JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 3, 2009

Pp. 407–426
DOI: 10.1089/cmb.2008.0081

# Analysis of Gene Sets Based on the Underlying Regulatory Network

ALI SHOJAIE and GEORGE MICHAILIDIS

## ABSTRACT

**Networks are often used to represent the interactions among genes and proteins. These interactions are known to play an important role in vital cell functions and should be included in the analysis of genes that are differentially expressed. Methods of gene set analysis take advantage of external biological information and analyze *a priori* defined sets of genes. These methods can potentially preserve the correlation among genes; however, they do not directly incorporate the information about the gene network. In this paper, we propose a latent variable model that directly incorporates the network information. We then use the theory of mixed linear models to present a general inference framework for the problem of testing the significance of subnetworks. Several possible test procedures are introduced and a network based method for testing the changes in expression levels of genes as well as the structure of the network is presented. The performance of the proposed method is compared with methods of gene set analysis using both simulation studies, as well as real data on genes related to the galactose utilization pathway in yeast.**

**Key words:** gene networks, gene set analysis, latent variable model, mixed linear model.

## 1. INTRODUCTION

IN STANDARD ANALYSIS OF DIFFERENTIAL EXPRESSION, statistical significance of each gene is assessed independently, and some method of multiple testing correction is then used to adjust the estimated $p$-values. Such methods are usually less sensitive in detecting genes that have smaller differences in mRNA abundance between different experimental conditions and may therefore be less powerful than desired. Furthermore, analyzing individual genes (single-gene analysis) often generates results that are not reproducible and lack meaningful biological interpretations. The focus of current research has thus shifted to analyzing *a priori* defined sets of genes (gene set analysis) and using external information to strengthen the analysis of differential expression. Analysis of gene sets results in increased power compared to single gene analysis. Furthermore, methods of gene set analysis can preserve the correlation among genes which may lead to more reliable inference. These methods however, do not directly incorporate the external information about the interactions among genes represented by the gene network. In this paper, we develop a model that directly incorporates the network information, and propose a general inference framework for testing the significance of genetic pathways.

Department of Statistics, University of Michigan, Ann Arbor, Michigan.

### *1.1. A motivating example*

In an interesting approach, Ideker et al. (2001) integrated gene expression and protein level data to study significant signaling and metabolic pathways in yeast *Saccharomyces cerevisiae*. They reported interactions among genes and proteins in different pathways along with information on the estimated correlation among genes in the network. The authors also grouped the genes into subnetworks (pathways) based on their biological functions. Figure 1, which was originally presented in Ideker et al. (2001), illustrates the network of genes under consideration. We also update the network of Ideker et al. (2001) based on newly

