) has been enabling comparative genome analysis and currently includes 451 complete genomes and offers a suite of analysis and tools including: a multi-species genome browser, operon and regulon prediction methods and results, a combined gene and species phylogeny browser, a gene ontology browser, a workbench for sequence analysis (including sequence motif detection, motif searches, sequence alignment and phylogeny reconstruction), and capabilities for community annotation of genomes.

VIMSS integrates functional genomic data and provides novel web-based viewing and mining tools for gene expression microarray, proteomic, and phenotype microarray data. Currently, these data are mostly project generated for wild-type and mutants of Desulfovibrio vulgaris and Shewanella oneidensis exposed to stress conditions found at DOE field sites. Selecting an organism or gene of interest in MicrobesOnline leads to information about and data viewers for VIMSS experiments conducted on that organism and involving that gene or gene product. It is also now possible to view microarray data from multiple stress conditions as an interactive heatmap and to analyze correlations between gene expression results from different experiments. Among the major new features is the ability to search any subset of experiments in the microarray data compendium for similar gene matches to a mean expression profile derived from an a priori determined group of genes (e.g., a known or predicted regulon). These new compendium-wide functionalities allow one to observe patterns in...
gene expression changes across multiple conditions and to search for similarities to these patterns. The information integration and analysis performed by VIMSS serves not only to generate insights into the stress responses and their regulation in these microorganisms, but also to document VIMSS experiments, allow contextual access to experimental data, and facilitate the planning of future experiments. VIMSS also is incorporating into MicrobesOnline publicly available functional genomics data from published research, so as to centralize and synergize data on microbial physiology and ecology in a unified comparative functional genomic framework.

Protein Complex Analysis Project (PCAP): Data Management and Bioinformatics Subproject

Adam P. Arkin,1,2,3 Ralph Santos,1,3 Y. Wayne Huang,1,3 Janet Jacobsen,1,2,3 Keith Keller,2,3 Steven S. Andrews,1,3 Steven E. Brenner,1,2,3 Max Shatsky,1,3 and John-Marc Chandonia1,3* (JMChandonia@lbl.gov)

1Lawrence Berkeley National Laboratory, Berkeley, California; 2University of California, Berkeley, California; and 3Virtual Institute for Microbial Stress and Survival, Berkeley, California http://vimss.lbl.gov

Project Goals: The Data Management and Bioinformatics component of the Protein Complex Analysis Project (PCAP) has two major goals: 1. to develop an information management infrastructure that is integrated with databases used by other projects within the Virtual Institute for Microbial Stress and Survival (VIMSS), and 2. to analyze data produced by the other PCAP subprojects together with other information from VIMSS to model stress responses relevant to the use of D. vulgaris and similar bacteria for bioremediation of metal and radionuclide contaminated sites. In addition to storing experimental data produced by the PCAP project, we will assess the quality and consistency of the data, and compare our results to other public databases of protein complexes, pathways, and regulatory networks. We will prioritize proteins for tagging, TAP, and study by EM based on analysis of VIMSS data and other bioinformatic predictions. All data we obtain on protein interactions will be analyzed in the context of the data currently stored in VIMSS. One of the primary goals of VIMSS is the creation of models of the stress and metal reduction pathways of environmental microbes. Ultimately, we wish to analyze PCAP data in such a way as to automatically generate hypothetical models of cellular pathways, which will be validated by comparison to experimental observations.

We are developing a modular LIMS system to store data and metadata from the high-throughput experiments undertaken by the other PCAP subprojects. Each module of the LIMS corresponds to a step in the experimental pipeline. Modules for tracking bioinformatic data, tagged constructs, biomass production, tagless purification, and single particle EM data have been deployed. We are also developing WIST (Workflow Information Storage Toolkit), a set of libraries and tools for rapid
LIMS development. WIST allows LIMS programmers to design multi-step workflows using modular core components, which can be added and arranged through a simple, intuitive configuration and template mechanism.

