Title: Probabilistic Annotation and Ensemble Metabolic Modeling in KBase

Authors: Patrik D'haeseleer[1*] (dhaeseleer2@llnl.gov), Jeffrey Kimbrel[1], Ali Navid[1], Chris Henry[2], Rhona Stuart[1]

Institutions: [1]Lawrence Livermore National Laboratory , Livermore, CA; [2]Argonne National Laboratory, Lemont, IL

Website URL: https://www.kbase.us/research/stuart-sfa/

Project Goals: We are developing tools for the DOE Systems Biology Knowledgebase (KBase) to give users a principled way to weight multiple sources of functional annotation against each other, enable better metabolic modeling of hard-to-annotate organisms and pathways, allow analysis of uncertainty in the resulting models network structure or behavior, and provide an infrastructure on which to build more sophisticated machine learning techniques in KBase.

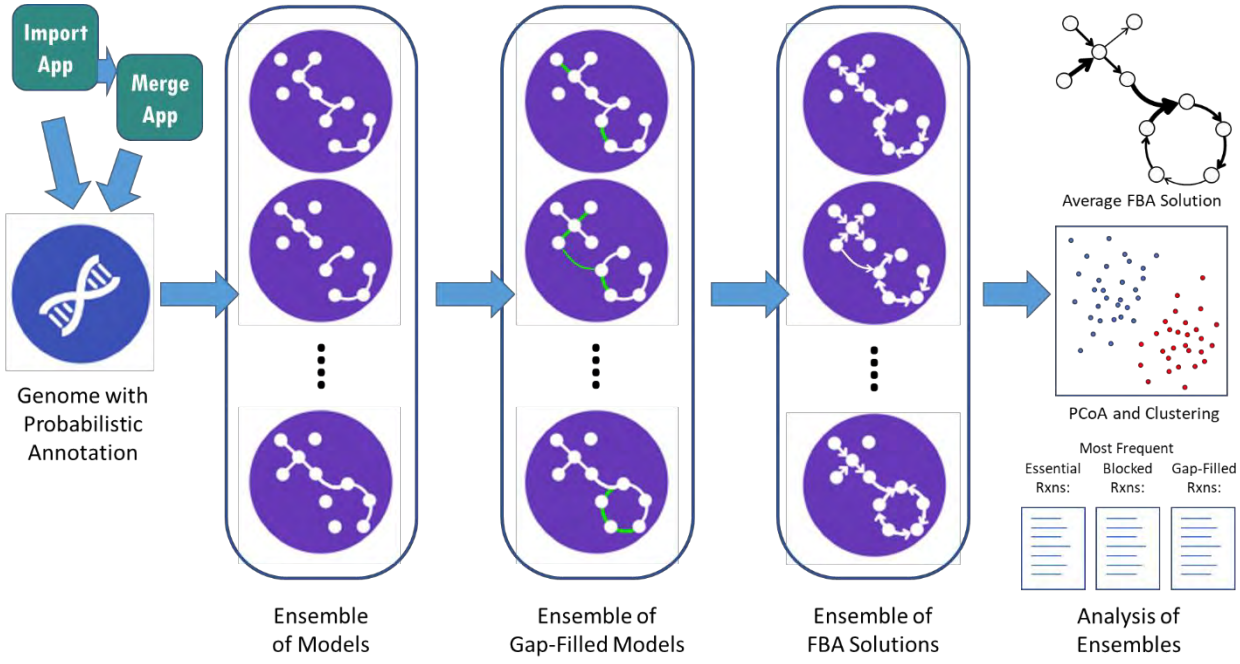
Abstract Text:

Our μBiospheres SFA investigates metabolic interactions in bioenergy-relevant microbial communities. A critical part of this research is development of genome-scale models of metabolism, which requires well-annotated genomes. We have found that combining annotations from multiple sources results in a more complete metabolic network reconstruction, greatly reducing the effort required to curate quality metabolic models (1). In a previous round of plus-up funding from DOE, we implemented a set of KBase Apps that allow users to upload metabolic annotations from multiple functional annotation tools, compare and merge these annotations, and to use them for model building and gapfilling to achieve significantly improved metabolic models. These Apps have proven to be very useful and are currently in daily use in our own SFA and several other research groups using KBase.

However, it is quite common for functional annotation tools to disagree on the function that should be assigned to certain genes, and this uncertainty can have significant consequences on the resulting metabolic networks and the behavior they predict for the organism. We are developing a set of KBase apps – and the underlying infrastructure to enable them – to allow the user to take advantage of these multiple inputs and explore the uncertainty and incompleteness of the functional annotations in their organism of interest in a probabilistic modeling framework. We will provide Apps to import annotation probabilities where available, implement Bayesian methods for merging annotation probabilities, generate an ensemble of models by sampling from the underlying probability distribution, sample alternative gapfilling solutions using the Medusa COBRApy package by Medlock and Papin (2), and then analyze the results of ensemble modeling. This ensemble approach will allow us to leverage all the existing KBase tools for importing and merging metabolic annotations, metabolic modeling, and gapfilling.

This work will provide our SFA and other KBase users a principled way to weight annotation sources against each other, enable better metabolic modeling of hard-to-annotate organisms and pathways, allow analysis of uncertainty in the resulting models network structure or behavior, and provide an infrastructure on which to build more sophisticated machine learning techniques in KBase.

These Apps will be broadly applicable to a wide range of users interested in model building using multiple annotation sources in KBase, as well as groups working on modeling challenging organisms and pathways, and those who would like to have access to state-of-the-art ensemble modeling tools.



Proposed workflow diagram, starting from imported annotation probabilities or probability estimates generated by our Merge app.

References/Publications

1. Griesemer M, Kimbrel JA, Zhou CE, Navid A, D'haeseleer P. Combining multiple functional annotation tools increases coverage of metabolic annotation. BMC Genomics. 2018 Dec 19;19(1):948.
2. Medlock GL, Moutinho TJ, Papin JA. Medusa: Software to build and analyze ensembles of genome-scale metabolic network reconstructions. PLoS Comput Biol. 2020 Apr 29;16(4):e1007847.

Funding Statement: This work was performed under the auspices of the U.S. Department of Energy at Lawrence Livermore National Laboratory under Contract DE-AC52- 07NA27344 and supported by the Genome Sciences Program of the Office of Biological and Environmental Research under the LLNL μBiospheres SFA, FWP SCW1039.