DOE/SC-0140

Section:
Knowledgebase and Computing for Systems Biology

U.S. DEPARTMENT OF
ENERGY
Office of Science

Joint Meeting 2011

Genomic Science Awardee Meeting IX

and

USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Meeting

[Revised: April 14, 2011]

Crystal City, Virginia
April 10-13, 2011

Prepared for the
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research
Germantown, MD 20874-1290

http://genomicscience.energy.gov

Prepared by
Biological and Environmental Research Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830
Managed by UT-Battelle, LLC
For the U.S. Department of Energy
Under contract DE-AC05-00OR22725

Knowledgebase and Computing for Systems Biology

248

Web API for Annotation, Access, and Analysis of Genomic Data in the SEED Framework

Ross Overbeek,[1]* Robert Olson,[2,3] Terrence Disz,[2,3] Bruce Parello,[1] Rob Edwards,[4] **Christopher Henry**,[2,3] Scott Devoid,[3] and **Rick Stevens**[2,3]

[1]Fellowship for Interpretation of Genomes, Ill.; [2]University of Chicago, Ill.; [3]Argonne National Laboratory, Ill.; and [4]San Diego State University, Calif.

Project Goals: This project aims to develop software and data infrastructure to support programmatic access to the SEED and Model SEED resources for the annotation of genomes, integration of omics data, and reconstruction of genome-scale metabolic models. Within this overarching objective, are four specific aims: (i) enhancing the computational infrastructure behind the SEED framework to improve extensibility, accessibility, and scalability; (ii) integrate extensions into the SEED data-model and software to accommodate new data types including regulatory networks, genome-scale metabolic models, structured assertions, eukaryotic genome data, and growth phenotype data; (iii) develop an application programming interface (API) to provide remote access to the SEED database and tools; (iv) apply SEED infrastructure, data, and APIs to annotate genomes and construct genome-scale metabolic models for organisms with applications to bioenergy, carbon cycle, and bioremediation.

The SEED environment for the integration, annotation, and comparison of genomic data now includes thousands of microbial genomes and many eukaryotic genomes, all linked with a constantly updated set of curated annotations embodied in a large and growing collection of encoded subsystems and derived protein families. Additionally, the Model SEED resource has been developed to translate SEED annotations into functioning genome-scale metabolic models. Recently we have developed a set of web-services for SEED that offer programmatic access to data and tools included within the SEED environment (find documentation at http://blog.theseed.org/servers/). The services include the ability to remotely submit genomes to RAST and Model SEED for annotation and modeling; they enable users to query the SEED database for genome features, functional annotations, gene orthologs, and sequence similarities; and they enable users to apply flux balance analysis with genome-scale metabolic models to simulate cell growth in a variety of media conditions and with a variety of mutations. We highlight the powerful features of the SEED web services by demonstrating how multiple functions may be combined together to answer important questions in biology. We apply the web services for accessing genomes, sequence similarities, metabolic neighborhood, subsystems, and metabolic models to identify genes that may be associated with gap filled reactions in metabolic models. Next, we combine genome feature queries, subsystems queries, expression data queries, and annotation queries to identify commonly clustered and co-expressed sets of functional roles across all known genomes. Finally, we combine flux variability analysis, gene knockout studies, essentiality data queries, and annotation queries to study redundancy of essential metabolic functions across all known genomes.

249

Enabling a Systems Biology Knowledgebase with Gaggle and Firegoose

J. Christopher Bare[1]* (cbare@systemsbiology.org), Karen Foley,[1] Tie Koide,[2] David J. Reiss,[1] Paul T. Shannon,[1] Dan Tenenbaum,[1] Steven M. Yannone,[3] Sung Ho Yoon,[1] Wei-Ju Wu,[1] and **Nitin S. Baliga**[1] (nbaliga@systemsbiology.org)

[1]Institute for Systems Biology, Seattle, Wash.; [2]Universidade de São Paulo, Ribeirão Preto, SP, Brazil; and [3]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://gaggle.systemsbiology.net
http://baliga.systemsbiology.net/enigma/

Project Goals: This project will build on previous iterations of the Gaggle framework to provide a means of interoperability between bioinformatics software and data sources that enables independently developed tools to be used together for exploratory analysis of high-throughput data. The framework will provide interoperability that is simple to implement and extensible to accommodate unforeseen applications.

Systems biology is made possible by technological advances in instrumentation on which thousands of measurements can be made simultaneously. Parallel advances in computing are critical to interpreting high-throughput biological data and developing an understanding of the fundamental molecular mechanisms that define the cell's regulatory circuitry.

In the Baliga lab at the Institute for Systems Biology we are reverse engineering biological circuits to understand how cells adapt to changing environments. Our long term vision is to build predictive mathematical models which can be applied to bioenergy, and bioremediation. This requires measuring and analyzing several different types of biological data: DNA sequence, transcription, protein abundance and protein-DNA binding.

* Presenting author

To this end, we have developed software tools for exploratory analysis of systems biological data and for inference of the regulatory networks that control gene expression. Network inference algorithms, cMonkey and Inferelator, discover coregulated modules of cellular functionality and their regulators based on gene expression and de novo regulatory motif detection.

The Gaggle framework and Firegoose provide interoperability between data sources, web applications and desktop software for exploration, analysis and visualization of the resulting large biological networks in the context of underlying raw data. The database integration and software interoperability enabled by Gaggle and Firegoose is going to be a key aspect of a Systems Biology Knowledgebase that takes into account the unanticipated advances in technologies for measuring new kinds of data.

Funding: This work conducted by ENIGMA was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231. Prior support provided by U.S. Department of Energy, Award No. DE-FG02-07ER64327 and DG-FG02-08ER64685.

250

The PhyloFacts Microbial Phylogenomic Encyclopedia: Phylogenomic Tools and Web Resources for the Systems Biology Knowledgebase

Kimmen Sjölander* (kimmen@berkeley.edu)

University of California, Berkeley

Project Goals: PhyloFacts is designed to improve the accuracy of functional annotation of microbial genomes through evolutionary reconstruction of gene families, integrating information from protein 3D structure, biological process, pathway association, protein-protein interaction and other types of experimental data to improve both the specificity and coverage of protein "function" prediction. Key computational challenges that we will address in this project include the development of a system for whole genome functional annotation and simultaneous taxonomic and functional annotation of metagenome datasets using HMMs (hidden Markov models) placed at internal nodes of gene family trees. A key component of our system is a novel phylogenomic approach to ortholog identification: Berkeley PHOG. Two genes are orthologous if they are descended from an ancestral gene by speciation. Orthologous genes are generally assumed to share a common function, while the functions of paralogous genes (related by duplication) can be anticipated to have diverged from that of the common ancestor.

Gene families evolve novel functions through diverse evolutionary processes, including point mutations, gene duplication, and changes in domain architecture resulting from gene fusion and fission events. While some of these evolutionary events produce relatively small changes to function (e.g., mutations at positions that are distant from an enzyme's active site), others (especially duplication events and domain architecture changes) can result in dramatic shifts in the molecular function and biological process. As a result, homology-based functional annotation, i.e., transferring the annotation of the top BLAST hit, has been shown to be associated with significant errors: upwards of 20-25% of sequences have been estimated to have errors in their functional annotations.

The combined use of evolutionary reconstruction and structural analyses helps avoid the errors in sequence functional annotation that are now known to be rampant. We take this approach, termed *structural phylogenomics* (Sjölander, "Getting started in Structural Phylogenomics", *PLoS Computational Biology*, 2010) in constructing the PhyloFacts Microbial Phylogenomic Encyclopedia (at http://phylogenomics.berkeley.edu/phylofacts/). PhyloFacts is designed to improve the accuracy of functional annotation of microbial genomes through evolutionary reconstruction of gene families, integrating information from protein 3D structure, biological process, pathway association, protein-protein interaction and other types of experimental data to improve both the specificity and coverage of protein "function" prediction.

Key computational challenges that we will address in this project include the development of a system for whole genome functional annotation and simultaneous taxonomic and functional annotation of metagenome datasets using HMMs (hidden Markov models) placed at internal nodes of gene family trees.

A key component of our system is a novel phylogenomic approach to ortholog identification: Berkeley PHOG (Datta *et al*, "Berkeley PHOG: PhyloFacts Orthology Group Prediction Web Server" *Nucleic Acids Research* 2009). Orthology is a phylogenetic term: two genes are orthologous if they are descended from an ancestral gene by speciation. Orthologous genes are generally assumed to share a common function, while the functions of paralogous genes (related by duplication) can be anticipated to have diverged from that of the common ancestor. We identify orthologs based on reconstructed evolutionary histories for proteins clustered on the basis of sharing a common domain architecture using the FlowerPower algorithm (Krishnamurthy et al, "FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function", *BMC Evolutionary Biology* 2007), and also for individual PFAM domains (in which cases proteins may have different overall domain architectures). The PHOG algorithm can be tuned to user-specified taxonomic distances, allowing highly precise predictions of orthology.

We are in the process of expanding PhyloFacts coverage of key microbial genomes, with extensions to the pipeline described in (Krishnamurthy *et al*, "PhyloFacts: An online structural phylogenomic encyclopedia for protein functional and structural classification", *Genome Biology* 2006) and will provide different ways of accessing the data for online (web-based) analysis and to download the core data for use at external sites.

