Journal of VLSI Signal Processing 38, 277–286, 2004

Identification of Shared Components and Sparse Networks in Gene Expression Time-Course Data

DEBASHIS GHOSH
Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Received January 14, 2003; Revised January 14, 2003; Accepted August 28, 2003

Abstract. High-throughput gene expression technologies such as microarrays have been utilized in a variety of scientific applications. In this article, we develop multivariate techniques for visualizing gene regulatory networks using independent components analysis (ICA) techniques. A desirable feature of the ICA method is that it approximates a biological model for the gene expression. The methods are outlined and illustrated with application to yeast gene expression data.

Keywords: high-dimensional data, network visualization, principal components, projection pursuit, singular value decomposition

1. Introduction

With the emergence of the systems biology approach to modelling the regulatory networks of organisms [1, 2], identification of coordinated patterns of gene expression in whole genomes of model organisms gives insight into the fundamental biological processes that underlie basic functions, such as development and growth. In this setting, we are not interested in studying the behavior of individual genes but rather that of groups of genes.

One major development that has allowed scientists to consider systems biology approaches to studying behavior of organisms is DNA microarrays. These high-throughput assays allow for the simultaneous monitoring of mRNA transcript levels for thousands of genes. Given the availability of gene expression datasets from microarray experiments, one natural goal in modelling the data has been to discover genetic networks that delineate regulatory mechanisms. Due to the current state of experimental variability in microarray data, it does not currently appear to be possible to construct genetic networks from the original data themselves unless external biological knowledge, such as information on upstream promoter elements [3] or functional annotation [4], is available. The aim is instead to identify a first-stage global topological map of regulation that can later be analyzed on a more local basis in greater detail.

There has been recent work describing methods for reverse engineering genetic networks using microarray data. Examples include [5–7]. These methods can be broadly grouped into two categories based on whether or not dimension-reduction methods are used. In this article, we focus on methods for studying networks which rely on dimension-reduction strategies.

Most genetic network reconstruction methodologies that have been developed have relied on principal components [7–10]. With this technique, the matrix representing the covariance matrix of gene expression profiles across the samples is decomposed using the singular value decomposition into three matrices, one representing the factor loadings, the second the matrix of singular values and the third the associated eigenvectors. While the method is computationally efficient and optimally reconstructs the data in an L_2 sense [11], there is no *a priori* reason biologically to believe that the Euclidean distance is the relevant metric in analyzing such data.

In this article, we propose a general approach for the analysis of gene expression datasets that arise in

time-course studies. The goal is to study the structure of coordinated gene expression profiles over time and to infer "network modules" from such data. We define this term later in the paper. A key assumption found in bioinformatics studies of network databases regarding the sparsity of networks, will also be utilized. This observation was used in [5] to develop a method for genetic network reconstruction using robust regression techniques. Our incorporation of sparsity into the modelling procedure leads to the development of alternative methods based on independent components analysis techniques [12, 13] and hard thresholding procedures. Independent components analysis was first proposed in [12] for the analysis of gene expression data; we extend this technique to a broader framework that yields some geometric intuition regarding network structure in gene expression data. We will be using data from a recent gene expression study in yeast published in [14] to illustrate the ideas. The format of the paper is as follows. In Section 2, we describe the relevant biological background and a model for the analysis of gene expression data from microarray time course experiments. One of the outputs of the model-based analysis in this section are estimated modes of expression, which we also define later. In Section 3, we then develop a method for studying the interrelationship between these modes that incorporates the sparse network topology found in biological databases. The proposed methods are then applied to the yeast time-course data in Section 4. We conclude with some discussion in Section 5.

2. Background and Model

2.1. Biological Motivation

The basis for the development of the methodologies in the paper is the study of biological regulatory networks. The focus of biology historically has been to understand observed phenomena on the level of individual molecules. However, most biological processes and activities arise from interactions among many different types of molecules. In [15], the concept of network modules was formulated. Network modules are groups of cells and molecules that interact among each other in order to perform various biological functions. The interactions can include protein-protein interactions and protein-DNA interactions, among others. Because actual cellular activity occurs at the level of the module, this is the unit that should be examined when studying biological phenomena [15].

