1 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.

2 Prior Work

Combining microarray data with molecular interaction networks was pioneered by Ideker et al [3]. In their work, they computed the z-scores of the differential expression for each gene and overlaid them on the corresponding node of a gene interaction network. They then summed the scores in each subnetwork, normalizing by the square root of the size of the subnetwork, yielding a scoring criteria for subnetworks which could be optimized by maximizing it. They showed that finding optimal subnetworks was NP-hard, motivating their provided heuristic based on simulated annealing.

Dittrich and Klau et al [2] built upon this work by providing an algorithm based on integer linear programming and the prize collecting Steiner tree problem which provided provably optimal subnetworks as well as “feasible” subnetworks which would differ from the optimal by at most a guaranteed amount. Despite
the fact that this algorithm could not be guaranteed to run in polynomial time in the general case, they claimed that for most practical instances, the running time would be reasonable.

Keller et al [4] provided a different means of computing differentially regulated subnetworks. They ranked the genes by their differential expression between two conditions and computed paths through the network, utilizing a Kolmorogov-Smirnov-like test on the ranks of the genes and returning subnetworks of a given size with maximal deviation of the obtained score from zero (minimal p-value). Poirel et al [5], upon which this work is based, used a network reconciliation method to find pathways with significant differential expression between two conditions. They took the differential expression of genes between two conditions and computed p-values for them, then overlaid these on a gene interaction network and reconciled the network by minimizing a cost function which was based on the difference between a gene’s differential expression and that of its neighbors as well as the difference between the reconciled expression level and the original one. In this way, subnetworks could be found in which one of the genes was not significantly differentially regulated, but neighbored many genes that were.

3 Methods

Given an undirected graph G(V, E) representing gene interactions and gene differential expression values, Poirel et al find differentially expressed subnetworks by minimizing the formula

\[ Cost = \alpha \sum_{v \in V} (r(v) - s(v))^2 + (1 - \alpha) \sum_{(u,v) \in E} (r(v) - r(u))^2 \]

where \( r(v) \) is the value of node v in the reconciled network and \( s(v) \) is its starting value.

We modify this formula by adding a second term corresponding to a second starting value (in this case we constrain the reconciled gene to be near its starting ex vivo value as well as its starting in vivo value). We also normalize the sums:

\[ Cost = \alpha_{ev} \frac{1}{|V|} \sum_{v \in V} (r(v) - s_{ev}(v))^2 + \alpha_{iv} \frac{1}{|V|} \sum_{v \in V} (r(v) - s_{iv}(v))^2 + (1 - \alpha_{ev} - \alpha_{iv}) \frac{1}{|E|} \sum_{(u,v) \in E} (r(v) - r(u))^2 \]

where ev and iv stand for ex vivo and in vivo, respectively.

This formula is minimized by taking all the derivatives with respect to \( r(v) \) and setting them equal to zero. The minimum can be computed iteratively by the following formula:

\[
\begin{cases}
  r(v)_{i+1} = \frac{\alpha_{ev} s_{ev}(v) + \alpha_{iv} s_{iv}(v) + (1 - \alpha_{ev} - \alpha_{iv}) \sum_{(u,v) \in E} r(u)}{\alpha_{ev} + \alpha_{iv} + (1 - \alpha_{ev} - \alpha_{iv}) d(v)} & \text{if } i = 0 \\
  r(v)_{i+1} = \frac{\alpha_{ev} s_{ev}(v) + \alpha_{iv} s_{iv}(v) + (1 - \alpha_{ev} - \alpha_{iv}) \sum_{(u,v) \in E} r(u)}{\alpha_{ev} + \alpha_{iv} + (1 - \alpha_{ev} - \alpha_{iv}) d(v)} & \text{if } i > 0
\end{cases}
\]

where \( d(v) \) is the degree of node v.

Subnetworks are extracted from the reconciled network by thresholding with a difference threshold (dt). Connected components, all of whose nodes differ
from their original values by more than the difference threshold, are taken as the subnetworks of interest. The parameters ($\alpha_{ev}$, $\alpha_{iv}$, and $dt$) are tuned by evaluating subnetwork size (number of nodes) and homogeneity score. The homogeneity score is computed as the percentage of genes in the subnetwork that are associated with a particular GO [1] term at a particular level in the GO hierarchy (being associated with a term at a lower/greater level is considered to be the same as being associated with the ancestor term at the level being considered)

4 Results

Figure 1 (next page) shows the average homogeneity of the returned subnetworks, weighted by subnetwork size, as the parameters $\alpha_{ev}$, $\alpha_{iv}$, and $dt$ are varied. Figure 2 shows the homogeneity of random subnetworks of certain sizes. The homogeneity generally increases with increasing difference threshold, though not outside the margin of error for smaller random subnetworks, at least not for GO level 2 terms. The higher difference thresholds (0.07 and above) may show significance when compared at GO level 4.

Network size was also taken into consideration when choosing the parameter settings. Figure 3 shows the network sizes versus scores for the chosen setting of $dt = 0.07$. At this setting there are at least a few medium size networks for each of the parameter settings for $\alpha_{ev}$ and $\alpha_{iv}$ while there are no networks that are too large. $\alpha_{ev}$ and $\alpha_{iv}$ were chosen to be 0.5 and 0.0 in one setting and 0.33 and 0.33 in another in order to compare the results of having just one network (the ex vivo one) with having two equally weighted. Figure 4 shows the distribution of sizes of subnetworks for each setting. In each case, there are far more singleton nodes than other sized networks, indicating that our thresholding method may not be the most appropriate way to extract subnetworks.

Figure 5 shows a visualization using Cytoscape of the subnetworks extracted from the two settings. At the top are the subnetworks from each setting showing their reconciled expression levels while at bottom are the subnetworks extracted from the equally weighted setting overlaid with ex vivo and in vivo expression levels, respectively. There does not appear to be a great deal of difference between the expression levels in vivo and ex vivo, indicating that perhaps the neighbors of the nodes in the network, which may have had differential expressions close to 0, played a larger role in influencing the differences of the nodes from their original values.
Figure 1: Figures showing that weighted average homogeneity of the returned subnetworks increases somewhat as the difference threshold increases. This trend is more pronounced when the homogeneity is computed at level 4 of the GO hierarchy (as opposed to level 2). The homogeneity is also generally greater when the ex vivo network is weighted higher in the reconciliation, perhaps because this simulates the effect of having a higher difference threshold (fewer nodes are further away from the ex vivo values).
Figure 2: Figures showing the homogeneity of random subnetworks. For the lower level, the homogeneity is substantially decreased for medium and larger sized networks.
Figure 3: Figures showing subnetwork size versus score for different parameter settings a) when the in vivo network is disregarded and b) when the ex vivo and in vivo networks are equally weighted with each other as well as the term representing neighbors in the reconciled network.
Figure 4: Figures showing subnetwork size versus number of subnetworks of that size for different parameter settings a) when the in vivo network is disregarded and b) when the ex vivo and in vivo networks are equally weighted with each other as well as the term representing neighbors in the reconciled network.
Figure 5: Figures visualizing the subnetworks returned by the reconciliation at difference threshold 0.07: a) and b): expression levels in the reconciled networks for $\alpha_{ev} = 0.5$, $\alpha_{iv} = 0.0$ and $\alpha_{ev} = 0.33$, $\alpha_{iv} = 0.33$, respectively. c) and d): ex vivo and in vivo, respectively, expression levels for the networks found with $\alpha_{ev} = 0.33$, $\alpha_{iv} = 0.33$.

References


