21. Towards a Comprehensive Knowledge Base for the Marine Diatom Phaeodactylum tricornutum

Jennifer Levering[1*](jlevering@ucsd.edu), Jared Broddrick,[1] Philip Miller,[1] Bernhard Ø Palsson,[1] Graham Peers,[2] Christopher L. Dupont,[3] Karsten Zengler,[1] Andrew E Allen (PI)[3,4]

[1]Bioengineering Department, University of California San Diego, La Jolla, CA, USA, [2]Colorado State University, Fort Collins, CO, USA, [3]J. Craig Venter Institute, La Jolla, CA, USA, [4]Integrative Oceanography Division, Scripps Institute of Oceanography, University of California San Diego, La Jolla, CA, USA

Project Goals: Genome-scale metabolic models are fundamental for the analysis of cellular processes at a systems level and represent an ideal organizational framework for analyses of functional genomics, experimental work and computational studies. In recent years, there has been an increasing interest in high-quality metabolic reconstructions of phototrophic organisms and robust computational tools to integrate 'omic' data from these organisms within genome-scale models. The approach of the project is to combine cutting- edge genome manipulation and physiological characterization with metabolic modeling. The ultimate goal is the exploration of next-generation biofuels through a comprehensive understanding of light-driven lipid metabolism in the model marine diatom Phaeodactylum tricornutum.

Bottom-up reconstructions are biochemically, genetically and genomically structured knowledge-bases that contain information such as reaction stoichiometry, reaction reversibility, and the association between genes, proteins and reactions. The first step in the genome-scale metabolic network reconstruction process involves the generation of a draft reconstruction based on the organism's genome and manually curated reference models. In the second step the draft reconstruction is manually refined using several resources such as genome annotation, primary organism-specific literature, experimental data and databases as KEGG and UniProt. During the manual curation phase the given information for each pathway is reviewed and added if missing.

One challenge of building metabolic network reconstructions for eukaryotic organisms is the prediction of subcellular protein localization. Based on previous work by Sunaga et al. (2014) we implemented a pipeline that combines several bioinformatic tools such as TargetP, SignalP, HECATR, and Mitoprot to generate a putative localization for each protein in the reconstruction. It is well known that protein localization in Phaeodactylum tricornutum (Pt) differs from other algae, such as Chlamydomonas reinhardtii, and pathways are spread among multiple compartments. Comparing the pipeline results with literature information facilitated the assignment of pathway localization.

The reconstructed metabolic network accounts for roughly 1,000 genes associated with approximately 4,800 reactions including over 200 intracellular transport reactions and about 2,500 metabolites distributed across six compartments, namely cytoplasm, extracellular space, chloroplast including thylakoid lumen, mitochondria and peroxisome. The reactions are mass and charge balanced and roughly 96% of the reactions have associated gene-reaction rules. Different identifiers such as KEGG compound and reactions IDs, ChEBIs and InChIs were used to annotate metabolites and reactions. These identifiers are also used in other metabolic network reconstructions and enable the comparison of the metabolic network of Pt to the ones of other microalgae, e.g. C. reinhardtii and Synechocystis sp. PCC6803.

Finally, using the COBRA Toolbox the manually curated reconstruction is converted into a mathematical model. This model will be evaluated and tested against well-known metabolic capabilities of Pt such as growth rate, by-product formation and secretion. An objective function is necessary to compute an optimal network state. To be able to predict growth rates the biomass objective function was chosen. This

objective accounts for both the composition of the cell and the energetic requirements necessary to generate biomass such as amino acids, lipids, carbohydrates, pigments and nucleotides. Currently, we are testing the models capabilities to produce all biomass components and fill gaps if necessary. The next step includes the exploration of the exact experimental derived biochemical composition of the cell and growth rates under different conditions to constrain the model and simulate cell growth under different conditions.

The database support for the model construction was based on two components. The first is the database repository of BiGG (bigg.ucsd.edu) in which highly curated metabolic models are stored within an Oracle system. Metabolite, reaction and gene tables allow cross model comparisons, as well as, links to external references such as KEGG and CAS. In development is an extensive revision and upgrade to be known as BiGG 2.0 and be based on a new Postgres database along with a python language backend system. Within this release, the number of curated models will be increased by an order of magnitude. New functionality will also be introduced. These include an open web service interface or API to allow binding to model, reaction and metabolite SQL queries. A versioning system will allow tracking of model changes over the development cycle. Lastly, the webpage has been redesigned to promote a user friendly, but powerful interface.

Besides reconstructing the metabolic network, a regulatory network is being built based on RNA sequencing data under multiple conditions. After normalizing the raw counts, the data set was filtered for all the genes present in the metabolic reconstruction. As a first step in the network inference, the dimensionality of the input data was reduced by grouping genes that are co- regulated into clusters. The resulting clustering is used to obtain a global regulatory network for Pt suggesting regulatory interactions. This network is used to explore Pt's global expression under novel perturbations.

Supported by DOE-DE-SC0008593, Optimization of Energy Flow Through Synthetic Metabolic Modules