There are several examples of network modules in the literature; *in vitro* experimentation has shown that processes such as DNA replication and glycolysis participate in such units. We make two points about these modules that will be relevant to the modelling approach taken later. First, they function in a relatively autonomous manner. This has been shown in the *in vitro* experiments referred to earlier in the paragraph. Another argument for the modules being independent is evolutionary. If the modules were not independent, then alterations or disruptions in individual cells and/or molecules would lead to global disruption of cellular processes, which is not evolutionary favorable. As argued in [15], the existence of modular structures leads to greater evolutionary plasticity. The second point about modules is that while they function in an autonomous fashion, cells and molecules can be involved in multiple modules.

Much recent work has focused on identification of network modules. We focus on two recent studies. In the first, reported in [16], computational methods based on microarray and sequence data were used to predict the existence of transcriptional regulatory modules that control gene expression in yeast. The second study [17], also performed on yeast, predicted the existence of 50 regulatory modules using both microarray and sequence data. The method we describe is fundamentally different from those in two respects. First, we will only be using the gene expression data and not rely on sequence data. These methods would be hindered by both the availability of sequence in higher eukaryotes as well as the relatively poor accuracy in predicting transcription factor binding sites in higher eukaryotes.

In addition to the biological literature described above, there have been theoretical studies that lead to the implication of modular structures in gene networks. One such work has been put forth by Thieffry and Romero [18]. They model gene regulatory networks using Boolean models. Based on the framework, they show analytically that regulatory networks should consist of small and relatively independent feedback circuits, analogous to the network modules described above.

2.2. Data and Model

Let a^T denote the transpose of the vector a. At the ith time point ($i = 1, \ldots, n$), we let $X_i = [X_{i1} \ldots X_{ip}]$ denote the $p \times 1$ gene expression profile vector (i.e. X_{ij} is the gene expression measurement of the jth gene,

$j = 1, \ldots, p$). We suppose that the data have already been preprocessed and normalized.

Based on the discussion in Section 2.1, we will assume that the same expression profiles are generated by a combination of independent latent network modules. Suppose further that the effect of the modules on gene expression can be approximated by linear functions. We can then formulate the following model for the gene expression profiles:

$$X = AS, \tag{1}$$

where X is the $n \times p$ matrix whose ith row is X_i, A is a $n \times n$ matrix of mixing coefficients and S is an $n \times p$ matrix. The components of A and S are assumed to be unknown. The intuition behind the model is that the observed gene expression profiles (X) represent a mixture based on linear combinations (A) of source signals (S) that are assumed to be independent. The source signals correspond to the network modules described in Section 2.1. The rows of matrix A represents the expression modes of the network at the separate time points.

The model in (1) is similar to factor analysis models. As has been shown elsewhere [13], the model in (1) is well defined if the components of S are nonnormal. This is because for the multivariate normal distribution, all linear transformations are multivariate normal as well. This makes A in (1) nonidentifiable. Thus, the ICA is a nonnormal factor analysis model, in contrast to principal components analysis, which is based on a Gaussian latent variable model [11]. The emphasis on interesting directions being based on nonnormality also links ICA with projection pursuit methods [19, 20]. Many authors have argued that the normality assumption is not adequate for gene expression data [21, 22].

There are several methods of estimation in the ICA model, including negentropy maximization, mutual information minimization, and maximum likelihood estimation. A survey of estimation methods available for (1) can be found in [13]. For numerical implementation of maximum likelihood estimation in (1), we will be utilizing the FastICA algorithm. Further details can be found in [23]. We first start by whitening the covariance of X such that the transformed data have mean zero and identity covariance matrix. The method is based on a fixed point iteration approach to finding a maximizer of $E\{G(W^T \tilde{X})\}$, where W corresponds to S, $\tilde{X}$ represents the whitened data and G is a function measuring independence or equivalently nongaussianness. The solution is found by a combination of Newton-Raphson-type search methods and Gram-Schmidt-type orthogonalization.

Based on the maximum likelihood estimation procedure, estimates of A and S are obtained. We now use these estimates to explore the structure of the gene expression time series. One quantity of interest is the similarity between the expression profiles of a gene in the original space (corresponding to the columns of X) to the structural axes defined by the reduced space (corresponding to the rows of A). For the gth gene ($g = 1, \ldots, p$) and ith structural axis of A, this is done by taking the absolute value of the dot product of the gth column of X and ith row of A, appropriately normalized. We will refer to this quantity as the response coefficient and denote it R_{gi}. We can use the response coefficient for the gth gene to calculate a measure of its overall contribution to the expression information over the time points:

$$V_g = \sum_{i=1}^{n} R_{gi}^2$$

Based on the values of V_g, we can determine what genes show the most similarity to the structural axes defined by model (1).

