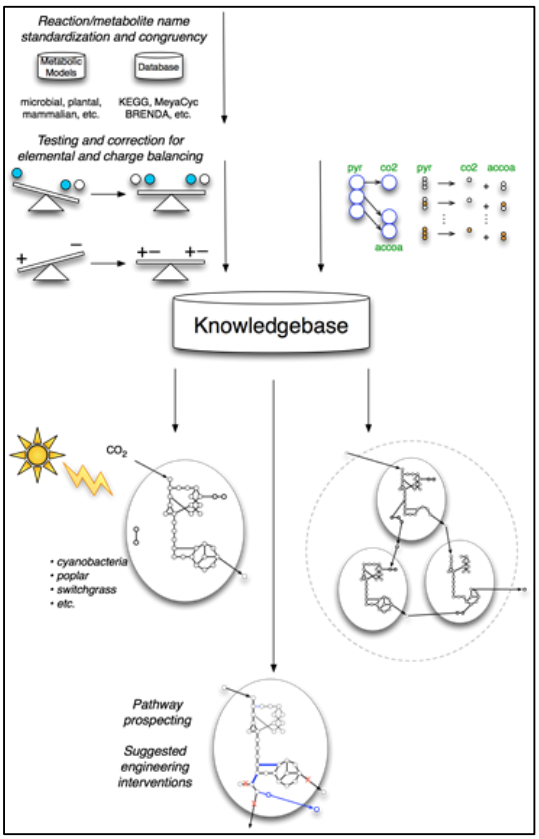
Development of a Knowledgebase (MetRxn) of Metabolites, Reactions and Atom Mappings to Accelerate Discovery and Redesign

Akhil Kumar[2]*(azk172@psu.edu), Saratram Gopalakrishnan[1]*(sxg375@psu.edu), Margaret Simons[1], Siu Hung Joshua Chan[1], and **Costas D. Maranas[1](costas@psu.edu)**

[1]Department of Chemical Engineering, The Pennsylvania State University, University Park, PA; [2]The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA

http://www.maranasgroup.com/

Project Goals: The project aims to expand genome-scale metabolic (GSM) models to include complex interactions through multi-tissues and multi-organisms or complex constraints based on atom-mapping information and Metabolic Flux Analysis (MFA). The whole-organism GSM model of maize will be used to determine bottlenecks in nitrogen metabolism and suggest genetic manipulations to improve nitrogen use efficiency. The multi-organism models will be used to explain experimental observations, analyze the physiological responses, and predict the interactions within microbial communities. We also developed tools and algorithms to organize and disseminate standardized metabolite and reaction information, map atom transitions from reactants to products, identify gene to protein to reaction relations, and design novel biotransformations to xenometabolites. This will facilitate the construction of a mapping model for flux elucidation using 13-C MFA at the genome-scale to provide insights into the role of model scale-up and biomass composition on metabolic flux resolution.



Genome-scale metabolic (GSM) models are a platform used to investigate the metabolic behaviors of complex organisms, microbial communities, and model organisms. Flux balance analysis (FBA) of reconstructed multi-tissue and multi-organism models can be used to determine interactions between different cell/tissue-types or organisms, resolve bottlenecks in limiting pathways, and study the metabolic trade-offs between species-level and community-level fitness functions. FBA often reports large solution spaces which can be further constrained using genome-scale MFA facilitated by isotope-labeling data and incorporating atom-mapping.

A whole plant model of maize was developed by reconstructing the root, stalk, leaf, kernel and tassel tissues using the phloem as the main form of metabolite transport among the tissues using transcriptomic data to determine the set of reactions included in each tissue. This model was simulated at three different growth conditions: vegetative leaf growth, tassel development, and kernel filling. Growth rate proportions were estimated for each growth stage and dry weight proportions were used to normalize the transport reaction flux between tissues. As the plant developed from the vegetative growth stage to the tassel development stage, 138 reactions had differing flux ranges. However, from the tassel growth stage to the kernel filling stage, only 77 reaction flux ranges differed.

