DOE/SC-0142

Section:
Knowledgebase and Computing for Systems Biology

U.S. DEPARTMENT OF ENERGY
Office of Science

2012

Genomic Science Awardee Meeting X

Bethesda, Maryland
February 26-29, 2012

[Revised: March 1, 2012]

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

220

KBase: An Integrated Knowledgebase for Predictive Biology and Environmental Research

Adam Arkin[1]* (aparkin@lbl.gov), Robert Cottingham,[2] Sergei Maslov,[3] Rick Stevens,[4] Dylan Chivian,[1] Paramvir Dehal,[1] Christopher Henry,[4] Folker Meyer,[4] Jennifer Salazar,[4] Doreen Ware,[5] David Weston,[2] Brian Davison,[2] and Elizabeth M. Glass[4]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [3]Brookhaven National Laboratory, Upton, N.Y.; [4]Argonne National Laboratory, Argonne, Ill.; and [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

http://kbase.us/
http://outreach.kbase.us

Project Goals: The Systems Biology Knowledgebase (KBase) has two central goals. The scientific goal is to produce predictive models, reference datasets and analytical tools and demonstrate their utility in DOE biological research relating to bioenergy, carbon cycle, and the study of subsurface microbial communities. The operational goal is to create the integrated software and hardware infrastructure needed to support the creation, maintenance and use of predictive models and methods in the study of microbes, microbial communities and plants.

KBase is a collaborative effort designed to accelerate our understanding of microbes, microbial communities, and plants. It will be a community-driven, extensible and scalable open-source software framework and application system. KBase will offer free and open access to data models and simulations, enabling scientists and researchers to build new knowledge, test hypotheses, design experiments, and share their findings to accelerate the use of predictive biology. Our immediate 18-month goal is to have a beta-version completed by February 2013.

The KBase microbial science domain will enable the reconciliation of metabolic models with experimental data with the ultimate aim of manipulating microbial function for applications in energy production and remediation. In order to accomplish this, we will enable users to expand on a strong foundation of quality genome annotations, reconstruct metabolism and regulation, integrate and standardize 'omics data, and construct models of genomes. In doing so we will vastly improve the planning of effective experiments, maximize understanding of microbial system function, and promote sharing of data and findings.

The plants science domain will initally target linking genetic variation, phenotypes, molecular profiles, and molecular networks, enabling model-driven phenotype predictions. We will also map plant variability onto metabolic models to create model-driven predictions of phenotypic traits. Initial work will focus on creating a workflow for rapidly converting sequencing reads into genotypes. We will also build tools for data exploration, and the linking of gene targets from phenotype studies such as genome-wide association studies, with co-expression, protein-protein interaction, and regulatory network models. Such data exploration will allow users to narrow candidate gene lists by refining targets, or be able to visualize a sub-network of regulatory and physical interactions among genes responsible for a phenotype. Users can also highlight networks or pathways impacted by genetic variation.

Through comparative analysis of metagenomes acquired over different spatial, temporal or experimental scales, it is now possible to define how communities respond to and change their environment. Our microbial communities team will build the computational infrastructure to research community behavior and build predictive models of community roles in the carbon cycle, other biogeochemical cycles, bioremediation, energy production, and the discovery of useful enzymes. We are building the next-generation metagenomic platform that provides scalable, flexible analyses, data vectors for models, tools for model creation, data quality control, application programing interfaces, and GSC-compliant standards for data collection. Initial efforts will target the development of bioprospecting and experimental design tools.

KBase will be composed of a series of core biological analysis and modeling functions, including an application programming interface that can be used to connect different software programs within the community. These capabilities will be constructed from the popular analysis systems at each of the KBase sites. Their integration into KBase will combine individual functions to create the next generation of biological models and analysis tools. The KBase application programming interface will also enable third-party researchers from our diverse community of users to design new functions. KBase will be supported by a computing infrastructure based on the OpenStack cloud system software, distributed across the core sites.

The success of the KBase project depends not only on producing a large-scale, open computational capability for systems biology research data management and analysis but also on positioning these tools to be used by the community. Our outreach program is designed to target different user groups: data providers, tool builders, and users of both data and tools. A significant effort will be made to connect the user groups and broader systems biology science communities to the KBase resources and efforts. We will provide

educational support, including providing access to outreach and technical staff and online venues in which to express questions, suggestions and needs to other users and our entire team. In addition to this, we will provide transparency to KBase, providing information about the project, team, and development with the scientific community.

New functionality will allow users to visualize data, create powerful models or design experiments based on KBase-generated suggestions. Although the integration of different data types will itself be a major offering to users, the project is about much more than data unification. KBase is distinguished from a database or existing biological tools by its focus on interpreting missing information necessary for predictive modeling, on aiding experimental design to test model-based hypotheses, and by delivering quality-controlled data.

This work is supported by the U.S. Department of Energy, Office of Biological and Environmental Research under Contract DE-AC02-06CH11357.

221

The DOE Systems Biology Knowledgebase: Microbial Communities Science Domain

Folker Meyer[1]* (folker@anl.gov), Dylan Chivian,[2] Andreas Wilke,[1] Narayan Desai,[1] Jared Wilkening,[1] Kevin Keegan,[1] William Trimble,[1] Keith Keller,[2] Paramvir Dehal,[2] Robert Cottingham,[3] Sergei Maslov,[4] Rick Stevens,[1] and **Adam Arkin**[2]

[1]Argonne National Laboratory, Argonne, Ill.; [2]Lawrence Berkeley National Laboratory, Berkely, Calif.; [3]Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [4]Brookhaven National Laboratory, Upton, N.Y.

http://kbase.us/

Project Goals: The Systems Biology Knowledgebase (KBase) has two central goals. The scientific goal is to produce predictive models, reference datasets and analytical tools and demonstrate their utility in DOE biological research relating to bioenergy, carbon cycle, and the study of subsurface microbial communities. The operational goal is to create the integrated software and hardware infrastructure needed to support the creation, maintenance and use of predictive models and methods in the study of microbes, microbial communities and plants. The microbial communities component will be focused on building the computational infrastructure to understand the community function and ecology through study of genomic and functional data and integration of community models with single-organism models. This will allow for researching community behavior and building predictive models of communities in their role in the environmental processes and the discovery of useful enzymes. The microbial communities infrastructure will support the overall KBase goal to provide a framework for experimental decision support and data interpretation.

The KBase microbial communities team will integrate both existing and new tools and data into a single unified framework that is accessible programmatically and through web services. The framework will allow the construction of sophisticated analysis workflows by facilitating the linkages between data and analysis methods. The standardization, integration and harmonization of diverse data types housed within the KBase and data located on servers maintained by the larger scientific community will allow for a single point of access ensuring consistency and quality-assurance/quality-control checks of data quality.

We have begun by creating KBase data and analysis services that will link our core resources: MG-RAST [1], metaMicrobesOnline [2], SEED [3], IMG/M [5] and ModelSEED [4]. These services will allow clients to access data and analysis methods across these tools without the burden of reconciling identifiers, learning different data access and programmatic access methods, ensuring data quality, and maintaining relevant metadata. New functionality, not currently available in our core tools, is being created within KBase using the programmatic interfaces.

Protoypical applications

Bioprospecting: Microbial diversity is a key element in the search for new, valuable compounds such as enzymes with novel properties. Elucidating novel proteins from microbial communities is a function that integrates metaMicrobesOnline functions with MG-RAST data through the KBase programmatic interface. It will allow for deep comparative analysis of protein families, expanding significantly the current functionality in MG-RAST. This includes detailed trees and alignments combining metagenomic sequences and sequences from complete genomes. The initial version of this tool will allow in-depth characterization of novel members of existing protein families; future versions will allow characterization of completely novel protein families. This will exercise the communities part of the API and also the microbes set of API calls and provide a useful, missing component to the combined tool suite.

Metagenomic Experimental Design Wizard: The "wizard" will assist in the design of metagenomic experiments. Using information on sequence data (including sequence quality) and the detailed community analysis, it will provide guidance on the analyses/experiments that are supported by the data. The wizard provides guidance on, for example, identifying microbial communities using sequence assembly, and providing a confidence value for community reconstructions obtained from metagenomic data.

References

1. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008 Sep 19;9:386.
2. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, et al. MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res. 2010 Jan;38(Database issue):D396-400.

3. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005 Oct 7;33(17):5691-702.
4. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 2010 Sep;28(9):977-82.
5. Victor M. Markowitz, I-Min A. Chen, Ken Chu, Ernest Szeto, Krishna Palaniappan, Yuri Grechkin, et al. IMG/M: the integrated metagenome data management and comparative analysis system, Nucleic Acid Res. 2011.

This work is supported by the U.S. Department of Energy, Office of Biological and Environmental Research under Contract DE-AC02-06CH11357.

222

The DOE Systems Biology Knowledgebase: Microbial Science Domain

Paramvir S. Dehal[1]* (psdehal@lbl.gov), Chris S. Henry,[2] Ben Bowen,[1] Steven Brenner,[1] Ross Overbeek,[2] John-Marc Chandonia,[1] Dylan Chivian,[1] Pavel S. Novichkov,[1] Keith Keller,[1] **Adam P. Arkin**,[2] Robert Cottingham,[3] Sergei Maslov,[4] and Rick Stevens[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Argonne National Laboratory, Argonne, Ill.; [3]Oak Ridge National Laboratory, Oak Ridge, Tenn.; and [4]Brookhaven National Laboratory, Upton, N.Y.

http://kbase.us

Project Goals: The Systems Biology Knowledgebase (KBase) has two central goals. The scientific goal is to produce predictive models, reference datasets and analytical tools and demonstrate their utility in DOE biological research relating to bioenergy, carbon cycle, and the study of subsurface microbial communities. The operational goal is to create the integrated software and hardware infrastructure needed to support the creation, maintenance and use of predictive models and methods in the study of microbes, microbial communities and plants. The microbes component of the KBase will be centered on an analysis pipeline that will include annotation of genome sequences, reconstruction of metabolic pathways and regulons, generation of metabolic and regulatory models, and reconciliation of models with existing 'omics datasets and new datasets uploaded by a user. KBase will provide access to this database and analysis pipeline via a powerful programmatic interface and an intuitive socially-enabled web interface.

The microbes component of the KBase project aims to unify existing 'omics datasets and modeling toolsets within a single integrated framework that will enable users to move seamlessly from the genome annotation process through to a reconciled metabolic and regulatory model that is linked to all existing experimental data for a particular organism. More importantly, we will embody tools for applying these models and datasets to drive the advancement of biological understanding and microbial engineering.

In order to drive the development of the microbes area and enable new science, we will focus on accomplishing prototype science workflows rather than general tasks. Work will be bootstrapped by leveraging data sets and tools developed and maintained by the MicrobesOnline, SEED, RegPrecise and ModelSEED resources. The initial microbes efforts will integrate prototype workflows for: (1) genome annotation and metabolic reconstruction, (2) regulon reconstruction, (3) metabolic and regulatory model reconstruction, and (4) reconciliation with experimental phenotype and expression data.

1. Evidence Based Genome Annotation and Metabolic Reconstruction: While the rate of genome sequencing continues to advance at an exponential pace, our ability to confidently assign structural and functional gene annotations has not kept pace. High quality gene annotations with confidence measures are a critical component of all genome scale modeling. Efforts to create genome scale regulatory and metabolic models are held back by the poor quality of existing gene models. To help resolve this, we are proposing a workflow that takes as input a genome sequence, RNASeq and/or high density tiling array data, and functional 'omics datasets. Through an iterative process combining the experimental data sets and comparative genomics, structural annotations will be improved and integrated into the RAST annotation server. High quality transcription start and operon predictions will be used to improve promoter and regulatory predictions. Gene functional annotations will be improved by combining model predictions to identify missing gene functions.

2. Regulon Reconstruction: Given accurate gene models, the KBase framework will provide integrated pipelines for building and refinement of higher level, regulatory and metabolic models. Reconstruction of genome-wide transcriptional regulatory network (TRN) is a necessary step toward the ultimate goal, building a ***predictive*** model of microbial organism.

3. Seamless Integration of Modeling into the Biological Research Process: Computational modeling of biological systems provides a mechanism for rapidly exploring a wide variety of alternative theories of complex phenomena observed in the wetlab. Yet modeling is restrained by a lack of interoperability of models and data, a high degree of mathematical and computational expertise required to utilize models, and an inability to rapidly build new models accurately capture the complete body of our current biological understanding. The KBase microbes team aims to integrate reconstruction and analysis algorithms for metabolic models, regulatory models, and ultimately many other modeling abstractions with a unified compilation of 'omics data to produce a framework that facilitates the process of building biological understanding. Algorithms will be included for simulating microbial behavior in a specified environment, predicting and modifying pathways according to data or design, and designing experiments to test biological theories.

4. Testing and Improving Consistency of Biological Understanding and Experimental Data: The microbes Kbase area will combine 'omics data and modeling algorithms within a single platform, making it possible to easily cross-validate data and models by comparing predictions with experimental observations. Additionally, tools and interfaces will be provided to guide the process of adapting the biological understanding that underlies our models to reconcile conflicts with experimental data. The results of this adaptation process will automatically feed back into the annotation algorithms built into KBase, globally improving annotations of all genomes. Initially this effort will focus on the interpretation and reconciliation of growth phenotype and gene expression data, but will ultimately be expanded to all types of 'omics data.

This work is supported by the U.S. Department of Energy, Office of Biological and Environmental Research under Contract DE-AC02-06CH11357.

223

The DOE Systems Biology Knowledgebase: Plant Science Domain

Doreen Ware[1,2]* (ware@cshl.edu), Sergei Maslov,[4] Shinjae Yoo,[4] Dantong Yu,[4] Michael Schatz,[1] James Gurtowski,[1] Matt Titmus,[1] Jer-ming Chia,[1] Sunita Kumari,[1] Andrew Olson,[1] Shiran Pasternak,[1] Jim Thomason,[1] Ken Youens-Clark,[1] Mark Gerstein,[5] Gang Fang,[5] Darryl Reeves,[5] Pam Ronald,[6] Chris Henry,[7] Sam Seaver,[7] David Weston,[3] Priya Ranjan,[3] Robert Cottingham,[3] Rick Stevens,[7] and **Adam Arkin**[8]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.; [2]United States Department of Agriculture–Agriculture Research Service; [3]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [4]Brookhaven National Laboratory, Upton, N.Y.; [5]Yale University, New Haven, Conn.; [6]University of California, Davis; [7]Argonne National Laboratory, Argonne, Ill.; and [8]Lawrence Berkeley National Laboratory, Berkely, Calif.

http://kbase.us

Project Goals: The Systems Biology Knowledgebase (KBase) has two central goals. The scientific goal is to produce predictive models, reference datasets, and analytical tools and to demonstrate their utility in DOE biological research relating to bioenergy, carbon cycle, and the study of subsurface microbial communities. The operational goal is to create the integrated software and hardware infrastructure needed to support the creation, maintenance, and use of predictive models and methods in the study of microbes, microbial communities and plants. The plant component of the KBase will allow users to model genotype-to-phenotype relationships using metabolic and functional networks as well as phenotype measurements and 'omic data. It will also support the reconstruction of new metabolic and functional networks based on expression profiles, protein-DNA, and protein-protein interactions. To accomplish this, we will provide interactive, data-driven analysis and exploration across multiple experiments and diverse data-types. We will provide users access to comprehensive collections of 'omics datasets together with relevant analytical tools and resources.

The major goal for the KBase plants area is to model genotype-to-phenotype relationships through analysis and integration of genomic, transcriptomic, methylomic sequencing data, metabolite and phenotype measurements, and the reconstruction of metabolic and functional networks based on expression profiles, protein-DNA, and protein-protein interactions. To accomplish this goal, KBase will provide interactive, data-driven analysis, and exploration across multiple experiments and diverse data-types. We will provide users access to comprehensive datasets from high-throughput experiments together with relevant analytical tools and resources. Users will be provided with a platform to analyze their own experimental data, integrate publicly available data from other 'omics' platforms, and have these results incorporated into a data exploration framework. We will aid in the translation of basic science through exploration across large 'omics experiments and the initiation of hypothesis-driven genetics studies without the overhead of data flow and management.

The KBase plants effort will consist of two major components, 1) genotyping workflows and 2) data exploration and prediction tools.

Genotyping workflows: Exponential growth in digital demands has motivated extensive research into improved algorithms and parallel systems, especially for genotyping samples, monitoring expression levels, and a host of other important biological applications.

Genotyping workflows will leverage our recent development of Jnomics, as our new Hadoop-based open-source package for rapid development and deployment of cloud-scale sequence analysis tools. Jnomics provides many pre-built tools out-of-the-box that accelerate common tasks, such as mapping, sorting, merging, filtering, and selection, to be performed as distributed tasks spread across a cluster. New tools can be easily created using an open-source Java API, especially for large-scale genotyping and expression analysis. Because it builds on Hadoop, Jnomics tools inherit Hadoop's efficiency and scalability for very large datasets such as the billions of short reads necessary for genotyping many large plant genomes. Furthermore, Jnomics is "file format agnostic", allowing it to seamlessly read and write most common sequence file formats (SAM, BAM, BED, fastq, fasta), making it easy for Jnomics to interface with other components.

