BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 21 2005, pages 4014–4020
doi:10.1093/bioinformatics/bti655

Gene expression

Network constrained clustering for gene microarray data

Dongxiao Zhu[1,3,*], Alfred O Hero[2,3,4], Hong Cheng[5,6], Ritu Khanna[5,6] and Anand Swaroop[5,6]

[1]Bioinformatics Program, [2]Department of Electrical Engineering and Computer Science (EECS), [3]Department of Statistics, [4]Department of Biomedical Engineering, [5]Department of Ophthalmology and [6]Department of Visual Science, University of Michigan, Ann Arbor, MI 48109, USA

Received on June 10, 2005; revised on July 27, 2005; accepted on August 30, 2005
Advance Access publication September 1, 2005

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 from gene microarray data. Based on the construction of co-expression networks that consists of both significantly linear and non-linear gene associations together with controlled biological and statistical significance, our approach tends to group functionally related genes into tight clusters despite their expression dissimilarities. We illustrate our approach and compare it to the traditional clustering approaches on a yeast galactose metabolism dataset and a retinal gene expression dataset. Our approach greatly outperforms the traditional approach in rediscovering the relatively well known galactose metabolism pathway in yeast and in clustering genes of the photoreceptor differentiation pathway.

Availability: The clustering method has been implemented in an R package 'GeneNT' that is freely available from: http://www.cran.org.

Contact: zhud@umich.edu

1 INTRODUCTION

An important area in microarray data analysis is to infer signaling pathways. The signaling pathway is a sequence of gene interactions leading to a specific biological endpoint function. Gene interactions are typically inferred through calculating the correlation between gene expression profiles over multiple relevant physiological/genetical conditions. Gene pairs with high correlation (e.g. >0.6) are hypothesized to be biologically relevant and to interact directly in the signaling pathways (Zhou *et al*., 2002; Stuart *et al*., 2003; Lee *et al*., 2004).

It is typical that only a few genes are experimentally confirmed to be in a signaling pathway. Gene clustering is a widely used approach that attempts to group all the genes in the pathway into a cluster such that functional prediction of unknown genes can be made based on the functionally known genes. Some of the more popular clustering methods include: hierarchical clustering (Eisen *et al*., 1998), *K*-means type clustering (Hartigan and Wong, 1979) and model-based clustering (Yeung *et al*., 2001). These methods have been successful in inferring many signaling pathways from gene microarray data.

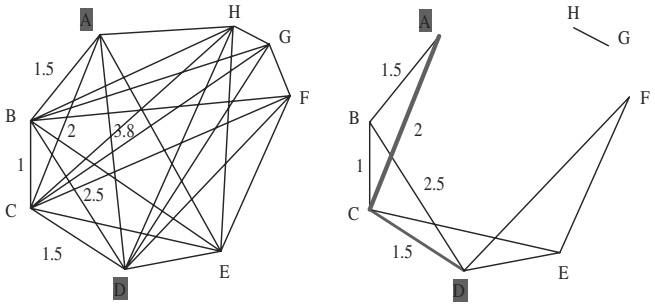
The ultimate goal of all gene clustering approaches is to group genes with similar functions into one single cluster. In practice, most approaches simply group genes with similar expression profiles (Eisen *et al*., 1998; Stuart *et al*., 2003; Lee *et al*., 2004), denoted as 'traditional clustering' throughout this paper. However, many genes in the same functional pathway may not have similar expression profiles as measured by correlation statistics or other pairwise expression similarity measure. This is especially true for pairs of genes that are not in the same region of a signaling pathway. These genes will not be discoverable using the traditional clustering methods. Thus, a well-known limitation of the traditional clustering approaches is that it only groups functional related genes with similar expression profiles, but misses out on many others with dissimilar expression profiles.

In a gene co-expression (also called 'relevance') network, graph vertices represent genes, and edges represent gene associations (Butte and Kohane, 2000; Butte *et al*., 2000). The traditional methods of clustering assume that the underlying network is fully connected, i.e. any biological function is executed through a direct interaction of a pair of genes (Fig. 1a). Direct pairwise gene interactions, represented by the fully connected subgraph (clique), only describes a small subset of gene interactions. In many cases, an endpoint biological function is more commonly executed through a series of inter-connected gene interactions (Fig. 1c, gene A, B, C, D, E, F). Consequently, for genes lying in a single pathway traditional clustering approaches often group these genes into several different clusters, e.g. each cluster determined by a similarly co-expressed clique. This breaking of a pathway across several clusters makes it more difficult for biologists to identify groups of genes having common functions. Thus approaches that are able to go beyond pairwise interactions to group the whole pathway into a single tight cluster are highly desirable.

A more realistic assumption for gene clustering may be that the underlying relevance network is only partially connected, i.e. the biological function is executed through either direct interaction or one or more intermediate genes (Fig. 1c). A gene clustering algorithm that accounts for such realistic network constraints is likely to be more powerful (Zhou *et al*., 2002; Zhou and Gibson, 2004). There are several challenges to developing such an approach: how to reliably extract the relevance network from microarray data and how to estimate the distance between two non-adjacent genes (genes that do not have similar expression profiles) in the network.

