Data Management Plan

The Chisholm Lab

ProPortal database: enabling data access, integration, and analysis

Our work on *Prochlorococcus* over the past 20 years has produced a lot of data on this organism, including genomics, transcriptomics, proteomics, and physiological data, and global ocean distributions. We also have meta-data for our ocean samples. Our long-term goal is to build a customized data management system to organize and integrate *Prochlorococcus* related data and to evolve and adapt our framework as new data types become available. We have begun this effort with the construction of a web site, *Prochlorococcus* Portal (http://proportal.mit.edu/) that houses access to completely sequenced genomes, microarray data, environmental cell distributions, metagenomic data from the Global Ocean Survey (GOS: http://www.jcvi.org/cms/research/projects/gos/overview/), and publications (including supplementary files) from our group. We will continue to build the Portal with the data emerging from the proposed work. A major addition will be to integrate the extensive metadata available for our samples as well as new physiological measurements of *Prochlorococcus* to address the connections our data can reveal between genomics, taxonomy, and function in this model system (Figure 1).

The ProPortal website is currently divided into four main sections for data access: Genomes, Environmental Cell Distribution, Microarrays, and Metagenome. This data is publicly available and the back-end database structure enables a visitor to the site to ask questions that integrate different data types, such as "Are genes over-expressed under phosphate-starvation conditions in microarray experiments in close proximity on sequenced genomes?"

The Microarray page contains individual experiments, which have information on gene expression under different stress conditions. By way of example, one can click on the gene or gene cluster for a gene of interest – for example the most upregulated gene under a particular stress condition – and, on the gene page, find the position of the gene and its neighbors on the genome using the Genome browser. One can also discover the unperturbed diel expression pattern of the gene, and submit the protein or DNA sequence of the gene to the BLAST tool at NCBI, or look at other genes in the same orthologous cluster.

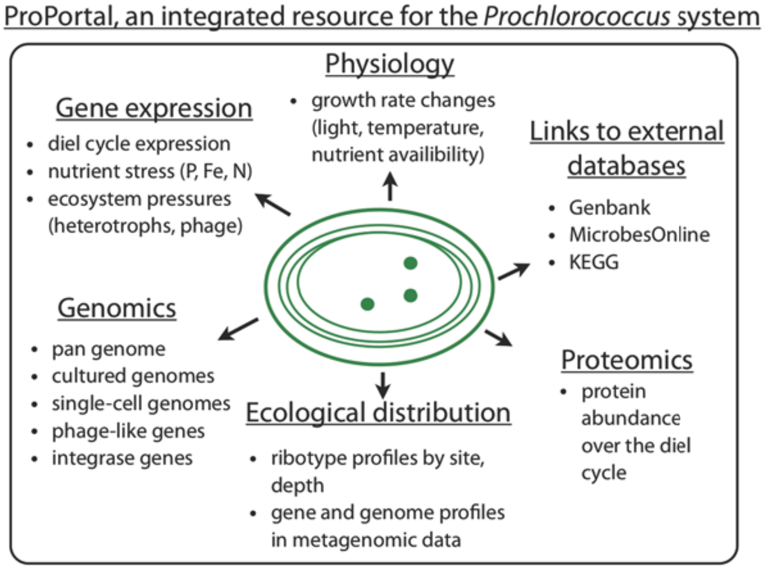
ProPortal, an integrated resource for the Prochlorococcus system
Physiology
• growth rate changes (light, temperature, nutrient availibility)
Gene expression
• diel cycle expression
• nutrient stress (P, Fe, N)
• ecosystem pressures (heterotrophs, phage)
Links to external databases
• Genbank
• MicrobesOnline
• KEGG
Genomics
• pan genome
• cultured genomes
• single-cell genomes
• phage-like genes
• integrase genes
Ecological distribution
• ribotype profiles by site, depth
• gene and genome profiles in metagenomic data
Proteomics
• protein abundance over the diel cycle

Figure 1, Data sources in the ProPortal database. ProPortal integrates genomic, taxonomic, and functional data into a public, searchable, and flexible resource. Some of these data are already in the Portal. Others will be added as part of the proposed research

<u>Future development of the ProPortal resource</u>

The utility of genomic data is greatly increased by connecting it to metadata. Some examples of metadata include: data related to the isolation site of a sample such as nutrient levels, temperature and light and experimental measurements such as the physiological response of different strains to light stress.

As part of another project, facilitated by Dr. Huiming Ding, a bioinformatician who just joined our team, we intend to improve data access and metadata integration in the following ways:

- Update orthologous gene clusters with the newly sequenced cultured genomes and the single cell genomes.
- Overlay RNASeq data on genomes.
- Augmenting gene annotations with pathway functionality by linking to the KEGG pathway database (http://www.genome.jp/kegg/pathway.html)
- Increase connections between metagenomic and genomic data. We currently map metagenomic reads from GOS onto *Prochlorococcus* genomes. We will link gene data to GOS reads to enable visitors to ask questions like: What *Prochlorococcus* gene groups are present at which sites?
- Add a Physiology section to the website to disseminate data such as growth rates by *Prochlorococcus* strain under light and dark stress.
- Improve access to phylogenetic information by providing alignments for rRNA sequences and individual genes.

Database development and public access to database schema

ProPortal is built and maintained with Django (http://www.djangoproject.com/), a high-level, Python-based web framework. While the data presented in ProPortal is specific to *Prochlorococcus*, *Synechococcus*, and cyanophage, the database framework is generalizable for any group working on microbial and phage genomes for which extensive metadata is available. We therefore provide access to our database schema (http://proportal.mit.edu/schema/) should other groups want to replicate the whole database or portions of it to maintain their own data.

Sequence data submission

All sequence data produced will be submitted to Genbank (http://www.ncbi.nlm.nih.gov/genbank/). Furthermore, we provide links *via* ProPortal to directly submit NCBI BLAST jobs for genes of interest, and to search for genes specific to particular NCBI-defined gene clusters. Gene identifiers are also linked to the MicrobesOnline (http://www.microbesonline.org/) database, which provides extensive annotation related to gene structure, function, and phylogeny.

Distribution of Cultures and DNA, and other information

We donate out cultures to the CCMP at the Bigelow Labs (https://ccmp.bigelow.org/) for distribution, and also supply them directly to researchers. We also supply DNA from our cultures upon request. Our public website (http://chisholmlab.mit.edu/) provides contact information for interested groups, as well as protocols, a list of cultures, publications, and genomes available.

Oceanographic data submission

All of our oceanographic data including environmental data an ecotype distributions will be submitted to the Biological and Chemical Oceanography Data Management Office. http://bcodmo.org/home, as we have done in the past.