We have also prioritized proteins for tagging, TAP, and study by electron microscopy based on analysis of gene expression data from the VIMSS Environmental Stress Pathway Project (ESPP) and bioinformatic predictions. To date, we have identified 403 *D. vulgaris* proteins as high-priority targets for tagging by the PCAP Microbiology Core. 127 of these proteins were chosen based on biological relevance (e.g., involvement in sulfate reduction pathways) and analysis of ESPP data (e.g., proteins for which the expression level is frequently observed to change in response to different stresses). In addition, *D. vulgaris* orthologs of *E. coli* proteins that were annotated as part of heteromeric or large (>250kD) homomeric complexes in the EcoCyc database or Butland TAP data were selected as targets. This was done in order to study the degree to which inter-protein interactions are conserved between orthologs, and to establish a baseline characterization of potential complexes to compare with the same proteins under stress conditions. 320 of the 403 genes are at or near the ends of their operons, and thus feasible to clone using our high throughput methods (further details are provided on Swapnil Chhabra’s poster).

**Gaggle: A Framework for Database Integration and Software Interoperability**

Christopher Bare, Paul Shannon, Michael Johnson, and Nitin S. Baliga* (nbaliga@systemsbiology.org)

Institute for Systems Biology, Seattle, Washington

Project Goals: Molecular Assemblies Genes and Genomes Integrated efficiently: Characterize conserved protein complexes from a systems perspective.

MAGGIE Component 3

Collaborations for Gaggle implementation within MAGGIE

Co-PIs: Nitin S. Baliga; Mike Adams; Steven M Yannone; Gary Siuzdak; John A. Tainer; Stephen R Holbrook

Institute for Systems Biology, Lawrence Berkeley National Laboratory (LBNL), The Scripps Research Institute, The Burnham Institute, University of Georgia (UGA), and University of California Berkeley (UCB)

A crucial challenge in systems biology is to combine the capabilities of diverse software tools and data resources to create an environment that promotes data exploration and analysis by a wide spectrum of users. A solution to this problem should recognize that data types, formats and software in this high throughput age of biology are constantly changing.

Gaggle is a simple, open-source Java software environment that helps to solve this problem of software and database integration. Guided by the classic software engineering strategy of separation of concerns and a policy of semantic flexibility, it integrates any new and existing popular programs and
web resources into a user-friendly, easily-extended environment by enabling sharing of four simple
data types (names, matrices, networks, and associative arrays). Gaggle uses Java RMI and Java Web
Start technologies and can be accessed at http://gaggle.systemsbiology.net. Gaggle is being routinely
used as a mechanism for data sharing among the various components of the DOE-funded multi-
institutional MAGGIE project.

119

Sensitivity Analysis on MS2 Viral Dynamics Using Interval
Mathematics

Ozlem Yilmaz,* Luke E. K. Achenie (achenie@engr.uconn.edu), and Ranjan Srivastava

University of Connecticut, Storrs, Connecticut

Project Goals: Parameter determination is the most critical step in developing a model. The para-
metric errors introduced and the required precision for the parameters play an important role.
The objective is to achieve parameter estimation using interval methods. This study represents
the first step in accomplishing this goal.

Validated solutions of initial value problems (IVPs) can be obtained using interval analysis. A
significant advantage over standard numerical methods is that an enclosure of the true solution is
obtained. Since there are often measurement errors in an experiment, parameters determined based
on the experimental data are also prone to errors. In this paper we employ a MATLAB version of
the interval ODE solver VNODE (Nedialkov, 1999) to identify the most sensitive parameters in a
biological model. The model under study explains how lytic RNA phage infects Escherichia coli C-
3000 and the viral dynamics between the phage and its host at the intercellular level. Experimental
data consisted of uninfected cell density (sensitive and resistant type), infected cell density, free phage
density and substrate (glucose) concentration. There were 9 parameters determined experimentally
and 6 parameters estimated using regression analysis (Jain et al., 2006). In our preliminary studies,
each parameter was defined over an interval. Among all, the parameter corresponding to the rate
of infection was found to be the most sensitive one. In the above studies, we note that the interval
Hermite-Obreschkoff (IHO) method converges better than the interval Taylor series (ITS). The
IHO unfortunately has high computational overhead and has convergence problems when the inter-
vals are large. To alleviate these problems we have been investigating hybrid approaches that combine
constraint satisfaction (Granvilliers et al., 2004 and Janssen et al., 2002) with IHO.