Multi-organism models also facilitate the investigation of synergy among microbiota, the interactions between the community and the host, and the effect of the diet on human health. A gut microbial model was developed to represent the metabolism in two major clades of bacteria (i.e. Firmicutes and Bacteroidetes), along with *Lactobacillus*. A reduced abundance of *Lactobacillus* in the gut microbiota and its bile salt hydrolase activity leads to the accumulation of the conjugated bile acid tauro-β-muricholic acid

and inhibits the intestinal farnesoid X receptor. By contrasting the minimum nutrients required in the presence and absence of bile salt, we found that the availability of a sulfur source is the single nutrient essentiality that differed between the two cases. In another study, we used a core microbial model of the representative Firmicutes and Bacteroidetes, as well as *Lactococcus garvieae* to model equol production in the gut community. It has been reported by Magee et al. that a higher fraction of the Asian population is able to produce equol compared to the Western population, however increased soy consumption did not convert a western non-producer into an equol producer. By modeling equol production at various growth ratios, we observed that equol production was indeed maximal at high *L. garvieae* abundance and low sugar availability to *L. garvieae*. In a low sugar environment, it appears that *L. garvieae* coverts diadzein (a soy ingredient) into equol as a means of replenishing the NAD^+ pool.

Reaction and metabolite data for construction of both maize and gut microbial metabolic model were primarily derived from the curated MetRxn database. The MetRxn knowledgebase is a unified repository of metabolite and reaction information from various metabolic models and databases. Overlapping information from 8 databases and 112 metabolic models was curated and standardized into 44,784 unique reactions and a million plus unique metabolites. During curation, incompatibilities related content representation, stoichiometric errors such as elemental or charge imbalances, and incomplete atomistic details were resolved and corrected for using a host of cheminformatics, lexicographic and phonetic algorithms. Users can access the standardized repository on a searchable web interface at www.metrxn.che.psu.edu. In addition, all charge and mass balanced reactions within the database are processed by our novel algorithm; Canonical Labelling for Clique Approximation (CLCA). CLCA leverages prime factorization to quickly generate unique molecular graphs, detect symmetries for all metabolites, and resolve bond transformation and atom transition information in each reaction. Bond transformation information was further leveraged to recognize the recurring schemes in substrate to product conversions across all reactions. The common substrate-product transformation schemes identified in 44,784 reactions have been encoded as 6,211 reaction rules within MetRxn. These reaction rules will enable users to annotate the otherwise regarded specialist enzyme with putative secondary activity; and design novel biotransformation schemes to various xenometabolites. Enzyme promiscuity annotations are expected to expand the metabolic potential and fill the many biochemical gaps between phenotype and model predictions.

Metabolic models used in 13-C MFA generally include a limited number of reactions from the central metabolic network. CLCA atom transition information is utilized for the construction of genome-scale mapping models (GSMM) to demonstrate the feasibility of genome-scale MFA and address the impact of model scale-up on prediction fidelity of metabolic fluxes using 13-C MFA. We have compared and contrasted fluxes and ranges estimated by minimizing the sum of square of differences between predicted and experimentally measured labeling patterns using a core model (75 reactions and 65 metabolites) based on central metabolism and a GSMM (697 reactions and 595 metabolites) obtained upon eliminating inactive reactions from *i*AF1260. While both the topology and estimated values of the metabolic fluxes remain largely consistent between the base and GSMM, 20 key reactions in central metabolism and transhydrogenase have wider flux ranges in the GSMM due to the possible activity of alternate routes and futile cycles. Inferred ranges for 81% of the reactions in the GSM model varied less than one-tenth of the basis glucose uptake rate because as many as 411 reactions in the GSM are growth coupled and determined by measured growth rate. Finally, the loss of information associated with mapping fluxes from MFA on a core model to a GSM model is quantified to reveal that the flux range of 295 reactions was narrower than supported by data, demonstrating that assumptions made during the core model construction propagate onto the GSM model leading to possibly erroneous conclusions about reaction flux identifiability.

Supported by funding from the U.S. Department of Energy by grant DE-FG02-05ER25684.