Gene co-expression networks typically use correlation statistics as pairwise similarity measures (a decreasing function of the distance for clustering) between gene expression profiles, followed by either direct correlation thresholding (Zhou *et al*., 2002) or a

*To whom correspondence should be addressed.



	A	B	C	D
A	0			
B	1.5	0		
C	2	1	0	
D	3.8	2.5	1.5	0

(b)

	A	B	C	D
A	0			
B	1.5	0		
C	2	1	0	
D	**3.5**	2.5	1.5	0

(d)

Fig. 1. Underlying network models and distance matrices for traditional clustering (**a**)(**b**) and network constrained clustering (**c**)(**d**). Obtained by removing some edges of weak correlations (long distances), e.g. distance longer than 3. The distance between two genes is a decreasing function of their correlation (see Equation 4). (a). Fully connected network; it assumes any two genes interact with each other directly in the network (connected). (b). Part of the distance matrix for the network model (a). (c). Partially connected network; it assumes only two genes with high correlation (e.g. 0.6) directly interacting with each other (connected). Grey edges represent the shortest-path from A to D. (d). Part of the distance matrix for the network model (c).

combination of significance level tests with correlation thresholding (Lee *et al*., 2004). While direct thresholding is useful in many cases it only controls biological significance but not error rate. Combining correlation thresholding with a level of significance test allows one to control biological and statistical significance, albeit in an *ad hoc* manner. A new network construction approach based on False Discovery Rate Confidence Interval (FDR-CI) was recently proposed to control biological and statistical significance simultaneously (Zhu *et al*., 2005). This approach was able to identify both linearly and nonlinearly co-expressed genes using the Kendall correlation coefficient combined with the Pearson correlation coefficient. The employment of nonlinear correlation measures is important when functionally related gene expression profiles are non-linearly correlated. Non-linear correlation can occur, for example, when gene expressions of different subunits of a whole enzyme are differentially regulated due to different enzyme efficiencies.

Regarding the estimation of distance between two non-adjacent genes in the relevance network, the shortest-path distance between them represents the most natural and parsimonious representation of biological interaction since genes along the shortest-path are likely to have similar functions (Zhou *et al*., 2002). Based on the new network construction algorithm and the shortest-path distance measure, we present a new clustering approach, called 'network constrained (NC) clustering' throughout this paper, that is able to group more functionally related genes into a single tight cluster even if their expression profiles are dissimilar.

2 METHODS

2.1 Constructing co-expression network

We formulate the network construction problem as a composite hypothesis test with multiple comparison. The algorithm proceeds as follows (Zhu *et al*., 2005).

2.1.1 Measuring the strength of association We use Γ to denote the true strength of association between a pair of gene expression profiles. Under a Gaussian linear hypothesis, the sample Pearson correlation coefficient $\hat{\boldsymbol{\rho}}$ is an appropriate metric (Bickel and Doksum, 2000). A robust distribution-free alternative is the sample Kendall rank correlation coefficient $\hat{\boldsymbol{\tau}}$ (Hollander and Wolfe, 1999). We define $\{g_p\}_{p=1}^{G}$ as the indices of G gene probes on the microarray; $\{X_{g_p}\}_{p=1}^{G}$ as normalized probe responses (random variables) and $\{\{x_{g_{p(n)}}\}_{p=1}^{G}\}_{n=1}^{N}$ as realizations of $\{X_{g_p}\}_{p=1}^{G}$ under N i.i.d. microarray experiments.

2.1.2 Hypothesis testing scheme For G genes on each microarray, we need to simultaneously test $\Lambda = \binom{G}{2}$ pairs of two-sided hypotheses:

$$H_0 : \Gamma_{g_i, g_j} \leq cormin \text{ versus } H_\alpha : \Gamma_{g_i, g_j} > cormin, \quad \text{for } g_i \neq g_j, \text{ and } g_i, g_j \in (1, 2, \ldots G) \tag{1}$$

where *cormin* is a Minimum Acceptable Strength (MAS) of correlation. The sample correlation coefficient $\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\rho}}$ or $\hat{\boldsymbol{\tau}})$ is used as a decision statistic to decide on pairwise dependency of two genes in the sample. For N realizations of any pair of gene probe responses, $\{x_{gi(n)}, x_{gj(n)}\}_{n=1}^{N}$, we first calculate $\hat{\boldsymbol{\tau}}$ or $\hat{\boldsymbol{\rho}}$. For large N, the Per Comparison Error Rate (PCER) P-values for ρ or τ are:

$$p_{\rho_{i,j}} = 2\left(1 - \Phi\left(\frac{\tanh^{-1}\left(\hat{\boldsymbol{\rho}}_{i,j}\right)}{(N-3)^{-1/2}}\right)\right) \tag{2}$$

$$p_{\tau_{i,j}} = 2\left(1 - \Phi\left(\frac{K}{N(N-1)(2N+5)/18^{1/2}}\right)\right) \tag{3}$$

where Φ is the cumulative density function of a standard Gaussian random variable and $K = \sum \sum_{1 \leq n \leq m \leq N} K_{nm}$. The above expressions are based on asymptotic Gaussian approximations to $\hat{\boldsymbol{\rho}}_{i,j}$ (Bickel and Doksum, 2000) and to $\hat{\boldsymbol{\tau}}_{i,j}$ (Hollander and Wolfe, 1999).