It is also important to graphically explore the columns of A as well. This shows the time profiles of the network modules, the hidden regulatory factors. Large positive and negative values of the modules suggest that they are in an activating or repressing state. Because of the multiplicative nature of the model (1), it is not possible to determine the correct sign on the estimates of A and S.

We mention in passing that this approach has some connections to that performed in [24]. However, our approach is fundamentally different in one major respect. The model we formulated was based on the biological considerations described in Section 2.1, and the independent components are interpretable as network modules. In [24], the authors use singular value decomposition to estimate structural axes without consideration of an underlying model based on biological observations. On a more statistical level, their method focuses on analyzing second moments in the data, while our estimation procedure involves higher order moments.

3. Graph-Theoretic Algorithm for Mode Analysis

The next stage of analysis involves relating the network mixing coefficients (the rows of A) to each other.

What has typically been used is a linear model approach [7, 10]. In particular, the following linear model has been formulated for the modes:

$$A(t + \Delta t) = MA(t), \tag{2}$$

where M is an $n \times n$ time-invariant matrix. The matrix M describes the connectivity between modes. Several versions of the model in (2) have been suggested by other authors [7, 10]. Estimation of a solution for M has been typically done using a numerical method such as simulated annealing. We take an alternative approach to that presented in (2). Before describing the algorithm, we first discuss the biological motivation behind it.

In a variety of bioinformatics investigations [25, 26], authors have found that the large-scale organization of biological networks mirrors systems found in other disciplines. Namely, the topology consists of a small percentage of highly connected nodes and a majority of sparsely connected nodes. Thus, we want to incorporate this inherent sparseness into the modelling procedure.

We first start by utilizing a graph-theoretic algorithm for modelling the dependencies among the expression modes. The distance between the modes can be represented by a $n \times n$ pairwise dissimilarity matrix $D = [d_{ij}]$ based on a distance metric defined on $R^n \times R^n$. We use the Euclidean distance here. The number d_{ij} represents the distance between modes i and j. From a graph-theoretic point of view, the matrix D is a representation of an N-complete graph with varying edge lengths between nodes of the graph.

We then incorporate the *a priori* biological sparsity in the network by use of the hard-thresholding method; we will set dissimilarities above a certain cutoff value to zero. To determine the critical values, we permute the rows of A and recompute the distance matrix D. We do this K times. We then set the d_{ij} value equal to zero if the observed distance is greater than the 90th percentile of the permutation distribution. When we zero out an entry in the dissimilarity matrix, it corresponds to deletion of the edge connecting the two vertices from the graph.

While the calculation of pairwise dissimilarity matrix is also important for hierarchical clustering of gene expression data, our purpose is for quantifying the dependencies among expression modes. Note that this is an algorithmic approach which does not correspond to an underlying model. However, our end product is a connectivity matrix for the expression modes, analogous to M in (2). Furthermore, the numerical algorithm used is much simpler than that proposed by previous authors and incorporates the observations of previous authors [25, 26] in a natural way.

4. Application: Yeast Cell-Cycle Data

The proposed methodologies are applied to data from a recently published microarray study examining the gene expression behavior of the yeast genome during the cell cycle [14]. We provide some brief details of the study. Three sets of techniques were used synchronizing the yeast culture samples: α-factor arrest, elutriation and arrest of a *cdc15* temperature-sensitive mutant. We analyzed the three datasets separately. These datasets consist of measurements taken on 6113 genes at 24, 18 and 14 time points, respectively. We took the following steps for preprocessing the data:

1. We removed genes with MAX/MIN $>$ 5 or with more than 10% missing data.
2. Missing values were imputed using median imputation across genes.
3. The remaining ratios were transformed using a logarithmic transformation (base 2).

For the three experiments, this yielded a total of 4785, 5564 and 5748 genes, respectively.

We first applied the independent components analysis to each dataset and determined the genes that had the largest response coefficient values for each microarray experiment. The modes for the three experiments are shown in Figs. 1–3. The plots show that there are some common aspects in the structures of the time profiles for the expression modes. For example, there appears to be a common activation around $t = 120$ minutes across the modes for the α-factor arrest data; this is seen in Fig. 1. Similar examples can be seen in the elutriation and cdc15 datasets as well.