References

University of Toronto, 1999.
2. Jain R., Knorr A.L., Bernacki J. and R. Srivastava, Investigation of Bacteriophage MS2 Viral Dynamics
Using Model Discrimination Analysis and the Implications for Phage Therapy, Biot. Prog., 22:6,
1650-1658.

* Presenting author
The BioWarehouse System for Integration of Bioinformatics Databases

Tom Lee, Valerie Wagner, Yannick Pouliot, and Peter D. Karp* (pkarp@ai.sri.com)
Bioinformatics Research Group, SRI International, Menlo Park, California

Project Goals: The goal of the BioWarehouse project is to allow scientists to integrate collections of DBs relevant to a genomics or systems-biology problem. BioWarehouse can integrate multiple public bioinformatics DBs into a common relational DB management system, facilitating a variety of DB integration tasks including comparative analysis, data mining, storage of locally generated data, and dissemination of data to the scientific community. All data are loaded into a common schema to permit querying within a unified representation.

BioWarehouse [1,2] 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, data mining, storage of locally generated data, and dissemination of data to the scientific community. All data are loaded into a common schema to permit querying within a unified representation.

BioWarehouse currently supports the integration of the following databases: BioCyc, BioPAX protein interactions datasets (such as BIND), CMR, Eco2Dbase, ENZYME, Genbank (microbial subset), Gene Ontology, KEGG, MAGE-ML gene expression datasets, NCBI Taxonomy, and UniProt. 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, gene expression data, protein interactions data, protein expression data, 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 [3]. 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


* Presenting author
A Cell Centered Database for Microbial Cells

Maryann E. Martone,1,*, Joy Sargis,1 Andrew McDonnell,2 Gerry McDermott,2 Carolyn Larabell,2 Mark Le Gros,2 Joshua Tran,1 Willy Wong,1 Vincent Ye,1 Harley McAdams1 (hmcadams@stanford.edu), and Mark H. Ellisman1

1Center for Research in Biological Systems and the National Center for Microscopy and Imaging Research, University of California, San Diego, California; 2The National Center for X-ray Tomography, Lawrence Berkeley National Laboratory, Berkeley, California; 3Developmental Biology Department, Stanford University, Palo Alto, California

Project Goals: Development of high-throughput methods to identify and characterize spatially localized multiprotein complexes in bacterial cells.

The overarching objective of the Dynamic spatial organization of multi-protein complexes controlling microbial polar organization, chromosome replication, and cytokinesis project is to identify and characterize the regulatory proteins and protein/DNA complexes that control development of the bacterial cell and to determine the cellular locations where these molecules and complexes perform their function. This type of multi-disciplinary project employs a diverse array of methodologies and generates large amounts of diverse data the usefulness of which increases if easily shared between the collaborative sites. Moreover, forming a unified understanding of microbial cell biology requires integration of data from different experimental techniques into a common frame of reference. Without an informatics framework in which to deposit and search data, the process of building comprehensive structural maps of a bacterial cell type among collaborative groups becomes difficult. In a component of this Genomics:GTL project, we are establishing a distributed Cell Centered Database and associated informatics infrastructure to house the 3D tomographic data generated using electron and X-ray tomography, and correlated light microscopies. Databases have become integral parts of data management and data dissemination in biology. As the output of the microscopy community continues to increase, the utilization of databases for standard EM practice and large scale data sharing is becoming more critical. The Cell Centered Database (CCDB) was created to address the need for additional databases for cell level light and electron microscopic data, particularly geared towards large 3D datasets such as those produced by electron tomography. Adaptation of the CCDB to serve as a framework for microbial cell-centered projects is a natural extension of the CCDB’s originally intended purpose.

