CSI 660 Project Proposal

Problem

Given microarray data of gene expression levels in different conditions at different times and network data showing how these genes interact, as well as annotations of different genes' roles within the cell, find groups of interacting genes that behave differently between the two conditions.

Motivation

Microarrays enable large volumes of data about gene expression to be gathered in a single experiment. However, the data is noisy and hard to interpret. Integrating interaction network information with the expression data can help uncover pathways of related genes that behave differently in the different conditions, which can lead to identification of targets for therapeutic intervention to treat disease or alter the course of development.

Related Work

1. One of the first papers, cited by many others as the originator of integrating network and microarray data, is Ideker, T. et al. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics, 18 (Suppl. 1), S233–S240. In it, the authors label nodes in the gene interaction network by z-scores of the expression levels of the genes and use a simulated annealing heuristic to extract high scoring subnetworks, calling them biologically relevant.

2. An exact approach was detailed in Dittrich et. al. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. There the authors transform the problem from (1) to a Prize Collecting Steiner Tree problem and solve it optimally. They assert that their formulation performs well on large datasets despite being NP-Complete, but don't give any details as to what characteristics of the data allow for this good performance.

3. The above methods suffer from the problem that if a subnetwork has a central gene that is not among the more highly differentially expressed genes, even though the genes around it are, they will fail to pick up on it and may miss the subnetwork entirely. In Poirel et. al., Reconciling differential gene expression data with molecular interaction networks, they attempt to rectify this by "reconciling" genes' p-values with those of the genes around them in the interaction network. They investigate 4 different methods that minimize a combination of the new p-value of a gene from its old one and the new p-value from the new p-values of the genes around it.

Formulation

This project will extend the method in 3. by also reconciling in-vivo and ex-vivo data. Thus, as compared to the former, for which new p-values were calculated by minimizing the formula

\[ E = q \cdot \sum (p_{\text{new}}(v) - p_{\text{old}}(v))^2 + (1 - q) \cdot \sum (p_{\text{new}}(v) - p_{\text{new}}(u))^2 \]

the new formula to minimize will be

\[ E = \alpha \cdot \sum (p_{\text{newexvivo}}(v) - p_{\text{oldexvivo}}(v))^2 + \beta \cdot \sum (p_{\text{exivivonew}}(v) - p_{\text{exivivonew}}(u))^2 + (1 - \alpha - \beta) \cdot \sum (p_{\text{newexvivo}}(v) - p_{\text{invivo}}(v))^2 \]
Data

Network data has been taken from GeneMania (http://genemania.org) for Mus musculus

Annotations were taken from Gene Ontology Consortium (http://geneontology.org)

In-vivo and ex-vivo gene expression data for submandibular salivary glands were provided by the Larsen lab (ex-vivo from their own experiments, in-vivo from the Salivary Gland Molecular Anatomy Project (http://sgmap.nidcr.nih.gov)