Grant Number: De-Sc0004916

251

Integrated Approach to Modeling Transcriptional Regulatory Networks

Pavel Novichkov[1]* (PSNovichkov@lbl.gov), Dmitry Rodionov,[2] Marcin Joachimiak,[1] David J. Reiss,[3] **Adam P. Arkin**,[1] **Inna Dubchak**,[1] and **Nitin Baliga**[3]

[1]Lawrence Berkeley National Laboratory Berkeley Calif.; [2]Burnham Institute for Medical Research La Jolla, Calif.; and [3]Institute for Systems Biology Seattle, Wash.

Project Goals: Integrate computational approaches for reconstruction of large-scale transcriptional regulatory networks and building predictive models in microbial genomes.

One of the major challenges facing the bioinformatics community in view of advances in DNA sequencing technology and constantly growing number of complete genomes is providing effective tools to enable high-quality reconstruction of transcriptional regulatory networks (TRN) and building a predictive models for transcriptional control of physiology in microbial organisms. This challenging problem can be addressed using two complementary approaches: i) a data-driven systems approach that relies on the integration of diverse data including high-throughput gene expression and protein-protein interactions data; ii) a genomic-driven approach based on the simultaneous analysis of several closely related genomes to decipher evolutionary conserved regulatory signals. Recently we were able to demonstrate that both approaches allow us to get deep insight into transcriptional regulation of microbial organisms and build large-scale TRNs and high-quality predictive models.

Using the data-driven systems approach we have constructed a large-scale predictive model for transcriptional control of physiology in *H. salinarum NRC-1*, which represents a class of poorly studied organisms (Archaea). Using relative changes in 72 transcription factors (TFs) and 9 environmental factors (EFs) this model accurately predicts dynamic transcriptional responses of ~80% genes in 147 newly collected experiments representing completely novel genetic backgrounds and environments.

A key aspect of the data-driven systems approach is the reduction of data complexity and the automated formulation of hypothesis based on multiple lines of evidence. We have developed an associative biclustering method to identify a variety of types of coherent patterns in combined functional genomics data. The method searches for statistical associations and estimates confidence across multiple data types. Results of searches incorporating data on transcription, such as gene expression microarrays, provide direct information on putative regulons. The method is placed in a computational framework, which allows rapid customization and deployment for new data sets and data types.

Finally, using comparative genomics approach we were able to build detailed large-scale TRNs in groups of closely related genomes representing diverse bacterial taxa, including *Shewanella* (16 genomes, 82 TFs), *Desulfovibrionales* (10 genomes, 32 TFs), *Bacillales* (11 genomes, 105 TFs), *Enterobacteriales* (12 genomes, 58 TFs), etc. The typical large-scale TRN covers a representative set of metabolic pathways and biological processes. For instance, *Shewanella* TRN encodes regulation of metabolism of carbohydrates, nitrogen and aminoacids, fatty acids and nucleotides, cofactors and metals, stress response, etc. The analysis gene expression profiling data of TF mutants, which have been recently done for three transcription factors FUR, PerR and Rex in *Desulfovibrio desulfuricans G20*, shows nice agreement with regulons reconstructed by comparative genomics.

TRNs built using data-driven systems approach contains a wealth of knowledge about the putative regulatory interactions that can be further analyzed in details by comparative genomics and automatically propagated to closely related species. On the other hand, regulons reconstructed from genomics-driven approach provide detailed description of the regulon content, assigned transcription factor and TF binding sites in promoter regions of target genes, and thus creates a solid ground for building predictive models. Thus we believe that tight integration of both strategies is required to achieve significant improvement in both coverage and quality of transcriptional regulatory networks. The iterative procedure of TRN and predictive model refinement will be implemented by means of establishing the central repository of putative regulons, which will serve as a source and sink of putative regulons predicted by different approaches.

252

Large-Scale Genomic Reconstruction of Transcriptional Regulatory Networks in Bacteria

Pavel Novichkov[1]* (PSNovichkov@lbl.gov), Alexey Kazakov,[1] Semen Leyn,[2] Dmitry Ravcheyev,[2] Andrei Osterman,[2] Inna Dubchak,[1] Adam Arkin,[1] and **Dmitry Rodionov**[2] (rodionov@burnham.org)

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [2]Sanford-Burnham Medical Research Institute, La Jolla, Calif.

Project Goals: The major goals of this project are to: (1) develop integrated platform for genome-scale regulon reconstruction; (2) Infer regulatory annotations and reconstruct transcriptional networks in several groups of microbial species important for DOE mission; (3) Develop RegKnowledgebase on microbial transcriptional regulation.

Genome-scale annotation of regulatory features of genes and reconstruction of transcriptional regulatory networks in a variety of diverse microbes is one of the critical tasks of modern Genomics and Systems Biology. It constitutes an important challenge, a prerequisite for understanding molecular mechanisms of transcriptional regulation in prokaryotes, identifying regulatory circuits, and interconnecting them with each other and with various metabolic, signaling, and other cellular pathways. A growing number of complete prokaryotic genomes allows us to extensively use comparative genomic approaches to infer *cis*-acting regulatory elements in regulatory networks of numerous groups of bacteria.

We developed a computational genomic-based approach implemented in the RegPredict web-server (regpredict.lbl.gov) facilitating fast and accurate inference and analysis of microbial regulons controlled by either DNA-binding transcription factors (TFs) or RNA regulatory elements (e.g., riboswitches). A key new concept of RegPredict is a cluster of co-regulated orthologous operons (CRON) that allows prediction of TF-binding sites (TFBSs) simultaneously in a set of taxonomically related genomes in a semi-automated way and provides a user-friendly GUI to perform comparative analysis of multiple CRONs. Major directions for genomics-based regulon inference are: (i) *regulon reconstruction* for a known regulatory motif in a set of reference genomes from a particular taxonomic group of bacteria; (ii) *ab initio prediction* of novel regulons using several scenarios for the generation of starting gene sets; (iii) *conservative propagation* of reconstructed regulons to all other microbial genomes in the same taxonomic group. The results of regulon inference performed by any implemented workflow are amenable for immediate deposition in the RegPrecise database (regprecise.lbl.gov) (Fig. 1).

We applied the integrative comparative genomics approach to infer transcriptional regulatory networks (TRNs) in various taxonomic groups of bacteria. A limited input of established regulon members in model species, e.g. *E. coli*, *B. subtilis*, *S. aureus*, was extracted from literature and publically available resources, such as RegulonDB and DBTBS databases. The obtained reference set of inferred regulons is available in the RegPrecise database and includes 420 regulons described in 94 genomes from the 9 taxonomic groups including Enterobacteria, *Bacillus*, *Shewanella*, *Streptococcus*, *Staphylococcus*, *Ralstonia*, *Desulfovibrio*, *Thermotoga*, and Cyanobacteria (Table 1). The reconstructed regulons control the key pathways involved in central metabolism, production of energy and biomass, metal homeostasis, stress response and virulence. Many of the *de novo* inferred regulons were experimentally validated in *S. oneidensis* and *T. maritima* models. The taxon-specific regulon collections will be further expanded to produce draft TRNs for other taxonomic groups of Proteobacteria and Firmicutes. To cover all sequenced genomes (e.g. strains of the same species) within the analyzed taxonomic groups (~1000 genomes), we will use automated conservative propagation procedure in the context of RegPrecise.

Overall, this project enables a wide spectrum of capabilities required by DOE Systems Biology KnowledgeBase incuding: (i) detailed TRNs for DOE-mission genomes; (ii) knowledgebase of regulatory interactions in a large set of microbial genomes; (iii) regulatory constraints for building predictive models; (iv) framework for validation and extension of gene regulatory networks reconstructed from gene expression profiling data.

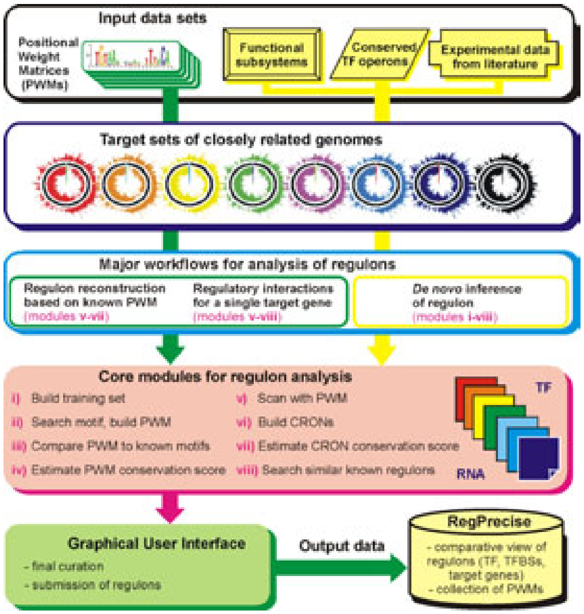
Input data sets
Positional Weight Matrices (PWMs)
Functional subsystems
Conserved TF operons
Experimental data from literature
Target sets of closely related genomes
Major workflows for analysis of regulons
Regulon reconstruction based on known PWM (modules v-vii)
Regulatory interactions for a single target gene (modules v-viii)
De novo inference of regulon (modules i-viii)
Core modules for regulon analysis
i) Build training set
ii) Search motif, build PWM
iii) Compare PWM to known motifs
iv) Estimate PWM conservation score
v) Scan with PWM
vi) Build CRONs
vii) Estimate CRON conservation score
viii) Search similar known regulons
TF
RNA
Graphical User Interface
- final curation
- submission of regulons
Output data
RegPrecise
- comparative view of regulons (TF, TFBSs, target genes)
- collection of PWMs

Figure 1. Regulon reconstruction workflow implemented in the RegPredict web-server and RegPrecise database. Two major components of this workflow are propagation of previously known regulons from model organisms to others, and *ab initio* prediction of novel regulons. RegPredict tool (http://regpredict.lbl.gov) applies the comparative genomic approach to the analysis of bacteria from a single taxonomic group of species. The inferred reference sets of regulons for various taxonomic groups of bacteria are collected and represented in the RegPrecise database (http://regpredict.lbl.gov).

