Title: Omics-enabled global gapfilling (OMEGGA) for phenotype-consistent metabolic network reconstruction of microorganisms and communities

Authors: Hyun-Seob Song,[1,2*] (hsong5@unl.edu), Joon-Yong Lee,[3] Firnaaz Ahamed,[1] Aimee K. Kessell,[1] William C. Nelson,[3] David M. L. Brown Jr.,[4] Christopher S. Henry,[5] Janaka N. Edirisinghe,[5] and **Kirsten S. Hofmockel[3]**

Institutions: [1]Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE; [2]Department of Food Science and Technology, University of Nebraska-Lincoln, Lincoln, NE; [3]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA; [4]IT/Research Computing Directorate, Pacific Northwest National Laboratory, Richland, WA; [5]Division of Mathematics and Computer Science, Argonne National Laboratory

Website URL: https://www.pnnl.gov/projects/soil-microbiome/research

Project Goals: PNNL's Phenotypic Response of Soil Microbiomes SFA aims to achieve a systems-level understanding of the soil microbiome's phenotypic response to changing moisture. We perform multi-scale examinations of molecular and ecological interactions occurring within and between members of microbial consortia during organic carbon decomposition, using chitin as a model compound. Integrated experiments address spatial and inter-kingdom interactions among bacteria, fungi viruses and plants that regulate community functions throughout the soil profile. Data are used to parametrize individual- and population-based models for predicting interspecies and inter-kingdom interactions. Predictions are tested in lab and field experiments to reveal individual and community microbial phenotypes. Knowledge gained provides fundamental understanding of how soil microbes interact to decompose organic carbon and enable prediction of how biochemical reaction networks shift in response to changing moisture regimes.

Abstract Text: Metabolic network models of microorganisms help us to understand cellular metabolic capabilities, evolution, and ecological principles, as well as aid in the biotechnological design and management of microbial strains and consortia with desired functions. Because of the fundamental importance of metabolic network models in such a wide range of applications, the DOE Systems Biology Knowledgebase (KBase, http://kbase.us) provides a suite of apps and modules supporting the reconstruction, prediction, and design of metabolic models for microorganisms.

Construction of metabolic network models is facilitated by iterative implementation of three key steps: draft model building, gapfilling, and manual curation. Draft metabolic network models (i.e., initial models constructed from genomic or metagenomic data) typically contain incomplete biochemical pathways (i.e., have gaps) due to underlying knowledge gaps in gene function. Gapfilling – adding reactions to a metabolic model to reconcile with phenotypic data – is an essential step in model building because it augments the completeness and functionality of metabolic networks. Typical gapfilling algorithms (including the process currently implemented in KBase) correct one erroneous prediction at a time by iteratively adding new reactions to the

network. This approach, however, often leads to 'false positives' for other growth conditions (i.e., the model predicts growth, but experimental data show non-growth). Occurrence of false positive predictions is a greater problem in modeling communities, compared to isolates, due to the substantially larger pool of reactions available as options for gapfilling.

Based on the hypothesis that false positives are caused by identifying a minimum number of reactions to add to the network (parsimony) without accounting for their broader biological relevance, we propose a new advanced optimization algorithm (termed OMics-Enabled Global GApfilling or OMEGGA) that uses multi-omics data profiles, including amplicon, transcriptomic, proteomic, and intracellular metabolomic data, to simultaneously fit a draft model to all available phenotype data. This novel integration of amplicon, transcript, protein, and metabolite data into model refinement will yield more precise and predictive models and increase the accuracy of identification of active reactions. We will demonstrate the effectiveness of OMEGGA using condition-specific multi-omics and phenotype data from the Model Soil Consortia-2 (MSC-2) and associated isolated organisms developed through PNNL's Soil Microbiome SFA. These organisms were isolated from chitin enrichment cultures of a native soil microbiome. Data generated from various combinations of MSC-2 isolates is extremely valuable in testing the proposed algorithm under diverse contexts. We will accordingly develop generalized KBase apps using the KBase Software Development Kit (SDK) for implementation of all required gapfilling processes in OMEGGA.

Our new optimization algorithms will enable constructing high-quality metabolic networks that best match both molecular and phenotypic observations by avoiding time-consuming manual troubleshooting, which generally does not guarantee a successful outcome. Integration of omics data for gapfree model construction through simultaneous fit to multiple phenotype datasets is a novel idea that will fundamentally change the way we annotate genomes and build metabolic networks by allowing us to consider a more conservative threshold in predicting gene functions towards higher accuracy but lower coverage. This is because those gaps resulting from the conservative threshold can be filled in through the integration of confident experimental evidence (i.e., omics data) as proposed in this work. Therefore, our computational tools and KBase Apps will significantly improve the accuracy in metabolic network models of all complex biological systems including microorganisms, microbial communities, plants, and fungi.

Funding Statement: PNNL is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830. This program is supported by the U. S. Department of Energy, Office of Science, through the Genomic Science Program, Office of Biological and Environmental Research, under FWP 70880 and FWP 78749.