KBase: An Integrated Knowledgebase for Predictive Biology and Environmental Research

Adam Arkin[1]* (aparkin@lbl.gov), Robert Cottingham[2], Sergei Maslov[3], Rick Stevens4, Dylan Chivian[1], Parmavir Dehal[1], Christopher Henry[4], Folker Meyer[4], Jennifer Salazar[4], Doreen Ware[5], David Weston[2], Brian Davison[2], Elizabeth M. Glass[4]

1 Lawrence Berkeley National Laboratory, Berkeley, CA; 2 Oak Ridge National Laboratory, Oak Ridge, TN; 3 Brookhaven National Laboratory, Upton, NY; and 4 Argonne National Laboratory, Argonne, IL; 5 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
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.