Data exploration and prediction: While the benefits of integrating large independent data sets for value-added research are vast, the influx of raw biological data presents unprecedented challenges in data management and representation. We believe that effective visualization will play a key role in mining and exploration of 'omics data sets. While genome

browsers are useful for anecdotal relationships between feature sets and genome maps—especially for bench scientists studying a region of interest—they are not ideal for correlational interrogation of the data. Effective visualization is crucial for generating hypotheses. Displays must be flexible insofar as they allow the end-user to represent highly dimensional data in various forms without compromising pairwise relationships among data points. We intend to develop aesthetic and interactive interfaces that enable researchers to intuitively draw relationships in their complex queries. The visualizations will allow end-users to transform massive data sets in real-time in order to explore latent system-wide relationships among the data. Data exploration tools will focus on integration and visualization of data from multiple 'omics studies, genetic variation, network and pathways models as well as phenotypic association. Users will, for example, map variability onto metabolic models, highlighting pathways predicted to be impacted by genetic variation. By overlaying genetic variation within the context of gene network models, we can predict pathways that are impacted by the pool of genetic diversity in the population.

This work is supported by the U.S. Department of Energy, Office of Biological and Environmental Research under Contract DE-AC02-06CH11357.

224

The KBase Architecture and Infrastructure Design

Tom Brettin*[1] (brettints@ornl.gov), Bob Olson,[2] Ross Overbeek,[2] Terry Disz,[2] Bruce Parello,[2] Shiran Pasternak,[5] Folker Meyer,[2] Michael Galloway,[1] Steve Moulton,[1] Dan Olson,[2] Shane Canon,[3] Dantung Yu,[4] Shiran Paternak, Pavel Novichkov,[3] Daniel Quest,[1] Narayan Desai,[2] Jared Wilkening,[2] Miriam Land,[1] Scott Deviod,[2] **Adam Arkin**,[3] Robert Cottingham,[1] Sergei Maslov,[4] and Rick Stevens[2]

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Argonne National Laboratory, Argonne, Ill.; [3]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [4]Brookhaven National Laboratory, Upton, N.Y.; and [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

http://kbase.us/

Project Goals: The Systems Biology Knowledgebase (KBase) has two central goals. The scientific goal is to produce predictive models, reference datasets and analytical tools and demonstrate their utility in DOE biological research relating to bioenergy, carbon cycle, and the study of subsurface microbial communities. The operational goal is to create the integrated software and hardware infrastructure needed to support the creation, maintenance and use of predictive models and methods in the study of microbes, microbial communities and plants. The driving objectives of the KBase architecture and infrastructure design focus on creating an unprecedented user experience. The integrated software and hardware infrastructure supporting the user experience comprises a continuously expanding collection of software and services. These are hosted on a physical infrastructure consisting of high speed wide area networking, cloud computing resources, and state of the art cluster computing resources.

Achieving our goal of an integrative architecture to support predictive modeling requires enabling a user experience that covers a range of users. These include senior biologists determined to understand and create biological models, computational biologists developing new algorithms and statistical models that form a basis for the biological models, and bioinformaticists chaining together complex workflows to generate, summarize and integrate data that feed into the biological models. These user activities are enabled at various levels of abstraction, including a) knowledge creation, reproduction and sharing; b) rich web applications; c) programmatic Application Programming Interface (API) libraries and scripts; and d) wire level communication access.

The development of the KBase Unified API is based on a service-oriented approach to deliver both functionality and data to the community. Behind this lies a set of services backed by servers. Initially, these services will be developed by the KBase infrastructure team and will support a long term goal of community developed and contributed services. Our initial set of services will be backed by the following servers:

1. **Genomic Servers** that provide access to a rapidly growing set of genomes, features of those genomes, and annotations of both genomes and features.
2. **Expression Data Servers** creating access to a growing body of expression data, along with the underlying encoding of metadata needed to support interpretation.
3. **Protein Family Servers** supporting access to a variety of the existing collections of protein families.
4. **Polymorphism Servers** capturing various genetic polymorphisms such as single nucleotide polymorphisms, tandem repeats, and copy number variations.
5. **Phenotype Servers** enabling the relationships between genotype and phenotype to be understood.
6. **Compound and Reaction Data Servers** supporting a unified and maintained representation of reaction networks.
7. **Metabolic Modeling Servers** that support the construction and maintenance of metabolic models.
8. **Regulatory Models Servers** that support the construction and maintenance of regulatory models.

Our KBase physical infrastructure builds on the successes of DOE investment in our national scientific cyber infrastructure and can therefore leverage enormous intellectual resources present in the DOE community.

Building on ESNet allows us to construct a wide area network between the partner labs that enables a virtual

hardware infrastructure. In the first quarter of the project we have established 10Gbit data transfer connectivity between KBase data transfer nodes.

Enabling cloud computing on Magellan will create new opportunities that range from rapid deployment of developer environments to highly scalable production servers. The acceptance of virtualization technology is growing, and the use of machine images produced by others is already visible in our core services. In the near-term, we are establishing the infrastructure to run existing images, both community based and internally created, on which the infrastructure is dependent. Examples include images supporting microbial community models and plant genome wide association studies. For the mid-term, we plan to contribute machine images to the community by creating snapshots of parts of our environment so that others can use them on the hardware of choice. For the long term objective, the ability to host running machine images on KBase hardware is a means for promoting collaboration and community support.

Cluster Computing has long been a critical part of biological data analysis. In collaboration with computing centers created by the Office of Advanced Computing Research such as NERSC, our underlying cluster services can leverage these resources and scale to meet needs.

KBase aims to power the next wave of biological research in DOE and beyond. Enabling these capabilities requires a software and hardware infrastructure that is integrated, extensible, and scalable. The architecture is designed to meet these needs and support user functionality to visualize data, create models or design experiments based on KBase-generated suggestions.

This work is supported by the U.S. Department of Energy, Office of Biological and Environmental Research under Contract DE-AC02-06CH11357.

225

The InterBRC Knowledgebase Data Registry

Dylan Chivian[1,2]* (DCChivian@lbl.gov), Guruprasad Kora,[3,4] Mustafa Syed,[3,4] Thomas Brettin,[4] Keith Keller,[2] Jason Baumohl,[2] Yury Bukhman,[5,6] Richard LeDuc,[5,6] Adam Arkin,[1,2] David Benton,[5,6] **Steve Slater**,[5,6] and **Edward Uberbacher**[3,4]

[1]Joint BioEnergy Institute, Emeryville, Calif.; [2]Lawrence Berkeley National Laboratory, Berkeley, Calif.; [3]Bioenergy Science Center, Oak Ridge, Tenn.; [4]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [5]Great Lakes Bioenergy Research Center, Madison, Wis.; and [6]University of Wisconsin, Madison

Project Goals: The InterBRC Knowledgebase project will provide a mechanism for the integration of data obtained by the DOE Bioenergy Research Centers (BRCs) with the DOE Systems Biology Knowledgebase (KBase).

The DOE Systems Biology Knowledgebase (KBase) is building a system that allows predictive modeling of Microbes, Microbial Communities, and Plant Systems Biology for the scientific community. The DOE Bioenergy Research Centers (BRCs) are large producers of systems biology data in all three areas, generating genome sequence, expression, proteomic, metabolomic, metabolic flux, growth, and phenotype data for microbes; metagenome, metatranscriptome, and metaproteome data for microbial communities; and genome, protein interaction, protein localization, allelic variation, and mutant phenotype data for plants. These data are rich sources for modeling via the KBase, but are currently either not easily accessible, or housed in a wide range of data repositories and thus challenging to bring together for comparative analysis. The InterBRC Knowledgebase will serve as the bridge between BRC data stored in dedicated systems and the common infrastructure of the KBase. Data will be indexed and searchable via the InterBRC Knowledgebase Data Registry, which will additionally provide the location and access protocol for retrieving data sets of interest from the dedicated services. These will be incorporated in turn into the KBase infrastructure and be available for comparative analysis and systems biology modeling by BRC researchers and the greater community.

This work is part of the InterBRC Knowledgebase project, the Joint BioEnergy Institute (JBEI), the Great Lakes Bioenergy Research Center (GLBRC), the Bioenergy Science Center (BESC), and the Systems Biology Knowledgebase project supported by the U.S. Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between the U.S. DOE and Lawrence Berkeley National Laboratory (LBNL), through contract DE-FC02-07ER64494 between the U.S. DOE and the Great Lakes Bioenergy Research Center (GLBRC) Cooperative Agreement, through contract DE-AC05-00OR22725 between the U.S. DOE and Oak Ridge National Laboratory (ORNL) administered by UT-Battelle, LLC, for the U.S. DOE, and through subcontract 4000105297 between ORNL and the GLBRC Cooperative Agreement.

226

GGKbase, a Portable Knowledgebase for Analysis and Integration of "Omic" Data From Microbial Communities

Brian C. Thomas,[1]* Ken-ichi Ueda,[1] Rahul Basu,[2] Chongle Pan,[3] Trent Northen,[3] Benjamin Bowen,[2]* and **Jill Banfield**[1]* (jbanfield@berkeley.edu)

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

http://genegrabber.berkeley.edu/

Project Goals: To develop a knowledgebase for analysis and integration of 'omics' data from microbial communities.

Cultivation-independent approaches provide access to the wide diversity of microorganisms in natural environments. Sequence data (metagenomic information) is foundational to most studies of natural microbial communities as it

enables functional analysis through proteomics (proteogenomics) and provides context for transcriptomic and metabolomics information.

Given the vast dataset sizes, extraction of biological and biogeochemical insight from 'omic' datasets is challenging. In the current project, we have constructed a workflow and knowledgebase that enables recovery, efficient and effective display, manipulation, and interrogation of such information. Most important, the structure is portable. Developed initially for analysis of data from acid mine drainage microbial communities, the structure has already been populated by information from multiple other projects, including the DOE Rifle IFRC metagenomics and proteomics efforts.

The pipeline includes all components required for analysis of next-generation sequencing information, from assembly through binning and functional annotation. An explicit goal of GGKbase is the recovery of near-complete genomes, a feature that distinguishes our approach from others (e.g., MGRAST). Curated and binned genome fragments are grouped into organismal "bins" from which inferences about metabolic potential can be made. Genes for which proteins have been identified from one or a series of samples using the open reading frames predicted from the metagenomic data are flagged at the "genome browsing" level, and detailed information about abundance and distribution can be accessed, gene-by-gene.

We have developed GeneGrabber (a component of GGKbase), a multi-user, list-based, social/sharing approach for analysis of the metabolism of individual organisms and comparative metabolic analysis at the community level. Individual genes or groups of genes belonging to a pathway can be assigned to one of more lists, as determined by the investigator, and these lists can be shared with other users, including the ability to invite new users to participate in curating a list. Because the lists are driven by a keyword search (or EC number, GO term etc.), genes can be identified and classified simultaneously across the entire dataset. This establishes metabolic profiles using tens, hundreds, and potentially thousands of genes at a time.

Understanding an ecosystem's metabolic potential is a complex task. Leveraging the extensive, content-based lists created for each metagenomics resource, we developed a tool within GeneGrabber for visualizing the extent of metabolic machinery present in the data. This visualization, termed "genome summary," is invaluable for exploring metabolic pathways and can identify which organisms in the community are responsible for a process. The genome summary is a useful tool for investigating the molecular underpinnings of ecosystem metabolic processes.

The GGKbase system has been engineered to access other 'omics' data resources, without having to resort to database federation. GGKbase uses representational state transfer (REST) to tap into other information sources and we have developed a caching system using Redis to accelerate user-centric data access. We have also developed an API to access the GGKbase resource both in Ruby as well as just using simple URL access. Currently, GGKbase includes both metagenomic and proteomic data. Additionally, we have expanded the GGKbase structure to now include metabolomics data and have developed a system for processing and displaying this data called MetaboliteAtlas. MetaboliteAtlas, like all components in the GGKbase utilizes a RESTful architecture and includes a separate front-end web display for in-depth metabolomics investigations.

GGKbase is under continual development, currently focused around addition of metabolomic and transcriptomic data. As more researchers begin using GGKbase, new ideas are captured into the structure. Current challenges in increasing the scale of the community metagenomics approach include automation of time-intensive binning steps and aspects of time series data analysis.

227

The Ribosomal Database Project: Tools and Sequences for rRNA Analysis

Benli Chai[1]* (chaibenl@msu.edu), Qiong Wang,[1] Jordan Fish,[1] Donna McGarrell,[1] C. Titus Brown,[2] Yanni Sun,[2] James M. Tiedje,[1] and **James R. Cole**[1]

[1]Center for Microbial Ecology, and [2]Computer Science and Engineering, Michigan State University, East Lansing

http://rdp.cme.msu.edu/
http://fungene.cme.msu.edu/

Project Goals: The Ribosomal Database Project (RDP; http://rdp.cme.msu.edu) offers aligned and annotated rRNA sequence data and analysis services to the research community. These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, greenhouse gas production, and bioremediation.

Our view of the evolutionary relationships among life forms on Earth has been revolutionized by the comparative analysis of ribosomal RNA sequences. Life is now viewed as belonging to one of three primary lines of evolutionary descent: Archaea, Bacteria and Eucarya. This shift in paradigm has not only challenged our understanding of life's origin, but also provided an intellectual framework for studying extant life—particularly the vast diversity of microorganisms. Ribosomal RNA diversity analysis using genes amplified directly from mixed DNA extracted from environments has demonstrated that the well-studied microbes described by classical microbial systematics represent only a small percentage of diversity. The use of rRNA to explore uncharacterized diversity had become such a relied-upon methodology that by 2008, 77% of all INSDC bacterial DNA sequence submissions described an rRNA sequence, and only 2% of these entries had a Latin name attached (valid or otherwise; Christen, 2008)! Examining the RDP's collection of quality rRNA sequences demonstrates that cultivated organisms represent only a fraction of observed rRNA diversity, and currently available genome sequences cover an even smaller slice of this cultivated fraction. Phylo-

genetically informed selection of sequencing candidates, as done in the GEBA Project, can help improve genome coverage of diversity represented by cultivated organisms (Wu et al., 2009), and single cell sequencing can provide partial genome data for uncultivated organisms; but it will be years before these techniques are able to make practical progress towards tackling the immense diversity represented by the collection of rRNA sequences. In fact, it is our knowledge of rRNA diversity that is guiding these efforts.

In the current release (August 2011), RDP offers 1,921,179 aligned and annotated quality-controlled public bacterial and archaeal rRNA sequences along with tools that allow researchers to examine their own sequences in the context of the public sequences (Cole et al., 2009). In addition, 11,892 researchers have over 7.4 million private pre-publication sequences in the *my*RDP account system. On average, over the past year, the RDP website was visited by over 10,100 researchers (unique ip addresses) in over 24,500 analysis sessions each month, an increase of 20% from the previous year, while Web Services (SOAP) interfaces to these RDP analysis functions process an additional 8.6 billion bases of sequence per month for high-volume users running their own analysis queues. In addition, during the past 12 months the RDP Pyro Pipeline has been used by over 446 researchers (unique e-mail addresses) to process their own high-throughput current-generation rRNA sequence data.

The popular RDP naïve Bayesian Classifier has been cited in over 380 publications in over 100 journals since its publication in 2007. The open-source command line RDP Classifier package is freely available from SourceForge. This package has been downloaded 2,745 times during the past 12 months and a total of 6,025 times since first offered in 2006. A new command line tool developed to meet the growing need for taxonomy-based analyses of large numbers of sequences in multiple samples, MultiClassifier, is also freely available from the RDP site.

We are now offering a fungal version of our RDP Classifier, in collaboration with Andrea Porras-Alfaro, Gary Xie, Cheryl Kuske and their co-workers (supported through a DOE Science Focus Area grant to Los Alamos National Laboratory). This Fungal Classifier is trained on 8,506 curated fungal 28S rRNA gene reference sequences, along with a hand-vetted fungal taxonomy including 1,702 genera plus higher-level taxa. This is an important addition to the RDP tools, as fungi play a major role in carbon cycling and plant health. We will continue to work with our LANL colleagues to improve and extend our set of fungal tools.

We are working with standards bodies, such as the Genomic Standards Consortium (GSC; http://gensc.org/gsc/) and the Terragenome Consortium (http://www.terragenome.org), to help define environmental annotation standards for rRNA and other environmental marker gene libraries, and to assure that RDP is ready for the new standards. Our work was incorporated into the new MIMARKS (Minimal Information about a MARKer gene Sequence) standard (Yilmaz *et al.,* 2011). RDP's Hierarchy Browser has been updated to allow searching on MIMARKS attributes.