We next determined the top 20 genes for each experiment based on the response coefficient value; these are given in Tables 1–3. The genes that we have found are different from those reported in [24] for the three experiments. However, their expression modes were based on principal components, while we use independent components analysis. This underscores the ability of the independent components analysis method of finding structures that are nonnormal. While many of the genes are of unknown function, there are genes listed

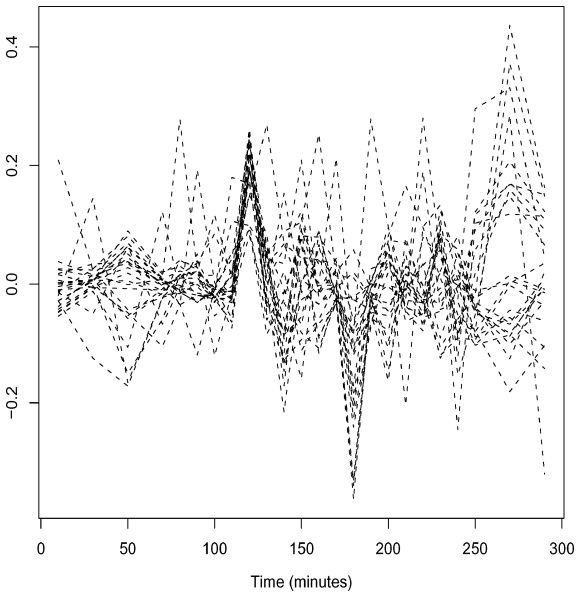

Figure 1. Time profiles of expression modes from independent components analysis of α-factor arrest yeast cell cycle data. These were estimated using independent components analysis and correspond to the columns of A.

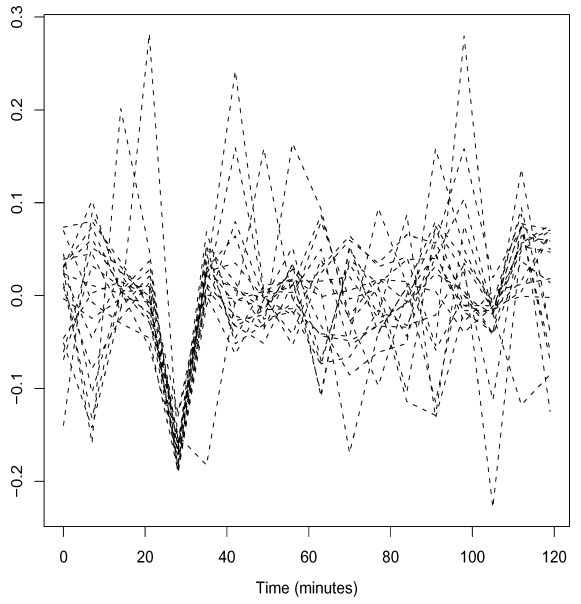

Figure 2. Time profiles of expression modes from independent components analysis of elutriation experiment with yeast cell cycle data (see Fig. 1).

that are involved in cell structure, translation (e.g., ribosomal proteins), and other cell functions.

To determine if the modes found were of biological interest and if they were consistent across

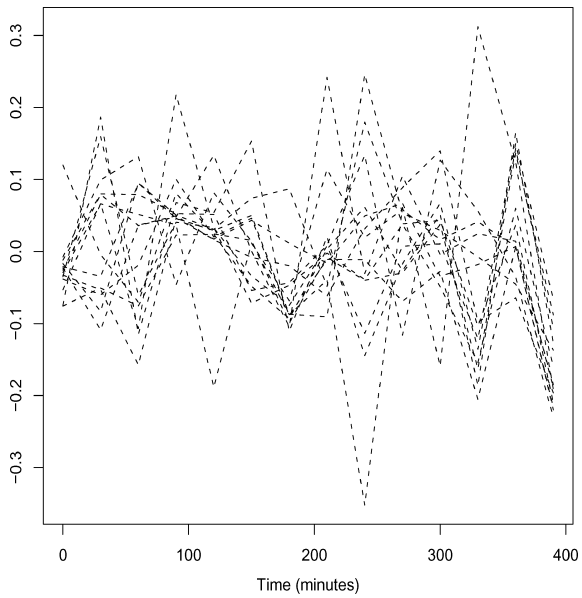

Figure 3. Time profiles of expression modes from independent components analysis of cdc15 yeast cell cycle data (see Fig. 1).

studies, we took the top 200 genes for each expression mode based on the response coefficient and then did a database search in the Stanford yeast microarray database (http://www.yeastgenome.org/) for the category of Gene Ontology [27] function term that appeared most frequently in the list. These results are summarized in Tables 4–6. Based on these results, we find that there is some consistency in the terms found across modes and across studies in terms of distributions of functions. Two limitations here is that we are considering the modal function keyword and that the search is on 200 genes.