Table 1. Taxonomic groups of bacteria targeted to genome-wide regulon reconstruction

Taxonomic group	Analyzed genomes	Inferred TF regulons	Example model species
Bacillaceae	11	105	*Bacillus subtilis*
Shewanella	16	82	*Shewanella oneidensis*
Enterobacteria	12	58	*Escherichia coli*
Staphylococcus	6	47	*Staphylococcus aureus*
Streptococcus	8	39	*Streptococcus pneumoniae*
Thermotogales	11	32	*Thermotoga maritima*
Desulfovibrionales	10	32	*Desulfovibrio vulgaris*
Ralstonia	6	15	*Ralstonia eutropha*
Cyanobacteria	14	10	*Synechococcus sp.* PCC7002
TOTAL	**94**	**420**	

253

Elucidation of the Transcriptional Response to Key Physiological Parameters in Bacteria: A Systems Biology Knowledgebase of Transcriptional Regulation

Stephen A. Federowicz* (sfederow@ucsd.edu), Young Seoub Park, Eric M. Knight, Donghyuk Kim, Byung-Kwan Cho, **Karsten Zengler**, and **Bernhard Ø. Palsson**

Department of Bioengineering, University of California – San Diego, La Jolla

http://systemsbiology.ucsd.edu

Project Goals: This project aims to: (1) create a fully curated, bottom up reconstruction of the transcriptional regulatory network in bacteria, using *Escherichia coli* as a model organism, (2) determine fundamental constraints on the regulatory response via network and sequence level features, (3) develop a non-Boolean constraints based modeling approach for regulation, (4) integrate the transcriptional regulatory network with metabolic and macromolecular synthesis models, and (5) provide a platform for genome scale metabolic engineering and synthetic design.

Project Description:
Global transcription factors represent one of the primary means by which an organism can sense and respond to stimuli. We experimentally determined the network level mechanisms of transcriptional response in *E. coli* by integrating genome wide binding analysis of five global transcription factors with associated expression profiling and transcription start site information. ChIP-chip experiments for ArcA, Fnr, FruR, Crp, Lrp, and ArgR revealed 143, 100, 45, 247, 144, and 63 unique binding regions, respectively. Binding peaks were mapped to gene expression data generated between a wild type strain and a gene deletion strain for each of the transcription factors to discern whether a binding event resulted in gene activation or repression. This allowed us to construct a functional network of transcriptional regulation that elucidates the response to major environmental stimuli. This includes the response to oxygen availability mediated by ArcA and Fnr, shifting carbon source mediated by FruR and Crp, and primary nitrogen source mediated by Lrp and ArgR. Subsequently, we were able to elucidate core connected feed back loops through which regulatory information flows and gene expression is systematically controlled. Specifically we found that arginine and leucine act as signaling molecules by shutting down their own transport and biosynthetic pathways in response to high levels of exogenous arginine or leucine. In contrast, tryptophan and tyrosine were shown to shut down only their own biosynthesis pathways, but not their own transport systems, suggesting that they function simply as nutrients to the cell. We thus propose the existence of specific network motifs containing multiple connected feedback loops that govern the stimulatory response in *E. coli.*

Furthermore, we integrated these binding regions with experimentally determined transcription start site (TSS) data using 5'-RACEseq in order to determine regulation of alternative transcripts and individual promoter architectures. First, we identified precise transcription factor binding locations via DNA sequence motifs. We were able to clearly identify previously well-determined binding motifs for each of the transcriptions factors and then used position weight matrices to rescan the entire genome. Aligning these sequence motifs to the TSS information generated a set of over 700 experimentally derived promoter architectures. Each of these contains one or more TF binding sites and a single TSS for each transcription unit. These were then clustered to reveal conserved patterns of promoter architecture and associated transcriptional regulation. Notably, we were first able to clearly recapitulate and expand upon the well-known promoter architectures for Crp. In depth analysis of other global transcription factors revealed a wealth of novel individual and shared distance preferences. Many of the most distinct preferences occurred directly over the TSS corresponding to direct repression in the case of ArcA, Lrp, and FruR. Others occurred around the -35bp region or further upstream around -60bp and were well correlated with transcriptional activation. Taken together these results suggest the presence of modular nucleoprotein complexes which confer transcriptional activation or repression.

Overall, using promoter level information in concert with network level motifs allows for the creation of fully curated, bottom up reconstructions of global transcriptional regulatory networks. These networks represent a curated systems biology knowledgebase of the transcriptional regulatory machinery within a cell and allow for physiologically meaningful computations of environmental perturbation.

254

Integrative Reconstruction of Carbohydrate Utilization Networks in Thermotogales

Dmitry Rodionov[1]* (rodionov@burnham.org), Vasiliy Portnoy,[2] Xiaoqing Li,[1] Irina Rodionova,[1] Dmitry Ravcheyev,[1] Olga Zagnitko,[3] Gordon Push,[3] **Yekaterina Tarasova,**[2] Kenneth Noll,[4] Robert Kelly,[5] Karsten Zengler,[2] **Andrei Osterman,**[1] and **Bernhard Palsson**[2]

[1]Sanford-Burnham Medical Research Institute, La Jolla, Calif.; [2]Department of Bioengineering, University of California San Diego, La Jolla, Calif.; [3]Fellowship for Interpretation of Genomes, Burr Ridge, Ill.; [4]Department of Molecular and Cell Biology, University of Connecticut, Storrs; and [5]Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh

Project Goals: The major goals of this project are to: (1) utilize bioinformatics and computational tools to enhance, refine and fill in the knowledge gaps in the current metabolic reconstruction of *T. maritima* with a particular emphasis on hydrogen production; (2) use comparative genomics to infer the key transcription factor

regulons for carbohydrate metabolism in *T. maritima* and other Thermotogales.

Bacteria of the deep-branched genus *Thermotoga* can produce hydrogen by fermenting a wide range of carbohydrates. A remarkable diversity of the Thermotogales *sugar diet* is matched by a large fraction of genes committed to carbohydrate degradation and utilization in their genomes. As a result the evolutionary plasticity of the sugar catabolic machinery in general, and due to a unique taxonomic position and lifestyle of Thermotogales, exact functions of many respective genes (and pathways) remained unclear even in the best-studied model species, *T. maritima*. To address this problem we applied an *integrative subsystems-based approach to the genomic reconstruction of metabolic and regulatory networks*. This approach established and validated in our previous studies (e.g. in the reconstruction of sugar catabolic machinery of the *Shewanella* genus) included three major levels of integration: (i) comparative analysis of 11 complete genomes from the Thermotogales order annotated by RAST server; (ii) parallel genomic reconstruction of biochemical transformations, uptake mechanisms and transcriptional regulation, and (iii) combining bioinformatic predictions with experimental testing in *T. maritima* model. Application of various bioinformatics tools implemented in the SEED (theseed.uchicago.edu) and RegPredict (regpredict.lbl.gov) web-sites allows us to substantially improve the accuracy of annotations as well as predict novel, previously uncharacterized genes and pathways. The developed detailed reconstruction integrates published and new results obtained by several research groups collaborating within the framework of the Biological Hydrogen Production Program (DOE DE-PS02-08ER08-12).