The PCER P-value refers to the probability of Type I error rate incurred in testing a single pair of hypothesis for a single pair of genes g_i, g_j. When considering the Λ multiple hypotheses for all possible pairs, as in previous studies, we adopt the False Discovery Rate (FDR) to control statistical significance of the selected gene pair correlations in our screening procedure (Reiner *et al*., 2003). The procedure guanantees that the false discovery rate associated with testing the hypotheses (1) does not exceed α.

2.1.3 Two-stage screening procedure Select a level α of FDR and a level *cormin* of MAS significance levels. We use a modified version of the two-stage screening procedure applied to gene screening (Hero *et al*., 2004). This procedure consists of:

Stage I. Test the simple null hypothesis.

$$H_0 : \Gamma_{g_i, g_j} = 0 \text{ versus } H_\alpha : \Gamma_{g_i, g_j} \neq 0$$

at FDR level α. The step-down procedure of Benjamini and Hochberg (1995) is used.

Stage II. Suppose Λ_1 pairs of genes pass the stage I procedure. In stage II, we first construct asymptotic PCER Confidence Intervals (PCER-CI's): $I^{\Lambda}(\alpha)$ for each $\Gamma(\rho$ or $\tau)$ in subset $\mathcal{G}_1$, and convert into FDR Confidence Intervals (FDR-CI's) : $I^{g}(\Lambda_1\alpha/\Lambda)$ (Benjamini and Yekutieli, 2004). A gene pair in subset $\mathcal{G}_1$ is declared to be both statistically significant and biologically significant if its FDR-CI does not intersect the MAS interval $[-cormin, cormin]$.

A relevance network composed of these screened gene pairs can then be constructed with the simultaneous control of statistical significance $\alpha \times 100\%$ and biological significance *cormin*.

2.2 Extract the giant connected component

Only gene pairs that are in the same Connected Component (CC) of the relevance network have finite distances and can be clustered. The largest connected component, usually of importance to both biological function and network topology (Ma *et al*., 2004b; Zhu and Qin, 2005), is called the Giant Connected Component (GCC) (Fig. 1c, genes A, B, C, D, E, F form a GCC). 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 the maximal set of vertices $U \subset V$ such that every pair of vertices u and v in U are reachable from each other. Our network constrained clustering method is applied to the GCC. Analogous to previous studies, we assume that almost all important genes are included in the GCC. The standard depth first search (DFS) algorithm [Cormem *et al*. (1990)] was used to extract the GCC from the gene microarray data.

2.3 Compute 'network constrained distance matrix'

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*. (2002), the w_{ij} is defined as:

$$w_{ij} = \left(1 - \text{abs}\left(\hat{\Gamma}_{ij}\right)\right)^p \tag{4}$$

The integer p is an exponential tuning parameter used to enhance the differences between low and high correlation. We define the matrix $W = [w_{ij}]$ as the 'Traditional distance matrix' (e.g. Fig. 1b).

We use the standard Floyd-Warshall algorithm to search among all-pairs for the shortest-paths within the GCC. Let $d_{ij}^{(k)}$ be the weight of a shortest-path from vertex i to vertex j such that all intermediate vertices on the path (if any) are in set $\{1,2,\ldots,k\}$. When $k = 0$, there is no intermediate vertex between vertices i and j, and we define $d_{ij}^{(0)} = w_{ij}$. A recursive definition of $d_{ij}^{(k)}$ is given by (Cormem *et al*., 1990):

$$d_{ij}^{(k)} = \begin{cases} w_{ij} & \text{if } k = 0, \\ \min\left(d_{ij}^{(k-1)}, d_{ik}^{(k-1)} + d_{kj}^{(k-1)}\right) & \text{if } k \geq 1, \end{cases} \tag{5}$$

where $d_{ij}^{(k-1)}$ is the length of the shortest-path when k is not a vertex on the path and $d_{ik}^{(k-1)} + d_{kj}^{(k-1)}$ is that k is a vertex on the path. We define the matrix $D = [d_{ij}]$ as the 'Network constrained distance matrix' (e.g. Fig. 1d). It can be used as input to many distance matrix based clustering software packages such as: hierarchical clustering (Eisen *et al*., 1998) and *K*-medoids (Hartigan and Wong, 1979). The calculation of matrix D can be easily extended to higher Eukaryote since the algorithm runs in polynomial time, i.e. $O(V^3 + V + E)$.

3 RESULTS

3.1 Sensitivity analysis

The FDR, MAS and exponential tuning parameter p are three parameters involved in calculating the network constrained distance matrix. It would be interesting to investigate the sensitivity of the results to variance in these parameters. The biological significance level MAS = 0.6 has been widely adopted as a correlation cut-off in the literature, e.g. Zhou *et al*., 2002, 2005. The selection of the FDR statistical significance level is intimately associated with the sample size and the underlying biological mechanism. Our selection of FDR = 5% imposes the stringent statistical criterion that on average only 5% of the genes discovered and included in the network will be false positives.