We then applied the permutation-based algorithm to find the graphs connecting the modes; they are presented in Figs. 4–6. The plots of the adjacency matrices show how we have incorporated sparsity in the network using hard thresholding of the distance matrix. Based on the plots, we find that most modes have few, if any connections, but that there are a few modes with multiple connections. Examples of such modes include mode 3 for the α-factor dataset, mode 15 in the elutriation dataset, and mode 14 for the cdc15 data. Combining the information presented here with that in Tables 4–6, we find suggestive evidence that the connections appear to be consistent across studies. In particular, the modes with the highest connections tend to have transcription-related function.

Table 1. Top 20 genes based on V_g for α-factor arrest data.

Yeast ORF	Gene name	Function
YGL241W	KAP114	Karyopherin (collective name for homologous family of nuclear transport receptors) of approximately 114 kD
YLR071C	RGR1	Affects chromatin structure, transcriptional regulation of diverse genes and sporulation
YIL146C		Extracellular mutant
YLR050C		Unknown
YPL143W	RPL33A	Ribosomal protein L33A
YDR177W	UBC1	Ubiquitin-conjugating enzyme
YGR047C	TFC4/PCF1	Transcription factor tau 131 kD subunit
YNL109W		Similar to cytochrome-c oxidase
YML029W		Unknown
YIL090W		Unknown
YCR031C	RPS14A	40S ribosomal protein S14e
YBR103W	SIF2	Sir4p interacting protein
YNR008W	LR01	Phospholipid:diacylglycerol acyltransferase
YDL138W	RGT2	Glucose permease
YNR037C		40S ribosomal protein S15e
YIL130W		Unknown
YIL043C	CBR1	Cytochrome-b5 reductase
YDR121W	DPB4	DNA polymerase II (epsilon) 4th subunit
YKL055C	OAR1	Putative 3-oxoacyl-(acyl carrier protein) reductase
YMR123W	PKR1	Hypothetical protein

Table 2. Top 20 genes based on V_g for elutriation data.

Yeast ORF	Gene name	Function
YIL045W	PIG2	Glycogen synthase 2 interacting protein
YGR171C	MSM1	Methionyl-tRNA synthetase, mitochondrial
YJR020W		Unknown
YLR170C	APS1	AP-1 complex subunit, sigma1 subunit
YPL147W	PXA1	ABC family long-chain fatty acid transporter
YPR073C	LTP1	Protein-tyrosine-phosphatase/acid phosphatase
YMR178W		Unknown
YNL224C		Unknown
YOL096C	COQ3	3,4-dihydroxy-5-hexaprenylbenzoate methyltransferase
YOL115W	TRF4	Topoisomerase 1-related protein
YHR157W	REC104	Meiosis-specific protein
YFR032C		Unknown
YDR231C	COX20	In the maturation and assembly of cytochrome oxidase involved protein
YGR057C	LST7	Hypothetical protein
YGR225W	SPO70	Unknown protein
YCR019W	MAK32	Necessary for structural stability of L-A double-stranded RNA-containing particles
YPR148C		Unknown
YGR017W		Unknown
YIL013C	PDR11	Putative member of the ABC family of membrane transporters
YKL206C		Unknown

Table 3. Top 20 genes based on V_g for cdc15 data.

Yeast ORF	Gene name	Function
YGR181W	TIM13	Translocase of the inner membrane
YIR027C	DAL1	Allantoinase
YMR253C		Unknown
YPR087W		Unknown
YGR028W	MSP1	MSP1 protein (tat-binding homolog 4)
YLR080W		Unknown
YDR279W		Unknown
YPL176C	YPL176C	Similarity to chinese hamster transferrin receptor protein
YPL173W	MRPL40	Mitochondrial ribosomal protein L40
YJL014W	CCT3	T-complex protein 1, gamma subunit
YBR078W	ECM33	Involved in cell wall biogenesis and architecture
YFL049W		Unknown
YEL014C		Unknown
YLR380W	CSR1	Unknown
YNL074C	MLF3	Unknown
YER049W		Unknown
YDL124W		Similar to aldose reductase
YMR251W		Hypothetical protein
YMR144W		Unknown
YML130C	ERO1	Required for protein disulfide bond formation in the ER

