The DOE Systems Biology Knowledgebase: Microbial Science Domain

Paramvir S. Dehal[1*], 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], Rick Stevens[1]

1 Lawrence Berkeley National Laboratory, Berkeley, CA; 2 Argonne National Laboratory, Argonne, IL; 3 Oak Ridge National Laboratory, Oak Ridge, TN; 4 Brookhaven National Laboratory, Upton, NY.
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 Reconstructi*on*:* 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 Reconstructio*n*:* 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.