To help our user community comply with these new community standards, RDP has added informative web pages to make our users aware of the GSC standards and developed tools to assist our users to collect and prepare compliant metadata (contextual data). RDP has created specialized MIMARKS templates using Google Docs office suite. These Google Spreadsheet templates for all 14 MIMARKS environmental packages provide embedded help and validation for MIMARKS attributes. By leveraging Google Docs, our users have a familiar tool that provides remote collaborative support, data storage, and a powerful user interface, all without need of local IT infrastructure. When researchers are ready to submit data, they can use the RDP's *my*RDP SRA PrepKit, which helps researchers prepare richly annotated sequence data in the complex XML documents required for submission to the NCBI and ENA SRA repositories. RDP GoogleSheets also contain macros that produce data formats compatible with standard ENA and NCBI sequence submission tools.

Beyond rRNA, RDP is leveraging its tools and services to provide support for analysis of coding for key environmental functions (functional genes). Like rRNA genes, protein-coding genes can also be selectively targeted for deep sequencing coverage. Genes that are important for environmental processes can thus be used to identify functional, ecological and evolutionary patterns. RDP's FunGene repository provides support for amplicon design and analysis for genes involved in a range of processes. Integrated tools are provided for primer testing, phylogenetic analysis and model refinement. FGPipeline (http://fungene.cme.msu.edu/FunGenePipeline/) has been recently developed to analyze this type of sequencing data. The major tools underpinning this pipeline include RDP FrameBot, which extends and implements an existing dynamic programming algorithm to detect and correct frameshift artifacts and filter out non-target reads, the HMMER3 aligner with preconfigured HMM models or user-supplied protein seeds, and RDP mcClust, which implements a memory-constrained hierarchical clustering algorithm for clustering large numbers of protein reads and includes a distributed computing option. This pipeline has been tested on important functional genes from the carbon and nitrogen cycles, and genes important for human-microbe interactions, including biphenyl dioxygenase (*bph*), important for bioremediation,

nitrogenase reductase (*nif*H), a key component in nitrogen fixation, and butyryl-CoA transferase (*but)* and butyrate kinase (*buk*), important for production of butyrate, the main energy source of human colonic epithelial cells.

The RDP's mission includes user support. Help is available online, through e-mail (rdpstaff@msu.edu), and by phone (517-432-4998). In the past year, RDP staff has helped users through over 900 emails and phone conversations.

References

1. Christen, R. 2008. Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. *Microbes and Environments* 23:253-268.
2. Yilmaz, P., Kottmann, R., Field, D., Knight R, Cole, J.R., Amaral-Zettler L. et al. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* 29: 415-420.
3. Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73:5261-5267.
4. Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056-1060.

The RDP is supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-99ER62848.

228

The NamesforLife Semantic Index of Phenotypic and Genotypic Data

Charles T. Parker,[1] Catherine Lyons,[1] and **George M. Garrity**[1,2*] (garrity@namesforlife.com)

[1]NamesforLife, LLC, East Lansing, Mich. and
[2]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing

http://www.namesforlife.com

Project Goals: Predictive models depend on high quality input data. But not all data are of similar quality nor are all of the data amenable to computational analysis without extensive cleaning, interpretation and normalization. Key among those needed to make the projects such as the DOE Knowledgebase (Kbase) operational are phenotypic data, which are more complex than sequence data, occur in a wide variety of forms, use complex and non-uniform descriptors and are scattered about the literature and specialized databases. Incorporating these data into the Kbase will require expertise in harvesting, modeling and interpreting the data. The *NamesforLife Semantic Index of Phenotypic and Genotypic Data* seeks to address this problem by taking the first steps toward building an ontology of bacterial and archaeal based on the taxonomic literature through the dvelopment of a draft vocabulary of phenotypic features of the taxonomic type strains.

Unlike sequence data, which are essentially universal, uniform and predictable, phenotypic data are inherently complex, noisy and "taxonomically parochial". The same trait may vary significantly under different conditions of growth, at different times during the life of a cell and under different environmental conditions. The language of phenotype is also complex and may be limited in taxonomic scope, requiring expert interpretation, as there is no equivalent to BLAST for searching for phenotypic data, and there is no central repository for such data. In some cases an entire language exists to describe the phenotypic features that apply to a single phylum, class or order (e.g., Cyanobacteria, Actinobacteria) or a particular class of features (e.g., lipids, structural carbohydrates). In addition, phenotypic data must be viewed in a historical perspective (time when the data were collected) to understand what was measured and how it was measured. As was the case with microbial physiology, which had fallen out of fashion as a field of research, so too has "classical" or polyphasic microbial taxonomy resulting in a dwindling community of experts who can readily interpret the existing information in the literature and in various specialized databases. How might that expertise be captured and applied to developing a standardized language and ontology of microbial phenotype?

In 2003, Garrity and Lyons proposed a novel approach to resolving ambiguity of biological nomenclature. Their approach provided a means of resolving the complex relationships that exist among names and the concepts and objects to which names apply. When coupled with Digital Object Identifiers (DOIs) their method provided a means by which names in digital content (e.g. journal articles, technical reports, web pages) and databases could be made actionable and directly linked to expertly curated information about the name, including its history of changes. NamesforLife, LLC has developed a suite of web services and applications based on this method that can be used to semantically enrich or enhance digital in a variety of formats. The Company has already demonstrated that vectors of names (Semiotic Fingerprints) can be used to index and cross-classify large corpora of scientific and patent literature based on the relationship between named organisms and the underlying subject matter of subsets of documents. The methods and tools are not, however, restricted to biological nomenclature and can be applied to terminologies of all types.

The long-term objective of this STTR project is to develop a semantic index of bacterial and archaeal phenotypes that can be used to augment annotation efforts and to provide a basis for predictive modeling of microbial phenotype. The index will be based on published descriptions of taxonomic type and non-type strains that have been the subject of ongoing genome sequencing efforts as this will provide a mechanism whereby hypotheses can be tested and verified, reproducibly. This project is tightly coupled with ongoing DOE projects (Genomic Encyclopedia of Bacteria and Archaea, the Microbial Earth Project, the Community Sequencing Project) and with two key publications, Standards in Genomic Sciences (SIGS) and the International Journal of Systematic and Evolutionary Microbiology

(IJSEM). The first step towards accomplishing this goal, and the primary objective of this Phase I project is the development of a draft vocabulary of phenotypic features.

Our approach towards developing a draft vocabulary of bacterial and archeal phenotype is based on a textual analysis of the richest source of descriptive information; the taxonomic literature. It follows a well-established path used for ontology construction based on derivation of domain-dependent hyponymy (is-a relationships) from a corpus and leverages tools, data resources and expertise that the Company has already developed. For this Phase I project, our target corpus consists of a subset of taxonomic literature of type strains from the IJSEM (2003-2009) and SIGS (2009-2011). These articles have been further subdivided into those pertaining to a single bacterial species or multiple species and higher taxa. The articles have been indexed with Apache Lucene to produce two separate indices; one of the complete documents and one of the descriptions of each new organism alone. We are currently developing a KWIC (KeyWord In Context) interface to permit location and display of a given word in the corpus in its surrounding context to understand usage variations within and across different taxa. Selection of terms for analysis is being done using Apache Luke, which provides facilities for determining usage frequency, coupled with curatorial review for relevance, categorization and synonymy. Then end goal of the Phase I project is to allow end users to view an article with all of the phenotypically relevant terms highlighted based on the KWIC index, to use the KWIC index to auto-populate phenotypic characteristics of a given strain based on the published literature and to allow endusers to adjust/negate/select phenotypic characteristics and their values for a strain, using a normalized set of terms.

Funding for the this project was provided through the DOE SBIR/STTR program (DE-SC0006191). Public funding for development of the NamesforLife infrastructure was received from the DOE SBIR/STTR program (DE-FG02-07ER86321), the Michigan Small Business Technology Developemt Corporation, the Michican Statgetic Fund, and the Michigan Universities Commercialization Initiative.

229

Development of Novel Random Network Theory-Based Approaches to Identify Network Interactions Among Nitrifying Bacteria

Ye Deng* (dengye@glomics.com), Zhili He, Feifei Liu, Yujia Qin, Joy Van Nostrand, and **Jizhong Zhou**

Institute for Environmental Genomics and Department of Botany and Microbiology, University of Oklahoma, Norman

Project Goals: Interactions among various microbial populations in a community play critical roles in determining the functioning of an ecosystem. However, little is known about such network interactions. Also, nitrification is an important step in the nitrogen cycle, and two groups of microorganisms, ammonia-oxidizing bacteria (AOB) and ammonia-oxidizing archaea (AOA) are thought to play essential roles in this microbe-mediated process. Thus the ultimate goal of this project is to develop a novel molecular ecological network approach and analysis pipeline to understand the genetic diversity and interaction of soil AOA and AOB in a grassland ecosystem. Three specific objectives have been pursued: (i) To develop a molecular ecological network approach and analysis pipeline; (ii) To determine the genetic diversity of soil AOA and AOB populations by pyrosequening of bacterial and archaea *amoA* genes; and (iii) To examine interactions of soil AOA and AOB populations by molecular ecological network analysis of *amoA* pyrosequening data.

Development of functional molecular ecological network approach and pipeline. Theoretically, the interactions among various microbial populations in a community play critical roles in determining the functioning of an ecosystem; however, these relationships remain unclear due to the lack of appropriate experimental data and computational analytic tools. To address such challenges, a mathematical approach for identifying molecular ecological networks (MENs) from high-throughput metagenomics sequencing data has been developed. Two major steps are involved: (i) construction of network based on the Random Matrix Theory; and (ii) network characterization using various mathematical approaches. Compared to other network reconstruction methods, the RMT-based approach is remarkable in that the networks are automatically defined and robust to noise, thus providing excellent solutions to several common issues associated with high-throughput metagenomics data. The MEN topological analyses have integrated the most recent findings in social and biological network analyses, such as general network structures (small-world, scale-free, modularity and hierarchy structures) and module-based eigengene analysis, for dissecting community organizations at the whole and module levels. Furthermore, we associated module characteristics with environmental traits to understand the importance of network interactions in determining community functions. To facilitate application by the scientific community, all of these methods and statistical tools have been integrated into a comprehensive Molecular Ecological Network Analysis Pipeline (MENAP), which is open-accessible through the internet.

Pyrosequencing analysis of bacterial and archaeal *amoA* genes. We have designed universal primers for amplifying archaeal or bacterial *amoA* genes (ammonia monoxygenase subunit A). Both bacterial *amoA* and archaeal *amoA* genes were amplified from soil samples in a grassland experiment site, BioCON (CO_2 concentrations, nitrogen fertilization and plant species), followed by 454 pyrosequencing. A total of 1.2M reads were obtained. After preprocessing, about 651K reads remained with 140K for bacteria and 511K for archaea. A total of 1911 OTUs for bacteria and 7922 OTUs for archaea were obtained at a similarity cutoff 0.95. Ammonium oxidizing bacteria (AOB) were dominated by species in *Nitosospiram, Nitrosovibrio* and *Nitrosomonas* genera (beta-proteobacteria) and Ammonium oxidizing archaea (AOA) were mostly from uncultured species in the *Crenarchaeota*

phylum. Further analysis showed that plant diversity and nitrogen fertilization had clear impacts on the community structures of the nitrifying communities, while elevated CO_2 did not have significant effects.

To understand the interactions of AOA and AOB populations in a community, the novel RMT-based network analysis were performed based on the sequencing data. Networks with expected characteristics were obtained. Further topological analysis of AOA and AOB interactions is in progress.

230

The PhyloFacts Phylogenomic Encyclopedia of Microbial Gene Families

Yaoqing Shen,[1] Bushra Samad,[2] Ajithkumar Warrier,[1] Ruchira Datta,[1] and **Kimmen Sjölander**[1,2,3*] (kimmen@berkeley.edu)

[1]QB3 Institute, [2]Department of Bioengineering, and [3]Department of Plant and Microbial Biology, University of California, Berkeley

http://makana.berkeley.edu/phylofacts/

Project Goals: To improve the precision of microbial (meta)genome functional annotation by providing phylogenomic analyses of microbial gene families, providing access to these analyses to the scientific community in the PhyloFacts Database. This pipeline includes: clustering all microbial genes from whole genomes into gene families including homologs from other species; constructing multiple sequence alignments and estimating protein family trees; ortholog identification; integrating experimental and annotation data; computationally scalable methods for HMM classification of (meta)genome sequences to PhyloFacts families and orthology groups.

PhyloFacts is a phylogenomic encyclopedia of gene family trees across the Tree of Life (1). Gene families are defined based on (1) agreement at the multi-domain architecture, and (2) on containing a single Pfam domain in common. The FlowerPower algorithm (2) is used to retrieve homologs from the UniProt database, parameterized separately for the two homology clustering criteria. For each gene family, we construct a multiple sequence alignment and phylogenetic tree; phylogenetic trees are used to identify orthologs in different species using the PHOG algorithm (3). This combination of single domain and domain-architecture clustering enhances the recall and precision of functional classification and orthology identification (4). We construct a hidden Markov model (HMM) for the family and use it to identify homologous protein structures by scoring the Protein Data Bank. Finally, we retrieve experimental and annotation data from various external resources, including UniProt, Gene Ontology, Pfam and KEGG and use these to provide informative descriptions of each family and orthology group, from which the functions of family and orthology-group members can be inferred.

We have focused during the last year on expanding our coverage of microbial gene families. More than 7M proteins are included in PhyloFacts families, representing a broad swath of species. For instance, within Archaea, >90% of *Halobacterium salinarum and* >87% of *Sulfolobus solfataricus* are represented by at least one PhyloFacts family. Within Bacteria, 100% of *Escherichia coli K12*, >94% of *Bacillus subtilis*, >91% of *Thermotoga maritime*, >90% of *Geobacter sulfurreducens*, >87% of *Sulfolobus solfataricus* and >79% of *Deinococcus radiodurans* are represented. Within Eukarya, >90% of *Saccharomyces cerevisiae* and >86% of *Arabidopsis thaliana* genes are included. Detailed coverage of representative species with whole genomes is presented at http://makana.berkeley.edu/phylofacts/coverage/.

Users can access the data in PhyloFacts in several ways, including sequence accession and inputting protein sequences in FASTA format for HMM classification. Data can be downloaded from individual PhyloFacts family pages and can also be downloaded in bulk from http://makana.berkeley.edu/phylofacts/downloads/.

We have also developed a prototype phylogenomic HMM classification system we call FAT-CAT (for Fast Approximate Tree Classification), to allow the functional classification of novel proteins and the simultaneous taxonomic and functional classification of metagenome reads using HMMs placed at internal nodes of gene trees. We will present the results of this analysis at the meeting.

References

1. Krishnamurthy, N., Brown, D.P., Kirshner, D. and Sjölander, K. (2006) PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome biology*, 7, R83.
2. Krishnamurthy, N., Brown, D. and Sjölander, K. (2007) FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC evolutionary biology*, 7 Suppl 1, S12.
3. Datta, R.S., Meacham, C., Samad, B., Neyer, C. and Sjölander, K. (2009) Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic acids research*, 37(Web Server issue): W84-W89.
4. Sjölander, K., Datta, R.S., Shen, Y. and Shoffner, G.M. (2011) Ortholog identification in the presence of domain architecture rearrangement. *Briefings in bioinformatics*. 12 (5): 413-422.

The PhyloFacts Microbial Encyclopedia is supported by the Office of Biological and Environmental Research in the DOE Office of Science.

231

Microbial ENergy Processes Gene Ontology (MENGO): New Gene Ontology Terms Describing Microbial Processes Relevant for Bioenergy

Trudy Torto-Alalibo[1]* (trudy@vbi.vt.edu), Endang Purwantini,[1] Joao C. Setubal,[1,2] Brett M. Tyler,[1,3] and **Biswarup Mukhopadhyay**[1]

[1]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg; [2]Department of Biochemistry, Institute of Chemistry, University of São Paulo, Brazil; and [3]Center for Genome Research and Biocomputing, Oregon State University, Corvallis

http://mengo.vbi.vt.edu

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.

The MENGO project 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 GO consortium was formed in 1998 to create universal descriptors, which can be used to describe functionally similar gene products and their attributes.

MENGO, an interest group of the GO consortium seeks to expand term development for microbial processes useful for bioenergy production. Currently, there are over 200 MENGO terms added to the GO. Areas covered include carbohydrate catabolic processes, oligosaccharide binding and transport, hydrogen production and methanogenesis. Additionally, over 200 GO annotations of bioenergy relevant gene products from microbes such as *Clostridium thermocellum*, *Methanosarcina barkeri*, *Bacteroides thetaiotaomicron* and *Chlamydomonas reinhardtii* have been made. A selection of terms and annotations will be highlighted in this presentation.

The MENGO interest group will also host a workshop right after the DOE contractor-grantee meeting on February 29th at the same venue. This workshop will highlight progress made in GO term development and microbial gene annotation as well as some of the challenges encountered. Additionally, we will have an open forum to hear from participants on other bioenergy areas to be targeted for further term development and microbial genomes to be annotated.

Funding for the MENGO project is provided by the Department of Energy as part of the Systems Biology Knowledgebase program- Grant# DE-SC000501.