The *genomic encyclopedia of sugar utilization* in Thermotogales includes more than 300 functional roles (isofunctional protein families or *FIGfams*) spanning at least 20 distinct pathways with mosaic distribution across 11 analyzed species. The detailed results of this analysis are captured in the subsystem "*Sugar utilization in Thermotogales*" available online from the SEED web-site. The current version of the subsystem comprises >130 cytoplasmic and extracytoplasmic sugar catabolic enzymes (including ~40 glycoside hydrolases), ~90 components of carbohydrate uptake systems (mostly ABC transporters), and 18 committed transcription factors. The developed metabolic model incorporates all published experimental data and inferences about enzyme activities, substrate specificities of transporters, and differential gene expression patterns on various carbohydrates (generated mostly for *T. maritima*). Our analysis revealed substantial differences in sugar catabolic pathways between Thermotogales and other previously studied bacteria. Most common are *nonorthologous gene replacements*, when a functional role is encoded by a gene, which is not orthologous (and, often, nonhomologous) to any of the previously described genes of the same function. The repertoire of transporters and regulators involved in sugar catabolism in *Thermotoga* demonstrates the most prominent differences in comparison with other taxa. We also discovered two novel pathways for utilization of inositol and galacturonate in *T. maritima*. These pathways include previously unknown biochemical transformations driven by 4 and 6 enzymes, respectively. Both pathways as well as most of the individual recombinant purified enzymes were experimentally characterized by enzymatic assays and genetic complementation methods. A performed substrate specificity profiling of the *T. maritima sugar kinome*, 15 predicted carbohydrate kinases over a panel of >40 diagnostic sugar substrates, provided a strong experimental support of the reconstructed pathways. Finally, the predicted catabolic capabilities of *T. maritima* were assessed by monitoring growth rates, substrate consumption and gene expression on a panel of various individual and mixed mono- and disaccharides.

A transcriptional regulatory network inferred from comparative genomic analysis of Thermotogales includes 32 transcription factors and their DNA binding sites unevenly distributed across 11 studied genomes. A current collection of regulons captured in the RegPrecise database (regprecise.lbl.gov) is centered on *T. maritima* and includes 18 transcription factors that were predicted to control expression of ~185 genes involved in sugar catabolic machinery of this model organism. Remarkably, a large fraction of these genes and operons are controlled by multiple transcription factors pointing to a complexity of regulatory responses to changing environmental conditions (Fig. 1). For example, we established partial overlaps between the xylose, glucuronate, and galacturonate regulons (XylR, KdgR, and UxaR, respectively); the glucose, trehalose and inositol regulons (GluR, TreR, and IolR); and the cellobiose, mannose, and glucooligosaccharide regulons (CelR, ManR, and GloR). The experimental assessment of the reconstructed regulatory network included *in vitro* analysis of selected individual regulons and *in vivo* gene expression profiling of *T. maritima* on various carbohydrate substrates. We used the first approach based on gel-shift mobility assays to validate all predicted DNA targets and identify small molecule effectors for six regulators from the ROK family (BglR, IolR, XylR, ChiR, TreR, and ManR). We are currently expanding this effort to characterize additional transcriptional regulators, Rex, UxaR, KdgR, CelR, and RhaR. Global gene expression profiles were obtained and analyzed for the growth on 12 different carbon sources using high-density oligonucleotide tiling arrays (Nimblegen). Gene induction patterns measured for tested mono- and disaccharides (trehalose, rhamnose, xylose, etc.) showed a strong correlation and provided additional information to refine respective regulons (TreR, RhaR, XylR, etc.) reconstructed by the genomic analysis.

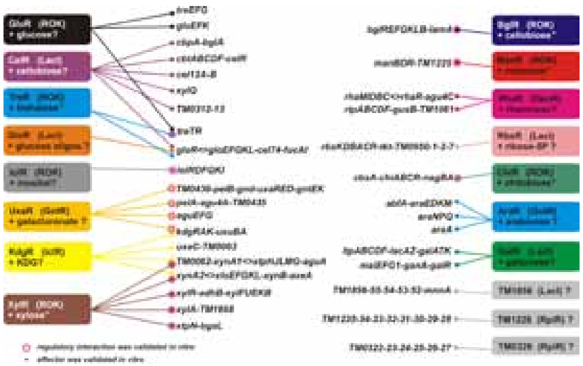

Figure 1. Reconstructed regulatory network for carbohydrate utilization genes and 18 transcription factors in *Thermotoga maritima.* Abbreviations: transcription factors (TFs) are shown in colored boxes; predicted regulatory interactions between a TF and its target operons are shown by lines; experimentally confirmed regulatory interactions are marked by magenta circles; TF protein family is denoted in parentheses; predicted small molecule effectors are listed after '+'; confirmed effectors are marked by asterisks.

Overall, this study yielded numerous insights into physiology, biochemistry and transcriptional regulation of the carbohydrate utilization machinery of *T. maritima* and other Thermotogales. It also provided additional validation of the established integrative approach to the genomics-driven reconstruction of metabolic and transcriptional regulatory networks, which can be applied to many other, yet unexplored, biological systems.

255

Variability of Transcriptional Regulatory Network in *Desulfovibrio* Species

Alexey Kazakov[1]* (AEKazakov@lbl.gov), Inna Dubchak,[1] **Dmitry Rodionov**,[2] and **Pavel Novichkov**[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; and [2]Sanford-Burnham Medical Research Institute, La Jolla, Calif.

Project Goals: Reconstruction of transcriptional regulatory network in DOE-mission genomes.

Use of sulfate-reducing bacteria for bioremediation of heavy metal-contaminated media has economical and environmental significance. Sulfate-reducing bacteria are a matter of big concern in engineering due to problems with corrosion of metal structures. Some bacteria of *Desulfovibrionales* order are also of clinical interest because they may act as opportunistic pathogens.

In this study we carried out large-scale comparative genomics analysis of regulatory interactions in *Desulfovibrio vulgaris* and 10 related species from *Desulfovibrionales* order using the integrated RegPredict web-server for regulon reconstruction (http://regpredict.lbl.gov/).

The resulting regulatory network contains 40 regulons controlled by transcription factors (TF) from 13 TF families, more than 1,300 binding sites and 4,000 target genes involved in stress response, amino acid metabolism, metal homeostasis. The analysis of gene expression profiling data of TF mutants, which have been recently done for two transcription factors FUR (iron homeostasis) and PerR (peroxide stress response) in *Desulfovibrio desulfuricans G20*, shows nice agreement with regulons inferred by comparative genomics. The analysis of correlations between multiple microarray expression profiles provides additional information for the assessment of regulon predictions as demonstrated by analysis of MetR regulon.

The detailed analysis of three large TF families ArsR, CRP, and GntR revealed substantial variations in regulons among *Desulfovibrionales* genomes. Phylogenetic analysis of these transcription factors showed that 72 regulators from 9 sub-families are well conserved in the analyzed genomes, whereas 60 regulators are poorly conserved (present in less than 3 genomes) and showing a mosaic distribution. This pattern is most likely a result of multiple evolutionary events such as horizontal gene transfer, gene loss, and gene duplication. For regulon reconstruction in the latter group of poorly conserved TFs, the standard orthology-based approach was inefficient, and thus we utilized a modified comparative genomics approach. We found closest TF homologs outside of the analyzed group of genomes, and applied motif detection procedure for set of genes in the conserved genomic neighborhood. As a result we were able to identify binding motifs for 40 poorly conserved TFs and for all well-conserved TF sub-families. Within each TF family we observed fair similarity between TF binding motifs. This observation can be utilized for development of automatic approach for large-scale inference of poorly conserved regulons.

An overall reference collection of 40 *Desulfovibrionales* regulogs can be accessed through RegPrecise database (http://regprecise.lbl.gov/).

256

Computational Modeling of Fluctuations in Energy and Metabolic Pathways of Microbial Organisms

Elijah Roberts*[1] (erobert3@illinois.edu), Piyush Labhsetwar,[2] **Nathan D. Price**,[2,3] **Carl R. Woese**,[4] and **Zan Luthey-Schulten**[1,2] (zan@illinois.edu)

[1]Department of Chemistry, [2]Center for Biophysics and Computational Biology, [3]Department of Chemical and Biomolecular Engineering, [4]Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL

http://www.scs.illinois.edu/schulten

Project Goals: In this project, we propose to develop a computational methodology for modeling individual microbes that connects stochastic reaction-diffusion dynamics of selected biochemical pathways under *in vivo* conditions with genome-scale modeling of metabolic and regulatory networks, techniques from systems biology. Using such a combined model, we propose that the dynamic behavior of a cell's entire biochemical network can be approximated in a way that will enable discovery of new phenomena regarding how cells respond to fluctuating environmental conditions.

Stochastic expression of genes produces heterogeneity in clonal populations of microorganisms under identical conditions. We analyze and compare the behavior of the inducible *lac* genetic switch using well-stirred and spatially resolved simulations for *Escherichia coli* cells modeled under fast and slow-growth conditions. Our new kinetic model describing the switching of the *lac* operon from one phenotype to the other incorporates parameters obtained from recently published *in vivo* single-molecule fluorescence experiments along with *in vitro* rate constants. For the well-stirred system, investigation of the intrinsic noise in the circuit as a function of the inducer concentration and in the presence/absence of the feedback mechanism reveals that the noise peaks near the switching threshold. Applying maximum likelihood estimation, we show that the analytic two-state model of gene expression can be used to extract stochastic rates from the simulation data. The simulations also provide mRNA-protein probability landscapes, which demonstrate that switching is the result of crossing both mRNA and protein thresholds. Using cryoelectron tomography of an *E. coli* cell and data from proteomics studies, we construct spatial *in vivo* models of cells and quantify the noise contributions and effects on repressor rebinding due to cell structure and crowding in the cytoplasm. Compared to systems without spatial heterogeneity, the model for the fast-growth cells predicts a slight decrease in the overall noise and an increase in the repressor's rebinding rate due to anomalous subdiffusion. The tomograms for *E. coli* grown under slow-growth conditions identify the positions of the ribosomes and the condensed nucleoid. The smaller slow-growth cells have increased mRNA localization and a larger internal inducer concentration, leading to a significant decrease in the lifetime of the repressor-operator complex and an increase in the frequency of transcriptional bursts.