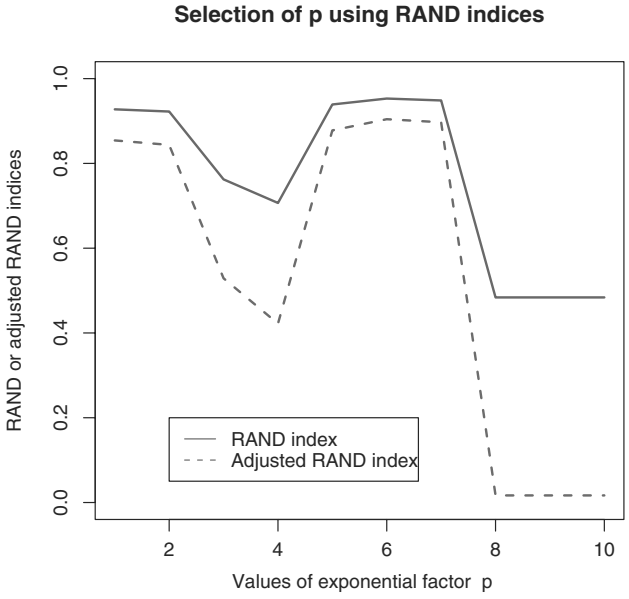


Fig. 2. Selection of p using RAND indices.

The parameter p in Equation (4) is an exponential tuning factor used to enhance the differences between expression similarity and dissimilarity. As pointed out by Zhou *et al*. (2002), for a fixed correlation threshold, as p is increased more transitive genes will be revealed at the expense of a higher false discovery rate. In Figure 2 we present results of an empirical study of the influence of p on clustering performance for a yeast galactose metabolism dataset (Ideker *et al*., 2000).

The galactose metabolism dataset represents approximately 6200 yeast gene expression levels on two-color cDNA microarrays over 20 physiological/genetic conditions (nine mutants and one wild type strains incubated in either GAL-inducing or non-inducing media) (http://www.sciencemag.org/cgi/content/full/292/5518/929/DC1). A subset of 205 gene expression profiles whose Gene Ontology (GO) annotation (Ashburner *et al*., 2000) falls into four functional classes were used (Yeung *et al*., 2003). We investigate the effect of p by examining how closely the clusters reproduce these functional classes as p varies. We used both the RAND index (Rand, 1971) and the adjusted RAND index (Hubert, 1985) as measures of consistency between the clustering results and GO annotations. Figure 2 shows that the network constrained clustering best conforms to the GO annotations when $p = 6$. Note that Zhou *et al*. (2002) also suggested using $p = 6$ to define the edge weight in their analysis.

3.2 Yeast galactose metabolism data

3.2.1 Data processing and network construction We empirically evaluated the performance of the proposed clustering approach by applying it to a relatively well-known yeast galactose metabolism signaling pathway and comparing it with the traditional clustering approaches. We used a subset of 997 genes that were identified by Ideker *et al*. (2000) using a standard generalized likelihood ratio testing procedure. Genes having a likelihood statistic $\lambda \leq 45$ were selected as differentially expressed and whose mRNA levels differed significantly from reference under one or more perturbations.

By measuring the pairwise gene correlations using both Pearson and Kendall correlation coefficients, we applied a two-stage algorithm to screen gene pairs with FDR $\leq$5% and MAS = 0.6 (Zhu *et al*., 2005). The resulting network is a mixed network within which edges are discovered with Pearson and Kendall correlation statistics. Our network construction algorithm and the screening criteria ensure false discovery of no more than 5% of the edges having strength of association >0.6 (Zhu *et al*., 2005).

3.2.2 Network constrained clustering We extracted the GCC from the co-expression network using a DFS type algorithm (see Methods). The GCC contains 772 genes within which almost all known structural genes in the pathway are included. This confirms the notion that GCC of the network has not only structural but also functional significance (Ma *et al*., 2004; Ma and Zeng, 2003; Zhu and Qin, 2005). The network constrained distance matrix for GCC was computed according to Equations (4) and (5) using GCC selected genes (see Methods) while the distance matrix for the traditional clustering method was computed according to Equation (4) only.

The yeast galactose metabolism pathway consists of at least three types of genes including transporter genes (GAL2, HXT1-10, the roles of other HXT genes are not entirely clear), enzyme genes (GAL1, GAL7, GAL10 etc.) and transcription factor genes (GAL3, GAL4, GAL80 etc.) (Wieczorke *et al*., 1999). Transcription factor genes are not discoverable from this microarray experiment as their expressions are typically time shifted and only one time sample was included. Since the pathway has been relatively well studied, we sought to compare our network constrained clustering approach with the traditional clustering approach through rediscovering the 14 important genes in the structural module (GAL2, HXT1-10, and enzyme genes: GAL1, GAL7, GAL10) of the yeast galactose metabolism pathway.