232

Gene Ontology Terms Describe Biological Production of Methane

Endang Purwantini[1]* (epurwant@vbi.vt.edu), Trudy Torto-Alalibo,[1] Joao C. Setubal,[1,2] Brett M. Tyler,[1,3] and **Biswarup Mukhopadhyay**[1]

[1]Virginia Bioinformatics Institute, Virginia Polytechnic Institute, Blacksburg; [2]Department of Biochemistry, Institute of Chemistry, University of São Paulo, Brazil; and [3]Center for Genome Research and Biocomputing, Oregon State University, Corvallis

http://mengo.vbi.vt.edu

Project Goals: The MENGO consortium in collaboration with the community of microbiologists engaged in bioenergy research and the Gene Ontology (GO) consortium aims to develop a comprehensive set of Gene Ontology terms that will describe bioenergy related biological processes and to annotate relevant microbial genomes with appropriate GO terms.

The MENGO project is a community-oriented multi-institutional collaborative effort that aims to develop new Gene Ontology (GO) terms to describe microbial processes of interest to bio-energy production. Such terms will aid in the comprehensive annotation of relevant genes in diverse microbial genomes. Among the 200 terms developed so far are a comprehensive set that describes processes involved in the biological production of methane/methanogenesis.

Biologically, methane is generated by methanogenic archaea from H_2 + CO_2, secondary alcohol + CO_2, formate, carbon monoxide, acetate, methanol, methylamines, and methanethiols. Pathways for methanogenesis from these substrates use unusual coenzymes such as coenzyme F_{420}, methanofuran, tetrahydromethanopterin, coenzyme M, cofactor F_{430}, and coenzyme B. Methanogenesis allows efficient mineralization of biological polymers in anaerobic niches of nature and thereby plays an important role in carbon cycle. This integrated process is leveraged for the production of methane from renewable resources and for waste treatment. The MENGO team has created some terms that are useful for describing the biological processes allowing methanogenesis from carbohydrates, including the biosynthesis of relevant coenzymes. Additionally, methanogenesis related gene products of certain methanogenic archaea such as *Methanocaldococcus jannaschii, Methanosarcina barkeri, Methanosarcina thermophilla, Methanosaeta concilii, Methanopyrus kandleri,* and *Methanothermobacter marburgenensis* have been manually annotated with GO terms.

Funding for the MENGO project is provided by the Department of Energy as part of the Systems Biology Knowledgebase program—grant# DE-SC000501.

233

Toward System Biology KnowledgeBase on Transcriptional Regulation in Bacteria

Pavel S. Novichkov[1]* (PSNovichkov@lbl.gov), Dmitry A. Ravcheev,[2,3] Alexey E. Kazakov,[1] Semen A. Leyn,[2,3] Adam P. Arkin,[1] Inna Dubchak,[1] and **Dmitry A. Rodionov**[2,3]* (rodionov@burnham.org)

[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [2]Sanford-Burnham Medical Research Institute, La Jolla, Calif.; and [3]Institute for Information Transmission Problems RAS, Moscow, Russia

http://regprecise.lbl.gov

Project Goals:

1. **Develop methods for genome-scale regulon reconstruction utilizing the comparative genomics approach and analysis of RNA and DNA regulatory sites.**
2. **Infer the regulatory interactions and reconstruct transcription regulatory networks (TRNs) in multiple groups of microbial genomes important for DOE mission.**
3. **Integrate the diverse sets of predicted and experimental data on microbial transcriptional regulation into a RegKnowledgeBase.**

Transcriptional regulation of gene expression in response to extracellular and intracellular signals is a key mechanism for successful adaptation of microorganisms to changing environmental conditions. Genes and operons directly co-regulated by the same transcription factor (TF) or by an RNA motif are considered to belong to a *regulon*. All regulons taken together form the transcriptional regulatory network (TRN) of the cell. Availability of complete genomes stimulated wide application of a computational genomics-based approach implemented in the RegPredict Web-server platform for fast and accurate inference of microbial regulons. During the past decade a large number of manually-curated high quality inferences of transcriptional regulons were accumulated for diverse taxonomic groups of bacteria.

We have developed the RegPrecise database (http://regprecise.lbl.gov) for capturing, visualization, and analysis of computationally predicted regulons in microbial genomes. The primary object of the database is a single regulon in a particular genome. Each regulon description contains a regulator, its effector, a set of target genes and operons, and their associated *cis*-regulatory sites. Each TF-operated regulon also has a DNA-binding site model (profile) represented as a nucleotide logo.

Bacterial TRNs are highly flexible in evolution of microbial genomes. The effective large-scale reconstruction of TRNs by comparative genomics requires selection of optimal sets of closely-related genomes. Therefore, the central strategy for regulon analysis in microbial genomes in RegPrecise is based on subdivision of all microbial species into small taxonomic groups that are analyzed independently. The highest level of regulon organization in the database is represented by taxonomic collections of regulons. The current version of RegPrecise contains 13 taxonomic collections of regulons covering major phyla of Bacteria (Proteobacteria, Firmicutes, Cyanobacteria, Actinobacteria etc.).

The total number of TF and RNA-motif regulons in the current release of RegPrecise database exceeds seven thousands. RegPrecise provides three classifications of regulons implemented as controlled vocabularies: (i) biological processes /metabolic pathways; (ii) effectors /environmental signals; (iii) TF protein families. Biological processes attributed to regulons in the database covers a wide spectrum of the cellular metabolism (Fig. 1).The current list of effectors of analyzed TFs includes more than 200 metabolites from the following major classes: amino acids, carbohydrates, nucleotides, lipids and fatty acids, co-enzymes, peptides and antibiotics, secondary metabolites, and inorganic chemicals. Regulons represented in RegPrecise includes ~5400 TFs from >50 TF protein families.

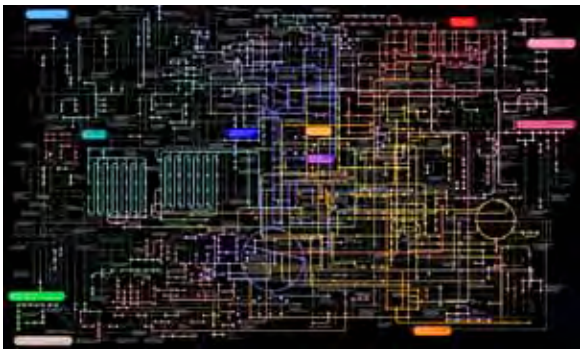


Figure 1. Overview of metabolic pathways covered by reconstructed TF regulons in RegPrecise DB.

The content of RegPrecise database is publically available through the RESTful web-services in JSON format.

In the next release of RegPrecise we will continue extending the regulon content to cover other diverse taxonomic groups. We are also planning to conduct the large-scale assignment of confidence levels to the predicted regulons based on available experimental evidences from literature and external web-resources (EcoCyc, CoryneRegNet, DBTBS, RegTransBase).

RegPrecise is a key component of the upcoming DOE Systems Biology KnowledgeBase. It will provide essential datasets of reference regulons in diverse microbes to enable automatic reconstruction of draft TRNs in newly sequenced genomes.

This research is supported by the Genomic Science Program (GSP), Office of Biological and Environmental Research (OBER), U.S. Department of Energy (DOE), under contract DE-SC0004999 with Sanford-Burnham Medical Research Institute and Lawrence Berkeley National Lab.

234

Reference Collection of Transcriptional Regulons in Bacillales

Semen A. Leyn[1,2]* (sleyn@burnham.org), Marat D. Kazanov,[2] **Pavel S. Novichkov**,[3] and **Dmitry A. Rodionov**[1,2]

[1]Sanford-Burnham Medical Research Institute, La Jolla, Calif.; [2]Institute for Information Transmission Problems RAS, Moscow, Russia; and [3]Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://regprecise.lbl.gov/

Project Goals:

1. **Develop integrated platform for genome-scale regulon reconstruction utilizing the comparative genomics approach and analysis of RNA and DNA regulatory sites.**
2. **Infer the regulatory interactions and reconstruct transcription regulatory networks in several groups of microbial species important for DOE mission.**
3. **Develop RegKnowledgeBase on microbial transcriptional regulation.**

Gram-positive facultative anaerobic bacteria from the Bacillales order were isolated from diverse habitats including soil, sea water, plants and animals. Bacillales use various strategies to respond and survive in a variety of stresses and environmental conditions including resistance to multiple antibiotics. *Bacillus subtilis* str. 168 is one of the best-characterized Gram-positive bacteria and a model organism for studying sporulation, cell differentiation, stress response and social behavior of bacteria. According to the DBD database, *B. subtilis* genome encodes 238 DNA-binding transcription factors (TFs) classified in 45 protein families. Of them, 120 TFs were studied experimentally and the respective regulatory interactions were captured in the DBTBS database. However, many of the previously studied TF regulons were studied insufficiently, providing an incomplete knowledge on the range of target genes and associated TF-binding sites (TFBSs).

Continuously growing number of available complete genomes allows successful application of the comparative genomics approaches for regulon analysis. We used a "knowledge-driven" approach, which combines the accumulated experimental information from literature and databases with novel bioinformatics tools for genomic reconstruction of regulatory interactions. We perform the comparative genomics reconstruction of regulons operated by either TFs or RNA-regulatory element using the RegPredict Web-server (http://regpredict.lbl.gov). RegPredict allows prediction of TFBS and RNA motifs in a group of selected genomes, with further identification and annotation of candidate members of the respective regulons. Functional analysis of target genes was based on annotations from the SEED database that were validated by the genomic context analysis in MicrobesOnLine.

In this study we carried out large-scale comparative genomics analysis of regulatory interactions in *B. subtilis* and 10 related species from the Bacillales order. For TF regulons, we first analyzed 59 regulons with previously known TFBS motifs according to literature and the DBTBS database on transcriptional regulation in *B. subtilis*. These known regulons were expanded by prediction of novel targets in *B. subtilis* and propagated to other studied genomes of Bacillales, resulting in refinement of TFBS motifs and identification of novel regulon members. Then we predicted novel TFBS motifs and reconstructed 32 TF regulons for which target genes have been previously defined in *B. subtilis* but whose TFBSs were unknown. Finally, we discovered novel TFBS motifs and reconstructed regulons *de novo* for 34 previously uncharacterized TFs. Novel regulons involve genes from the following biological processes: utilization of various carbohydrates; metabolism of glutamate, histidine and thiamine; stress responses; drug/metabolite transport. Totally, more than 3500 TFBSs have been predicted in the *Bacillales* group (from 200 to 600 sites per genome).

For RNA-operated regulons, we used bacterial RNA regulatory motifs collected from the Rfam database, scanned the studied genomes with these motifs to identify new occurrences of each RNA family, and annotated the respective target operons. Among 37 reconstructed RNA regulons there are 11 families of metabolite-sensing riboswitches, 17 types of aminoacyl-tRNA-responsive T-boxes, one regulon controlled by the RNA-binding protein PyrR, five predicted ribosomal protein leader RNA structures and five regulons for *in silico* predicted RNA motifs of unknown function. The reconstructed RNA motif-operated regulons in Bacillales control key metabolic pathways including biosynthesis of vitamins and cofactors (cobalamin, riboflavin, thiamine, nucleoside queuosine), biosynthesis of glucosamine, metabolism of most amino acids, biosynthesis and salvage of purines and pyrimidines, and magnesium homeostasis.

The reference collection of transcriptional regulons in the Bacillales group of bacteria is available in the RegPrecise database (http://regprecise.lbl.gov/). Currently, this collection constitutes the biggest transcriptional regulatory network among various taxonomic groups with reconstructed TRNs in RegPrecise.

This research is supported by the Genomic Science Program (GSP), Office of Biological and Environmental Research (OBER), U.S. Department of Energy (DOE), under contract DE-SC0004999 with Sanford-Burnham Medical Research Institute and Lawrence Berkeley National Lab.

235

Community-Based Approach for Genome-Wide Regulon Annotation in Bacteria

Dmitry A. Ravcheev[1,2]* (dravcheev@sanfordburnham.org), Aaron A. Best,[3] Mikhail S. Gelfand,[2,4] **Pavel S. Novichkov**,[5] and **Dmitry A. Rodionov**[1,2]

[1]Sanford-Burnham Medical Research Institute, La Jolla, Calif.; [2]Institute for Information Transmission Problems RAS, Moscow, Russia; [3]Hope College, Holland, Mich.; [4]Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia; and [5]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.

http://regprecise.lbl.gov

Project Goals:

1. **Develop integrated platform for genome-scale regulon reconstruction utilizing the comparative genomics approach and analysis of RNA and DNA regulatory sites.**
2. **Infer the regulatory interactions and reconstruct transcription regulatory networks (TRNs) in several groups of microbial species important for DOE mission.**
3. **Develop RegKnowledgeBase on microbial transcriptional regulation.**

In the light of the constantly growing number of complete genomic sequences accurate genome-scale annotation of regulatory features is a one of the critical task of modern genomics and system biology. Inference of regulatory interactions in newly sequenced genomes allows reconstruction of the transcriptional regulatory networks (TRNs) and effective modeling of cellular metabolic and signal pathways. TRN is a fine-tuned system including controllers, such as transcription factors (TFs), their binding sites (TFBSs), and RNA regulatory elements (e.g., riboswitches), and sets of target genes whose expression is regulated by these controllers. Genome-scale reconstruction of TRNs in bacteria requires accurate prediction of a large number of regulatory interactions between controllers and their targets. A regulon, defined as a set of genes under the direct control of a certain controller, constitutes a 'building block' of each TRN. The comparative genomics approach was successfully applied for reconstruction of multiple regulons in diverse bacterial groups. A regulog concept is used to represent a regulon inferred and projected in a group of closely-related bacterial genomes. High-quality reconstruction of each regulog constituting the combined TRN in a group of taxonomically-related genomes is a labor intensive work that can be carried out by the scientific community.

We developed a workflow for coordinated reconstruction of multiple bacterial regulons by the community annotators (Fig. 1). The workflow utilizes the comparative genomics-based approach implemented in the RegPredict Web-server (http://regpredict.lbl.gov). This reconstruction pursues two main tasks: (i) propagation of the known regulon to new genomes; (2) expansion of the known regulon by prediction of new regulon members. Initial datasets of regulatory interactions including information about TFs, the previously mapped TFBSs and sets of regulated genes (mostly experimentally determined) are extracted from literature and public databases, such as DBD, EcoCyc, CoryneRegNet, DBTBS, RegTransBase, and RegPrecise. These data are mapped to complete genomes forming initial sets of tasks for community annotators. Each regulon is reconstructed by a community annotator in the reference set of taxonomically-related genomes using the computational platform RegPredict. After the completion of regulon annotation task, each output regulog undergoes quality control by curators. Finally approved regulogs include information on regulators and their targets in a group of taxonomically-related genomes.

The community-based workflow was efficiently used for initial TRN inference in four taxonomic groups of bacteria, including Enterobacteriaceae, Lactobacillaceae, Streptococcaceae, and Corynebacteriaceae. The reference set of inferred transcription factor regulogs is available in the RegPrecise database (http://regprecise.lbl.gov) and includes 173 regulogs described in 50 genomes.

In the Enterobacteriaceae lineage (12 species including *Escherichia coli*), 64 regulons were reconstructed by a community of 27 undergraduate students from Moscow State University (MSU) in Russia. These include the global regulons ArgR, Crp, Fur, FruR, LexA, and PurR that control from tens to a hundred of targets per genome. The obtained collection of the Enterobacteriales regulogs contains more than 600 target genes, including ~100 new regulon members. Another community of 36 MSU students performed annotation of 43 regulons in the Corynebacteriaceae lineage (8 species including *Corynebacterium glutamicum*). The final set of reconstructed regulons comprises more than 150 target operons, including multiple novel members of regulons. Streptococcaceae and Lactobacillaceae are two closely related lineages of the Firmicutes phylum. Regulon reconstruction in 30 selected genomes from these two taxonomic groups was carried by a community of 18 undergraduate students from Hope College (Holland, MI). As a result of this community effort, 33 regulons were inferred in each of these two lineages. Among them, 22 regulons are shared between the two lineages, whereas the remaining 18 regulons are lineage-specific. Final collections of regulons in Streptococcaceae and Lactobacillaceae contain 200-300 regulated operons per genome with multiple novel predicted regulon members.

The community-based approach for *in silico* reconstruction of multiple regulons applied in this work to four taxonomically diverse groups of bacterial genomes is a promising approach for large-scale annotation of regulatory features. Detailed TRNs obtained by this approach for DOE-mission genomes will constitute an important dataset for the forthcoming DOE Systems Biology KnowledgeBase.

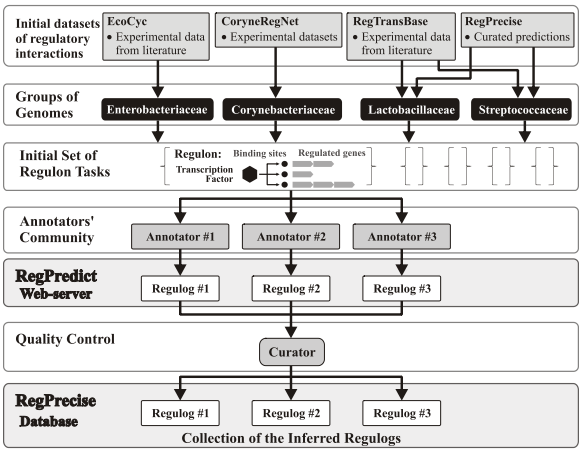


Figure 1. Workflow for genome-wide regulon annotation by scientific community.