Funding: This work was supported by U.S. Department of Energy (Office of Biological and Environmental Research) under award DE-FG02-10ER6510.

257

Microbial Data Integration and Metagenomic Workflows and the Development of the Biological Resource Network

Brian C. Thomas[1]* (bcthomas@berkeley.edu), Chongle Pan,[2] Itai Sharon,[1] Vincent Denef,[1] Robert Hettich,[2] Trent Northen,[3] Ben Bowen,[3] and **Jill Banfield**[1]

[1]University of California, Berkeley; [2]Oak Ridge National Laboratory; and [3]Lawrence Berkeley National Laboratory

Project Goals: Our goal is to construct a network of integrated metagenomic, proteomic, metabolomic, and transcriptomic data from microbial communities. Additionally, we will develop methods for analyzing, visualizing, and understanding the complex and highly redundant nature of these systems. Our research will generate a scalable microbial community knowledge base that addresses the challenges of complex systems.

Microbial systems biology relies upon the coordination of multiple data types, such as genomic, proteomic, metabolomic, and transcriptomic data. Commonly, metagenomic workflows begin with just sequence information, but quickly expand to include a variety of data sources, including metadata such as geochemical measurements, time of sample collection etc. Integration of these other data streams is challenging, yet their dynamic interconnection is a critical goal in bioinformatics and will help provide a functional understanding of the microbial systems. Effective data integration and handling must also address the reality that data (especially gene calls, functional annotations, assignment of genes to organisms within a community, etc.) can exist in many states – live (updated), stale (out of date), dead (wrong), lost, etc. The goal of our DOE Kbase project is to make sense of the breadth and diversity of biological data from a single system by providing a knowledge base and workflows to facilitate integration of 'live' data and promote data access.

Towards this goal, we present an update on our development of a microbial data integration system, called the **B**iological **R**esource **N**etwork (BRN). BRN is a network of data resources—spanning genomic, proteomic, metabolomic and transcriptomic data types—interlinked via an underlying software system based on representational state transfer (REST, or "RESTful") services. RESTful software architectures simplify component implementation, reduce the complexity of component semantics, improve the effectiveness of performance tuning, and increase the scalability of components. BRN components are separated by data type:

our metagenomic service is separate from our proteomic service or metabolomic service etc. This modular architecture allows us to scale individual components as data content or complexity demands change, and allows key design choices for a resource to be made by component experts. This feature is well suited for cloud-based computing, where new nodes can be allocated to BRN components as demand increases (for example, during a data update stage or during a normalization calculation step). Integration of the various BRN components is achieved with a unified application-programming interface (API). Requests can be made to the individual BRN components if a specific type of data is desired, however, using the API any service in the BRN can make proxy requests to other BRN-aware services. For example, the metagenomics resource is informed of the proteomics resource: It can therefore respond to requests such as "retrieve proteins from metagenomic contig 'Lepto2_contig42' with uniquely expressed proteins found in sample 'AMD_FloatingBiofilm_A'". This simple query represents a joining of two diverse data sets originating from the same sample—the BRN imposes no restrictions on the individual resources with regard to data integrity or format. This is the opposite of what is commonly found in federated data repositories, where data from diverse sources are combined into a single large entity, separated from the original, "live" data. This splintering of data from its maintainers immediately renders it, at the very least, "stale." Federated databases do not benefit from updates made to the original data source in real time. BRN components can be individually scaled and are not required to be served from the same location: BRN web services are connected via the Internet, allowing resources located anywhere in the world to be linked into the Network. Currently the BRN contains services for metagenomics and proteomics resources (hosted at U.C. Berkeley) and initial progress has been made incorporating a metabolomics resource (hosted at LBNL).

GeneGrabber is a database, web application and workflow for processing, analyzing, visualizing and serving metagenomic and linked data. As a BRN component, data are easily accessible using simple REST requests via HTTP (either from a web browser or programmatically using the BRN API). GeneGrabber has workflows that assist in data processing, import, and annotation updates. GeneGrabber is a "gene-centric" view of metagenomic data. As mentioned above, GeneGrabber can communicate with any other resource that has a RESTful interface (other BRN services or non-BRN resources, such as UniRef etc).

A critical component of any resource is access and incorporation of new data. Metagenomics begins with tens of millions of sequencing reads that must be assembled, refined, searched, annotated, and assigned to organisms. Workflows greatly help this endeavor. We have developed a metagenomics workflow for handling next-generation sequencing data. The approach has been refined during assembly of genomes from short read Illumina data for several strains of *Sulfobacillus*, all of them with little or no similarities to previously sequenced *Sulfobacillus* strains. New tools and strategies have been developed, including a novel iterative assembly strategy that optimizes assembly of genomes of different coverage depths from the same sample. Most importantly, we have established a scaffolding method that automates the manual assembly curation step—this new method has performed well in preliminary analyses. We have also refined an unsupervised clustering strategy for the purpose of data binning that is based on emergent self-organizing maps (ESOM). The approach has been extended to incorporate data types other than sequence signatures (e.g. temporal distribution patterns), enabling much improved genome classification.

Additionally, we have developed a workflow to process proteomics data using a large computer cluster at ORNL. We use the Sequest and DTASelect algorithms to process standard shotgun proteomics data. A new open-source algorithm, Sipros, was developed to process proteomic stable isotope probing data (Pan et al. In press). The results are integrated with sample metadata and integrated into the BRN using the ProteomeDB service. ProteomeDB enables analyzing proteomic results using information from other resources in the BRN. For example, protein co-expression patterns can be correlated with operon information retrieved from GeneGrabber

Initial progress towards our DOE Kbase project goal can be seen in GeneGrabber and ProteomeDB and their underlying RESTful API and we will provide demonstrations of the underlying data access. Future work will focus on expanding the BRN to include metabolomics (both GCMS and MS/MS), transcriptomics data resources and continued development of workflows facilitating data integration.

258

Insights from the Reconstruction of 3000 Metabolic Models

Christopher Henry,[1,2]* **Ross Overbeek**,[3] Matt DeJongh,[4] Aaron Best,[4] Fangfang Xia,[1,2] Scott Devoid,[2] and **Rick Stevens**[1,2]

[1]University of Chicago, Ill.; [2]Argonne National Laboratory, Ill.; [3]Fellowship for Interpretation of Genomes, Ill.; and [4]Hope College, Mich.

Project Goals: This project aims to develop software and data infrastructure to support programmatic access to the SEED and Model SEED resources for the annotation of genomes, integration of omics data, and reconstruction of genome-scale metabolic models. Within this overarching objective, are four specific aims: (i) enhancing the computational infrastructure behind the SEED framework to improve extensibility, accessibility, and scalability; (ii) integrate extensions into the SEED data-model and software to accommodate new data types including regulatory networks, genome-scale metabolic models, structured assertions, eukaryotic genome data, and growth phenotype data; (iii) develop an application programming interface (API) to provide remote access to the SEED database and tools; (iv) apply SEED infrastructure, data,

and APIs to annotate genomes and construct genome-scale metabolic models for organisms with applications to bioenergy, carbon cycle, and bioremediation.

Genome-scale metabolic models have emerged as a valuable resource for generating predictions of global organism behavior based on the sequence of nucleotides in the genome. These models can accurately predict essential genes, organism phenotypes, organism response to mutation, and metabolic engineering strategies. Recently we developed the Model SEED resource (http://seed-viewer.theseed.org/models/) for the high-throughput reconstruction of new genome-scale metabolic models for microbial genomes. We demonstrated the ability of this resource to produce 130 new genome-scale models that are comparable in scale to existing published models. We also validated and optimized these models against available Biolog and gene essentiality data. Now we have applied the Model SEED to producing draft metabolic models for over 3000 microbial genomes, representing nearly all complete microbial genomes currently available in GenBank. New algorithms were developed for the gap-filling of these models to enable the activation of every possible reaction in the models; new algorithms were applied for the generation of biomass reactions based on completeness of annotated pathways for biomass precursors; and finally, new algorithms were applied for using the SEED tools to identify gene candidates that may be associated with the gap-filled reactions. This work reveals insights into the diversity of microbial genomes, the completeness of our knowledge of these genomes, and the areas of our knowledge where more gaps presently exist.

259

Toward the Genomic Organization of Superorganisms: Trends in the Functional Content of Metagenomics Samples

Koon-Kiu Yan[1,2]* (koon-kiu.yan@yale.edu), **Sergei Maslov**,[3] and **Mark Gerstein**[1,2,4]

[1]Program in Computational Biology and Bioinformatics, [2]Department of Molecular Biophysics and Biochemistry, and [4]Department of Computer Science, Yale University, New Haven, Conn.; and [3]Department of Condensed Matter Physics and Materials Science, Brookhaven National Laboratory, Upton, N.Y.

