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

Jennifer Levering[1*](jlevering@ucsd.edu), Alessandra A. Gallina[1], Philip Miller[1], Adam Feist[1], Karsten Zengler[1], Graham Peers[2] (Co-PI), Bernhard Ø Palsson[1] (Co-PI), Christopher L. Dupont[2] (Co-PI), **Andrew E Allen (PI)**[3,4]

[1]Bioengineering Department, University of California San Diego, La Jolla, CA, USA and [2]Colorado State University, Fort Collins, CO, [3]J. Craig Venter Institute, La Jolla, CA, USA, [4]Integrative Oceanography Division, Scripps Institution of Oceanography, University of California, San Diego, CA 92093, United States

Project Goals: Genome-scale metabolic models are fundamental for the analysis of cellular processes at a system level and represent an ideal organizational framework for analyses of functional genomics experimental data 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 more comprehensive understanding of light-driven lipid metabolism in a model marine diatom.

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 organisms genome annotation and manually curated reference models. To obtain an automated reconstruction the RAVEN Toolbox was used. The draft reconstruction accounts for 589 genes associated with 835 reactions and 1027 metabolites distributed across 5 compartments, namely cytosol, chloroplast, endoplasmic reticulum, mitochondria and peroxisome. In the second step the draft reconstruction is manually refined. The functions in the automated draft reconstruction will be evaluated against organism-specific literature and data. 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 *Phaeodactylum tricornutum* (Pt) such as growth rate, by-products and secretion.

The genome-scale model (GEM) requires an organism-specific biomass objective function which accounts for both the composition of the cell and the energetic requirements necessary to generate biomass. An initial effort has been made in order to gather Pt-specific biochemical data from the available literature. Some of the main differences with the used references models include lipid content (e.g. better representation of C14 and C16 acyl chains), photosynthetic pigments and carbohydrate storage (i.e. chrysolaminarin). The main conclusion has been the need to generate additional data to account for variability of cellular biochemical composition in different growth conditions. Currently, experimental determination of the biochemical composition is being carried out. Given the importance of a comprehensive organization of the available data and information to be used during the model curation, an effort is being made to generate a Pt- specific bibliome database. This type of database represents a useful concept and tool likely to become increasingly used. The up-to date available literature has been collected and currently 1212 publications are being manually screened and categorized. A web-based interface has been implemented in order to provide a user-friendly tool for the scientific community.

A Postgres relational database was developed to provide storage and retrieval capabilities for the GEM. The relational database stores the components of the model each of which can be queried for comparative analyses. Management of the model versions and development history are also made possible through the relational database. The database supplies the model as an exported SBML or through an SQL interface for direct retrieval and data management. A web interface will provide flexible public access and a graphical report generation facility.

Besides reconstructing the metabolic network, a regulatory network is being built based on RNA sequencing data under multiple conditions. The first step in the network inference is the reduction of dimensionality by grouping genes that are co-regulated into clusters using cMonkey. 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