Unsupervised posterior analysis of signaling pathways from gene microarray data

Dongxiao Zhu, [1,2] Alfred O Hero[2] and Anand Swaroop[3]
[1]Bioinformatics Program, [2]Departments of EECS, Biomedical Engineering and Statistics,
[3]Departments of Ophthalmology and Human Genetics, University of Michigan, Ann Arbor, MI 48105

***Abstract*— Most gene clustering algorithms only group similarly co-expressed genes into clusters. In light of gene regulation network, many transitively co-expressed genes are also likely to be functionally related. We propose a new clustering approach that is able to group both similarly co-expressed genes and transitively co-expressed genes into tight clusters of interest.**

I. Introduction

Two widespread and complementary procedures to estimate gene co-expression pathways from microarray data are: similar co-expression analysis [1] and transitive co-expression analysis [2]. Similar co-expression analysis uses clustering to assign each gene into a group based on co-expression profile similarity. Since it ignores constraints of the underlying gene regulation network, this can be viewed as a "prior" approach. It is very simply implemented but has the drawback that many irrelevant genes to the biological process are falsely classified. Similar co-expression analysis is also confounded by the fact that genes in the same signaling pathway do not necessarily have similar expression profiles. Transitive co-expression analysis was proposed to overcome these shortcomings [2]. By considering underlying gene regulation network constraints, it can be viewed as a "posterior" approach. Consequently, it can frequently discover novel transitive genes in the pathway that otherwise would be missed by similar co-expression analysis. However, this approach depends completely on the availability of terminal genes that are biologically known to be lying in the same pathway. Moreover, the linear manner of discovery is not very efficient for discovering interconnected pathway components.

Directly inspired by the metabolic network decomposition analysis in Ma et al., 2004 [3], we propose an efficient and powerful posterior approach that integrates the complementary features of similar co-expression and transitive co-expression analyses to cluster genes from co-expression network based on "shortest-path distance". We illustrate the approach and compare with previous similar co-expression and transitive co-expression analyses on a yeast hextose metabolism dataset [4].

II. Section2

We extract a network from microarray data using a FDR Confidence Interval (FDR-CI) based two-stage algorithm with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS) [5]. This technique improves upon previous network extraction methods [2] because our constructed network simultaneously controls error rate and strength of association. Furthermore, it is able to incorporate both linearly and non-linearly co-expressed genes by using non-Euclidean inter-profile distance measures.

The Giant Connected Component (GCC) of an undirected graph $G = (V, E)$, where V is the set of all vertices and E is the set of all edges, is a maximal set of vertices $U \subset V$ such that every pair of vertices u and v in U are reachable from each other. We designed a simple algorithm to extract the GCC from the undirected graph:

- Calculate marginal degree for each vertex in the graph, denoted as K.
- Sort K in the decreasing order,i.e.$K_{(1)}, K_{(2)}, \ldots, K_{(n)}$.
- Start from the best connected to least connected vertices, greedily grow the GCC until the newly formed giant component is not a GCC.

The vertices of the extracted GCC are ordered by connectivity, which facilitates network based analysis since highly connected vertices are often of biological interest. To obtain the same list of vertices but in the original order, the standard depth first search (DFS) algorithm can be used as described in [6]. In both cases, the algorithm runs in polynomial time [6].

Let $\hat{\Gamma}_{ij}$ be the sample correlation coefficient between gene i and j, e.g. estimated from a gene microarray sequence by Pearson or Kendall correlation statistic. Let w_{ij} be the weight of the edge between gene i and gene j. Similar to Zhou et al [2], the w_{ij} is defined as:

$$w_{ij} = (1 - \mathrm{abs}(\hat{\Gamma}_{ij}))^p \tag{1}$$

The integer p is an exponential factor to enhance the differences between low and high correlation. We systematically studied impact of choosing p on the clustering results. The clustering results do not change with a different p for prior clustering, but change mildly for posterior clustering (supplemental table). We determine p based on the biological prior knowledge that known functionally related genes form a tightest cluster.

We use the standard Floyd-Warshall algorithm to search among all-pairs for the shortest-paths within GCC. Let $d_{ij}^{(k)}$ be the weight of a shortest-path from vertex i to vertex j passing through k intermediate vertices. When $k = 0$, there is only one edge between vertex i and vertex j, and we define $d_{ij}^{(0)} = w_{ij}$. A recursive definition of $d_{ij}^{(k)}$ is given by [6]:

$$d_{ij}^{(k)} = \begin{cases} w_{ij}, & k = 0; \\ \min(d_{ij}^{(k-1)}, d_{ik}^{(k-1)} + d_{kj}^{(k-1)}), & k \geq 1. \end{cases} \quad (2)$$

The matrix $D = (d_{ij})$ is called the "shortest-path distance matrix". It can be used as input to many distance matrix based clustering software such as: hierarchical clustering and K-medoids.

III. SECTION3

We sought to compare our posterior clustering approach with prior clustering approach and shortest-path analysis in discovering the structural module (transporter genes and enzyme genes) of yeast hextose metabolism pathway. We constructed a co-expression network using a subset of 997 differentially expressed genes in the yeast dataset [4] with a FDR constraint of 5% and a MAS constraint of 0.6. We then extracted a GCC of 644 genes. The shortest-path distance matrix for GCC was computed according to equation (1) and equation (2), while the distance matrix for all the 997 genes was computed according to equation (1) only.