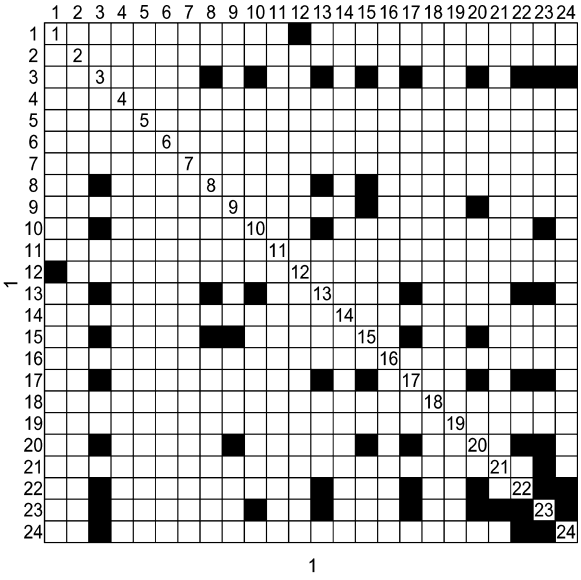

Figure 4. Plot of adjacency matrix for expression modes in α-factor dataset using permutation-based method; solid cell denotes connection, while an empty cell indicates no connection. This was found using the method in Section 3.

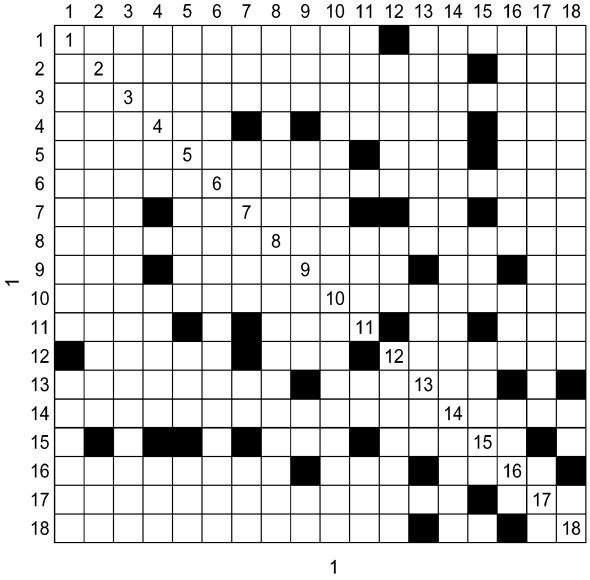

Figure 5. Plot of adjacency matrix for expression modes in elutriation dataset using permutation-based method; solid cell denotes connection, while an empty cell indicates no connection (see caption to Fig. 4).

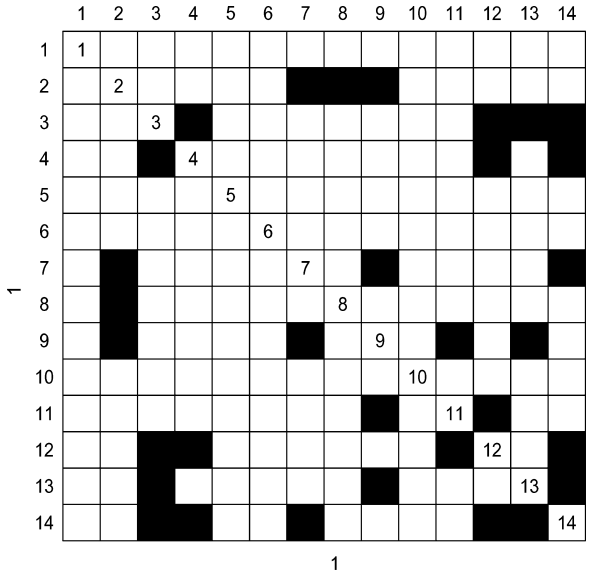

Figure 6. Plot of adjacency matrix for expression modes in cdc15 dataset using permutation-based method; solid cell denotes connection, while an empty cell indicates no connection (see caption for Fig. 4).

5. Discussion

With the explosion of transcription information available in microarray experiments, one goal researchers have taken is to reconstruct genetic networks using these data. However, because of the limited sample sizes and the experimental variation, reconstructing entire gene regulatory pathways seems like an overly ambitious task. We focus on a more immediate task, identifying dominant dynamic modes and studying their interrelationships.

