NETWORK CONSTRAINED CLUSTERING FOR GENE MICROARRAY DATA

Dongxiao Zhu [a,b], *Alfred O Hero* [b]

[a]Bioinformatics Program,[b]Departments of EECS, Biomedical Engineering and Statistics,
University of Michigan, Ann Arbor, MI 48105

1. ABSTRACT

Many bioinformatics problems can be tackled from a fresh angle offered by the network perspective. Directly inspired by metabolic network structural studies, we propose an improved gene clustering approach for inferring gene signaling pathways. Based on the construction of co-expression networks that consists of both significantly linear and nonlinear gene associations together with controlled biological and statistical significance, we can make accurate discovery of many transitively co-expressed genes and similarly co-expressed genes. Our approach tends to group functionally related genes into a tight cluster. We illustrate our approach and compare it to the traditional clustering approaches on a retinal gene expression dataset. The clustering method has been implemented in an R package "GeneNT" that is freely available from: http://www-personal.umich.edu/˜ zhud/genent.htm/.

2. INTRODUCTION

The problem of biological network analysis has attracted much interest and curiosity from the biomedical research community in recent years. Much of this interest can be attributed to the ability of a biological network to capture relationships among biological entities, such as small molecules, genes and proteins, and on the patterns and implications of these relationships [1]. Many researchers have realized that the network perspective provides additional leverage for making biological prediction and discovery that can improve inference from experimental data such as microarrays.

There are three main types of intracellular networks that correspond to the three biological entities: the metabolic network, the gene regulation network, and the protein-protein interaction network. Metabolic networks were of interest even before the emergence of the high throughput technology because the central pathways that dominate the network structure were specified by biochemical experiments on stoichiometries of many reactions. These initial applications of metabolic network inference used a non-statistical framework (Boolean networks) since no replicates were available from reaction data. Many theoretical approaches have been implemented to analyze metabolic networks including the network decomposition and isomorphism approaches. The inference of large-scale gene regulation and protein-protein interaction networks require high throughput techniques (e.g. gene microarray, yeast two-hybrid system and *in vivo* pull-down assay), and thus are subject to statistical uncertainty. The application of network analysis techniques to these networks have been hindered by at least two facts: 1) There are few error control algorithm available. 2) Researchers in the microarray data analysis field did not start to appreciate these approaches before the successful application to inferring gene pathways [2].

Inferring gene pathway from microarray is a relatively recent area in microarray data analysis. The gene pathway is a sequence of gene interactions leading to a specific biological endpoint function. Traditional gene clustering approaches that group similarly co-expressed genes into pathways have been widely accepted [3]. On the one hand, clustering is powerful, computationally efficient, and often gives rise to biologically meaningful discoveries. On the other hand, clustering assigns all genes into clusters while completely relaxing the constraints of the underlying gene regulation network, i.e., some of the genes may not be relevant to the underlying biological process and hence should not be forced into the cluster [4]. Moreover, clustering is limited by the fact that genes in the same biological pathway do not necessarily have similar expression profiles.

Transitive co-expression analysis can be viewed as one possible way to apply network analysis techniques for inferring functional relationships among genes [2]. This approach allows discovery of functionally related genes in a pathway that do not have similar expression profiles. It nicely complements the traditional gene clustering approach. However, the linear manner of network discovery leads to unreliable inferences of interconnected pathway components. Therefore, approaches which integrate features of traditional gene clustering and transitive co-expression analysis are high-

This work was supported by grants from the National Institute of Health (EY01115), The Foundation Fighting Blindness, Sramek Foundation and Research to Prevent Blindness.

ly desirable.

Directly inspired by the metabolic network decomposition analysis in Ma et al., 2004 [5], and based on an error-control algorithm for extracting gene regulation network from microarray data developed in our group [6], we propose an improved gene clustering approach (denoted as “NC clustering” throughout this paper) and demonstrate its advantages over traditional approaches (denoted as “Traditional clustering” throughout this paper) using mouse retinal development gene expression data. The data represents a total of 45,101 gene expression profiles over 14 conditions (including knock-outs of several key transcription factors and time).

