Computational Methodologies for Identification of Phenotype-specific Biological Processes in Microbial Communities

Nagiza F. Samatova[1,2] (samatovan@ornl.gov), **Chongle Pan**[2] (pcl@ornl.gov), **Robert (Bob) L. Hettich**[2] (hettichrl@ornl.gov), **Jill Banfield**[3] (jbanfield@berkeley.edu), Matthew C Schmidt[1,2] (mcschmid@ncsu.edu), William Hendrix[1,2] (wthendri@ncsu.edu), Andrea M Rocha[4] (amrocha@mail.usf.edu), and Kanchana Padmanabhan[1,2,*] (kpadman@ncsu.edu)

[1]North Carolina State University, Raleigh, NC 27695; [2]Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831; [3]University of California, Berkeley, CA 94720; [4]University of South Florida, Tampa, FL 33620;
*Presenting author

http://freescience.org/cs/gtl/

Project Goals: Develop computational methodologies to identify phenotype-related biological systems and their interplays. The crux of the project is a procedure for: (1) identification and characterization of phenotype-related genes; (2) identification of phenotype-biased cellular subsystems; (3) reconstruction of phenotype-specific metabolic pathways; and (4) elucidation of symbiotic/competing crosstalks between these pathways.

Phenotypes that certain microorganisms express assist in activities like breaking down the lignocellulosic barrier of biomass, and the biodegradation of various environmental contaminants. Tackling the various bioremediation and bioenergy problems with the help of genetic engineering requires the understanding of the cellular subsystems that help with the phenotype-expression in the organism. To supplement experimentation methods, computational methodologies need to be used. These methods could reveal phenotype-related "signals'" and their combinatorial interplay by comparing potentially hundreds of microorganisms with millions of genes organized into thousands of cellular subsystems.

We developed **graph-theoretical & statistical methods and released open-source software for *in silico* prediction of cellular subsystems related to the expression of a target phenotype**.

The Network Instance-Based Biased Subgraph Search (**NIBBS)** [5] is capable of comparing hundreds of genome-scale metabolic networks to identify ***metabolic subsystems that are statistically biased toward phenotype-expressing organism***. NIBBS identifies the set of all phenotype-*biased* metabolic network motifs. From the results obtained, for example, for bio-hydrogen production phenotype, NIBBS was able to identify metabolic subsystems like acetate and butyrate fermentation, fatty acid biosynthesis, amino acid metabolism, and nitrogen metabolism. The validation for those results was found in literature. In addition, NIBBs was also able to provide clues about pathway cross-talks, including those involved in production of Acetyl-CoA

The ***α,β*-motif finder** [2] and **bi-clustering** [3] approach allow for identification of ***phenotype-related functional modules*** that, in addition to metabolic subsystems, could include their regulators, sensors, transporters. The functional modules identified are present across a set of phenotype-expressing organisms. This approach can identify conserved modules across any subset of the input organisms and hence can identify sub-phenotype-specific modules as well. From the results obtained, for example, for the light fermentation hydrogen production phenotype, the modules associated with N-fixation, iron regulation, and ammonia uptake were found. When applied to the bio-hydrogen production phenotype, this method was able to identify modules responsible for synthesis, metal insertion, or regulation of hydrogenase and nitrogenase enzymes complexes. Within hydrogen producers, these two complexes

play important roles in production of hydrogen. On further analysis, various pathway crosstalks including those between iron and nitrogen related metabolic pathways and iron uptake and ammonia assimilation were predicted.

The Dense ENriched Subgraph Enumeration (**DENSE**) [1] algorithm allows for ***incorporating partial prior knowledge about the proteins involved in a phenotype-related process into the identification of sets of functionally associated proteins in a phenotype-expressing organism***. This method, when applied to the protein functional association network of the *Clostridium acetobutylicum,* a dark fermentative, hydrogen producing bacterium was able to predict known and novel associations including those with regulatory, signaling, and uncharacterized proteins

We supplemented the computational methodologies discussed so far with a method to computationally analyze the results of the methods for biological significance and improve functional annotation [4]. This method is different from other existing significance analysis techniques in that it takes into account an inherent design principle of biological networks, *hierarchical modularity*. Functionally homogenous modules combine in a hierarchical manner into larger, less cohesive subsystems. The method quantifies biological significance both with a score known as **Hierarchical Modularity Score (HMS)** and a confidence of the score via a *p*-value. Additionally, it provides the hierarchical functional annotation of the modules.

References

1. Willam Hendrix, Andrea M Rocha, Kanchana Padmanabhan, Alok Choudhary, Kathleen Scott, James R Mihelcic and Nagiza F Samatova, *DENSE: efficient and prior knowledge-driven discovery of phenotype-associated protein functional modules*, BMC Systems Biology 2011, **5**:172.
2. Mathew C Schmidt**,** Andrea M Rocha**,** Kanchana Padmanabhan**,** Zhengzhang Chen**,** Kathleen Scott**,** James R Mihelcic and Nagiza F Samatova**,** *Efficient alpha,beta-motif Finder for Identification of Phenotype-related Functional Modules*, BMC Bioinformatics 2011, **12**:440.
3. Kevin Wilson, Andrea Rocha, Kanchana Padmanabhan, Kuanyu Wang, Zhengzhang Chen, Ye Jin, James R Mihelcic and Nagiza F Samatova, *Detecting Pathway Cross-talks by Analyzing Conserved Functional Modules across Multiple Phenotype-Expressing Organisms*, IEEE International Conference Bioinformatics & Biomedicine, 2011.
4. Kanchana Padmanabhan, Kuanyu Wang, Nagiza Samatova, *Functional Annotation of Hierarchical Modularity,* (under review).
5. Mathew C Schmidt**,** Andrea M Rocha**,** Kanchana Padmanabhan**,** Yekaterina Shpanskaya **,** Jill Banfield, Kathleen Scott**,** James R Mihelcic and Nagiza F Samatova, *NIBBS-Search for Fast and Accurate Prediction of*
6. *Phenotype-Biased Metabolic Systems,* (under review).
7. Zhengzhang Chen, Kanchana Padmanabhan, Andrea M. Rocha, Yekaterina Shpanskaya, James R. Mihelcic and Nagiza F. Samatova, *SPICE: Discovery of Phenotype-Determining Component Interplays,* (under review).

This research is supported by both the Office of Biological and Environmental Research and by the Office of Advanced Scientific Computing Research of the U.S. Department of Energy.