For the first goal, we have utilized independent components analysis. The utility of this method is founded on the notion that the interesting linear structures exist in nonnormal directions. In addition, the modes have direct biological interpretation as latent variables that regulate expression. The use of independent components-based methods can provide complementary results to those given by singular value based techniques.

In order to study the dependencies among modes, we have utilized a graph-theoretic algorithm. The advantage of this approach is that it provides a natural approach to incorporate an *a priori* biological assumption of network sparsity through hard thresholding of pairwise dissimilarities. This method offers simpler alternatives to previous methods for addressing issues of inferring latent genetic networks and their dependencies from gene expression data [7, 10].

Table 4. Enrichment of Gene Ontology (GO) terms for expression modes for α-factor arrest data.

Mode	Function
1	Transport
2	Biological process unknown
3	Transcription
4	Biological process unknown
5	DNA replication and chromosome cycle
6	Signal transduction
7	Cell growth
8	Cell cycle
9	Biological process unknown
10	Cell cycle
11	Metabolism
12	Amino acid metabolism
13	Transport
14	Transport
15	Transcription
16	Biological process unknown
17	Cell cycle
18	Metabolism
19	DNA replication and chromosome cycle
20	Transcription
21	Metabolism
22	Biological process unknown
23	Transcription
24	Biological process unknown

The thresholding method described here is not model-based. One alternative we are currently developing for studying the dependencies between expression modes over time (i.e. the columns of A) involves a curve registration algorithm [28]. Here, our data consists of the trajectories for the n expression modes, $A_i(t)$ $(i = 1, \ldots, n)$, observed at n time points. We formulate the following model:

$$y(t) = A_i\{h_i(t)\} + \epsilon_i(t), \tag{3}$$

where $y(t)$ is a template function to which A_i $(i = 1, \ldots, n)$ will be aligned to, and $\epsilon_i(t)$ is a zero-mean stochastic process. The $h_i(t)$ represent time-warping functions to which the features of $A_i(t)$ will be aligned to; they represent the objects that are of interest. Modes

Table 5. Enrichment of Gene Ontology (GO) terms for expression modes for elutriation data.

Mode	Function
1	Cell cycle
2	Amino acid metabolism
3	Biological process unknown
4	Transport
5	Biological process unknown
6	Biological process unknown
7	Cell growth and/or maintenance
8	Transport
9	Cell cycle
10	DNA replication and chromosome cycle
11	Amino acid metabolism
12	Transcription
13	Metabolism
14	Biological process unknown
15	Transcription
16	Biological process unknown
17	Transcription
18	Signal transduction

Table 6. Enrichment of Gene Ontology (GO) terms for expression modes for cdc15 data.

Mode	Function
1	Amino acid metabolism
2	Biological process unknown
3	Transport
4	Signal transduction
5	Cell growth and/or maintenance
6	Cell cycle
7	Amino acid metabolism
8	Metabolism
9	DNA replication and chromosome cycle
10	Transport
11	Biological process unknown
12	Biological process unknown
13	Signal transduction
14	Transcription

with similar time-warping functions are likely to correspond to groups that are coregulated. We will communicate these results in a separate report.

While permutation-based approaches have been used for hypothesis testing here, it would also be desirable to develop model-based approaches for inference using the independent components model (1). This is also currently an area of ongoing research.

Acknowledgments

The author would like to thank the editor and three reviewers, whose comments substantially improved the presentation of the manuscript.