The CCDB encompasses and hosts many types of data in the dimensional range that lies between gross morphology and macromolecular structure – the so-called “mesoscale”. Thus, unlike many structural databases on the web, the CCDB takes a “cell centered” approach in that it involves imaging of the cell by multiple techniques. These data and the software infrastructure that hosts and serves them are unique in scope, and provide the community with data that may have been technically difficult to obtain but is very rich in information content. This is particularly valuable resource for further examination, data mining and for use in development of computational models of structures and physiological processes that occur in cells and tissues. From its beginnings in 1998, the CCDB was envisioned as a grid and/or web-based federated database system. The CCDB pioneered
the production of a distributed, connectable repository system for managing and sharing data for a
growing research community of microscopists.

During the past year, the CCDB completed a comprehensive set of data input forms, available
through a secure grid-portal which can be launched from any web-browser. The portal architecture
builds upon the Telescience framework (http://telescience.ucsd.edu), a grid-based portal architecture
that simplifies and abstracts away the complexity of coordinating distributed resources, data, applica-
tions, and collaborations. Through the portal, users may upload, search and display their imaging
data. Users manage their own groups and permissions, allowing them to share data with selected
colleagues. Programmatic interfaces were created for microscopes so that data can be automatically
acquired from the electron microscope and deposited into the CCDB. Similar interfaces are under
construction for the X-ray microscope at LBNL and the laser-scanning multiphoton light micro-
scopes at UCSD/NCMIR.

Although the original CCDB was created around imaging of eukaryotic cells, the CCDB models
the process of reconstruction from a set of micrographs or images, including analysis and segmenta-
tion. Thus, although some of the tables were specific to eukaryotic organisms, most of the CCDB is
generic for 2D, 3D and 4D microscopic imaging and was readily adapted for prokaryotes. Over the
past year, we have been working with our collaborators on this Genomics:GTL project to modify
tables such that they capture essential specimen preparation details for working with microbial cells.
In addition, we have established a complete version of the CCDB at our partner site at Lawrence
Berkeley National Laboratory. Tomography data generated from UCSD on Caulobacter are now
housed in the CCDB. A newly constructed soft X-ray microscope at this site will be used to collect
tomographic data of intact Caulobacter cells. The data model of the CCDB was designed to be exten-
sible, to allow the incorporation of data from new microscopic imaging technologies such as soft
x-ray tomography, by extending the CCDB schema.

The CCDB provides the backbone of a data management system for GTL microbial projects.
Groups can either establish their own CCDB on site, or utilize the web-based portal version of
CCDB hosted by UCSD. The ultimate goal is to allow the construction of representations of cells
from data about their molecular constituents using a framework obtained from whole cell electron
tomography, soft X-ray microscopy and high-resolution light microscopy. Future work will involve
the development of spatial coordinate systems and ontologies to situate molecular constituents within
their cellular contexts in order to bridge different scales of resolution and microscopy techniques.

122 GTL

Developments in the Systems Biology Workbench

Frank Bergmann,2* Anastasia Deckard,1 and Herbert M. Sauro1 (hsauro@u.washington.edu)

1Department of Bioengineering, University of Washington, Seattle, Washington and 2Keck Graduate
Institute, Claremont, California

Project Goals: To develop a modular software framework (Systems Biology Workbench - SBW)
for integrating the diverse software applications developed by the systems biology community.

In this abstract we describe some of the accomplishments achieved in 2006, both with respect to
software and modeling. We have continued to improve the portability of the software across the
three main platforms, Windows, Linux and Mac OS, to develop new modules and pursue integra-

* Presenting author
tion with other third-party tools (Such as Copasi and Oscill8). In addition we have developed a new Wiki page where all documentation will reside (http://sbw.kgi.edu/sbwWiki/). To ensure that our software efforts are useful to the community we have also collaborated with three experimental groups to develop realistic biological models (Stem cell model, P53 model and MAPK model). The results of all three projects have been published.

**Progress in Software Provision**

We have made numerous minor improvements and some major changes in our software provision. Some of these developments are described in four papers we published this year. Here we only describe some of the major developments.

**JDesigner:** JDesigner (Visual editing tool) has undergone many improvements as a result of user feedback. The interface has been redesigned to include better visualization, better support for SBML, autolayout of networks and provision for storing multiple parameter sets within a given model.

