

A knowledge base for computational pathway reconstruction

Peng Wang, Zhengchang Su, Phuongan Dam, Ying Xu

Computational Systems Biology Laboratory

Department of Biochemistry and Molecular Biology, University of Georgia

Abstract

The Synechococcus WH8102 knowledge base (<http://www.csbl.bmb.uga.edu/WH8102>) is a web based relational database developed to facilitate computational effort to reconstruct regulatory pathways and serve as a gateway for biologist to access the data. It is the repertoire that integrates a variety of knowledge derived both from literature and computational prediction. Those data are organized in hierarchical fashion. The basic building blocks are functional annotation and structure prediction of individual molecule. Those data are then organized into clusters based on computationally predicted operon, regulon and molecular complexes. Finally all data are compiled into pathways derived from combined efforts of literature mining and computational prediction. A number of tools have been developed to facilitate the data retrieval including a SQL query engineer and several viewers to browse genome, molecular complexes and pathways.

1. Introduction

Marine unicellular cyanobacteria are responsible for an estimated 20–40% of chlorophyll biomass and carbon fixation in the oceans [1]. Cyanobacteria have been selected as one of the model organisms for computational

pathway reconstruction in department of energy's genome to life project. Pathway modeling sits at the frontier of bioinformatics research and draws the resource of large varieties of data. We established the knowledge base to serve as the central resource to coordinate the computational pathway reconstruction efforts.

2. Data content and database implementation

Data in the Synechococcus knowledge base centers on those contents that will facilitate the reconstruction of regulatory pathways, including: (1) the complete genomic sequence and its functional annotation; (2) structural models for the entire genome; (3) computationally predicted operons; (4) computationally predicted protein-protein interaction map of the entire genome; (5) reconstructed pathways combining literature mining and computational prediction. The database was implemented as a relational database using MySQL and php.

3. Querying and browsing software

We have implemented a series of tools to facilitate data retrieval and visualization. A SQL based query engine was provided to search the database via gene id, genbank id or keywords. The results are displayed as a list to further direct user to their interested results. A java applet was

provided to let the user visualize the entire protein interaction network. The user can center the graph at desired protein, display interactions at different scale and choose how many interactions to display. A graphic genome browser was also provided to facilitate the visualization of genomic information.

4. Future Directions

We will further expand the database to include regulons and all the major regulatory pathways in *Synechococcus*. Tool will also be provided to integrate the database with other web based tools such as protein structure prediction pipeline.

5. References

- [1] B. Palenik, B. Brahamsha, F.W. Larimer, M. Land, L. Hauser, P. Chain, J. Lamerdin, W. Regala, E.E. Allen, J. McCarren, I. Paulsen, A. Dufresne, F. Partensky, E.A. Webb, J. Waterbury. The genome of a motile marine *Synechococcus*. *Nature*. 2003 424:1037-1042.

Acknowledgment

DOE Office of Biological/Environmental Research, Genome to Life Project, "Carbon sequestration in *Synechococcus* sp: from molecular machines to hierarchical modeling"