References

1. H. Kitano, *Foundations of Systems Biology*, Cambridge, MA: MIT Press, 2001.
2. T. Ideker, T. Galitski, and L. Hood, "A New Approach to Decoding Life: Systems Biology," *Annu. Rev. Genom. Hum. Genet.*, vol. 2, 2001, pp. 343–372.
3. H. Bussemaker, H. Li, and E.S. Siggia, "Regulatory Element Detection Using Correlation with Expression," *Nature Genetics*, vol. 27, no. 2, 2001, pp. 167–171.
4. X. Zhou, M.C. Kao, and W.H. Wong, "Transitive Functional Annotation by Shortest-Path Analysis," *Proceedings of the National Academy of Sciences USA*, vol. 99, no. 20, 2002, pp. 12783–12788.
5. P. D'haeseleer, S. Liang, and R. Somogyi, "Linear Modeling of mRNA Expression Levels During CNS Development and Injury," in *Pacific Symposium on Biocomputing*, R. Altman et al. (Eds.), Singapore: World Scientific Press, 1999, pp. 41–52.
6. A.J. Hartemink, D.K. Gifford, T.S. Jaakkola, and R.A. Young, "Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks," in *Pacific Symposium on Biocomputing*, R. Altman et al. (Eds.), Singapore: World Scientific Press, 2001, pp. 422–433.
7. M.K.S. Yeung, J. Tegnér, and J.J. Collins, "Reverse Engineering Gene Networks Using Singular Value Decomposition and Robust Regression," *Proceedings of the National Academy of Sciences USA*, vol. 99, no. 9, 2002, pp. 6163–6168.
8. O. Alter, P.O. Brown, and D. Botstein, "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling," *Proceedings of the National Academy of Sciences USA*, vol. 97, no. 18, 2000, pp. 10101–10106.
9. S. Raychaudhuri, J.M. Stuart, and R.B. Altman, "Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series," in *Pacific Symposium on Biocomputing*, R. Altman et al. (Eds.), Singapore: World Scientific Press, 2000, pp. 452–463.
10. N.S. Holter, A. Maritan, M. Cieplak, N.V. Fedoroff, and J.R. Banavar, "Dynamic Modeling of Gene Expression Data," *Proceedings of the National Academy of Sciences USA*, vol. 98, no. 4, 2001, pp. 1693–1698.
11. M. Tipping and C. Bishop, "Probabilistic Principal Component Analysis," *Journal of Royal Statistical Society (B)*, vol. 61, no. 3, pp. 611–622.
12. W. Liebermeister, "Linear Modes of Gene Expression Determined by Independent Component Analysis," *Bioinformatics* vol. 18, no. 1, 2002, pp. 51–60.
13. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York: Wiley and Sons, 2001.

14. P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, 1998, pp. 3273–3297.
15. L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature*, vol. 402S, 1999, pp. C47–C52.
16. J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, "Revealing Modular Organization in the Yeast Transcriptional Network," *Nature Genetics*, vol. 31, 2002, pp. 370–377.
17. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module Networks: Identifying Regulatory Modules and Their Condition-Specific Regulators from Gene Expression Data," *Nature Genetics*, vol. 34, no. 2, 2003, pp. 166–176.
18. D. Thieffry and D. Romero, "The Modularity of Biological Regulatory Networks," *Biosystems*, vol. 50, 1999, pp. 49–59.
19. J.H. Friedman and J.W. Tukey, "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions of Computers*, vol. 9, pp. 881–890.
20. J.H. Friedman, "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, vol. 82, no. 397, 1987, pp. 249–266.
21. M.K. Kerr, M. Martin, and G.A. Churchill, "Analysis of Variance for Gene Expression Microarray Data," *Journal of Computational Biology*, vol. 7, no. 6, 2000, pp. 819–837.
22. S. Dudoit, Y.H. Yang, M. Callow, and T.P. Speed, "Statistical Methods for Identifying Differentially Expressed Genes Replicated cDNA Microarray Experiments," *Statistica Sinica*, vol. 12, no. 2, 2002, pp. 111–140.
23. A. Hyvärinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, 1999, pp. 626–634.
24. S.A. Rifkin, K. Atteson, and J. Kim, "Constraint Structure Analysis of Gene Expression," *Functional and Integrative Genomics*, vol. 1, no. 3, 2000, pp. 174–185.
25. H. Jeong, B. Tabor, R. Albert, Z.N. Oltvai, and A.L. Barabási, "The Large-Scale Organization of Metabolic Networks," *Nature*, vol. 407, no. 6804, 2000, pp. 651–654.
26. H. Jeong, S.P. Mason, A.L. Barabási and Z.N. Oltvai, "Lethality and Centrality in Protein Networks," *Nature*, vol. 411, no. 6833, 2001, pp. 41–42.
27. Gene Ontology Consortium, "Creating the Gene Ontology Resource: Design and Implementation," *Genome Research*, vol. 11, 2001, pp. 1425–1433.
28. J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, New York: Springer-Verlag, 1997.

Debashis Ghosh is Assistant Professor of Biostatistics. He received his Ph.D. in biostatistics in 2000 from the University of Washington. Dr. Ghosh's research interests are in various areas of statistics and biostatistics including failure time and survival analysis, longitudinal data, design and analysis of high-throughput gene expression studies, clustering and classification problems and semiparametric inference.
ghoshd@umich.edu