ModelSEED release 2: High Throughput Genome-Scale Metabolic Model Reconstruction and Analysis in KBase

José P. Faria[1*], Filipe Liu[1], Janaka N. Edirisinghe[1], Samuel M.D. Seaver[1], James G. Jeffryes[1], Qizh Zhang[1], Pamela Weisenhorn[1], Boris Sadkhin[1], Nidhi Gupta[1], Tian Gu[1] and Christopher S. Henry[1], Robert Cottingham[2], and **Adam P. Arkin[3]**

[1]Argonne National Laboratory, Lemont, IL; [2] Oak Ridge National Lab, Oak Ridge, TN; [3]Environmental Genomics and Systems Biology, Lawrence Berkeley National Lab, Berkeley, CA.
*presenting author

http://www.kbase.us

Project Goals: Short statement of goals. (Limit to 1000 characters)

The Department of Energy Systems Biology Knowledgebase (KBase) is a platform designed to solve the grand challenges of Systems Biology. KBase has implemented bioinformatics tools that allow for multiple workflows, including genome annotation, comparative genomics, and metabolic modeling. First released in 2010, the ModelSEED [1] genome-scale model reconstruction pipeline has now built over 200k draft metabolic reconstructions and supported hundreds of publications. Here, we describe the first major update to this model reconstruction tool with important new features including: (1) a dramatically improved representation of energy metabolism ensuring models produce accurate amounts of ATP per mol of nutrient consumed; (2) new templates for Archaea and Cyanobacteria; and (3) greatly improved curation of all metabolic pathways mapping to RAST and other annotation pipelines.

Abstract text

KBase has made several key improvements to the ModelSEED model reconstruction tool. ATP production was improved in our model reconstruction procedure by constructing core models, testing for proper ATP production from this core, then ensuring that ATP production does not incorrectly explode when expanding the core model to a genome-scale model. We similarly improved our gapfilling approach to ensure that gapfilling does not cause a model to start over-producing ATP. While other approaches aim to correct ATP overproduction in models, these new procedures in the ModelSEED pipeline aim to ensure that ATP overproduction does not happen in the first place. To handle the necessary expansion of templates, we developed machine learning (ML) classifiers to determine automatically which template most correctly applies to a new genome being modeled. The classifiers will also produce new template and biomass objective functions specific to archaea and cyanobacteria (modeled after existing published metabolic models of these species). This ML approach allows for the rapid introduction of additional modeling templates, enabling researchers working with unclassified species or metagenome-assembled genomes extracted to achieve more specific reconstructions.

To improve metabolic pathway annotation completeness and accuracy in ModelSEED models, we first updated our biochemistry database to include the latest reaction data from KEGG, MetaCyc, BIGG, and published models. Next, we manually curated the major pathways in our

reconstruction templates to reconcile pathway representation across these multiple databases. Finally, we curated our mapping of RAST functional roles to this reconciled biochemistry based on data mined from KEGG and published metabolic models.

Within the KBase platform, we demonstrate our improved model reconstruction pipeline on a phylogenetically diverse set of approximately ~6400 genomes (all Bacteria and Archaea complete genomes in KEGG) and constructed draft genome-scale metabolic models (GEMs). We show how the gene counts and modeling metrics (ATP production, biomass yields, reaction classification, pathway representation) are improved with this new release of the ModelSEED. We also selected genomes for which comprehensive Biolog data are available, and we compared model predictions of all data with experimental results. The comparisons showed significant improvement compared to models generated by the original ModelSEED. Finally, we compare the pathway content, gapfilling, size, and gene counts in our models with models constructed by the CarveMe method.

The listed improvements will be available as an update to the ModelSEED reconstruction pipeline. The new release will be made available across all web platforms currently supported; KBase (kbase.us), ModelSEED (modelseed.org), and PATRIC (patricbrc.org) resources.

References

1. Henry, Christopher S., et al. "High-throughput generation, optimization and analysis of genome-scale metabolic models." Nature biotechnology 28.9 (2010): 977-982.

This work is supported as part of the Genomic Sciences Program DOE Systems Biology Knowledgebase (KBase) funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.