Bioinformatics methods for metabolomics based biomarker detection in functional genomics studies

Preeti Bais
Major Professors – Dr. Julie Dickerson and Dr. Basil Nikolau

Abstract

The biochemical and physiological function of a large proportion of the approximately 27,000 protein-encoding genes in the Arabidopsis genome is experimentally undetermined using sequence homology techniques alone. This thesis presents a set of bioinformatics resources including a software platform for data visualization and data analysis that address the key issues in incorporating the metabolomics data for functional genomics studies.

Since a single metabolomics technique cannot cover the whole metabolome, multiple mass spectrometry based metabolomics platforms are integrated together to get biomarkers across a wide range of metabolite families. The use of different metabolomics platforms increases the coverage of the metabolome, but multiple platforms present significant challenges on integrating data across the platforms. Different strategies for integrating the metabolomics abundance data from multiple platforms are compared to find the ideal method for biomarker discovery.

One of the biggest challenges in metabolomics is to understand the role of structrully unknown metabolites which constitute a major part of the detected metabolites in any large scale metabolomics study. A new method of putatively identifying unknown metabolites by partial correlation networks is proposed. Gaussian graphical models which are based on first order partial correlation networks built across a large range of mutant lines are sparse and mimic the actual biochemical topology. These correlation networks are preserved across many different mutations and can provide a cost effective alternative for getting more insight into the underlying biochemical network and form hypotheses about the novel pathways. A comprehensive study of 70 single gene knock mutants vs. wild type samples is performed using Random Forest machine learning algorithm and a biomarker database for the key metabolites including the putative identifications of unknown metabolites is built.

A proof-of-concept analysis on the oxoprolinase (OXP1) and gamma-glutamyl transpeptidase (GGT1 and GGT2) single gene knock-out mutants in the glutathione degradation (GSH) pathway of the *Arabidopsis* confirms the known biology that OXP1 is responsible for conversion of 5-oxoproline (5-OP) to glutamic acid. In addition, GGT1/GGT2 analysis supports the hypothesis that the GGT genes may not be major contributors for the 5-OP production. Also, the GGT2 mutation does not appear to alter the biochemical profile of the cells in comparison to the wild type samples, suggesting that it may be a redundant gene.

The metabolomics database, the biomarker database and the data mining tools are implemented in a web based software suite at www.plantmetabolomics.org.