3. METHODS

3.1. Constructing co-expression network

We extract a network from microarray data using a False Discovery Rate Confidence Interval (FDR-CI) based two-stage algorithm with simultaneously controlled FDR and Minimum Acceptable Strength (MAS) [6]. The algorithm provides an initial co-expression discovery that controls only FDR, which is then followed by a second stage co-expression discovery which controls both FDR and MAS. This technique improves upon previous network extraction methods, e.g. [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 co-expression network, and more specifically the GCC described below will specify the gene pairs that should be included in the clusters.

3.2. Search for Giant Connected Component

Only those pairs of genes in the Giant Connected Components (GCC) of the co-expression network are relevant to the signaling pathway and should be included in the clusters. The 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 have designed and implemented a simple but effective algorithm to extract the GCC from the undirected graph: 1) Calculate marginal degree for each vertex in the graph, denoted as K. 2) Sort K in decreasing order, i.e. $K_{(1)}, K_{(2)}, \ldots, K_{(n)}$. 3) 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 high ly 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 [7].

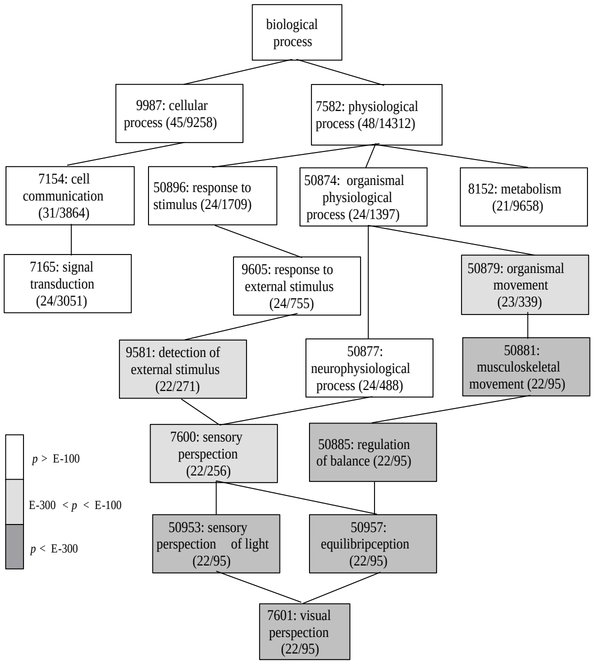
biological process
9987: cellular process (45/9258)
7582: physiological process (48/14312)
7154: cell communication (31/3864)
50896: response to stimulus (24/1709)
50874: organismal physiological process (24/1397)
8152: metabolism (21/9658)
7165: signal transduction (24/3051)
9605: response to external stimulus (24/755)
50879: organismal movement (23/339)
9581: detection of external stimulus (22/271)
50877: neurophysiological process (24/488)
50881: musculoskeletal movement (22/95)
7600: sensory perspection (22/256)
50885: regulation of balance (22/95)
50953: sensory perspection of light (22/95)
50957: equilibripception (22/95)
7601: visual perspection (22/95)
p > E-100
E-300 < p < E-100
p < E-300

Fig. 1. The GO *biological process* graph of the NC2 cluster (see Table 1). In each node, the four-digit or five-digit number is the GO ID for the GO annotation that follows. The decimal in parentheses is the ratio of probesets read-in from the clustering results (having the GO annotation) versus the total number of probesets with that annotation on the chip. The color scheme of a node corresponds to statistical significance level of the node (vocabulary). Note: the GO graph has been pruned by a thresholding count of 20 in order to display only most significant nodes.

3.3. Compute ”shortest-path” distance matrix for GCC

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 gene pairs varying low and high correlation.

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 [7]:

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

The matrix $D = (d_{ij})$ is called the "shortest-path distance matrix". It can be used as input to distance matrix based clustering software such as: hierarchical clustering [3] and K-medoids [8].

4. RESULTS

We obtained a mouse retinal gene expression dataset from our collaborators at Kellogg Eye Center at University of Michigan. The aim of retinal gene expression experiment is to investigate the gene regulation hierarchy of photoreception differentiation during retinal development and to discover unknown genes related to this pathway. (Since the dataset has not been published, we must omit more detailed biological description of the dataset, which will be reported elsewhere).