For comparison with a widespread clustering algorithm we used agglomerative hierarchical clustering [implemented in R function hclust()]. We expect that other traditional clustering methods such as *K*-means or *K*-medoids would give similar results. For calculating distance between clusters, we implemented a 'complete' method in which the longest geodesics between genes in the two clusters are used as distance between clusters. As empirically demonstrated in (Speed, 2003), the 'complete' method gives rise to better cluster separation.

Figure 3 shows the traditional clustering approach using all 997 genes and Figure 4 shows the traditional clustering approach using the 772 genes in the GCC. In both cases, the 14 structural genes are separated into three subclusters (Figs 3 and 4). In Figure 3, all GAL genes are nicely grouped in a cluster, but not the HXT genes, In Figure 4, all HXT genes are grouped into a single cluster, but the algorithm fails to combine GAL gene clusters with HXT gene clusters. Figures 3 and 4 show that the GCC gene selection process has some desirable effects on clustering by removing a few unrelated genes (Tseng and Wong, 2005) that are not relevant to the biological pathway. However, using the GCC gene selection procedure alone does not significantly improve clustering performance.

We think that this undesirable separation of genes in the pathway is due to the presence of gene expression dissimilarity between subclusters and gene expression similarity within subclusters. To test this hypothesis, we plotted the correlation matrix of 14 genes in the structural module and did hierarchical clustering (Fig. 5). The color intensities in Figure 5 correspond to the levels of correlations

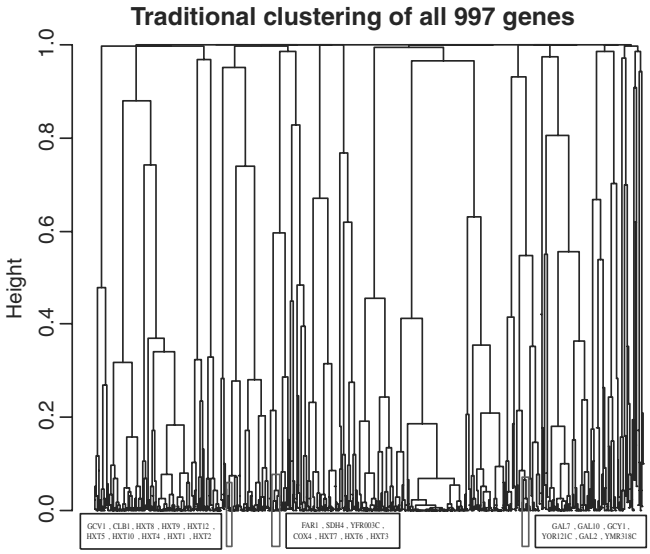


Fig. 3. Traditional clustering: agglomerative hierarchical clustering using all 997 differentially expressed genes. The 14 structural genes are separated into three clusters (grey rectangular).

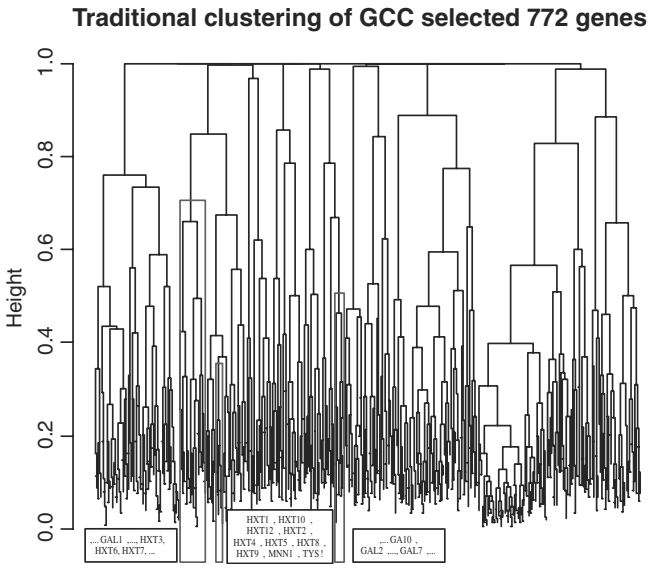


Fig. 4. Traditional clustering: agglomerative hierarchical clustering using the GCC selected 772 genes. The 14 structural genes are separated into three clusters (grey rectangular). Dots indicate incomplete clusters are shown due to space limitation.

