Introduction

Microarrays make it easy to gather large volumes of gene expression data in a single experiment. However, these data are noisy and hard to interpret. A recent trend is to integrate microarray data with gene interaction networks in order to pick out subnetworks in which the expression profiles of the genes within them seem to be changing in a similar way. In doing this, it is hoped that one will be able to identify pathways that are important in various biological processes, such as diseases like cancer or developmental processes like branching morphogenesis.

In prior work, this is usually done by analyzing the different conditions (diseased versus non-diseased) separately. In this work we make the contribution of finding differentially expressed subnetworks between two conditions by allowing both conditions to contribute to the difference in the subnetworks. We do this by a network reconciliation process.

Prior Work

Combining microarray data with gene interaction networks was pioneered by Ideker et al. In their work, they overlaid z-scores of differential expression values on a gene interaction network. They scored subnetworks by adding up the z-scores within them, normalized by the square root of the size of the subnetwork. Since they found that finding the optimal high scoring subnetworks was an NP-hard problem, they developed a heuristic based on simulated annealing to find an approximate solution. Dittrich et al. found an exact solution based on integer linear programming and the prize collecting Steiner tree problem, which they claimed worked well in practice, though they provided no details on what characteristics were required of the data for the solution to be computed in reasonable time. The work on which this project is based, by Poirel et al., borrowed the idea of reconciling label information of nodes in a network with the labels of their neighbors until convergence from Zhou et al., and applied it to microarray data and gene interaction networks.

Methods

From Poirel et al., the original formulation of the problem is to minimize

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

The new formula to minimize will be to minimize

\[ E = \alpha_{\text{EV}} \cdot \sum (r_{\text{new}}(v) - s_{\text{exvivo}}(v))^2 + \alpha_{\text{IV}} \cdot \sum (r_{\text{new}}(v) - s_{\text{invivo}}(v))^2 + (1 - \alpha_{\text{EV}} - \alpha_{\text{IV}}) \cdot \sum (r_{\text{new}}(v) - r_{\text{new}}(u))^2 \]

This is done by taking the derivative with respect to each \( r_{\text{new}}(v) \) and iteratively changing values of each one until the maximum change at any iteration is sufficiently small, or until the maximum number of iterations has been reached. Algorithm pseudocode is as follows:
Set all values in the reconciled network to an initial value (say the starting ex vivo value)

While the maximum number of iterations has not been reached

    For each node in the reconciled network

        Get its current value

        Get the values of its neighbors

        Adjust the value of the node according to the formula

        Check the difference between the new value and the old value

    If no node differed from its previous value by more than the iteration threshold, break

If looking up a node's value takes $O(1)$ time, then the algorithm takes $O(IVE)$ time, where $I$ is the number of iterations, $V$ is the number of nodes, and $E$ is the number of edges.

Results

![Figure 1. Heatmaps of subnetwork homogeneity; average (a) and maximum (b), showing that the returned subnetworks tend to be more homogenous as one of $\alpha_{iv}$ or $\alpha_{ev}$ gets close to 1, at which point only the starting values of in-vivo or ex-vivo matter in the formulation.](image)
Figure 2. Subnetwork size and size vs. score for three different values of the parameters αiv and αev. The subnetworks returned were of genes that differed either all positively or all negatively from their original ex-vivo values. Most of the returned subnetworks were very small (1-4 genes); however, each set of parameters also returned a couple very large subnetworks, indicating that the difference threshold may need to be raised.

References


Ideker, T. et al. 2002 Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics, 18 (Suppl I), S233-S240