The dataset was preprocessed using "rma" method [9], and it was subjected to an initial screening. A total of 837 genes whose $\log_2$-transformed expression indices vary more than 3.5 folds were kept for further analysis. We constructed a co-expression network using relatively relaxed constraints (FDR $\leq 20\%$ and MAS = 0.7) to retain a sufficiently large number of gene pairs in the network [5]. A GCC of size 764 genes were extracted (see Methods). These 764 genes were used in NC clustering according to equation(1) and equation(2) and the total 837 genes were used in traditional clustering according to equation(1) only. The exponent was set to 3 according to biological prior knowledge that all known genes in the pathway form a tight cluster.

We used Gene Ontology (GO) [10] annotation as the objective criteria to compare the two clustering approaches. GO is a set of standard hierarchical vocabularies to describe the *biological process*, *molecular function* and *cellular component* of genes. It is conveniently represented as a graph where nodes represents standard vocabularies and edges represent the relationship (either "is-a" or "part of") between vocabularies. A child node is the more specific vocabulary than its parent node(s). A list of probesets from either NC clustering or traditional clustering can be mapped to a GO graph (e.g. *biological process* graph), the appearance counts of all nodes of the GO graph can be calculated as well as their p-values of chi-square statistics. The most significant node(s)(corresponding to the smallest p-value(s)) usually describe(s) the biological functions of the probeset list. For example, in the GO *biological process* graph (Fig. 1) of the NC2 cluster (a probeset cluster identified by NC clustering approach, see Table 1), the leaf node "visual perspection" (GO ID: 0007601) may be most suitable to describe the function of NC2 cluster. Assuming similar cluster size, the smallest p-value of the same leaf node over different GO graphs corresponding to different clusters of probesets also identifies the best (tightest) cluster (see Table 1).

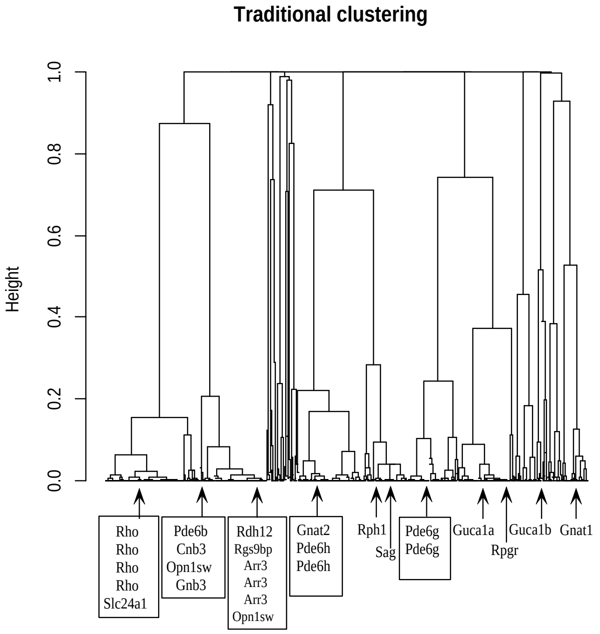

Fig. 2. Traditional clustering: Dendrogram obtained by agglomerative hierarchical clustering using all differentially expressed genes. Some genes appear more than once because Affymetrix chips replicate probesets for some genes to avoid problems caused by gene homologues (same in Fig. 3).

Fig. 2 and Fig. 3 demonstrate the relative advantages of NC clustering compared to traditional clustering. All 26 probesets having the leaf node annotation "visual perspection" are mapped to the leaves of the dendrograms. In Fig. 2, these 26 probesets either form scattered small clusters or become standalone genes. In contrast, 23 out of the 26 probesets are incorporated into a relative small and tight cluster A of 101 probesets as shown in Fig. 3. Table 1 presents a more detailed comparison of the two clustering methods based on statistics of the leaf node vocabulary "visual perspection" in several clusters. Rows represent NC or traditional clusters of different sizes, and columns represents statistics of the leaf node vocabulary. Overall, "visual perspection" is much better represented in NC clusters than in traditional clusters as indicated by the p-values (column