This research is supported by the Genomic Science Program (GSP), Office of Biological and Environmental Research (OBER), U.S. Department of Energy (DOE), under contract DE-SC0004999 with Sanford-Burnham Medical Research Institute and Lawrence Berkeley National Lab.

236

Computational Methodologies for Identification of Phenotype-Specific Biological Processes in Microbial Communities

Nagiza F. Samatova[1,2] (samatovan@ornl.gov), **Chongle Pan**[2] (pcl@ornl.gov), **Robert (Bob) L. Hettich**[2] (hettichrl@ornl.gov), **Jill Banfield**[3] (jbanfield@berkeley.edu), Matthew C. Schmidt[1,2] (mcschmid@ncsu.edu), William Hendrix[1,2] (wthendri@ncsu.edu), Andrea M. Rocha[4] (amrocha@mail.usf.edu), and Kanchana Padmanabhan[1,2*] (kpadman@ncsu.edu)

[1]North Carolina State University, Raleigh; [2]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [3]University of California, Berkeley; and [4]University of South Florida, Tampa

http://freescience.org/cs/gtl/

Project Goals: Develop computational methodologies to identify phenotype-related biological systems and their interplays. The crux of the project is a procedure for: (1) identification and characterization of phenotype-related genes; (2) identification of phenotype-biased cellular subsystems; (3) reconstruction of phenotype-specific metabolic pathways; and (4) elucidation of symbiotic/competing crosstalks between these pathways.

Phenotypes that certain microorganisms express assist in activities like breaking down the lignocellulosic barrier of biomass, and the biodegradation of various environmental contaminants. Tackling the various bioremediation and bioenergy problems with the help of genetic engineering requires the understanding of the cellular subsystems that help with the phenotype-expression in the organism. To supplement experimentation methods, computational methodologies need to be used. These methods could reveal phenotype-related "signals" and their combinatorial interplay by comparing potentially hundreds of microorganisms with millions of genes organized into thousands of cellular subsystems.

We developed **graph-theoretical and statistical methods and released open-source software for *in silico* prediction of cellular subsystems related to the expression of a target phenotype**.

The Network Instance-Based Biased Subgraph Search (**NIBBS)** [5] is capable of comparing hundreds of genome-scale metabolic networks to identify ***metabolic subsystems that are statistically biased toward phenotype-expressing organism***. NIBBS identifies the set of all phenotype-*biased* metabolic network motifs. From the results obtained, for example, for bio-hydrogen production phenotype, NIBBS was able to identify metabolic subsystems like acetate and butyrate fermentation, fatty acid biosynthesis, amino acid metabolism, and nitrogen metabolism. The validation for those results was found in literature. In addition, NIBBs was also able to provide clues about pathway cross-talks, including those involved in production of Acetyl-CoA.

The ***α,β*-motif finder** [2] and **bi-clustering** [3] approach allow for identification of ***phenotype-related functional modules*** that, in addition to metabolic subsystems, could include their regulators, sensors, transporters. The functional modules identified are present across a set of phenotype-expressing organisms. This approach can identify conserved modules across any subset of the input organisms and hence can identify sub-phenotype-specific modules as well. From the results obtained, for example, for the light fermentation hydrogen production phenotype, the modules associated with N-fixation, iron regulation, and ammonia uptake were found. When applied to the bio-hydrogen production phenotype, this method was able to identify modules responsible for synthesis, metal insertion, or regulation of hydrogenase and nitrogenase enzymes complexes. Within hydrogen producers, these two complexes play important roles in production of hydrogen. On further analysis, various pathway crosstalks including those between iron and nitrogen related metabolic pathways and iron uptake and ammonia assimilation were predicted.

The Dense ENriched Subgraph Enumeration (**DENSE**) [1] algorithm allows for ***incorporating partial prior knowledge about the proteins involved in a phenotype-related process into the identification of sets of functionally associated proteins in a phenotype-expressing organism***. This method, when applied to the protein functional association network of the *Clostridium acetobutylicum,* a dark fermentative, hydrogen producing bacterium was able to predict known and novel associations including those with regulatory, signaling, and uncharacterized proteins.

We supplemented the computational methodologies discussed so far with a method to computationally analyze the results of the methods for biological significance and improve functional annotation [4]. This method is different from other existing significance analysis techniques in that it takes into account an inherent design principle of biological networks, *hierarchical modularity*. Functionally homogenous modules combine in a hierarchical manner into larger, less cohesive subsystems. The method quantifies biological significance both with a score known as **Hierarchical Modularity Score (HMS)** and a confidence of the score via a *p*-value. Additionally, it provides the hierarchical functional annotation of the modules.

References

1. Willam Hendrix, Andrea M Rocha, Kanchana Padmanabhan, Alok Choudhary, Kathleen Scott, James R Mihelcic and Nagiza F Samatova, *DENSE: efficient and prior knowledge-driven discovery of phenotype-associated protein functional modules*, BMC Systems Biology 2011, 5:172.
2. Mathew C Schmidt, Andrea M Rocha, Kanchana Padmanabhan, Zhengzhang Chen, Kathleen Scott, James R Mihelcic and Nagiza F Samatova, *Efficient alpha,beta-motif Finder for Identification of Phenotype-related Functional Modules*, BMC Bioinformatics 2011, 12:440.
3. Kevin Wilson, Andrea Rocha, Kanchana Padmanabhan, Kuanyu Wang, Zhengzhang Chen, Ye Jin, James R Mihelcic and Nagiza F Samatova, *Detecting Pathway Cross-talks by Analyzing Conserved Functional Modules across Multiple Phenotype-Expressing Organisms*, IEEE International Conference Bioinformatics and Biomedicine, 2011.
4. Kanchana Padmanabhan, Kuanyu Wang, Nagiza Samatova, *Functional Annotation of Hierarchical Modularity,* (under review).
5. Mathew C Schmidt, Andrea M Rocha, Kanchana Padmanabhan, Yekaterina Shpanskaya, Jill Banfield, Kathleen Scott, James R Mihelcic and Nagiza F Samatova, *NIBBS-Search for Fast and Accurate Prediction of Phenotype-Biased Metabolic Systems,* (under review).
6. Zhengzhang Chen, Kanchana Padmanabhan, Andrea M. Rocha, Yekaterina Shpanskaya, James R. Mihelcic and Nagiza F. Samatova, *SPICE: Discovery of Phenotype-Determining Component Interplays*, (under review).

This research is supported by both the Office of Biological and Environmental Research and by the Office of Advanced Scientific Computing Research of the U.S. Department of Energy.

237

OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities

Ali R. Zomorrodi* (zomorrodi@engr.psu.edu) and **Costas D. Maranas**

Department of Chemical Engineering, The Pennsylvania State University, University Park

Project Goals: The goal of this project is to develop an efficient and comprehensive computational framework for the flux balance analysis of microbial communities using genome-scale metabolic models.

Microorganisms rarely live isolated in their natural environments but rather function in consolidated and socializing communities. Despite the growing availability of high-throughput sequencing and metagenomic data, we still know very little about the metabolic contributions of individual microbial players within an ecological niche and the extent and directionality of interactions among them. This calls for development of efficient modeling frameworks to shed light on less understood aspects of metabolism in microbial communities. Here, we introduce OptCom, a comprehensive flux balance analysis framework for microbial communities, which relies on a multi-level/objective optimization formulation to properly describe trade-offs between individual vs. community level fitness criteria (see Figure 1). In contrast to earlier approaches that rely on a single objective function, here, we consider species-level fitness criteria for the inner problem while relying on community-level objective maximization for the outer problem. OptCom is general enough to capture any type of interactions (positive, negative or combinations thereof) and is capable of accommodating any number of microbial species (or guilds) involved.

To quantify the deviation of community members from their optimal behavior, we introduce a metric called 'optimality level' (c^k) for each species involved. The optimality level for each one of the microorganisms is quantified using a variation of OptCom which we refer to as *descriptive* through incorporating all available experimental data for the entire community (e.g., community biomass composition) as constraints in the outer problem and all data related to individual species as constraints in the respective inner problems while allowing the biomass flux of individual species to fall below (or rise above) the maxima of the inner problems. An optimality level of less than one for a microorganism k implies that it grows sub-optimally at a rate equal to $100c^k$ % of its maximum to optimize a community-level fitness criterion while matching experimental observations. Alternatively, an optimality level of one implies that microorganism k grows exactly optimally, whereas a value greater than one indicates that it achieves a higher biomass production level than the community-specific maximum (i.e., super-optimality) by depleting resources from one or more other community members.

We applied OptCom to quantify the syntrophic association between *D. vulgaris* and *M. maripaludis* and assess the optimality levels of growth in phototrophic microbial mats of Octopus and Mushroom Springs of Yellowstone National Park. We also used OptCom to elucidate the extent and direction of inter-species metabolite and electron transfer in a model microbial community and examine the possibility of adding a new member to this community. Our study demonstrates the importance of trade-offs between species- and community-level fitness driving forces and lays the foundation for metabolic-driven analysis of various types of interactions in multi-species microbial systems using genome-scale metabolic models.

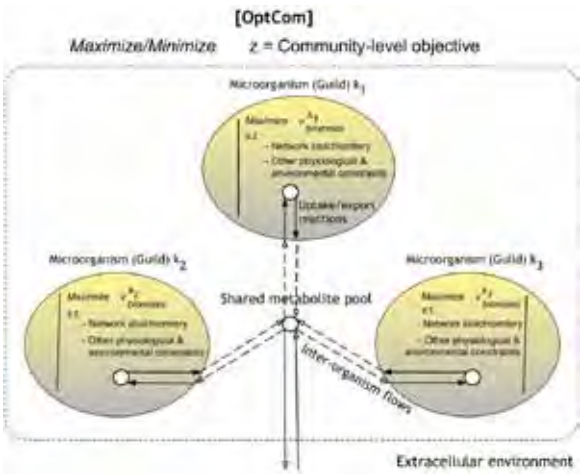


Figure 1. **Pictorial illustration of OptCom.** OptCom relies on a multi-level optimization structure where a separate biomass maximization problem is defined for each species as inner problems. These inner problems are then integrated in the outer stage through the inter-organism flow constraint to optimize a community-level objective function.

238

MetRxn: A Knowledgebase of Metabolites and Reactions Spanning Metabolic Models and Databases

Akhil Kumar[1]* (azk172@psu.edu), Patrick F. Suthers,[2] and **Costas D. Maranas**[2]

[1]Bioinformatics and Genomics, [2]Department of Chemical Engineering, The Pennsylvania State University, University Park

Project Goals: The goal of this project is to develop a comprehensive knowledgebase that standardizes metabolite and reaction entries consolidated from a wide range of databases and existing genome-scale metabolic models.

Increasingly, metabolite and reaction information is organized in the form of community, organism, or even tissue-specific genome-scale metabolic reconstructions. These reconstructions account for reaction stoichiometry and directionality, gene to protein to reaction associations, organelle reaction localization, transporter information, transcriptional regulation and biomass composition. A key bottleneck in the pace of reconstruction of new high quality metabolic models is our inability to directly make use of metabolite/reaction information from biological databases (e.g., BRENDA, KEGG, MetaCyc, EcoCyc, BioCyc, BKM-react, UM-BBD, Reactome.org, Rhea, PubChem, ChEBI etc.) or other models due to incompatibilities of representation, duplications and errors.

A major impediment is the presence of metabolites with multiple names across databases and models, and in some cases within the same resource, which significantly slows down the pooling of information from multiple sources. Therefore, the almost unavoidable inclusion of multiple replicates of the same metabolite can lead to missed opportunities to reveal (synthetic) lethal gene deletions, repair network gaps and quantify metabolic flows. Moreover, most data sources inadvertently include some reactions that may be stoichiometrically inconsistent and/or elementally / charge unbalanced, which can adversely affect the prediction quality of the resulting models if used directly. Finally, a large number of metabolites in reactions are partly specified with respect to structural information and may contain generic side groups (e.g., alkyl groups -R), varying degree of a repeat unit participation in oligomers, or even just compound class identification such as "an amino acid" or "electron acceptor".

MetRxn is a knowledgebase that includes standardized metabolite and reaction descriptions by integrating information from BRENDA, KEGG, MetaCyc, Reactome.org and 44 metabolic models into a single unified data set. All metabolite entries have matched synonyms, resolved protonation states and are linked to unique structures. All reaction entries are elementally and charge balanced. This is accomplished through the use of a workflow of lexicographic, phonetic, and structural comparison algorithms (see Figure 1). MetRxn allows for the download of standardized versions of existing genome-scale metabolic models and the use of metabolic information for the rapid reconstruction of new ones.

We describe the development and highlight applications of the web-based resource MetRxn that integrates, using internally consistent descriptions, metabolite and reaction information from 8 databases and 44 metabolic models. The MetRxn knowledgebase contains over 76,000 metabolites and 72,000 reactions (including unresolved entries) that are charge and elementally balanced. By conforming to standardized metabolite and reaction descriptions, MetRxn enables users to efficiently perform queries and comparisons across models and/or databases. For example, common metabolites and/or reactions between models and databases can rapidly be generated along with connected paths that link source to target metabolites. MetRxn supports export of models in SBML format. New models are being added as they are published or made available to us. MetRxn uses relational database models (MySQL) and is available as a web-based resource at http://metrxn.che.psu.edu.

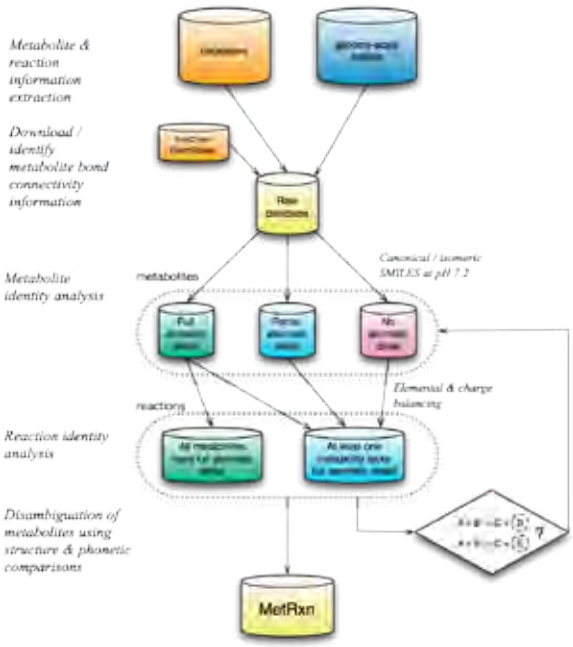


Figure 1: Flowchart outlining the construction of MetRxn.

239

Computational Design of Bioenergy-Related Metabolic Pathways

Mario Latendresse* (latendre@ai.sri.com), Mike Travers, and **Peter Karp**

SRI International, Menlo Park, California

http://metacyc.org
http://brg.ai.sri.com/ptools

Project Goals: The general goal of our project is to support computational design of metabolic pathways for metabolic engineering applications. Our approach combines the following elements. (1) Users will specify a target metabolite, a feedstock compound, and other constraints on their design problem. (2) A pathway search algorithm will construct alternative pathways by combining reactions from the MetaCyc database. MetaCyc is a highly curated multi-organism metabolic database that currently contains more than 9500 metabolic reactions. The search algorithm will rank pathways according to multiple criteria. (3) Users will view the output of the pathway search algorithm using a graphical interface that facilitates user comprehension and evaluation of the pathways. The resulting software will be an add-on module to SRI's Pathway Tools software.

Pathway Search
We have implemented an initial version of a graph search algorithm to design novel pathways, given a source and target metabolite. This software is capable of integrating a variety of metrics and filters that affect the search. Metrics implemented to date include atom conservation, molecular similarity, avoiding certain molecules, penalties for using a reaction that is outside of the taxonomic range of the target organism, and penalties for generating or consuming certain kinds of side products/reactants. The search engine is capable of using reactions from any PGDB including MetaCyc. The results of the search can be displayed using the web-based reaction atom-mapping viewer, a new capability of Pathway Tools that can display the trajectories of individual atoms through a sequence of reactions.

Future planned work includes adding new search metrics (such as thermodynamic constraints), and an enhanced user interface that allows scientists to interactively view and control the search process.

Automated Computation of Reaction Atom Mappings
An atom mapping describes explicitly the one to one transfer of each atom from the reactants of a reaction to its products. A reaction might have several chemically valid mappings due to the symmetries of reactants and products or for other reasons. These mappings allow the computation of the flow of essential atoms from source to target metabolites in a pathway. Note that a typical reaction equation used by biologists or chemists does not describe the atom mapping, as such details are difficult to provide and might be overwhelming to the reader. When provided, atom mappings are typically described as supplementary data for each reaction.

Some previous work on computing atom mappings does not compute all possible atom mappings and is therefore incomplete. For example, the KEGG RPAIR database describes only one atom mapping per reaction.

