Combining multiple functional annotation tools increases completeness of metabolic annotation

Marc Griesemer[1], Jeffrey A. Kimbrel[1], Carol Zhou[2], Ali Navid[1], **Patrik D'haeseleer[1]** (dhaeseleer2@llnl.gov)

[1]Lawrence Livermore National Laboratory, Livermore CA, USA

https://bio-sfa.llnl.gov/

Project Goals: The LLNL Bioenergy SFA seeks to support sustainable and predictable bioenergy crop production through a community systems biology understanding of microbial consortia that are closely associated with bioenergy-relevant crops. We focus on host-microbial interactions in algal ponds and perennial grasses, with the goal of understanding and predicting the system-scale consequences of these interactions for biomass productivity and robustness, the balance of resources, and the functionality of surrounding microbial communities. Our approach integrates 'omics measurements with quantitative isotope tracing, characterization of metabolites and biophysical factors, genome-enabled metabolic modeling, and trait-based representations of complex multi-trophic biological communities, to characterize the microscale impacts of single cells on system scale processes.

Genome-scale metabolic modeling is a cornerstone of our genome and omics-enabled computational analysis of microbial communities and algal-bacterial interactions. However, genome-scale systems biology modeling efforts are invariably based on very incomplete functional annotations. Annotated genomes typically contain 30-50% of genes with little or no functional annotation (Hanson *et al.*, 2009), severely limiting our knowledge of the "parts lists" that the organisms have at their disposal. In metabolic modeling, these incomplete annotations are often sufficient to derive a reasonably complete model of the core metabolism consisting of well-studied (and thus well-annotated) metabolic pathways sufficient for growth in pure culture. Modern metabolic modeling efforts, however, are moving beyond studying core metabolic pathways in a single organism towards multi-species models, real-world communities and ecosystems.

In Flux Balance Analysis (FBA), the issue of missing metabolic annotations is dealt with by "gap filling" - the addition of a set of metabolic reactions beyond those that were derived directly from the genome annotation. However, gap filling typically requires the addition of several dozen reactions to allow production of biomass in simple defined nutrient media (Henry *et al.*, 2010); the reactions added by different gap filling algorithms may have little or no overlap with each other (Krumholz *et al.*, 2015); and even after gap filling, enough pathway holes remain to block on average one-third of the reactions in each model. Clearly, a more complete identification and annotation of metabolic reactions would be preferable to the addition of dozens of poorly supported reactions just to patch the holes in the network. Recent genome-scale modeling of *Clostridium beijerinckii* NCIMB 8052 (Milne *et al*, 2011) demonstrated that the total number of genes and reactions included in the final curated model could be almost doubled by incorporating multiple annotation tools.

Here, we present results on a comprehensive reannotation of 27 bacterial reference genomes from BioCyc, focusing on enzymes with EC numbers annotated by KEGG, RAST, EFICAz, and the BRENDA enzyme database. Our analysis shows that in comparison to metabolic annotation by one of the widely used metabolic modeling platform such as KEGG and RAST, annotation using multiple tools results in a drastically larger metabolic network reconstruction, adding on average 40% more EC numbers, and 37% more metabolic genes. These results are even more pronounced for bacterial species that are more distantly related to well-studied model organisms such as *E. coli* and *B. subtilis*.

Accurate metabolic models also rely on determination of substrate transport between the bacterium and its environment. Because of the computational difficulty in predicting substrates, transporter annotations have been rarely used in genome-scale metabolic modeling. Instead, most metabolic modeling methods simply assume that a transporter exists for the cellular import and export of any necessary metabolite. Better prediction tools such as TransportDB's Transporter Automatic Annotation Pipeline (TransAAP) now allow us to generate substrate predictions that are sufficiently detailed to be included in metabolic pathways, and that could give insights into growth or metabolite exchange phenotypes that are not readily apparent from the metabolic enzymes present in the genome. Our analysis across 27 reference genomes shows that transporter annotation using TransAAP adds 4 to 8 times more transporters with sufficiently detailed substrate annotation to be included in metabolic modeling, compared to annotation by KEGG or RAST alone, adding to each genome on average 212 to 233 additional transport reactions respectively.

The combination of multiple metabolic annotation tools allows us to achieve a significantly more complete genome annotation and metabolic network reconstruction, especially for non-model organisms, and for non-core pathways. We expect that these enhanced modeling abilities will be essential to study newly sequenced algal symbionts in complex real-world interactions. We are developing a computational pipeline to allow us to integrate metabolic annotations from IMG, KEGG, RAST, EFICAz and TransportDB, using SRI's Pathway Tools platform. Applying this approach to draft genomes of algal symbiotic bacteria, we demonstrate the effect of this more comprehensive annotation on the size and completeness of the metabolic reconstruction, and on our increased understanding of algal-bacterial metabolic interactions. This analysis, coupled with growth assays on minimal media with defined carbon and nitrogen substrates, has led to the identification of candidate substrates metabolized by the bacteria, and potential growth factors they may be exchanging with the algae.

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 Biofuels SFA, FWP SCW1039.