Project Goals: To better understand how natural habitats shape the organization of bacterial communities from a system-wide analysis of the trends between the average genome size and the proportion of different functionally categorized genes in metagenomic samples.

In prokaryotes the proportion of genes attributed to particular cellular processes varies with the size of the genome. More specifically, it has been reported that in many high level functional categories the number of genes in each category scales N_c as a power-law of the genome size (N), i.e. $N_c \sim N^{\alpha}$. These scaling laws, shaped by the underlying evolutionary processes, shed light on the organization of prokaryotic genomes. Nevertheless, in natural environments, bacterial species do not live in an isolated fashion but in forms of bacterial communities whose organizations are shaped by the interplay between species, for instance bacterial symbiosis. To a certain extent, a community of many bacterial species behaves as a superorganism, where individual species work collectively to survive. Toward this end, it is instructive to examine the functional content of the "genomes" of these superorganisms.

In this work, we explore the idea of superorganisms using metagenomic samples collected in different environmental habitats. We download metagenomic samples from sources such as the CAMERA database, and map the reads to curated functional categories such as the Clusters of Orthologous Groups (COG), and the Kyoto Encyclopedia of Genes and Genomes (KEGG). The ratio between, for instance, the number of reads corresponding to transcription factors and the total number of reads, is related to the scaling relationships described and we could infer parameters such as the average genome size of the sample. The inferred average genome size sheds light on whether large genomes or small genomes are favored in the corresponding habitat. We can further estimate, for each sample, the average genome size of species and the proportion of reads mapped to different functional categories and look for systematic trends similar to scaling laws demonstrated among individual genomes. Trends in superorganisms will be compared with the corresponding trends among individual genomes.

260

Numerical Optimization Algorithms and Software for Systems Biology: An Integrated Model of Macromolecular Synthesis and Metabolism of *Escherichia coli*

Ines Thiele[1]* (ines.thiele@gmail.com), R.M.T. Fleming,[1] Stefan Gretar Thorleifsson,[1] A. Bordbar,[2] R. Que,[2] and B.O. Palsson[2]

[1]Center for Systems Biology, University of Iceland, Reykjavik, Iceland; and [2]Bioengineering Dept. University of California, San Diego, La Jolla

Project Goals: This project aims to reconstruct genome-scale models of metabolism and macromolecular synthesis and to develop algorithms capable of solving the resulting large, stiff and ill-scaled matrices. We aim to combine state of the art reconstruction and constraint-based modeling and analysis tools with high-end linear optimization solvers and convex flux balance analysis. The incorporation of thermodynamic information in addition to environmental constraints will allow an accurate assessment of feasible steady states. While we will prototype the reconstruction and algorithm developments with *Escherichia coli*, we will employ the resulting networks to

determine thermodynamically favorable pathways for hydrogen production by *Thermotoga maritima*.

Systems biology aims to understand the mechanisms of the genotype-phenotype relationship by investigating interactions between cellular components. We constructed a high-resolution, gene-sequence dependent network of an *E. coli* cell accounting for major cellular processes, e.g., metabolism, transcription and translation using a bottom-up approach relying on genomic and biochemical information. Using a constraint-based modeling approach (COBRA), we found that changes in codon usage altered the fitness of the in silico *E. coli* in various environmental conditions. This result agrees with experimental observations that not only gene content but also gene sequence determine the environment an organism can occupy. A gene's nucleotide composition is governed by the demand-supply principle of codon usage and tRNA abundance. For the first time, we were able to model the genotype-environment-phenotype relationship while previous efforts suggested such relationship merely based on statistical analysis of codon bias and environmental niche preference.

261

Thermotoga maritima Systems Biology Knowledgebase: A Computational Platform for Multi-Subsystem Reconstruction and Multi-Scale Stoichiometric Modeling

Joshua A. Lerman[1]* (jalerman@ucsd.edu), Daniel R. Hyduke,[1] Haythem Latif,[1] Vasiliy A. Portnoy,[1] Alexandra C. Schrimpe-Rutledge,[2] Joshua N. Adkins,[2] Richard D. Smith,[2] Ines Thiele,[3] **Michael A. Saunders,**[4] Karsten Zengler,[1] and Bernhard Ø. Palsson[1]

[1]Department of Bioengineering, University of California – San Diego, La Jolla; [2]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Wash., [3]Center for Systems Biology, University of Iceland, Reykjavík, Iceland., and [4]Department of Management Science and Engineering, Stanford University, Stanford, Calif.

http://systemsbiology.ucsd.edu

Project Goals: This project aims to: (1) develop a genome-scale model of macromolecular synthesis for *Thermotoga maritima*, (2) fully integrate it with the metabolic model, (3) guide the development of algorithms capable of finding optimal steady-state solutions to the combined model despite its large, sparse, and ill-scaled constraint matrix, (4) achieve phenotypic predictions with respect to gene expression, (5) classify the failure modes of the model to prioritize the reconstruction of regulatory circuits, and (6) enable the direct mapping of omics data to aid in the determination of the governing constraints on biohydrogen production from organic waste products.

Project Description:

Over the past decade, the process of reconstructing metabolic networks at the genome scale has become prevalent in molecular systems biology. There is growing interest in using these genome-scale models of metabolism as contextual frameworks for the analysis of the vast amount of omics data currently generated. Transcriptome and proteome data have been integrated with metabolic network models and have provided insight into regulation, the frugality of the genomic program under growth selection pressure, host-pathogen interactions, and drug-responses. However, the omics integration has so far only been accomplished in a Boolean or indirect fashion. This is because the metabolic network reconstructions in wide use employ a Boolean mapping between genetic loci and enzymatic activity termed the gene-protein-reaction (GPR) relationship.

We have previously shown that it is possible to construct a genome-scale model of RNA and protein expression based on a set of basic biochemical reactions. This process was first completed in *Escherichia coli*, and the resulting model was called the 'E'-matrix, which stands for the reconstruction of gene Expression.

Here, we completed a model of macromolecular synthesis for *T. maritima*, which has a relatively small genome of 1.8 Mb. This reconstruction relied heavily on an accurate transcription unit architecture (see abstract by Latif et al.), which detailed the machinery and regulatory conditions required for RNA synthesis. The metabolic network and the macromolecular synthesis network were subsequently merged into a single multi-scale model (termed the 'ME' matrix, for Metabolism and Expression). This ME matrix is shown to be computable and has unique capabilities compared to standard metabolic reconstructions. We found that the incorporation of the biochemical reactions underlying the expression of gene products within a metabolic network reconstruction allowed for the removal of the artificial Boolean GPRs and facilitated the simulation of variable enzyme concentrations, opening a range of new applications of flux balance analysis. For example, the explicit representation of transcription and translation provided an opportunity to directly employ quantitative transcriptomic and proteomic measurements as model constraints.

Ultimately, after adding mathematically derived constraints to couple certain reactions in this model, we achieved a model fully linking the functions of over 650 genes in *T. maritima*, representing around 35% of the entire genome. Note that the current metabolic reconstruction for *T. maritima* covers only 25% of the total genome. This model represents one of the most comprehensive models of any organism to date. Our formulation leads to a reduced dependence on artificial objective functions, such as the biomass objective function, which do not have a mechanistic biochemical basis. For example, nucleotides and amino acids are no longer drawn out of the cell in bulk. Instead, individual RNA and protein synthesis fluxes are decision variables in the optimization problem and the *in silico* cell must decide how it should invest its building blocks given a finite capacity to synthesize them.

We show promise of long-term applications of this type of model, including the prediction of the governing constraints on biohydrogen production from organic waste products, thermostable protein engineering, interpretation of adaptive evolution, and minimal genome design. Incorporating expression of macromolecules into a genome-scale model of metabolism resulted in a quantitative model for transcription and translational processes, but it did not approximate the regulatory and signaling mechanisms that control expression. Analysis of the failure modes of our model will help us prioritize regulons to add to the model to increase predictive power. When no such information is available, the model will lead to discovery in terms of the suggested potential for novel regulatory circuits. Interestingly, this may be immediately tractable in *T. maritima* as its genome supports relatively few transcriptional regulatory states, with only a limited number of transcription factors (see abstract by D. Rodionov et al.), some of which have been experimentally characterized.

Grant Information: Numerical Optimization Algorithms and Software for Systems Biology (DOE Award ER25917 / DE-PS02-08ER08-01)

262

Deciphering the Growth Dynamics of *Shewanella oneidensis* by Integrating Metabolite and Gene Expression Profiles with Stoichiometric Modeling

Ed Reznik,[1]* Sara Baldwin,[1]* Qasim Beg,[2] and **Daniel Segrè**[1,2,3] (dsegre@bu.edu)

[1]Program in Bioinformatics, [2]Department of Biomedical Engineering, [3]Department of Biology, Boston University, Boston, Mass.

http://prelude.bu.edu

Project Goals: This project is aimed at combining stoichiometric genome-scale modeling of metabolism in *Shewanella oneidensins* MR-1 with gene expression data and metabolic profiles to understand how metabolic regulation affects growth and byproduct secretion.