As far as we know, all computational approaches to atom mapping published so far, are combined with a post-processing step involving manual curation. That is, a scientist reviews the computed atom mappings for possible errors and appropriate corrections are applied when necessary. But the necessary corrections are not applied to the computational approach itself to avoid future computed errors.

In the approach we have taken, all the chemically valid mappings of each reaction of MetaCyc are computed. Moreover, we aim to have a computational approach that does not require manual post-processing.

Technically, our approach is based on mixed integer linear programming: for each reaction, a linear program is generated from all possible valid mappings of the reaction where the objective is to minimize the sum of the costs of the bonds broken and made. We currently use bond costs that have been determined by a chemist. A linear solver solves the linear program, giving all possible optimal solutions, that is, all possible correct mappings for one reaction. This technique has been applied to almost all the reactions of MetaCyc. It is more computationally efficient than all other known techniques published so far.

The computation of correct mappings depends on the bond costs used. In the near future we intend to systematically

validate these bond costs using a linear program based on a sample of correct mappings (positive examples) and incorrect mappings (negative examples). We also plan to include atom mappings in future versions of MetaCyc and of other Pathway/Genome Databases.

This work was funded by the Department of Energy under grant DE-SC0004878.

240

Bioenergy Curation in the MetaCyc Database of Metabolic Pathways

Ron Caspi* (caspi@ai.sri.com), Deepika Weerasinghe, and **Peter Karp**

Bioinformatics Research Group, SRI International, Menlo Park, Calif.

http://metacyc.org
http://biocyc.org

Project Goals: Two goals of our project are (1) To expand the coverage of bioenergy-related metabolic information in the MetaCyc database and (2) To generate within the BioCyc database collection organism-specific PGDBs for all energy-relevant organisms sequenced by JGI.

MetaCyc (metacyc.org) is a literature-curated database containing more than 1,800 metabolic pathways, collected from a wide variety of organisms, with an emphasis on microorganisms and plants. The goal of MetaCyc is to contain a representative sample of every experimentally elucidated pathway, thereby cataloging the universe of known pathways and enzymes. MetaCyc contains rich, detailed, and high quality data, including minireview summaries with extensive literature citations, enzyme information, evidence codes, and links to other databases.

In addition to its role as an encyclopedic resource for metabolic pathways, MetaCyc also serves as a reference database for the Pathway Tools software, which predicts the metabolic pathway complement of an organism from its annotated genome, creating a Pathway/Genome Database (PGDB) for that organism.

Since only pathways that already exist in MetaCyc can be inferred by Pathway Tools in organism-specific databases, the pathway content of MetaCyc strongly influences the metabolic networks predicted by Pathway Tools. Two goals of our project are (1) To expand the coverage of bioenergy-related metabolic information in MetaCyc, and (2) To generate within our BioCyc database collection organism-specific PGDBs for all energy-relevant organisms sequenced by JGI.

During the past year we have done the following:

- Greatly expanded the coverage in MetaCyc of compounds found in cellulosic biomass, with an emphasis on natural cellulosic and hemicellulosic polymers.
- Created 8 pathways for the degradation of important biomass compounds such as celluloses, rhamnogalacturonans, xylans, arabinans, xyloglucans, and carrageenans.
- Curated many enzymes that are involved in cellulosic biomass degradation.
- Curated eight naturally occurring and/or bioengineered hydrogen production pathways and related enzymes.
- Curated ten naturally-occurring and/or engineered biosynthetic pathways for potential biofuel compounds, including 1-butanol, 3-methyl butanol, isopropanol, all-trans-farnesol, long chain fatty acid esters, the algal lipid diacylglyceryl-N,N,N-trimethylhomoserine, the algal fatty acids docosahexanoate, arachidonate, and eicosapentaenoate, and the algal triterpinoid botryococcene.

In the remainder of the second year of this project we will add a new type of pathway diagram that will facilitate the presentation of enzymatic degradation of complex polymers, and will use this new tool to create many pathway diagrams for lignocellulose-degradation pathways.

Reference

1. Caspi et al (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. Database Issue[Epub ahead of print] PMID 22102576.

MetaCyc and BioCyc are funded by grant GM80746 from the NIH National Institute of General Medical Sciences. Bioenergy curation is funded by grant DE-SC0004878 from the Department of Energy.

241

Enhancing the SEED Framework for Curation and Analysis of Genomic Data and Genome-Scale Metabolic Models

Christopher Henry*[1] (chenry@mcs.anl.gov), **Ross Overbeek**,[2] Robert Olson,[1] Terrence Disz,[1] Bruce Parello,[2] Rob Edwards,[3] Matt DeJongh,[4] Aaron Best,[4] Fangfang Xia,[1] Scott Devoid,[1] and **Rick Stevens**[1]

[1]University of Chicago, Ill.; [2]Fellowship for Interpretation of Genomes, Ill.; [3]San Diego State University, Calif.; and [4]Hope College, Mich.

Project Goals: The objective of the SEED Knowledgebase project is encompassed by four goals: (1) enhance the computational infrastructure behind the SEED framework to improve extensibility, accessibility, and scalability; (2) integrate extensions into the SEED database to accommodate new computational and experimental data

types including regulatory networks, biochemistry and models, eukaryotic genome data, and growth phenotype data; (3) develop web services providing access to all SEED data and algorithms including genome annotation, model reconstruction, flux balance analysis, and data query and access; and (4) apply the SEED framework to annotate and model organisms with applications to bioenergy, carbon cycle, and bioremediation. This project will provide users with programmatic access to SEED data and algorithms, it will produce new models of bioenergy organisms, and it will enable integration of expression data and regulation with annotations and models.

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. In the SEED Knowledgebase project, we have endeavored to enhance the SEED environment in three areas: (1) development of a web API offering programmatic access to SEED data and algorithms (including trees and expression data); (2) extension of the SEED interface to enable curation of public genome annotations by registered users; and (3) reconstruction and analysis of metabolic models for all available prokaryotic genome sequences.

1. Web API for Programmatic Access to Data and Algorithms
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; enabling users to query the SEED database for genome features, functional annotations, gene orthologs, and sequence similarities; and enabling 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. Specifically, we apply the web services to: (1) identify new genome annotations based on model gapfilling; (2) identify commonly clustered and co-expressed sets of functional roles across all known genomes; and (3) to study redundancy of essential metabolic functions across all known genomes.

2. Extension of the SEED to Enable Curation of Public Genomes
We recently launched a new version of the SEED website called the Public SEED (pubseed.theseed.org). This new site now contains over 3000 annotated prokaryotic genomes, and it is continuously updated to include every complete prokaryotic genome sequence that is available in GenBank. In addition to offering a greatly enhanced database of genomes, the Public SEED also offers improved access to genomes, through a powerful new search feature. This feature enables users to rapidly search Public SEED content for organism names, gene names, locus IDs, and many other queries and returns results sorted by object type. The Public SEED also offers a unique ability enabling registered users to alter any genome annotation in the Public SEED, providing a unique resource for the community annotation of genomic data.

3. Reconstruction and Analysis of Metabolic Models for all Prokaryotic Genomes
Over the past decade, 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. We have applied the Model SEED framework to produce 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 and improve the automated generation of biomass composition reactions. These algorithms were applied to assess the quality of annotations in the SEED framework, and to identify high-priority gaps to filled in these annotations. Finally, we applied 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.

242

Tools and Approaches for Integrating Multiple Genetic and Cellular Networks

Gang Fang,[1] Darryl Reeves[1]* (darryl.reeves@yale.edu), Koon-Kiu Yan,[1] Nitin Bhardwaj,[1] Declan Clarke,[1] Chao Cheng,[1] Pedro Alves,[1] Sergei Maslov,[2] and **Mark Gerstein**[1]

[1]Computational Biology and Bioinformatics, Medical School, Yale University, New Haven, Conn. and [2]Department of Condensed Matter Physics and Materials Science, Brookhaven National Laboratory, Upton, N.Y.
http://networks.gersteinlab.org/

Project Goals: Our overall goals of this project are the development of tools for the analysis of networks and pathways in plants and microorganisms related to enable the Systems Biology Knowledgebase proposed by the DOE.

One important task of the future KBase is to provide a platform to help users analyze gene biological function and inspire experiments for the purpose of biofuel development. A gene's biological function is essentially its relationships or interactions with other biological objects within the cell and around the environment. It is necessary to understand

gene function on a genomic scale, and from the integration of genetic and cellular networks. We approach this through the prediction and analysis of biological networks, focusing on protein-protein and transcription-factor-target interactions. We describe how these networks can be determined through integration of many genomic features and how they can be analyzed in terms of various topological statistics. In particular, we will report a number of recent analyses: (1) Improving the prediction of molecular networks through systematic training-set expansion; (2) Showing how the analysis of pathways across environments potentially allow them to act as biosensors; (3) Analyzing the structure of the regulatory network which indicates that it has a hierarchical layout with the "middle-managers" acting as information bottlenecks; (4) Showing these middle managers tend be arranged in various "partnership" structures giving the hierarchy a "democratic character"; (5) Comparing the topology and variation of the regulatory network to the call graph of a computer operating system; (6) Developing a framework to integrate together various kinds of biological networks (e.g. relating to TFs and miRNAs) into an integrated meta-network; (7) Integrating this meta-network with actual molecular structures; and (8) Creating practical web-based tools for the analysis of these networks (DynaSIN and tYNA).

We acknowledge funding from the DOE DE-SC0004856 grant.

243

Biological Significance of Gene Modules in an *Arabidopsis thaliana* Co-Expression Network

Darryl Reeves[1]* (darryl.reeves@yale.edu), Gang Fang,[2] and **Mark Gerstein**[1,2,3]

[1]Program in Computational Biology and Bioinformatics, [2]Department of Molecular Biophysics and Biochemistry, and [3]Department of Computer Science, Yale University, New Haven, Conn.

Project Goals: Construction of a biologically meaningful co-expression network for *Arabidopsis thaliania* for use in the Systems Biology Knowledgebase (KBase).

Gene expression is subject to environmental and cell growth conditions. Gene co-expression modules which are associations between gene expression and experimental perturbations and/or cell phenotypes provide important clues for understanding gene biological function. Such modules can assist researchers in the design of additional experiments for identifying favorable genes and metabolic pathways for biofuel development. Thousands of gene expression datasets, mainly derived from microarray experiments, presenting a large range of conditions are publicly available now, and will be integrated into KBase. It is necessary to develop an online toolset to allow users to query, display and make *in silico* analysis of expression data. As one of the major packages for analyzing gene expression, WGCNA[1] (weighted correlation network analysis) will be used to build gene co-expression modules and to cluster cell phenotypes in KBase. In this poster, we present a work-flow showing the steps and options to perform this analysis.

Based on the test run of a number of *Arabidopsis thaliana* datasets, we showed how the choice of tree cutting algorithm affects co-expression module generation and, ultimately, functional annotation of the modules. The clustered co-expression network was grouped into modules using three tree-cutting methods provided by the WGCNA package: static, dynamic, and hybrid. We tested each method and optimized their module generation. The criteria for optimization were how enriched modules are for Gene Ontology (GO) terms and KEGG pathways. The goal is to help users determine biological significance of the modules generated by each set of parameters tested.

Reference

1. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.

244

Dynamic Metabolic Model Building Based on the Ensemble Modeling Approach

Jimmy G. Lafontaine Rivera[1]* (lafonj@gmail.com), Ali R. Zomorrodi[2]* (zomorrodi@engr.psu.edu), Thomas Wasylenko[3]* (tmwasylenko@gmail.com), **Costas D. Maranas**,[2] **Greg Stephanopoulos**,[3] and **James C. Liao**[1]

[1]Department of Chemical and Biomolecular Engineering, University of California, Los Angeles; [2]Department of Chemical Engineering, The Pennsylvania State University, University Park; [3]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge

Project Goals: The goal of this project is to develop a novel modeling approach to describe the dynamic behavior of metabolic systems (in particular, flux changes upon enzyme tuning) by integrating multiple data platforms including flux, metabolite, transcriptome, and enzyme tuning data. Although the utility of such models is undeniable, their development has been impaired by inadequate modeling approaches, the sheer size of the problem, and difficulties in accessing the intracellular environment. As a result, little progress has been made in realizing such dynamic models despite the continuously increasing number of intracellular measurements that are becoming available by high throughput methods. The resulting models from the proposed research will account for pathway enzyme kinetics and aim to predict the effects of genetic manipulations designed to bring about changes in metabolic flux and overproduction of metabolites, such as tuning various enzyme levels or the Michaelis-Menten constants (K_m) of key enzymes. In this context, such models will be instrumental for constructing microbial strains to produce various biofuels such as ethanol, 1-butanol, and isobutanol from renewable resources. We will use

production of these fuels in *Escherichia coli* as a model system, because of *E. coli*'s central role as a test bed in systems biology, the wealth of kinetic and regulatory information available and its successful usage for the production of biofuels. While the *E. coli* focus will facilitate model development, the approach developed will be general and applicable to other microorganisms and eventually plants. The project is based on the Ensemble Modeling (EM) approach, robust flux and metabolite measurements, and an efficient optimization scheme developed in the PIs' laboratories.

Presently, there are no satisfactory dynamic models of cellular function. This unique deficiency persists despite recent advances in the areas of high throughput measurement of cell-wide intracellular biomolecules and molecular level simulations of various systems. Current approaches for creating dynamic models of cellular function attempt to do so by fitting transient metabolite concentration data to various kinetic rate expressions. These data are difficult to obtain and often have large experimental errors, making it impossible to scale up to the levels required by our current understanding of cellular metabolism. The Ensemble Modeling (EM) framework was proposed to address this issue by relying only on steady-state measurements (although transient measurements can also be used) to create accurate kinetic models of cellular metabolism. Here we introduce the EM framework and show how it can be integrated with robust flux measurement techniques and efficient optimization schemes in order to arrive at such models.

245

Reliable Numerical Methods for FBA and FVA

Yuekai Sun (yuekai@gmail.com)*,[1] Ronan M.T. Fleming,[2] and **Michael A. Saunders**[3]

[1]ICME, Stanford University; [2]Center for Systems Biology, University of Iceland, Iceland; and [3]Department of Management Science and Engineering, Stanford University

http://www.stanford.edu/group/SOL/

Project Goals:

1. **Develop algorithms for solving optimization problems involving large stoichiometric matrices. (a) Extend existing sparse linear programming algorithms to enable the solution of such systems, in which the matrix coefficients represent reactions at multiple timescales and thus vary over many orders of magnitude. (b) Develop a convex optimization algorithm for computing thermodynamically feasible reaction fluxes in a general instance of a genome-scale integrated metabolic and macromolecular biosynthetic network. (c) Implement a parallel convex optimizer in to enable sampling of the thermodynamically feasible set. (d) Disseminate software to the systems biology community.**
2. **Investigate cyclic dependency between metabolic and macromolecular biosynthetic networks. (a) Predict the material and energy cost of macromolecular synthesis in an integrated metabolic, transcriptional and translational model of *Escherichia coli*. (b) Reconstruct and analyze the macromolecular synthesis network of *Thermotoga maritima*.**
3. **Quantify the significance of thermodynamic constraints on prokaryotic metabolism. (a) Simultaneous prediction of metabolic fluxes and concentrations in *Escherichia coli*. (b) Validate and interpret flux and concentration prediction in *Escherichia coli* and *Thermotoga maritima*. (c) Predict thermodynamically favorable pathways for hydrogen production by *Thermotoga maritima* on a range of substrates. (d) Numerically sample mass conserved, thermodynamically feasible steady state fluxes and concentrations in *Escherichia coli*.**

Concerning project goal 1(a):
Integrated networks of organism metabolism and expression are inherently multi-scale because typical fluxes can vary over eight orders of magnitude. Such networks require special methods to analyze them accurately, and naive use of off-the-shelf optimization software for flux balance analysis (FBA) can produce severely inaccurate solutions. We describe methods for obtaining greater reliability.

The multi-scale nature of integrated networks is also problematic for flux variability analysis (FVA). The traditional FVA formulation sacrifices sharpness in the calculated bounds to ensure feasibility of the sequences of linear programs. In practice the bounds can be off by orders of magnitude. We describe an FVA formulation that guarantees both feasible linear programs and sharp calculated bounds.

