Genomics:GTL Program Projects

Harvard Medical School

1

Metabolic Network Modeling of *Prochlorococcus marinus*

George M. Church* (g1m1c1@arep.med.harvard.edu), Xiaoxia Lin, Daniel Segrè, Aaron Brandes, and Jeremy Zucker

Harvard Medical School, Boston, MA

The marine cyanobacterium *Prochlorococcus marinus* dominates the phytoplankton in the tropical and subtropical oceans and contributes to a significant fraction of the global photosynthesis. Several strains of *Prochlorococcus* have been sequenced, which provides us a promising starting point for investigating the relationship between genotype and phenotype at a genome scale and with a comparative approach. To achieve the ultimate goal of understanding the metabolism at a systems level, we are developing and utilizing new metabolic network models in several directions.

Comparison and connection of day-night metabolisms

Day-night cycles are known to play a central role in the metabolism of *Prochlorococcus*. We are exploring two approaches to model the difference and connection between day and night. One is to take the full metabolic network and formulate two separate models assuming different nutrient conditions and optimality criteria. Then the flux predictions can be compared to mRNA and protein expression data. In the other approach, we make use of the protein expression data, which helps to reduce the feasible flux space and leads to finer flux predictions.

Construction of metabolic networks

One major challenge in constructing complete and accurate *in silico* metabolic networks for quantitative analysis such as flux balance analysis (FBA) is to identify reactions that are "missed" in the annotation. We have been mainly using Pathway Tools software suite developed by SRI to identify metabolic reactions and are developing new algorithms to construct the "functional" metabolic network from a network perspective. Biochemical reactions with identified enzymes are included and then an "optimal" set of reactions are added such that the network produces the specified growth phenotype given corresponding nutrient conditions. Identification of the missing links will also help to refine the genome annotation. Another problem is that there exist "orphaned enzymes" — experimentally elucidated biochemical reactions whose enzyme has never been sequenced. To address this problem, we are utilizing a pathway hole-filling algorithm developed by SRI and developing bioinformatics techniques to identify candidate genes for these orphaned enzymes.

Analysis of metabolic networks with mass balance and energy balance

Conventional flux balance analysis (FBA) only considers mass balance. We are incorporating constraints representing the second law of thermodynamics, which eliminates thermodynamically infea-

sible fluxes. A subset of the additional constraints exhibits non-convexity, giving rise to substantial difficulty in the solution of the resulting optimization problem. We are developing new methods to overcome this challenge to make full use of combined FBA and EBA (energy balance analysis).

Construction and comparative study of whole-cell metabolic networks of MED4 and other strains

By combining a bioinformatics pipeline for generating metabolic network models from genome annotations and manual inspection/modification, we have constructed the *in silico* metabolic network of central carbon metabolism and amino acid biosynthesis for *Prochlorococcus* MED4, a high-light-adapted strain. We are extending it towards the genome-wide network. In addition, we will construct metabolic network models for the other sequenced strains, including the low-light-adapted MIT9313. Comparison of the structures of their metabolic networks and the calculated flux distributions under varying conditions will enable us to understand at a systems level how these different strains adapt their metabolisms to the different environments.

Project Web site: http://arep.med.harvard.edu/DOEGTL/

2

Quantitative Proteomics of *Prochlorococcus marinus*

Kyriacos C. Leptos[1]* (leptos@fas.harvard.edu), Jacob D. Jaffe[1], Eric Zinser[2], Debbie Lindell[2], Sallie W. Chisholm[2], and George M. Church[1]

[1]Harvard Medical School, Boston, MA and [2]Massachusetts Institute of Technology, Cambridge, MA

With the capability of performing whole-cell proteome analysis, a need to extent the above capability to whole-cell protein quantitation has proven to be a necessity. For this purpose we developed MapQuant, a platform-independent open-source software, which given large amounts of mass-spectrometry data, outputs quantitation for any organic species in the sample. We have previously applied MapQuant in the study of standardization samples at different concentrations on both LCQ and LTQ-FT spectrometers and also in the content of protein mixture of medium complexity and have showed linearity of signal with respect to the quantity of protein introduced.