A major challenge in systems biology is the integration of multiple types of data in order to understand the dynamics of complex biological processes. To this end, we present a flexible method for combining time-series measurements of growth rate, metabolite concentrations in the supernatant, and gene expression data with genome-scale metabolic modeling. We apply our approach to a batch growth experiment in which the bacterium *Shewanella oneidensis* MR-1 was observed during the transition from exponential to stationary phase, in minimal lactate medium.

In our algorithm, we implement a modified, time-dependent version of the GIMME approach, previously developed for mapping gene expression data onto flux balance models (1). GIMME searches for a set of fluxes that minimizes the inconsistency with gene expression data, based on a scoring function that considers "on" any gene with expression above a given threshold.

In the original version of GIMME, a universal threshold value was used to determine reaction activity across all genes. In our new method we use a large compendium of microarrays to determine a unique threshold for each gene, and we demonstrate that, by doing so, we improve the predictive power of the algorithm. In addition, taking advantage of our time series data, we link together successive optimizations using a dynamic flux balance analysis framework (2). This enables us to investigate how the metabolism of *S. oneidensis* adjusts over a prolonged period of time as nutrients become depleted in the environment.

Initial versions of our algorithm used biomass data and expression data to determine the distribution of fluxes over time. The flux predictions were compared with experimental measurements of metabolite concentrations in the media to assess the predictive capabilities our algorithm. We next wanted to assess which internal fluxes would need to change in order to optimally align our external flux predictions with experimental measurements. Towards this goal, we implemented a quadratic programming minimization of a global inconsistency score which includes deviations from both expression data and nutrient concentrations in the supernatant. This procedure allowed us to make specific biological hypotheses regarding which metabolic pathways would need to carry more flux in order to produce the observed metabolite profiles.

Our algorithm produces flux predictions that are qualitatively better than those obtained by a standard dynamic flux balance algorithm. In particular, we capture previously uncharacterized time-dependent profiles for pyruvate and acetate utilization and excretion. Based on our current results, we propose that this method can be extended to inform future efforts in metabolic modeling, with potential applications in metabolic engineering and microbial ecology.

References

1. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. *PLoS Computational Biology*. (2008) 4(5): e1000082.
2. Mahadevan R, Edwards J, Doyle FJ. Dynamic Flux Balance Analysis of Diauxic Growth in *Escherichia coli*. *Biophysical Journal*. (2002) 83(3) pp.1331-1340: doi:10.1016/S0006-3495(02)73903-9.

263

Building and Testing an Open Source Platform for Spatio-Temporal Stoichiometric Modeling of Metabolism in Microbial Ecosystems

William J. Riehl[1]* (briehl@bu.edu), Niels Klitkord,[1] William Harcombe,[2] Christopher J. Marx,[2] Nathaniel C. Cady,[3] and **Daniel Segrè**[1,4] (dsegre@bu.edu)

[1]Graduate Program in Bioinformatics, Boston University, Boston, Mass.; [2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Mass; [3]College of Nanoscale Science and Engineering, University at Albany, Albany, N.Y; and [4]Department of Biology and Department of Biomedical Engineering, Boston University, Boston, Mass.

http://prelude.bu.edu

Project Goals: The goal of this project is to build a framework for spatio-temporal flux balance modeling of multiple interacting microbial species. The platform will be used to study syntrophic interactions in microbial ecosystems. Predictions will be compared to experimental measurements performed using cell microprinting technology.

Genome-scale models of microbial metabolism are becoming increasingly important in addressing questions related to bioenergy production, bioremediation, and carbon and nitrogen cycling in the biosphere. As automated gene annotation pipelines, network gap-filling algorithms, and high throughput experimental methods improve, it will become gradually feasible to model virtually any sequenced microbe using this approach. Yet, some of the most fundamental properties of natural microbial ecosystems crucially depend on aspects that are beyond the stoichiometries of individual biochemical species. These include contact- or metabolite-mediated interactions between different microbes, dynamical changes of the environment, spatial structure of the underlying geography and evolutionary competition between distinct subpopulations.

As part of the DOE Systems Biology Knowledgebase we are developing COMETS (**C**omputation **O**f **M**icrobial **E**cosystems in **T**ime and **S**pace), an open source, broadly applicable and user-friendly platform for performing spatially distributed time-dependent flux balance simulations of microbial ecosystems. By taking advantage of the computational efficiency of flux balance model calculations, we implement a spatially structured lattice of interacting metabolic subsystems. These subsystems represent a level of detail that is intermediate between a fine-grained single-cell modeling approach, and a global mean-field approximation. The COMETS simulation engine combines a modified version of dynamic flux balance analysis (dFBA[1]) with a finite differences approximation of diffusion dynamics, to simultaneously track the spatio-temporal fate of multiple environmental molecules and microbial species. The COMETS platform has the capacity to bridge multiple spatial and temporal scales, making it possible to observe long term dynamics of microbial populations growing in a given environmental setting, based on constant updates of local nutrient availabilities and exchanges, and ultimately determined by the metabolic activity of individual microbial species. Thus, it can be used as a platform for modeling the growth of a single bacterial species in a Petri dish, biofilm formation on complex substrate morphologies, seasonality of microbial communities in a specific geographical setting, or the growth and diffusion of a microbe that has been genetically engineered toward bioremediation in a contaminated body of water.

We present a first fully working first version of our platform, which uses the open-source GNU Linear Programming Kit (GLPK) for performing the dFBA calculations, and a Java-based language (Processing) for coordinating simulations and rapid visualization. COMETS can run on individual CPUs, as well as on a dedicated high performance 48-core machine. We have started applying COMETS to the analysis of several different examples, including the growth of a single species in a 2-dimensional environment and syntrophic growth of natural and engineered small microbial consortia. In ongoing work, computer simulations will be integrated with experiments, allowing us to (i) calibrate the simulation parameters towards faithful representation of microbial growth patterns, and (ii) perform pilot studies on microbial ecosystem dynamics. Specifically, we are currently employing quill-pen microprinting technology to print patterns of cells onto solid substrates in a variety of patterns and species combinations. The ability to print such combinations of cells on a surface allows us to measure several quantities used or predicted by the model, including nutrient uptake rate, cell growth rate, and metabolic byproduct diffusion in a microscale environment. These initial microscale models and experiments will be gradually scaled up to larger environments, and more complex microbial ecosystem.

Reference

1. Mahadevan R, Edwards JS, and Doyle FJ. *Dynamic flux balance analysis of diauxic growth in Escherichia coli.* Biophys J. 2002, Sep; 83(3):1331-40.

264

Modularity in Biological Systems Affects the Phenotypic Outcome of Multiple Genetic Perturbations

Hsuan-Chao Chiu[1]* (hcchiu@bu.edu), Christopher J. Marx,[2] and **Daniel Segrè**[1,3] (dsegre@bu.edu)

[1]Graduate Program in Bioinformatics, Boston University, Boston, Mass.; [2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Mass.; and [3]Departments of Biology and Biomedical Engineering, Boston University, Boston, Mass.

http://prelude.bu.edu

Project Goals: The goal of the project is to understand how multiple genetic perturbations jointly affect microbial phenotypes (such as fitness or metabolic fluxes).

An important application of systems biology is to provide a predictive understanding of complex biological network to help generate modified microbial strains useful for metabolic engineering applications. Both evolutionary and rational design strategies towards increasing productivity of modified strains often face the challenge of understanding how individual or combined multiple genetic perturbations affect phenotypes, such as growth rate, or the rate of byproduct secretion.

The degree of interdependency between two alleles in determining phenotypes, referred to as epistasis, is the subject of active experimental and computational research. For example, epistasis has been recently shown to be useful for identifying genes that belong to mutually dependent pathways or functional units, providing valuable information both about specific processes and about the global organization of cellular functions.

Here we propose a new theoretical framework for studying epistasis. Specifically, we study how the modular organization of a biological system affects the magnitude or sign of epistasis. Based on the assumption that different modules in a biological system can be associated with measurable phenotypes, we ask whether it is possible to infer the degree of epistasis relative to a given phenotype (e.g. fitness) based on how this phenotype depends on other phenotypes. In particular, we start by analyzing an explicit example of fitness function, in which fitness is expressed as the difference between a benefit and a cost component. Following a null assumption that mutations affect each component in a multiplicative manner, our model analytically predicts a prevalent negative epistasis distribution between pairs of mutations that are either both beneficial or both deleterious. Our results suggest that antagonistic interaction may prevail among beneficial mutations, providing diminishing returns payoff on multiple perturbations, and potentially slowing down microbial adaptation. We next provide a general expression for epistasis relative to a phenotype that has an arbitrary functional dependence on other phenotypes, under the assumption of small perturbations. This expression successfully reflects the lack-of-epistasis intuition for perturbations that act on independent modules, and provides a general "epistasis propagation law" describing how epistasis on a given high level trait can be quantified based on epistasis on lower-level traits.

Our results suggest a possible connection between the architecture of genetic interaction networks and the modular organization of biological systems, and provide general insight on the inherent benefits and limitations of multi-perturbation microbial strain design strategies.