246

A Systems Biology Knowledgebase and Analysis Platform for *De Novo* Phenotype Predictions, Integrated Omics Analysis, and Iterative Model Improvement

Edward J. O'Brien[1]* (ejobrien@ucsd.edu), Joshua A. Lerman,[1] Roger L. Chang,[1] Daniel R. Hyduke,[1] Karsten Zengler,[1] Bernhard Ø. Palsson,[1] and **Michael A. Saunders**[2]

[1]Department of Bioengineering, University of California–San Diego, La Jolla and [2]Department of Management Science and Engineering, Stanford University, Stanford, Calif.

http://systemsbiology.ucsd.edu

Project Goals: This project aims to: (1) reconstruct and refine genome-scale models of macromolecular synthesis and metabolism for *Thermotoga maritima* and *Escherichia*

***coli* (2) develop modeling conventions and simulation procedures for phenotype prediction with the integrated models, (3) guide the development of algorithms capable of finding optimal steady-state solutions despite the large, sparse, and ill-scaled constraint matrices, (4) use these prototype models to design a software platform and standard operating procedures to reconstruct general multi-subsystem stoichiometric models, (5) develop and enable methods to parameterize and constrain the network with biophysical models and the direct mapping of diverse omics data, and (6) classify the failure modes of the model to prioritize subsystems and regulatory circuits for model expansion.**

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 metabolic models for both *de novo* phenotype prediction and analysis of omics datasets. To this end, constraint-based algorithms and model-driven omics analysis have provided insight into gene regulation, adaptive evolution, microbial communities, metabolic engineering, and drug response and design. While metabolic models have proven to be a powerful tool, the genetic content of these models is formalized by a Boolean mapping between genetic loci and metabolic reactions, termed the gene-protein-reaction (GPR) relationship. This modeling paradigm only allows for heuristic analysis of transcriptomic and proteomic data and Boolean predictions of genetic requirements.

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 gene Expression). An analogous macromolecular synthesis reconstruction has also been completed for *Thermotoga maritima*, which relied heavily on experimental refinement of the transcription unit architecture. The metabolic and the macromolecular synthesis networks have subsequently been merged into integrated models (termed the 'ME' matrix, for Metabolism and Expression), which allow for explicit analysis and simulation of transcriptomes and proteomes in the context of the underlying reaction network. Not only does inclusion of gene expression increase the scope of the model, but the interdependency of gene expression and metabolism also constrains and refines the metabolic solution space, leading to more accurate predictions.

The ME model formulation additionally 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 to invest its finite resources to synthesize them. This framework is shown to capture known trends in the cellular composition of RNA and protein at various growth rates. Allowing for variable cell composition makes the model more generally applicable to diverse environments whereas the biomass objective function tunes the model to a particular condition.

Our experience reconstructing and analyzing these prototype models made it clear that future model refinement, expansion to other subsystems, omics analysis, and reconstructions for other organisms requires a new software platform and standard operating procedures (SOPs). Analogous software and SOPs exist for metabolic models, but certain features of multi-subsystem models necessitate a redesign of the reconstruction framework including: 1) an order of magnitude larger number of reactants and reactions, 2) many different types of molecules (e.g. rna, proteins, metabolites), 3) 'template' reactions for common cellular processes, 4) efficiency parameters coupling different subsystems, 5) the need for direct integration with growing omics datasets, and 6) future expansions to include other cellular processes and subsystems. Due to these challenges and opportunities, we have developed a database-driven solution to reconstruct, query, and generate these multi-subsystem models. We have designed this software to directly couple with model simulation and analysis software, and have included features and flexibility to address the points listed above.

Future efforts will focus on refining model parameters and constraining condition-specific flux variables with omics datasets and biophysical models. Model improvement will then be mediated by the identification of failure modes to prioritize scope expansions (including signaling and regulatory interactions). We demonstrate the promise for long-term applications of this type of model for metabolic and protein engineering, interpretation of adaptive evolution, and analysis of cellular regulation and optimality.

Grant Information: Numerical Optimization Algorithms and Software for Systems Biology (DOE Award DE-SC0002009)

247

Elucidation of Distinct Transcriptional Regulatory Logic in Bacteria: Applications of a Constraints-Based Systems Biology Knowledgebase

Ali Ebrahim* (aebrahim@ucsd.edu), Stephen A. Federowicz, 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 tran-

scriptional regulatory network with metabolic and macromolecular synthesis models, and (5) provide a platform for genome scale metabolic engineering and synthetic design.

Constructing a systems biology knowledgebase requires the synthesis of a number of critical components into a single platform to allow for diverse computations and analyses. These components include 1) an underlying data model and database 2) tools for integrating, comparing, and analyzing diverse sets of data, and 3) a computational model which mathematically relates the underlying biochemical information. The knowledgebase is built using an iterative workflow: biological experiments are performed to generate data, data is analyzed and its results are integrated into a model, and the model is used to direct future experiments. Here we detail the process of all three steps and display the biological insight gained through the successful utilization of steps one and two.

We first performed an integration and analysis of ChIP-chip, gene expression, and transcription start site (TSS) data obtained at the genome scale for *Escherichia coli*. These specific experiments allow for the elucidation of distinct logical programs and genome scale regulatory mechanisms. These two pieces can then be combined to build a comprehensive model of transcriptional regulation. Logical programs are executed by bacteria in response to common environmental signals or physiological shifts and often include a small molecule signal and associated transcription factor. Initial studies of amino acid metabolism and the transcription factors ArgR and Lrp revealed that arginine and leucine can act as signaling molecules to regulate the transport, biosynthesis, or utilization of 16 amino acids[2]. Similar network motifs governing the flow of an effecter molecule were also shown for purine metabolism and the transcription factor PurR[1]. This has led us to investigate the aerobic-anaerobic shift regulated by ArcA and Fnr, along with the phenomenon of catabolite repression regulated by Crp and Cra. Obtaining a holistic understanding of these systems along with amino acid metabolism allows us to gain an understanding of regulation in response to carbon, nitrogen, and electron acceptor shifts.

In addition to systems level regulation it is possible to gain an understanding of specific regulatory mechanisms at the genome scale in the form of diverse promoter architectures and transcription factor mediated bidirectional transcription. Here we show how promoter architectures occurring in both a unidirectional and bidirectional fashion confer patterns of activation and repression on associated transcription units. Combining this information with systems level principles forges a tie between network and sequence level mechanisms to provide a powerful modeling framework. Utilization of this framework for engineering and synthetic approaches promises to enable a new era of molecular engineering. Overall, the integrated knowledgebase enables a wide range of analysis and is poised to guide future experiments in a model driven fashion towards a comprehensive understanding of transcriptional regulation.

References

1. Cho B-K, Federowicz S, Embree M, Park Y-S, Kim D, Palsson BØ. 2011. The PurR regulon in *Escherichia coli* K-12 MG1655. Nucleic acids research. 39(15):6456-64
2. Cho B-K, Federowicz S, Park Y-S, Zengler K, Palsson BØ. 2011. Deciphering the transcriptional regulatory logic of amino acid metabolism. Nature Chemical Biology. 8(1):65-71

248

Multi-Scale Spatially Distributed Simulations of Microbial Ecosystems

William J. Riehl,[1] William Harcombe,[2] Niels Klitgord,[3] Amrita Kar[1]* (akar@bu.edu), Pankaj Mehta,[4] Nathaniel C. Cady,[5] Christopher J. Marx,[2] and **Daniel Segrè**[1,6]* (dsegre@bu.edu)

[1]Graduate Program in Bioinformatics, Boston University, Boston, Mass.; [2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Mass.; [3]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, Calif.; [4]Department of Physics, Boston University, Boston, Mass.; [5]College of Nanoscale Science and Engineering, University at Albany, Albany, N.Y.; and [6]Department of Biology and Department of Biomedical Engineering, Boston University, Boston, Mass.

http://comets.bu.edu

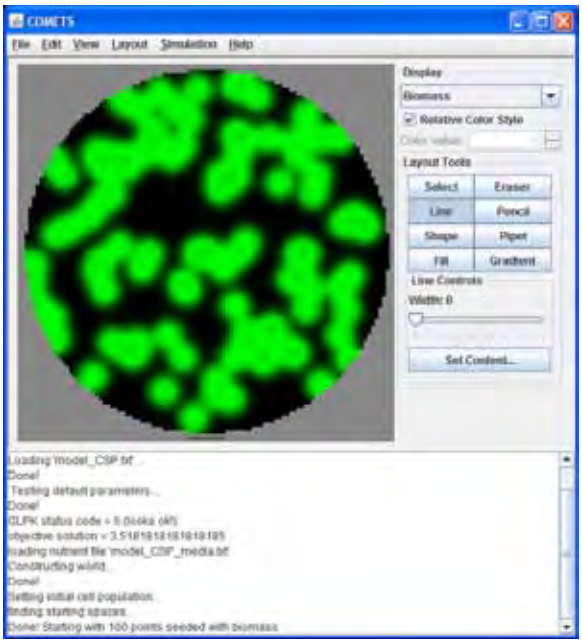
Project Goals: The goal of this project is to develop a tool for facilitating simulation, validation and discovery of multiscale dynamical processes in microbial ecosystems. Our Computation Of Microbial Ecosystems in Time and Space (COMETS) is a an open-source platform for performing spatially distributed time-dependent flux balance based simulations of microbial metabolism. Our plan involves building the software platform itself, calibrating and testing it through comparison with experimental data, and integrating simulations and experiments to address important open questions on the evolution and dynamics of cross-feeding interactions between microbial species.

COMETS (Computation Of Microbial Ecosystems in Time and Space) is a broadly applicable and user-friendly platform for modeling metabolic interactions between microbial species. This platform builds on dynamic flux balance analysis (dFBA [1]) to perform time-dependent metabolic simulations of microbial ecosystems, bridging the gap between stoichiometric and environmental modeling. Simulations occur on a spatially structured lattice of interacting metabolic subsystems, representing a level of detail that is intermediate between fine-grained single-cell modeling and a global population modeling approach.

The current version of COMETS incorporates three fundamental steps: *(i)* Implementation of cellular growth (increase of biomass), using a hybrid kinetic-dFBA solver for every point in the 2D lattice. Upper bounds on uptake

fluxes for the dFBA calculation are estimated based on a concentration-dependent saturating function, in analogy with Michaelis-Menten kinetics. Each grid point may contain biomass for an arbitrary number of different species; *(ii)* Advance of the front of biomass, which we treat as an incompressible fluid (in analogy to [2]). This step involves the solution of the Laplace equation, followed by a calculation of the new biomass front using a level set method; *(iii)* Implementation of a finite differences approximation of the diffusion equation for modeling the diffusion of extracellular small molecules, i.e. environmentally available nutrients and secreted byproducts.

To correctly perform this last part of the simulation, we exploit the natural separation of time scales between growth and diffusion of small molecules. The typical time scale associated with growth, t_{growth}, is set by cell doubling times and is of order 10^3 seconds. The dependence of growth rates on the external nutrients is incorporated by solving a spatially dependent FBA on a time scale $t_{FBA} << t_{growth}$, and is typically of order 10^2 seconds. Diffusion is then performed on time scales t_D ~10 seconds, an order of magnitude smaller than the FBA update times. This separation of time scales allows us to efficiently model the complex dynamics in a computationally tractable manner while ensuring that our scheme is physically consistent. An important consequence of this scheme is that it associates a length scale l with each point in our lattice. The scale is set by the smallest diffusion constant, D_{min}, and is given by $l^2 \sim D_{min}/t_D$. Typically, for small metabolites D_{min} ~10^{-5} cm^2s, implying that we have a spatial resolution of about 10^{-2} cm per grid point.



Our prototype of COMETS uses the open-source GNU Linear Programming Kit (GLPK) for performing the dFBA calculations, and a Java platform for coordinating the simulations and for rapid visualization. Several microbial species have been imported into COMETS, including *Escherichia coli*, *Salmonella typhimurium*, *Shewanella oneidensis*, *Lactococcus lactis*, and *Saccharomyces cerevisiae*. New species and new environmental settings can be easily incorporated through a macro language, or using a custom graphical user interface (see Figure). COMETS is being tested by comparing computational simulations to available and newly measured spatial distributions of biomass in individual *E. coli* colonies on agar [3]. In addition, we are applying this platform to study the population dynamics of two syntrophic bacteria [4], with special attention to the effects of initial density in a diverse spatially organized system that contains a population of cheaters.

References

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.
2. Xavier J, Martinez-Garcia E, and Foster KR. *Social evolution of spatial patterns in bacterial biofilms: when conflict drives disorder.* Am Nat 2009, Jul;174(1):1-12.
3. Pipe LZ and Grimson MJ. *Spatial-temporal modeling of bacterial colony growth on solid media.* Mol. BioSyst. 2008, Mar;4(3):192-198.
4. Harcombe W. *Novel cooperation experimentally evolved between species.* Evolution, Jul;64(7):2166-72.

249

A Multi-Scale Approach to the Simulation of Lignocellulosic Biomass

Goundla Srinivas,[1] Glass Dennis,[1] Sergiy Markutsya,[2] Yana Kholod,[2] Ajitha Devarajan,[2] John Baluyut,[2] Monica Lamm,[2] Theresa L. Windus,[2] Xiaolin Cheng[1,3]* (chengx@ornl.gov), **Mark S. Gordon**,[2] and **Jeremy C. Smith**[1,3]

[1]Oak Ridge National Laboratory, Oak Ridge, Tenn.; [2]Ames Laboratory, Iowa State University, Ames; and [3]University of Tennessee, Knoxville

http://cmb.ornl.gov/

Project Goals: In concert with the imminent increase in the Department of Energy's leadership supercomputing power to the petaflop range, the objective of this project is to develop multiscale methods for extending the time- and length-scales accessible to biomolecular simulation on massively parallel supercomputers. This project also aims to apply the developed multiscale approaches to obtain an understanding of the structure, dynamics and degradation pathways of extended cellulosic and lignocellulosic materials. Information from multiscale simulation, when closely integrated with experiment, will provide fundamental understanding needed to overcome biomass recalcitrance to hydrolysis.

The multiscale simulation methods, ranging from highly accurate quantum mechanical (QM) methods to coarse-grained molecular dynamics (MD), have been used to obtain an understanding of the structure, dynamics and degradation pathways of extended cellulosic and lignocellulosic materials using capability high-performance simulation. Treating solvent implicitly is a critical multiscale concept, and to

this end we have developed a parallel order-*N* Poisson-Boltzmann solver (1,2) and a treecode-based Generalized Born electrostatic solvation method (3). Furthermore, a statistical mechanical multiscale approach was derived that was found to describe the temperature dependence of cellulose fiber stability (7), and complementary Fragment Molecular Orbital (FMO) and all-atom MD simulations have been performed of cellulose crystal structures.

A range of coarse-grained models have been developed on various length scales (8). Based on the effective fragment potential, coarse-grained models have been developed for benzene and methanol and for glucose in solution. The coarse graining has also involved development and application of Boltzmann inversion techniques and of the "REACH" (Realistic Extension Algorithm via Covariance Hessian) methodology, which maps results obtained from atomistic MD simulations onto models for larger-scale, coarse-grained MD.

The physical properties of lignocellulosic biomass derived using the multiscale methodologies serve as a basis for interpreting an array of biophysical experiments (4-5), and, in particular, the simulation models derived will be used to calculate and interpret a variety of neutron-scattering properties. To aid in the interpretation we have developed "dynamical fingerprinting" as a means of reconciling the multiple time scales accessed by experiment and simulation (6). This combination of simulation and experiment will eventually lead to a description of the physicochemical mechanisms of biomass recalcitrance to hydrolysis, and thus will aid in developing a strategy as to how rationally to overcome the resistance.

References

1. Lu B*, Cheng X*, Huang J, McCammon JA. AFMPB: An Adaptive Fast Multipole Poisson-Boltzmann Solver for Calculating Electrostatics in Biomolecular Systems. *Comput Phys Commun.* 2010, 181, 1150
2. Zhang B, Lu B, Cheng X, Huang J*, Pitsianis N, Sun X, and McCammon JA. Mathematical and Numerical Aspects of the Adaptive Fast Multipole Poisson-Boltzmann Solver. *Communications in Computational Physics* 2011, in press
3. Xu Z*, Cheng X* and Yang H. Treecode-based Generalized Born method. *J Chem Phys.* 2011 134(6):064107
4. Prinz, JH, Chodera, J, Pande V., Swope W, Smith JC and Noe F. Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics. *J. Chem. Phys.* 2011, 134, 244108
5. JH Prinz, M. Held, JC Smith and F. Noe. Efficient Computation, Sensitivity and Error Analysis of Committor Probabilities for Complex Dynamical Processes. *Multiscale Mod. & Sim*. 9, 545-567 (2011).
6. F. Noe, S. Doose, I. Daidone, M. Lollmann, M. Sauer, J.D. Chodera and JC Smith. Dynamical Fingerprints: Probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proc. Natl. Acad. Sci. U.S.A.* 2011, 108(12):4822-7
7. Klein H*, Cheng X*, Smith JC* and Shen T*. Transfer matrix approach to the hydrogen-bonding in cellulose Iα fibrils describes the recalcitrance to thermal deconstruction. *J Chem Phys.* 2011, 135, 085106
8. Srinivas G, Cheng X and Smith JC*. A Solvent-Free Coarse Grain Model for Crystalline and Amorphous Cellulose Fibrils. *J. Chem. Theo. Comp.* 2011, 7 (8), 2539–2548

This research is funded by the Genomic Science Program, Office of Biological and Environmental Research, and the Scientific Discovery through Advanced Computing program, U. S. Department of Energy, currently under FWP ERKJE84. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by a DOE INCITE award from the Office of Science of the U.S. Department of Energy.