(increasing correlations are represented from yellow to red). It is evident from Figure 5 that expression correlations within GAL genes and HXT genes are much higher than the correlations between the two groups. This explains the separation of these two gene clusters in the associated clustering dendrogram (Figs 3 and 4). Among the HXT gene clusters, HXT3, HXT6 and HXT7 are highly correlated [red (dark) zone in Fig. 5]. It explains the actual separation of these three genes from the remaining HXT genes shown in the clustering dendrogram (Fig. 3). Figs 3, 4 and 5 show that

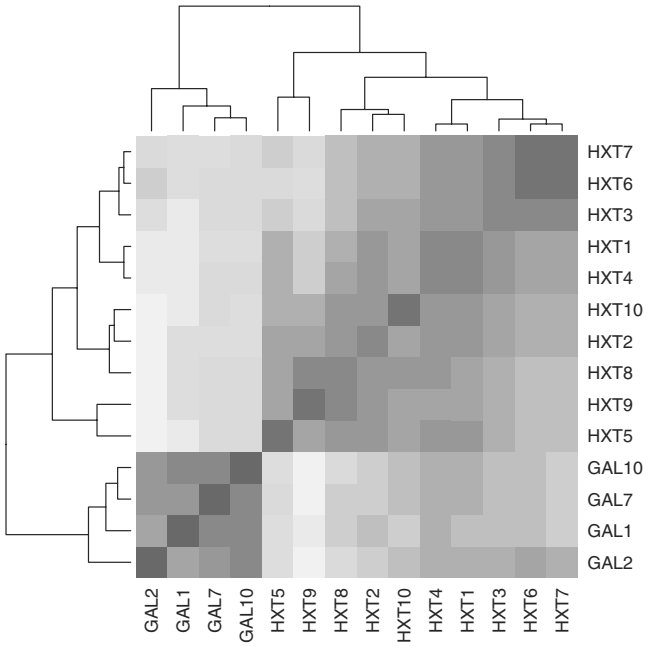


Fig. 5. Correlation matrix of 14 structural genes with clustering dendrogram. White to grey corresponds to the low correlations to high correlations.

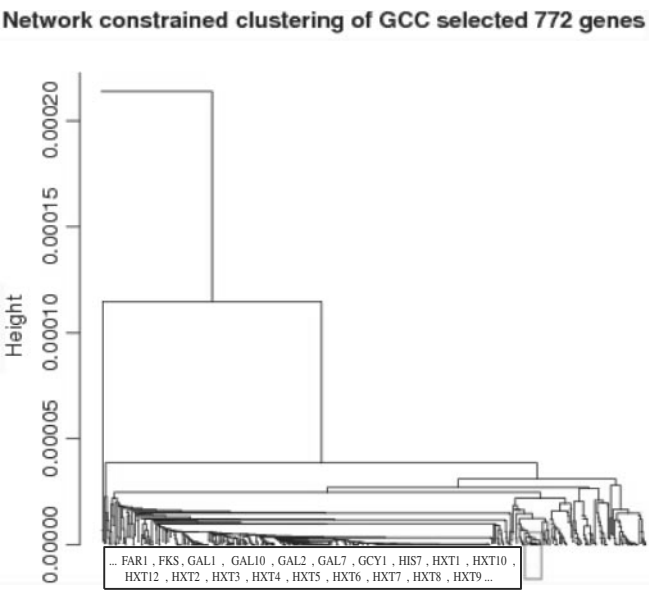


Fig. 6. Network constrained clustering: agglomerative hierarchical clustering using network constrained distance matrix calculated from relevance network [Equation (5)].

traditional clustering methods failed to group functionally related genes with dissimilar expression profiles (low correlations) into one cluster.

Figure 6 presents results of applying our network constrained clustering algorithm to the 772 genes selected by GCC extraction. Note that all 14 structural genes that failed to be clustered together by the traditional approach (Fig. 3) are grouped into a single tight cluster by the network constrained clustering approach. As has been shown, the GCC selection process contributes only moderately to the apparent success. This demonstrates that employment of the network constrained distance matrix can lead to significant improvement in clustering performance.

3.3 Retinal gene expression data

The aim of the retinal gene expression experiment is to investigate the gene pathway of photoreception differentiation during retinal development and to discover unknown genes related to this pathway. The retinal data represent a total of 45 101 gene expression profiles over five time points measured in both wide type and Nrl (Swaroop *et al*., 1992) (the Maf-family transcription factor, the key regulator of photoreceptor differentiation in mammals) knockout mice: submitted. The data will soon be available for download through the NCBI Gene Expression Omnibus (GEO).

The data were preprocessed using the 'rma' method (Bolstad *et al*., 2003), and it was subjected to an initial screening using the two-stage screening method proposed by Hero *et al*. (2004) in which the top 1000 genes ranked by FDR and Fold Change are kept for further analysis. We constructed a co-expression network similar to the yeast analysis (FDR $\leq$5% and MAS = 0.6) in the last subsection. A GCC of size 790 genes was extracted. These 790 genes were used in our NC clustering algorithm according to Equations (4) and (5) while the total 1000 genes were used in the traditional hierarchical clustering algorithm according to Equation (4) only.

As above we used GO 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 represent 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 obtained from any clustering method can be mapped to a GO graph (e.g. biological process graph), the appearance counts of all nodes of the GO graph as well as their *P*-values of chi-square statistics can be calculated. The most significant node(s)[corresponding to the smallest *P*-value(s)] usually describe(s) the biological functions of the probeset list. Specifically, all genes having GO annotation 'visual perception [GO:0007601]' are expected to belong to the photoreceptor differentiation pathway.

