The DOE Systems Biology Knowledgebase: Microbial Communities Science Domain

Folker Meyer[1]*(folker@anl.gov), Dylan Chivian[2], Andreas Wilke[1], Naryan Desai1, Jared Wilkening[1], Kevin Keegan[1], William Trimble[1], Keith Keller[2], Paramvir Dehal[2], Robert Cottingham[3], Sergei Maslov[4], Rick Stevens[1], **Adam Arkin[2]**

1 Argonne National Laboratory, Argonne, IL; 2 Lawrence Berkeley National Laboratory, Berkely, CA; 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 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.