Student Oral Presentation–Tuesday

265

Algorithms for Synthetic Ecology: Microbial Cross-Feeding Induced by Environmental Transformations

Niels Klitgord[1]* (klitgord@gmail.com), and **Daniel Segrè**[1,2]

[1]Program in Bioinformatics, [2]Department of Biology and Department of Biomedical Engineering, Boston University, Boston, Mass.

http://prelude.bu.edu

Project Goals: We use stoichiometric genome-scale models of metabolism to identify environmental conditions that induce cross-feeding interactions between different microbes.

Interactions between microbial species can be mediated by the exchange of small molecules, secreted by one species and metabolized by another. Both one-way (commensal) and two-way (mutualistic) interactions may contribute to complex networks of interdependencies. Understanding these interactions constitutes an open challenge in microbial ecology, with important applications in metabolic engineering and environmental sustainability. In parallel to natural communities, it is possible to explore interactions in artificial microbial ecosystems, e.g. pairs of genetically engineered mutualistic strains. Here we computationally generate artificial microbial ecosystems without re-engineering the microbes themselves, but rather by predicting their growth on appropriately designed media. We use genome-scale stoichiometric models of metabolism to identify media that can sustain growth for a pair of species, but fail to do so for one or both individual species, thereby inducing putative symbiotic interactions. We first tested our approach on two previously studied mutualistic pairs, and on a pair of highly curated model organisms, showing that our algorithms successfully recapitulate known interactions, robustly predict new ones, and provide novel insight on exchanged molecules. We then applied our method to all possible pairs of seven microbial species, and found that it is always possible to identify putative media that induce commensalism or mutualism. Our analysis also suggests that symbiotic interactions may arise more readily through environmental fluctuations than genetic modifications. We envision that our approach will help generate microbe-microbe interaction maps useful for understanding microbial consortia dynamics and evolution, and for exploring the full potential of natural metabolic pathways for metabolic engineering applications.

For more information: http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1001002

266

Integration of Flux-Balance Analysis and Pathway Databases

Peter D. Karp* (pkarp@ai.sri.com), and **Mario Latendresse**

SRI International

We describe new computational techniques for generating metabolic flux models from pathway databases. The Pathway Tools[1] software is a software package for creating, updating, visualizing, and analyzing Pathway/Genome Databases (PGDBs) for organisms with sequenced genomes. We have recently developed software for generating a linear programming model of a metabolic reaction network that is stored within a PGDB. Pathway Tools automatically invokes the SCIP solver on that model. The resulting optimized fluxes are then displayed on a metabolic pathway map for the PGDB by Pathway Tools, to accelerate a user's understanding of the predicted fluxes.

Benefits of this approach are that the metabolic flux model is closely coupled with an integrated genomic/metabolic knowledge base, and with other computational tools for manipulating that knowledge base. For example, users can visualize reactions, metabolites, pathways, and genome information using the rich visualization capabilities of Pathway Tools. Users can update metabolic reactions, substrates, and pathway definitions using the interactive editors within Pathway Tools, and those updates are reflected in the flux-balance model that is generated from the PGDB. In addition, metabolic model debugging tools within Pathway Tools can be applied to the metabolic flux model. Example debugging tools include tools for element balancing of metabolic reactions, and for detecting dead-end metabolites in the metabolic network.

In addition we have developed novel methods for completing a metabolic model. We have extended the gap-filling work of Maranas and colleagues to yield a multiple gap-filling approach. Using a meta-optimization procedure that is also automatically generated from a PGDB, our software will extend an incomplete metabolic model by postulating reversals of unidirectional reactions in the metabolic model, and by postulating additions of new reactions to the metabolic model from the MetaCyc database. These two approaches extend metabolic models to produce biomass compounds that they were previously unable to synthesize. Additionally, our software will gap-fill the nutrient compounds, that is, adding additional nutrient compounds that will produce biomass compounds that could not be produced. Finally, the software will identify which biomass compounds can still not be produced even after the preceding types of gap filling, thus further focusing the user's model debugging efforts. Taken together, these techniques can radically shorten the time required to develop FBA models from months to days.

Reference

1. Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M., and Caspi, R. (2010) "Pathway Tools version 13.0: Integrated Software for Pathway/Genome Informatics and Systems Biology," *Briefings in Bioinformatics* 11:40-79.

267

Curation and Computational Design of Bioenergy-Related Metabolic Pathways

Peter D. Karp* (pkarp@ai.sri.com), **Mario Latendresse**, and **Ron Caspi**

SRI International

Project Goals: Goal 1: Enhance the MetaCyc metabolic pathway database to contain bioenergy-related enzymes and pathways. Goal 2: Develop computational tools for engineering metabolic pathways that satisfy specified design goals, in particular for bioenergy-related pathways.

Pathway Tools[1] is a systems-biology software package written by SRI International (SRI) that produces Pathway/Genome Databases (PGDBs) for organisms with sequenced genomes. Pathway Tools also provides a wide range of capabilities for analyzing the predicted metabolic networks, and analyzing user-generated omics data. More than 1,500 academic, industrial, and government groups use Pathway Tools. An integral part of the Pathway Tools software is MetaCyc[2], a large, multiorganism database of metabolic pathways and enzymes that is manually curated by SRI and its collaborators. Our project has two goals:

1. Enhance MetaCyc to contain bioenergy-related metabolic enzymes and pathways.
2. Develop computational tools for engineering metabolic pathways that satisfy specified design goals, in particular for bioenergy-related pathways.

We are significantly expanding the coverage of bioenergy-related metabolic information in MetaCyc. Version 14.6 of MetaCyc contains 1,600 pathways and 9,000 bioreactions from 2,100 organisms. The information in MetaCyc has been curated from more than 26,000 publications. We are curating additional bioenergy-related metabolic pathways into MetaCyc from the biomedical literature. By adding these additional pathways to MetaCyc, the pathway prediction component of the Pathway Tools software will be able to recognize these pathways in newly sequenced genomes. We plan to use Pathway Tools to generate organism-specific

Pathway/Genome Databases (PGDBs) for all energy-relevant organisms sequenced at JGI. SRI will make these databases freely available to the public via its BioCyc website

Our second goal is to develop an efficient computational tool for the engineering of metabolic pathways. The tool will satisfy design goals specified by the user and will offer two operational modes: a fast, approximate mode that will quickly find a near-optimal pathway and an exact mode that will find the optimal pathway. Design goals will include starting and ending compounds for the pathway, and will enable the specification of constraints such as preferred or disallowed intermediates. The tool will utilize the large collection of enzymes in the MetaCyc database as a reference, and its suggestions will be validated with the bioenergy-related pathways we are curating into MetaCyc. The pathways generated by this tool will be ranked and clustered according to several optimality criteria.

References

1. Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M., and Caspi, R. (2010) "Pathway Tools version 13.0: Integrated Software for Pathway/Genome Informatics and Systems Biology," *Briefings in Bioinformatics* 11:40-79.
2. Caspi, R., Altman, T., Dale, J.M., Dreher, K., Fulcher, C.A., Gilham, F., Kaipa, P., Karthikeyan, A.S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Paley, S., Popescu, L., Pujar, A., Shearer, A.G., Zhang, P., and Karp, P.D., (2010) "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases," *Nuc. Acids Res.* 36:D623-31.

268

Gene Ontology for Microbial Processes Useful for Bioenergy Production

Trudy Torto-Alalibo[1]* (trudy@vbi.vt.edu), Endang Purwantini,[1] Biswarup Mukhopadhyay,[1] T.M. Murali,[2] Brett M. Tyler,[1] and **Joao C. Setubal**[1]

[1]Virginia Bioinformatics Institute, Blacksburg; and [2]Virginia Polytechnic Institute and State University, Blacksburg

Project Goals: In collaboration with the community of microbiologists interested in energy-related processes and with the Gene Ontology (GO) consortium, develop a comprehensive set of Gene Ontology terms that describe biological processes relevant to energy-related functions. Annotate microbial genomes relevant to bioenergy-production with appropriate GO terms Develop a database and web interface for storing and displaying manual annotations.

The MENGO project (http://mengo.vbi.vt.edu/) is a community-oriented multi-institutional collaborative effort that aims to develop new Gene Ontology (GO) terms to describe microbial processes of interest to bioenergy. Such terms will aid in the comprehensive annotation of gene products from diverse energy-related microbial genomes. The Gene Ontology consortium was formed in 1998 to create universal descriptors, which can be used to describe functionally similar gene products and their attributes across all organisms. MENGO, an interest group of the GO consortium, solicits help from the bioenergy community in developing GO terms relevant for bioenergy-production related processes such as biomass deconstruction, solventogenic fermentation, H_2 production, methanogenesis, and synthesis of hydrocarbons. Currently, based on community input, five potential working groups have been formed. These working groups will work on GO term development in focal areas that include biomass deconstruction, central metabolism and protein structure-function relationships. A few general concepts associated with the selected focal areas have been developed. These and related GO terms will be presented. The MENGO interest group will host a workshop right after the DOE contractor-grantee meeting from April 13-14 at the same venue. This workshop will introduce participants to the Gene Ontology. Additionally, we will have an open forum to seek input from participants on general concepts relevant to bioenergy-production related processes, which will subsequently inform MENGO term development. Funding for the MENGO project is provided by the Department of Energy as part of the Systems Biology Knowledgebase program.