**Jarnac:** Jarnac is our scripting tool for developing models. We have ported a 'lite' version of Jarnac to Linux and Mac OS, called JarnacLite, which permits users to describe a pathway very rapidly in text form, it can then be submitted to other SBW enabled editors such as JDesigner, Copasi or cellDesigner, where they can be refined.

**3D Tool:** The most exciting project we are working on is the 3D tool which allows network models to be visualized as 2D planes from which rise translucent pillars representing the concentrations of species. The visual representation can be fed in realtime from a SBW compatible simulator or data file so that the pillars will rise and fall as the model evolves. This gives a very interesting and different perspective on the model. In collaboration with Virginia Tech (VT), we are now able to visualize pathway simulations in real-time on the GigaPixel displays at VT. GigaPixel displays are physically large, high resolutions displays (>50 screens) that enable large quantities of data to be displayed simultaneously.

**Composition Standard:** We have published a proposal for a human readable format for building large models from small submodels. Current standards (SBML) can only describe a single model at a time. With the development of large models and in particular the rise of synthetic biology, there is a growing need to develop a compositional framework analogous to similar initiatives in the electronics field where standards such as Verilog have greatly stimulated the ability to share components and circuits (Details can be found at our Wiki: http://sbw.kgi.edu/sbwWiki/)

**User Base:** Our software continues to show increased usage, now running at around 600 downloads per month, with a total of 6500 downloads since the last GTL meeting. A recent review in Nature Biotechnology (2006) highlighted our GTL software as the best systems biology software currently available. On a number of occasions in 2006 we were ranked the number one bioinformatics project on source forge (out of ~700 projects).
The Ribosomal Database Project II: Introducing *myRDP* Space and Quality-Controlled Public Data


Center for Microbial Ecology, Michigan State University, East Lansing, Michigan

**Project Goals:** The Ribosomal Database Project II (RDP) offers aligned and annotated rRNA sequence data and analysis service to the research community. These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

Through its website (http://rdp.cme.msu.edu), The Ribosomal Database Project II (RDP) offers aligned and annotated rRNA sequence data and analysis service to the research community (Cole et al., 2006. Nucleic Acids Research; doi: 10.1093/nar/gkl889). These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, and bioremediation.

Updated monthly, the RDP maintains 286,257 aligned and annotated quality-controlled rRNA sequences as of November 2006 (Release 9.45; Fig. 1). As a major quality improvement, all sequences are now tested for sequence anomalies, including chimeras, using Pintail from the Cardiff Bioinformatics Toolkit (Ashelford et al., 2005. Appl. Environ. Microbiol. 71:7724-7736).

**myRDP Space:** This new feature allows users to maintain their own private sequence collection on the RDP servers aligned in sync with the RDP public alignment. With a *myRDP* account, researchers upload private rRNA sequences in Sequence Groups, which can range from a single sequence to thousands of sequences. Any combination of *myRDP* sequences and RDP public sequences can be selected for download or further analysis with the RDP tool suite. Sequences can be downloaded in formats ready for input to a wide variety of third-party phylogenetic and ecological tools.

After upload, *myRDP* sequences are automatically placed into the bacterial taxonomy using the RDP Classifier and aligned to match the RDP public alignment using RNACAD. Since the alignment remains in sync with the RDP public alignment, there is no need for the alignment compromises necessary to maintain compatibility between physically separated alignments. The RDP Classifier is accurate on sequences as small as 100 bases, although the information content varies along the molecule (Fig. 2).

**myRDP Pipeline:** The new *myRDP* release incorporates a high-throughput sequence processing pipeline tailored to the requirements of single-read environmental sequence projects. The *myRDP* Pipeline consists of an integrated suite of publicly available and in-house developed programs. It provides the researcher with a simple path from sequencer output to quality-controlled, aligned sequences and analysis.

**Video Tutorials:** New short video tutorials demonstrate some of the more complex analytical tasks, including use of the new *myRDP* Pipeline. These tutorials average three minutes in length. They capture the screen as the tasks are performed, while the narrator explains the tasks and the choices available to the user.