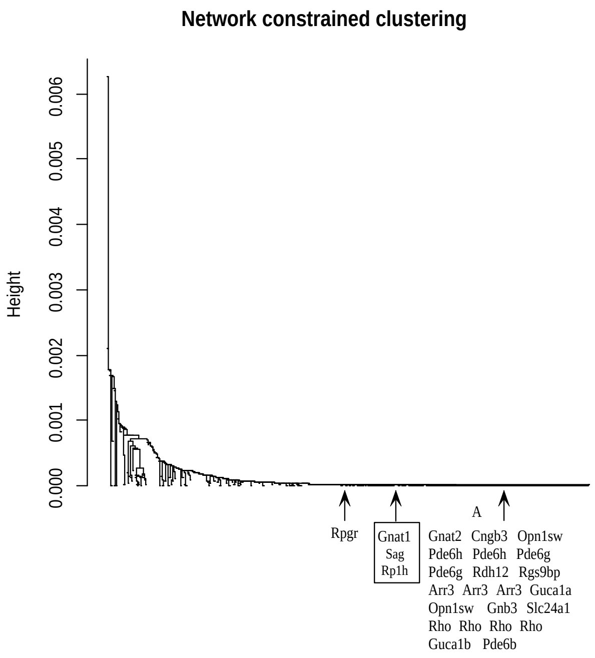

Fig. 3. Network constrained clustering: dendrogram obtained by agglomerative hierarchical clustering from relevance network.

6). These results show that our NC clustering method tends to form a tighter cluster of interest. Our collaborators are now investigating the pathway suggested by the NC2 cluster, which includes 44 unannotated probesets.

5. CONCLUSION

The application of network analysis techniques to networks extracted from high throughput data is limited due to inadequate replicates of probesets leading to high uncertainty (low p-values). In this paper, using simultaneously controlled biological and statistical significance [6], we have applied network clustering techniques, and demonstrated significant advantages over the traditional clustering approa ches that do not account for network constraints. The advantages of our method are: 1) It tends to group functional related genes into tight clusters despite the lack of similarity between expression profiles. 2) It includes constraints on statistical and biological significance and generates p-values on clustered genes. 3) It is sufficiently flexible because the calculated network constrained distance matrix can be fitted to popular distance-based clustering software.

	# Probsets	# Annotated probsets	Counts	Chi-square	p-value
NC1	204	100	25	1.0E3	1.3E-226
NC2	101	57	22	1.4E3	1.9E-311
NC	764	411	26	2.4E2	5.4E-54
TD1	204	114	13	2.3E2	3.0E-56
TD2	57	43	10	3.8E2	5.6E-85
TD3	15	14	7	5.9E2	3.3E-130
TD	837	436	26	2.2E2	1.8E-50

Table 1. Comparison of clustering results. Rows NC1, NC2 and NC presents the NC clustering results evaluated by appearance counts of the leaf node "visual perspection" in the GO *biological process* graph. Rows TD1, TD2, TD3 and TD presents the traditional clustering results evaluated by appearance counts of the same leaf node. These results are calculated using data mining tools from Affymetrix Netaffy Center [11].

6. REFERENCES

[1] Wasserman, S. and Faust, K. (1994) Social network analysis: methods and applications. *Cambridge Press*.

[2] Zhou,X.J., Kao,M. et al. (2002) Transitive functional annotation by shortest path analysis of gene expression data. *Proc Natl Acad Sci USA*, **99**, 12783-12788.

[3] Eisen,M., Spellman,P. et al. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, **95**, 14863-8.

[4] Tseng, G.C., Wong, W.H. (2005) Tight clustering: A Resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**, 10-16.

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

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

[7] Cormem, T.H., Leiserson, C.E., Rivest, R.L. (1990) Introduction to Algorithm. Cambridge: *MIT Press*.

[8] Kaufman, L., Rousseeuw, P. (1990) Finding groups in data: An introduction to cluster analysis. New York: John Wiley & Sons.

[9] Bolstad, B.M., Irizarry, R.A. et al. (2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics*, **19**, 185-193.

[10] Ashburner, M., Ball, C.A. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, **25**, 25-29.

[11] Cheng, J., Sun, S. et al. (2004) NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics*, **20**, 1462-1463.