We thoroughly compared the two clustering results with respect to three criteria (appearance counts, separation and *P*-values of the GO category: visual perception) at each cluster number ranging from 1 to 20. Only the largest 20 clusters were investigated as the remaining clusters contained fewer than 5 genes. The first two clustering criteria measure stability of the 'visual perception' cluster as a function of cluster numbers, and the third criterion measures the enrichment of the interested GO vocabulary as a function of cluster numbers. Figures 7 and 8 demonstrate that the 'visual perception' cluster acquired by NC clustering is quite stable over different cluster numbers but not that acquired by traditional clustering. Figure 9 demonstrates that the interested GO vocabulary 'visual perception' is much more enriched by NC clustering over different cluster numbers. In Figure 7, the initial (cluster number = 1)

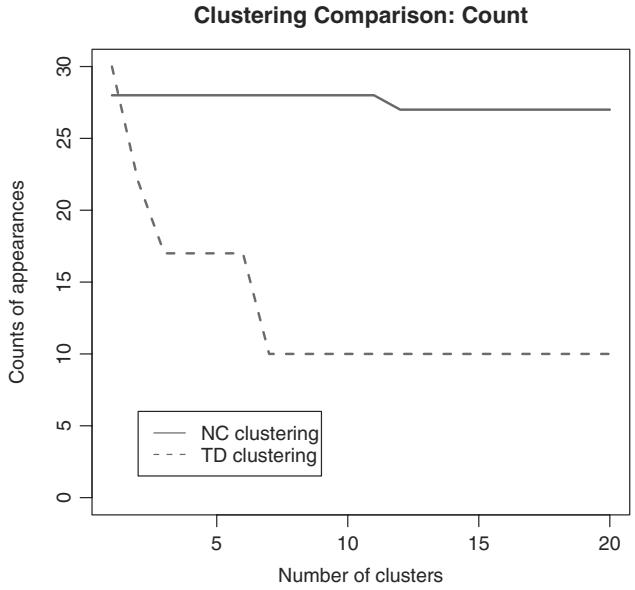


Fig. 7. Clustering comparison - GO vocabulary 'visual perception' counts.

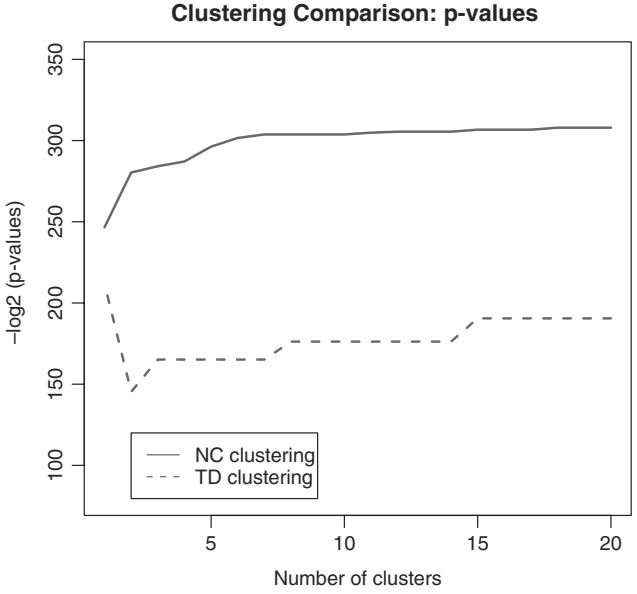


Fig. 9. Clustering comparison - GO vocabulary 'visual perception' *P*-values.

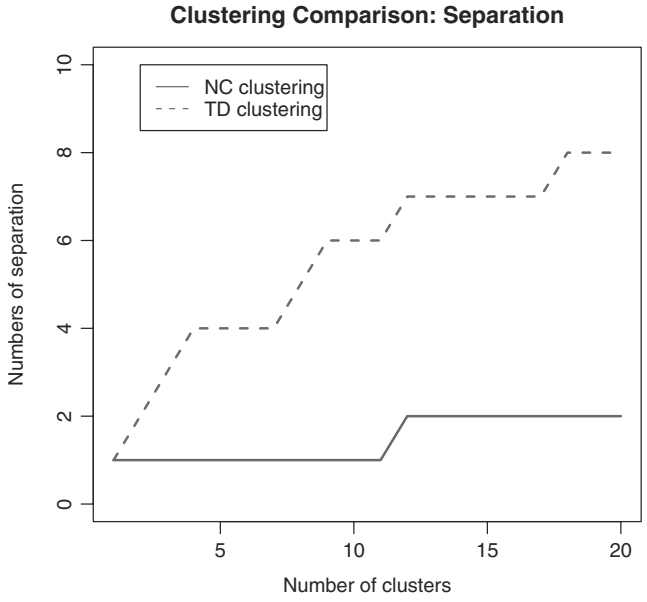


Fig. 8. Clustering comparison - GO vocabulary 'visual perception' separation.

count difference (28 versus 30) is due to the GCC gene selection criterion.