* Presenting author
Figure 1. Increase in number of publicly available bacterial small-subunit rRNA sequences. Suspect quality sequences were flagged as anomalous by Pintail in testing with two or more reference sequences from different publications.

Figure 2. Classification accuracy rate for 16S rRNA sequence fragments of 100 bases. The gray bars on the x axis define the hypervariable regions. The V2 and V4 regions may make attractive targets for in-depth taxonomic analysis of environmental samples by the new short-read sequencing technologies.
Milestone 3

An Integrated Knowledge Resource for the *Shewanella* Federation

Nagiza F. Samatova1,* (samatovan@ornl.gov), Denise Schmoyer,1 Tatiana Karpinets,1 Guruprasad Kora,1 Sergey Passovets,1 Michael Leuze,1 and Ed Uberbacher1 (ube@ornl.gov)

Collaborators from Shewanella Federation: Timothy S. Gardner2, Gyorgy Babnigg3, Carol S. Giometti4, Margrethe Serres4, Anna Obraztsova5, Grigoriy E. Pinchuk6, Alexander Beliaev6, Margaret F. Romine6, Kenneth Nealson5, and James K. Fredrickson6

1Oak Ridge National Laboratory, Oak Ridge, Tennessee; 2Boston University, Boston, Massachusetts; 3Argonne National Laboratory, Argonne, Illinois; 4Marine Biology Laboratory, Woods Hole, Massachusetts; 5University of Southern California, Los Angeles, California; and 6Pacific Northwest National Laboratory, Richland, Washington

**Project Goals:** This project is a component of the *Shewanella* Federation and as such contributes to the overall goal of applying the tools of genomics, leveraging the availability of genome sequence for 18 additional strains of *Shewanella*, to better understand the ecophysiology and speciation of respiratory-verse members of this important genus. To understand these systems the SF is using genome-based approaches to investigate *Shewanella* as a system of integrated networks; first describing key cellular subsystems — those involved in signal transduction, regulation, and metabolism — then building towards understanding the function of whole cells and, eventually, cells within populations. As a general approach, the SF is collectively employing complimentary “top-down” bioinformatics-based genome functional predictions, high-throughput expression analyses, and functional genomics approaches to uncover key genes as well as metabolic and regulatory networks. The “bottom-up” component employs more traditional approaches including genetics, physiology and biochemistry to test or verify predictions. This information will ultimately be linked to analyses of signal transduction and transcriptional regulatory systems and used to develop a linked model that will contribute to understanding the ecophysiology of *Shewanella* in redox stratified environments.

Critical to the success of the *Shewanella* Federation (SF) project and, arguably, to the overall success of the Genomics: GTL program is the sharing and integration of various types of information and data. The *Shewanella* Knowledge Base (http://www.shewanella-knowledgebase.org/) is a data and knowledge integration environment that allows *Shewanella* investigators (1) to capture, integrate and retrieve diverse ‘omics’ data for systems biology studies; (2) to navigate and superimpose information across gene-, protein-, expression- and pathway-levels; and (3) to perform com-
parative visual analyses in a cell systems context. The ultimate goal is to facilitate the generation of new hypotheses and knowledge about *Shewanella* systems behavior, while minimizing the researchers’ effort, time and complexity.

The *Shewanella* Knowledge Base takes advantage of existing databases, resources and tools via direct linkages to avoid duplication of efforts occurring elsewhere. Its open architecture allows anyone interested to contribute and access information and data available for *Shewanella* species.

During the first year of the project the following progress has been made. We addressed the major problems associated with (a) data synthesis into a common data model from a distributed group of investigators, (b) data standards, and (c) experimental protocols representation. The system currently provides data models (compliant with community accepted data standards if available) for representing metabolic pathways, genome- and gene-level information, and experiment metadata. Initial schema design from the MIT ExperiBase project was extended to formally describe experimental campaigns, protocols, and computational analysis results for the following data types: mutants, cell culture, transcriptomics, and proteomics.