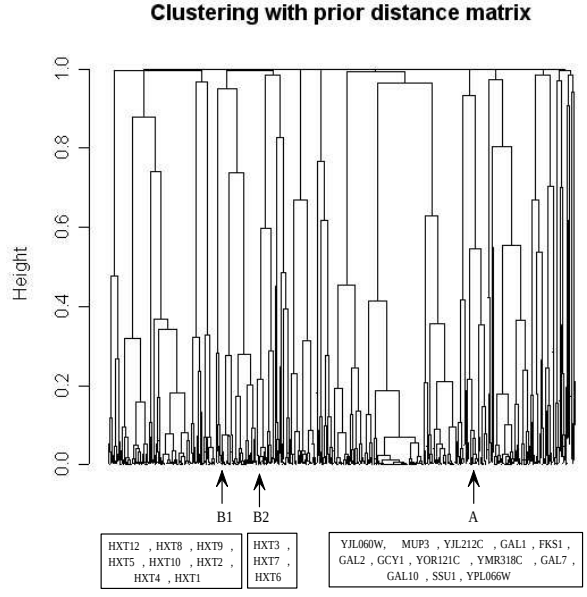

Fig. 1. *Prior clustering: Dendrogram obtained by agglomerative hierarchical clustering using all differentially expressed genes.*

We use the largest geodesic between genes in the clusters for hierarchical clustering [7]. The structural module is separated into three subclusters (A, B1 and B2) in prior clustering (Fig.1) but is integrated into one tight cluster (A) in posterior clustering (Fig.2). Furthermore, enzyme genes (A) form a quite loose cluster in prior clustering (Fig.1, A), e.g. GAL1 and GAL10 are separated by 6 genes while in fact, they should be quite close to each other as shown by our posterior clustering procedure (Fig.2, A).

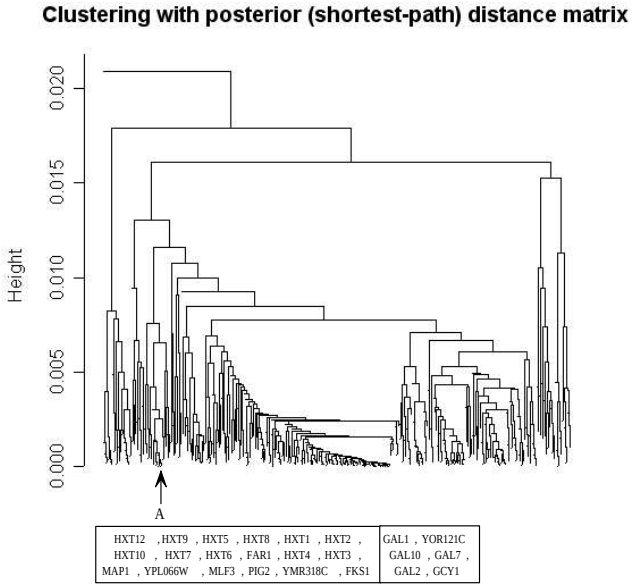

Fig. 2. *Posterior clustering: Dendrogram obtained by agglomerative hierarchical clustering from relevance network.*

In comparison to the shortest-path analysis, the supervised transitive co-expression analysis depends on availability of terminal genes, while our unsupervised method does not. For example, the hypothetical ORF "YPL066W" does not lie in any shortest-path and hence no functional prediction can be made. In contrast, our posterior clustering method strongly supports its role in the hextose transport (Fig.2). Checking the literature more carefully, we found that a system deletion of "YPL066W" exhibits growth defect on a non-fermentable (respiratory) carbon source [8]. This seems to support our prediction. Further experimental work will be necessary to characterize the biological functions of this gene.

REFERENCES

[1] Eisen,M., Spellman,P., Brown,P.O., Botstein,D. "Cluster analysis and display of genome-wide expression patterns." *Proc Natl Acad Sci USA*, 95: 14863-8, 1998.

[2] Zhou,X., Kao,M. and Wong,W.H. "Transitive functional annotation by shortest path analysis of gene expression data." *Proc Natl Acad Sci USA*, 99:12783-12788, 2002.

[3] Ma, H.W., Zhao, X.M. et al. "Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph." *Bioinformatics*, 20, 1870-1876, 2004.

[4] Ideker, T., V. Thorsson, et al. "Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data." *J Comput Biol*, 7(6): 805-17, 2001.

[5] Zhu, D., Hero, A.O. "Gene co-expression network discovery with controlled statistical and biological signifi cance." To appear in *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, 18-23 March, 2005.

[6] Cormem, T.H., Leiserson, C.E., Rivest, R.L. "Introduction to algorithms." *MIT Press*, Cambridge, MA, 1990.

[7] Wasserman, S. and Faust, K. "Social network analysis: methods and applications." *Cambridge Press*, Cambride, UK, 1994.

[8] Saccharomyces Genome Database (SGD). http://www.yeastgenome.org/