The *Prochlorococcus* species is an abundant marine cyanobacterium that contributes significantly to the primary production of the ocean and whose life cycle is synchronized to the solar day (the "diel cycle"). In this study we leverage previously obtained protein identification data and the capabilities of MapQuant to quantify the proteins in a time-series dataset which includes 25 time points distributed along a 48-hour period (two diel cycles) of the strain MED4 of *Prochlorococcus marinus*. Protein samples from the growing culture were collected in duplicate and digested into peptides using trypsin, each time-point sample subjected to liquid chromatography coupled to hybrid linear ion trap-FTICR mass spectrometry, giving rise to a total 150 LC/MS experiments. The data acquisition took place on a Finnigan LTQ-FT mass spectrometer and it involved the acquisition of maximum two MS/MS spectra per MS spectrum. MS/MS spectra were interpreted using the program SEQUEST. The cross-correlation scores assigned to peptides that scored were filtered using thresholds to take into account false-positive results and the peptides were compiled into a summary list. This list of highly scored peptides was used as landmarks for evaluating MapQuant performance. MapQuant algorithms include morphological operations, noise filtering, watershed segmentation, peak finding and fitting, peak clustering and isotopic-cluster deconvolution and fitting using binomially distributed clusters of gaussioid peaks.

* Presenting author

MapQuant outputs a list of potential organic species, by reporting four physical attributes for each isotopic cluster that it deconvolves. Those attributes are the m/z and the retention time (RT) of the monoisotopic peak, its charge and its carbon content. We have employed an m/z, RT and charge matching approach to assigning MapQuant Isotopic Clusters (MQIC) to the landmark peptides identified by SEQUEST in the same run with 91% success. However, MQICs that were assigned to a peptide using SEQUEST constitute 3% of the total MQIC found in a 2-D map. We are in the process of developing a matching algorithm that will be able to assign identities to unassigned MQICs. This approach will utilize SEQUEST peptides identified in the same organism *Prochlorococcus marinus* MED4 in five LC/LC/MS/MS experiments performed in the past, which correspond to five different environmental conditions. The matching algorithm should enable mapping of many of the remaining (97%) of the unidentified MQICs.

Our end goal is to be able to perform quantitation for most peptides found in the 25 time-points of the two diel cycles and hope to understand how carbon fixation, light-response and cell division are coordinated throughout the daily cycle.

Project Web site: http://arep.med.harvard.edu/DOEGTL/

3

Genome Sequencing from Single Cells with Ploning

Kun Zhang[1]* (kzhang@genetics.med.harvard.edu), Adam C. Martiny[2], Nikkos B. Reppas[1], Sallie W. Chisholm[2], and George M. Church[1]

[1]Harvard Medical School, Boston, MA and [2]Massachusetts Institute of Technology, Cambridge, MA

Currently genome sequencing is performed on cell populations because of the difficulty in preparing sequencing template from single cells. This makes the genome sequences of many difficult-to-culture organisms inaccessible or poorly assembled. We have developed a method that enables genome sequencing from a single cell by performing polymerase cloning (ploning). In this method, we prepare sequencing templates from single cells with real-time multiple displacement amplification (rtMDA), which allows us to tackle the big technical challenge in single-cell whole genome analysis: to detect and suppress spurious amplification while targeting a single molecule of a microbial chromosome.

Experiments on *Escherichia coli* show that, (1) an amplification magnitude of 10^9 was achieved by rtMDA, (2) strain-specific genetic signatures were preserved, (3) neither spurious amplification product nor chimeric sequence was detected, (4) an estimated 97% of the target genome could be recovered from a polymerase clone (plone) at the 10X sequencing depth. The remaining regions are not missing, but present at lower copy numbers, and easily recovered by PCR. Since the low-coverage regions seem random, genome coverage can be improved by pooling the sequencing reads from two or more plones of the same type of cells during the assembly stage. Furthermore, we successfully performed ploning on both fresh and frozen *Prochlorococcus* cells, and obtained nearly complete coverage on both strains (MED4 and MIT9312) we tested. Plones of single cells from an ocean sample (from the Hawaii Ocean Time-series) are being screened for *Prochlorococcus* cells for genome sequencing. Initial results indicate successful amplification of single *Prochlorococcus* cells from this sample. After further screening of genome coverage, whole genome shot-gun sequencing will be performed on a few selected plones.