The project has also developed interfaces for datasets selection for comparative visual analysis of gene and protein expression profiles superimposed with publicly available biological information. For example, the investigator might be interested in identifying genes and subsystems (networks) required for anaerobic respiration of *S. oneidensis* MR-1 with various electron acceptors. The figures depict example questions the investigators may exploit, specifically:

1) What experimental data is available for FedEx1 campaign dealing with aerobic and anaerobic growth of *S. oneidensis* MR-1 under steady state conditions? (Fig. A)

2) How do proteomics and transcriptomics expression compare across gene clusters of interest? (Fig. B)

3) How do operon predictions by BioCyc match the expression predictions, or what proteins in the cluster are missed in proteomics experiments? (Fig. C)

4) How does the activity of enzymes involved in anaerobic and aerobic respiration change at the pathway level during the transition of *S. oneidensis* from aerobic to anaerobic growth? (Fig. D)

5) Are cytochrome mutants of interest available and what phenotypes do they have? (see web-site)

In the outgoing years the data model will be extended to support other types of SF-related data, various visual interfaces and analysis tools for comparative “omics” studies in the systems biology context. Capabilities for automatics data uploads and data access controls will be provided.

Figure B. Consistent changes in the gene and protein abundance of some respiration related enzymes under the same growth transition. Several mono- and diheme cytochromes and enzymes involved in their biogenesis show down-regulation in anaerobic conditions versus aerobic ones. Several multi-heme cytochromes and related enzymes show consistent up-regulation.
Figure C. Consistency of expression for Ni-Fe hydrogenaze genes with the operon predictions by BioCyc. The predicted hypAED operon shows a consistent up-regulation both at gene and protein levels, while the other operon shows an overall up-regulation, but the hypF gene shows a significant down-regulation, also certain genes in the operon have missing protein expression values.

Figure D. The fragments of “anaerobic respiration – electron donors reaction list” (top) and “aerobic respiration – electron donor II” (bottom) pathways in ShewCyc from SRI with gene expression during the transition of *S. oneidensis* from aerobic to anaerobic growth.

This research sponsored by the U.S. DOE-BER, Genomics:GTL Program. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.
Informatics Strategies for Large-Scale Novel Cross-linker Analysis

Gordon A. Anderson* (Gordon@pnl.gov), Nikola Tolic, Xiaoting Tang, and James E. Bruce


Project Goals: This project is focused on the development of novel cross-linker molecules to allow detection of protein-protein interactions. This work is done in collaboration between Washington State University (WSU) and the Pacific Northwest National Laboratory (PNNL). The researchers at WSU are developing the cross-linker molecules and methodologies while the PNNL researchers are focusing on the development of informatics tools to enable data analysis and identification.

The analysis of protein interactions in biological systems represents a significant challenge for today’s technology. Chemical cross-linking provides the potential to impart new chemical bonds in a complex system that result in mass changes in the analysis of system tryptic peptides. However, system complexity and cross-linker product heterogeneity have precluded wide-spread chemical cross-linking use for large-scale protein interaction identification. The development of mass spectrometry identifiable cross-linkers called Protein Interaction Reporters (PIRs) has enabled on-cell chemical cross-linking experiments with product type differentiation. However, the complex datasets resultant from PIR experiments demands new informatics capabilities to allow interpretation. This presentation details our efforts to develop such capabilities and describes the program X-links which allows PIR product type differentiation. Furthermore, we also present the results from Monte Carlo simulation of PIR-type experiments to provide false positive identification rate estimates for the PIR product type identification through observed precursor and released peptide masses. Our simulations also provide false positive estimations based on accurate peptide mass measurements and database complexity. Overall, the calculations show a low rate of false positive identification of PIR product types due to random mass matching at under 8% at 10 ppm mass measurement accuracy. In addition this presentation illustrates the effect of database complexity on the PIR strategies ability to uniquely identify peptides using the constraints from this methodology. The PIR strategy includes a concept of development of a constrained protein database that increases the ability to uniquely identify cross-linked peptides and thus proteins. This presentation illustrates this methodology and quantifies the effect on unique peptide identification.

This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-04ER63924.

* Presenting author