250

A Multiscale Approach to the Simulation of Lignocellusic Biomass

Monica H. Lamm[1,2]* (mhlamm@iastate.edu), John Baluyut,[1,2] Ajitha Devarajan,[1] Yana Kholod,[1] Sergiy Markutsya,[1] Theresa L. Windus,[1,2] and **Mark S. Gordon**[1,2]

[1]Ames Laboratory, Ames, Iowa and [2]Iowa State University, Ames

Project Goals: Multiscale methods in theoretical chemistry and molecular simulation are applied to obtain an understanding of the structure, dynamics, and degradation pathways of cellulosic biomass. Using methods that range from accurate quantum chemical calculations to all-atom and coarse-grained molecular dynamics simulation, this project aims to use theory and simulation as a guide for overcoming the recalcitrance to hydrolysis in the production of fuel from biomass.

Cellulosic ethanol production is a two-stage process that involves the hydrolysis of cellulose to form simple sugars and the fermentation of these sugars to ethanol. Hydrolysis of cellulose is the rate-limiting step, and there is a great need to characterize the process with theoretical chemistry and molecular simulations to better understand the complex mechanisms that are involved. The ultimate goal is to generate accurate coarse-grained molecular models that are capable of predicting the structure of lignocellulose before and after pretreatment so that subsequent *ab initio* calculations can be performed to probe the degradation pathways.

Current computational studies include: 1) determining the energy barrier to rotation of free hydroxyl groups in cellulose Iα, 2) characterizing the interfragment and interchain interaction energies in cellulose Iβ with fragment molecular orbital (FMO) calculations of, 3) developing coarse-grained models for crystalline and amorphous cellulose fibers from all-atom molecular dynamics simulations using classical force fields and the effective fragment potential (EFP).

This research is supported under FWP AL-08-330-039 by the Genomic Science Research Program, Office of Biological and Environmental R esearch and the Scientific Discovery through Advanced Computing program in the U.S. Department of Energy Office of Science.

251

Combining Whole Cell Stochastic Simulations with Systems Biology Approaches

Elijah Roberts,[1] Piyush Labhsetwar,[2] John Cole,[3] Nathan Price,[1,2] and **Zaida Luthey-Schulten**[1,2,3]* (zan@illinois.edu)

[1]School of Chemical Sciences, [2]Center for Biophysics and Computational Biology, and [3]Department of Physics, University of Illinois, Urbana

In this work we compute the stochastic reaction-diffusion dynamics of selected biochemical pathways to show how individual cells vary expression of a set of genes in response to an environmental signal. The whole cells simulated under *in vivo* conditions include ribosomes, DNA, and large protein complexes which take up 30-50% of the cell volume and are placed according to data from cryoelectron tomography and proteomics. Using GPU processors, we simulate the dynamics for an entire cell cycle and compare the mRNA/protein distributions to those observed in single molecule experiments. We show how such distributions can be used to derive additional kinetic parameters and integrate effects of cell to cell variations into flux balance analysis of genome scale models of metabolic networks. The distribution of growth rates calculated for a colony of bacteria are analyzed and correlated to changes in fluxes through the metabolic network.

Publications

1. "Long time-scale simulations of *in vivo* diffusion using GPU hardware", E. Roberts, J. Stone, L. Sepulveda, Wen-mei Hwu, and Z. Luthey-Schulten, in *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing*, 2009.
2. "Noise contributions in an inducible genetic switch: A whole cell simulation study", E. Roberts, A. Magis, J. Ortiz, W. Baumeister, and Z. Luthey-Schulten, *Plos Comput. Biol.* 7(3), e1002010 March (2011).
3. "Determining the stability of genetic switches: Explicitly accounting for mRNA noise" Michael Assaf, Elijah Roberts, and Zaida Luthey-Schulten, *Phys.Rev.Lett.* 24, 248102, (2011).

252

Toolbox Model of Evolution of Metabolic Pathways on Networks of Arbitrary Topology

Tin Yau Pang[1,2] and **Sergei Maslov**[1]* (maslov@bnl.gov)

[1]Biology Department, Brookhaven National Laboratory, Upton, N.Y. and [2]Department of Physics and Astronomy, Stony Brook University, N.Y.

Project Goals: The biological functioning of a living cell involves coordinated actions of its metabolic and regulatory networks. Metabolic networks are composed of many semi-autonomous functional units – metabolic pathways. These pathways are routinely controlled by dedicated transcription factors, and the activities of individual pathways need to be well coordinated with each other. The project goal is to investigate general principles behind such coordination in prokaryotic genomes. To this end we carry out dynamical and evolutionary modeling of the integrated network encompassing metabolic and regulatory interactions.

It has been previously reported[1] for prokaryotic genomes that the number of Transcription Factors (TFs) is proportional to the *square* of the total number of genes. As a consequence of this trend the fraction of TFs (the so-called "regulatory overhead") is less than 0.5% in small (<500 genes) bacterial genomes, while in large genomes (~10,000 genes) it can be as high as 10%. We recently proposed[2] a general explanation of this empirical scaling law and illustrated it using a simple model in which metabolic and regulatory networks co-evolve together. In our model prokaryotic organisms acquire new metabolic functions by the virtue of horizontal gene transfer of entire co-regulated metabolic pathways from a shared gene pool (the "universal metabolic network" or bacterial metabolic pan-genome). This transfer is followed by removal of redundant enzymes and assignment of a dedicated TF regulating the newly acquired pathway[3]. The whole process can be compared to a homeowner buying sets of tools from a hardware store and later returning duplicate items. We view the full repertoire of metabolic enzymes (or more generally all non-regulatory proteins) encoded in the genome of an organism as its collection of tools. Adapting to a new environmental condition (e.g. learning to utilize a new nutrient source) involves acquiring new tools as well as reusing some of the tools that are already encoded in the genome. As the toolbox of an organism grows larger, it can reuse its existing tools more often and thus needs to acquire fewer new enzymes to master each new functional task. From this argument it follows that, in general, the number of metabolic pathways and their regulators should always scale faster than linearly with the total number of genes in a genome. The empirically observed quadratic scaling between these two numbers can be mathematically derived for a broad range of universal network topologies[4]. Furthermore, the sizes of evolutionary conserved pathways in our model have a long-tailed power-law distribution that agrees with empirical observations. This offers a conceptual explanation for the empirically observed broad distribution of regulon sizes or TFs out-degrees in regulatory networks.

References

1. E van Nimwegen, "Scaling laws in the functional content of genomes", Trends Genet 19, 479-84 2003
2. S Maslov, S Krishna, T Y Pang, K Sneppen, "Toolbox model of evolution of prokaryotic metabolic networks and their regulation", PNAS 106, 9743-9748 2009
3. J Grilli, B Bassetti, S Maslov, and M Cosentino Lagomarsino, "Joint scaling laws in functional and evolutionary categories in prokaryotic genomes", Nucleic Acids Research doi: 10.1093/nar/gkr711 2011

4. TY Pang, S Maslov, "Toolbox model of evolution of metabolic pathways on networks of arbitrary topology" PLoS Comp. Bio 8, e1001137 2011.

Support of this work was provided by the DOE Systems Biology Knowledgebase project "Tools and Models for Integrating Multiple Cellular Networks."

253

Enabling the Use of Externally-Built Alignments and Trees in ARB for Evolutionary Analysis

Steve Essinger,[1] Erin Reichenberger,[1] **Chris Blackwood**,[2] and **Gail Rosen**[1]* (gailr@ece.drexel.edu)

[1]Electrical and Computer Engineering Department, Drexel University and [2]Biology Department, Kent State University

In order to investigate gene evolution, gene sequences from various organisms are commonly aligned to form a phylogenetic tree. Besides viewing the taxonomic information on the tree, a user may want to visually inspect how the gene product and KEGG pathway with the associated sequence has evolved, giving greater power to evolutionary hypothesis testing. A software package, such as ARB, has the power to pool this information from Genbank records, but ARB uses the local computer resources to perform the alignment. Therefore, a user may want to use external resources (such as the CIPRES portal on Terragrid), to perform the alignment and tree construction, and then import and link that information back into ARB to manipulate the data.

To accomplish this, we have created a pipeline that integrates external alignment and de novo tree construction for an arbitrary protein family (even one that contains over 10,000 member sequences). We have developed custom python scripts and an ARB import filter to extract meta-data from Genbank records and import this info with an externally-built alignment and phylogenetic tree. Using our scripts, a custom database, that includes all of the sequences and associated meta-data in the study, is imported into an ARB database using uniqueIDs. The user can then use the ARB suite of tools to manipulate the phylogenetic tree and display the associated metadata.

We demonstrate the use of our tool by examining a protein family of interest to the "Tracking down the cheaters" project. All code will be made available on our website that will allow other groups to view custom fields extracted from Genbank records on phylogenetic trees using externally-built trees and alignments.

254

Numerical Optimization Algorithms and Software for Systems Biology: A Globally Convergent Algorithm for Computing Stable Non-Equilibrium Steady State Concentrations in Genome-Scale Networks

Ronan M.T. Fleming[1]* (ronan.mt.fleming@gmail.com), Ines Thiele,[1] and **Michael A. Saunders**[2]

[1]Center for Systems Biology, University of Iceland, Iceland and [2]Department of Management Science and Engineering, Stanford University

http://www.hi.is/~rfleming

Project Goals: 1. Develop algorithms for solving optimization problems involving large stoichiometric matrices. (a) Extend existing sparse linear programming algorithms to enable the solution of such systems, in which the matrix coefficients represent reactions at multiple timescales and thus vary over many orders of magnitude. (b) Develop a convex optimization algorithm for computing thermodynamically feasible reaction fluxes in a general instance of a genome-scale integrated metabolic and macromolecular biosynthetic network. (c) Implement a parallel convex optimizer in to enable sampling of the thermodynamically feasible set. (d) Disseminate software to the systems biology community. 2. Investigate cyclic dependency between metabolic and macromolecular biosynthetic networks. (a) Predict the material and energy cost of macromolecular synthesis in an integrated metabolic, transcriptional and translational model of *Escherichia coli*. (b) Reconstruct and analyze the macromolecular synthesis network of *Thermotoga maritima*. 3. Quantify the significance of thermodynamic constraints on prokaryotic metabolism. (a) Simultaneous prediction of metabolic fluxes and concentrations in *Escherichia coli*. (b) Validate and interpret flux and concentration prediction in *Escherichia coli* and *Thermotoga maritima*. (c) Predict thermodynamically favorable pathways for hydrogen production by *Thermotoga maritima* on a range of substrates. (d) Numerically sample mass conserved, thermodynamically feasible steady state fluxes and concentrations in *Escherichia coli*.

Concerning project goal 3(a): At the core of computational systems biology lies a paradox. All of the currently available genome-scale modeling methods can only model chemical reaction rates, but not the abundance (or concentration) of the molecules involved in these reactions. At the same time, the vast majority of experimental omics data are measures of the abundance of some molecule, rather than the rate. The reason for this paradox is that modeling steady state reaction kinetics has been limited to small systems of chemical reactions as the inherently nonlinear systems of equations at the core of such models have been intractable to solve. We present the first globally convergent algorithm for simultaneously computing stable non-equilibrium steady state molecule concentrations and reaction rates. We leverage this

algorithm to simultaneously predict stable steady state concentrations and reaction rates in *E. coli*, which are numerical consequences of various hypotheses regarding the manner in which evolved kinetic parameters may be optimal with respect to optimization of various cellular system objectives.

submitted post-press

259

Development of Predictive Software Tools to Construct and Analyze Large Dynamical Networks for Systems Biology Knowledgebase

Ravishankar R. Vallabhajosyula* (rrv@cfdrc.com), B. Prabhakarpandian, and Kapil Pant

CFD Research Corporation, Huntsville, Ala.

http://www.cfdrc.com

Project Goals: Recent biotechnological advances have accelerated the generation of 'omic' data. This has driven the development of computational tools to model the biological systems by inferring mechanisms responsible for response to external stimuli. However, lack of kinetic information for most biochemical interactions limits the predictive capabilities of these tools. CFD Research Corporation (CFDRC) is developing predictive modeling toolkit to overcome this limitation, thereby facilitating rapid and accurate characterization of the effects of the environment on phenotypes. In particular, our toolkit will enable (1) identification of significant biological features from omic datasets, (2) construction of a comprehensive network model of cellular pathways, and (3) simulation of this pathway model using a kinetics-free algorithm to predict the altered phenotypes when selected targets in the network are modified. This methodology is being validated using well-characterized organisms (e.g., yeast) as well as selected microbe-based biosystems of DoE interest (e.g., identification of higher quantity and quality biofuel yielding algal strains).

Recent developments in genetic engineering and biotechnology have enabled the modification of genes in an organism or the introduction of genes from other organisms towards achieving the desired phenotypes. However, these experimental procedures are often carried out without adequate systems-level knowledge of the cellular biology, which can lead to unexpected outcomes. Well-designed computational methodologies can be used to prevent such scenarios with the aid of predictive software tools. A key goal of the DoE Systems Biology Knowledgebase (Kbase) is to facilitate analysis of vast omic datasets for characterizing the response of organisms to various environmental stimuli towards predicting phenotypes. For example, such tools will be able to identify algal strains with improved attributes of biofuel production, while simultaneously overcoming slow growth rates associated with some of these strains. Such computational approaches should be based on a comprehensive understanding of the cellular biology of the organisms of interest, and will be significantly aided by the adaptation and application of novel algorithms and software that can analyze multi-omic data related to the observed response to various external stimuli.

Under DOE sponsored research, CFD Research Corporation (CFDRC) is currently developing predictive computational tools to address the goals of Kbase towards characterizing the response biological organisms to environmental stimuli that serve as inputs and predicting phenotypes most likely to be observed. Figure 1 shows a schematic of the framework being developed by CFDRC. Drawing upon available databases, our approach relies on the construction of mechanistic Systems Biology based and data-driven models of the differentially regulated cellular pathways. The complex pathway models are then analyzed without requiring information on the kinetics of various biochemical interactions. This enables the discovery and ranking of targets (for example genes, proteins or metabolites) for potential modification and the prediction of their response when these modifications are implemented. This approach thus offers the potential to inform experiments for the development of strains efficient at generating the desired phenotype such as algae strains that can produce biofuels at a higher rate.

Figure 1: Schematic Detailing the use of Omic Data to Identify Targets towards Predicting Phenotypes

As part of the ongoing Phase I study, we are developing a prototype of the software toolkit using transcriptional data to construct and analyze complex pathway networks in an extensible SBML format (Hucka et al., 2003) that will be enhanced to analyze other omic data types in future. Development of these tools will enable researchers to analyze pathways that play important roles in sensing and responding to the external conditions in an integrated manner. These tools are important to understand the organism's behavior in the modified environment including its survival and in predicting the associated phenotypes. Towards this goal, we are studying the yeast environmental stress response to various external conditions as a test case to test and validate the model. We are also in active discussions with different organizations to demonstrate the technology for microbial systems of DoE interest e.g., identification of targets for genetic engineering of algal strains for higher quantity and quality biofuel production.

References

1. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., et al., The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models, *Bioinformatics*, **19**(4): 524-531, 2003.

This work is being supported through the DOE Office of Biological and Environmental Research under an SBIR Phase I grant (DE-SC0006190).

submitted post-press

262

The GreenCut Resource, a Phylogenomically Derived Inventory of Proteins Specific To the Plant Lineage

Rikard Fristedt,[1] HiroakiYamasaki,[1] Steven Karpowicz,[1] Arthur Grossman,[2] and **Sabeeha Merchant**[1]* (merchant@chem.ucla.edu)

[1]Department of Chemistry and Biochemistry and Institute for Genomics and Proteomics, University of California, Los Angeles and [2]Department of Plant Biology, Carnegie Institution for Science, Stanford, Calif.

http://www.chem.ucla.edu/dept/Faculty/merchant.html

Project Goals: (see abstract)

The plastid is a defining structure of photosynthetic eukaryotes and houses many plant specific processes, including the light reactions, carbon fixation, pigment synthesis, and other primary metabolic processes. Identifying proteins associated with catalytic, structural, and regulatory functions that are unique to plastid-containing organisms is necessary to fully define the scope of plant biochemistry. We performed phylogenomics on 20 genomes to compile a new inventory of 597 nucleus-encoded proteins conserved in plants and green algae but not in non-photosynthetic organisms. At the time of analysis, 286 of these proteins were of known function, whereas 311 are not characterized. This inventory was validated as applicable and relevant to diverse photosynthetic eukaryotes using an additional eight genomes from distantly related plants (including Micromonas, Selaginella, and soybean). Manual curation of the known proteins in the inventory established its importance to plastid biochemistry. To predict functions for the 52% of proteins of unknown function, we used sequence motifs, subcellular localization, co-expression analysis, and RNA abundance data. About 18% of the proteins in the inventory have functions outside the plastid and/or beyond green tissues. Although 32% of proteins in the inventory have homologs in all cyanobacteria, unexpectedly, 30% are eukaryote-specific. Finally, 8% of the proteins of unknown function share no similarity to any characterized protein and are plant lineage-specific. We have initiated functional analyses of the eukaryote-specific proteins and we present phenotypes for loss of function mutations in some of the unknown GreenCut genes.