4 DISCUSSION

Estimating signaling pathways from gene expression data is one of the most active research areas in microarray data analysis. Co-expression analysis is one of the most popular approaches. While at this stage many functional predictions made through co-expression analysis are based on the assumption of 'Guilt-by-Association,' there are still few methods for functional predictions from dissimilar expression profiles. Transitive co-expression analysis (Zhou *et al*., 2002) is a systematic method to accomplish functional prediction from dissimilar gene expression profiles (Zhou and Gibson, 2004; Zhou *et al*., 2005).

Systematic network analysis approaches have been widely applied to many biological networks such as metabolic networks e.g. Gagneur *et al*. (2003). Many theoretical approaches have been implemented to analyze metabolic networks including network decomposition and isomorphism methods. For example, Ma *et al*. (2004) presented a network decomposition approach to analyze metabolic pathways, by considering the global network structure rather than local marginal connectivity. They showed that chemical reactions in the same cluster are indeed functionally related. Our approach extends this to gene co-expression networks extracted from microarray data. Our network constrained clustering differs significantly from the traditional clustering approach in at least two aspects: (1) it uses GCC selected genes instead of all differentially expressed genes for clustering; (2) it uses a hybrid distance matrix that is composed of both direct distances and shortest-path distances for clustering instead of the traditional distance matrix that is composed of only direct distance matrix. The latter has been shown to lead to clustering improvements.

Gene co-expression networks differ from metabolic networks and Protein–protein interaction networks in that the edges are inferred from hypothetical rather than physical interactions. Statistical methods are more useful in dealing with inherent uncertainties. The method we adopted constructs the co-expression network by simultaneously controlling biological and statistical significance. Our network constrained clustering method has the following features: (1) it tends to group functionally related genes into a tight cluster disregarding whether these genes have similar expression profiles; (2) it is sufficiently flexible because the calculated network constrained distance matrix can be fitted into many popular distance-based clustering software packages and (3) the algorithm runs in polynomial time.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their constructive comments. This work was partially supported by grants from the National Institute of Health (EY01115), The Foundation Fighting Blindness, Sramek Foundation and Research to Prevent Blindness.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.

Benjamini,Y. and Yekutieli,D. (2004) False discovery rate adjusted multiple confidence intervals for selected parameters. *J. Am. Stat. Assoc*., **100**, 71–80.

Bickel,P.J. and Doksum,K.A. (2000) *Mathematical Statistics: Basic Ideas and Selected Topics*. 2nd edn. Prentice Hall, Upper Saddle River, NJ.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.

Butte,A. and Kohane,I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput*., **5**, 415–426.

Butte,A. *et al*. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. USA*, **97**, 12182–12186.

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

Eisen,M. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Gagneur,J. *et al*. (2003) Hierarchical analysis of dependency in metabolic networks. *Bioinformatics*, **19**, 1027–1034.

Hartigan,J.A. and Wong,M.A. (1979) A *k*-means clustering algorithm. *Applied Statistics*, **28**, 100–108.

Hero,A.O. *et al*. (2004) Multicriteria gene screening for analysis of differential expression with DNA microarrays. *EURASIP JASP*, **1**, 43–52.

Hollander,A. and Wolfe,D. (1999) *Nonparametric statistical methods*. Wiley-Interscience, Hoboken, NJ.

Hubert,A. (1985) Comparing partitions. *J. Classif*., **2**, 193–198.

Ideker,T. *et al*. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol*., **7**, 805–817.

Lee,H. *et al*. (2004) Coexpression analysis of human genes across many microarray datasets. *Genome Res*., **14**, 1085–1094.

Ma,H.W. and Zeng,A.P. (2003) The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, **19**, 1423–1430.

Ma,H.W. *et al*. (2004a) Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC. Bioinformatics*, **5**, 199.

Ma,H.W. *et al*. (2004b) Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics*, **20**, 1870–1876.

Rand,W. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc*., **66**, 846–850.

Reiner,A. *et al*. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 386–375.

Speed,T. (ed.) (2003) *Statistical analysis of gene expression microarray data*. Chapman & Hall/CRC Press, Boca Raton, FL.

Stuart,J.M. *et al*. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Swaroop,A. *et al*. (2002) A conserved retina-specific gene encodes a basic motif/leucine zipper domain. *Proc. Natl Acad. Sci. USA*, **89**, 266–270.

Tseng,G.C. and Wong,W.H. (2005) Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**, 10–16.

Wieczorke,R. *et al*. (1999) Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in *Saccharomyces cerevisiae*. *FEBS Lett*., **464**, 123–128.

Yeung,K.Y. *et al*. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **10**, 977–987.

Yeung,K.Y. *et al*. (2003) Clustering gene-expression data with repeated measurements. *Genome Biol*., **4**, R34.

Zhou,X.J. and Gibson,G. (2004) Cross- species comparison of genome-wide expression patterns. *Genome Biol*., **5**, 232.

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

Zhou,X.J. *et al*. (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol*., **23**, 238–243.

Zhu,D. and Qin,Z.S. (2005) Structural comparison of metabolic networks in selected single cell organisms. *BMC. Bioinformatics*, **6**, 8.

Zhu,D. *et al*. (2005) High throughput screening of co-expressed gene pairs with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS). *J. Comput. Biol*., **12**, 1029–1045.