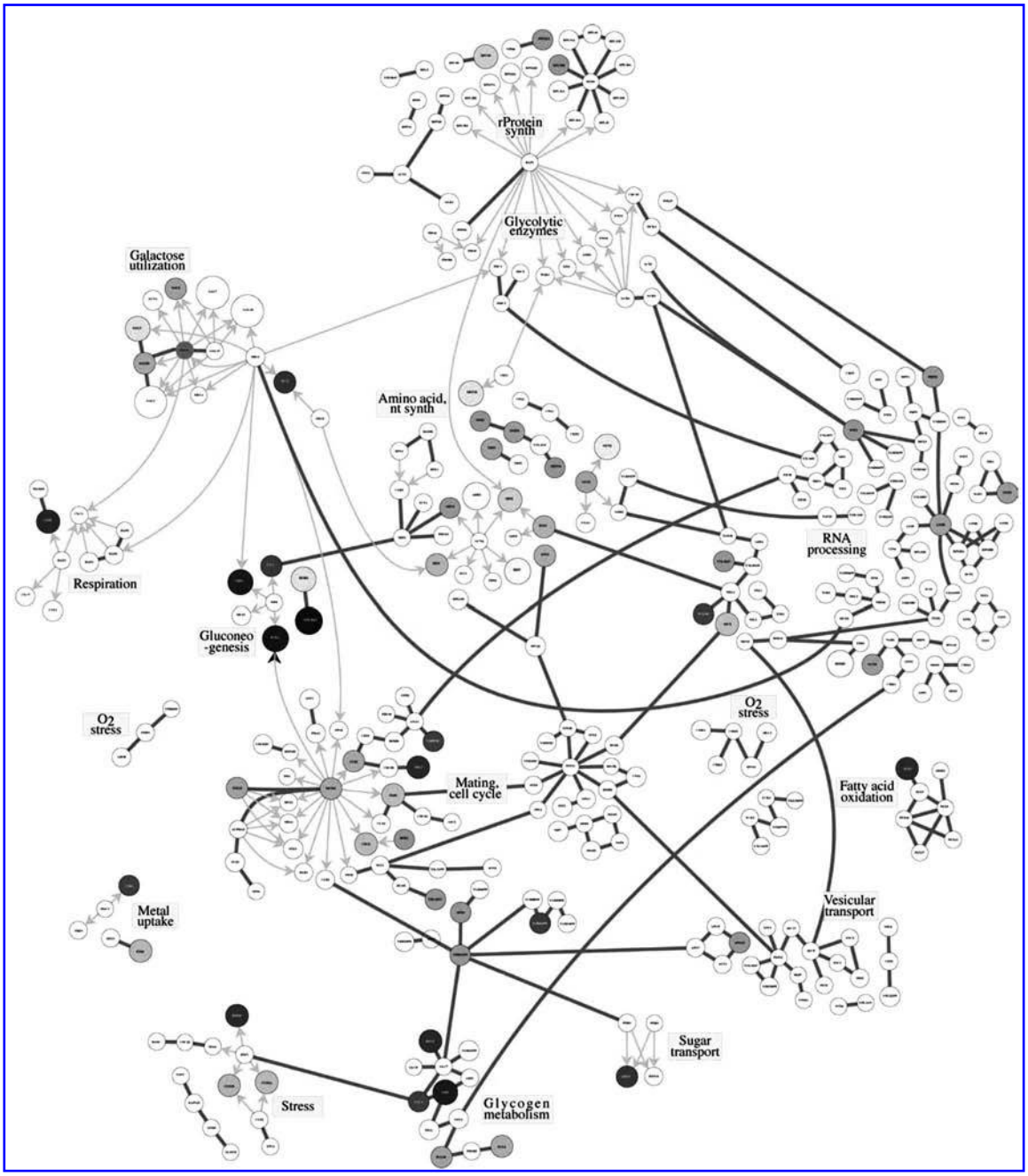


**FIG. 1.** Yeast galactose utilization pathway published in Ideker et al. (2001). (Printed with permission from *Science and the American Association for the Advancement of Science*.)

TABLE 1. ANALYSIS OF IDEKER 2001 DATA USING GSEA

| *Pathway* | *Size* | *NOM p-val* | *FDR q-val* | *FWER p-val* | *Involved in gal (+/−)* |
|---|---|---|---|---|---|
| Galactose utilization | 12 | 0.0020 | 0.00114 | 0.003 | + |
| Amino acid synthesis | 30 | 0.1853 | 0.21562 | 0.676 | + |
| rProtein synthesis | 28 | 0.5261 | 0.44938 | 0.972 | + |
| Stress | 12 | 0.02004 | 0.19283 | 0.108 | − |
| Vesicular transport | 19 | 0.07243 | 0.54138 | 0.489 | − |
| Glycogen metabolism | 12 | 0.1321 | 0.41115 | 0.538 | − |
| Respiration | 9 | 0.1878 | 0.39508 | 0.637 | − |
| $O_2$ stress | 13 | 0.2384 | 0.6601 | 0.906 | − |
| Fatty acid oxidation | 7 | 0.4694 | 0.82373 | 0.963 | − |
| Mating, cell cycle | 58 | 0.3583 | 0.71842 | 0.968 | − |
| Sugar transport | 2 | 0.7358 | 1 | 0.993 | − |
| Metal uptake | 4 | 0.8374 | 1 | 0.997 | − |
| Gluconeogenesis | 7 | 0.8455 | 0.98853 | 0.997 | − |
| RNA processing | 75 | 0.9879 | 1 | 1 | − |
| Glycolytic enzymes | 16 | 0.9683 | 0.98189 | 1 | − |

The first two columns illustrate the pathway considered and the number of genes in the pathway. For each gene set, the Nominal $p$-value, FDR $q$-value, and FWER $p$-value are reported along with the involvement of the pathway in gal+/gal− conditions.

defined interactions among genes reported in Bader et al. (2004). This results in a network of 343 genes with 419 interactions for which estimates of correlations among genes are also available (this data is referred to as the *Ideker data* henceforth).

The mRNA expression levels of genes in the Ideker data are measured in 9 different perturbations of GAL genes along with the wild type yeast. For each perturbation, two samples of data are available. The first set of samples represents the expression levels of genes in cells grown in presence of galactose (gal+), while the second set includes expression levels for cells grown in absence of galactose (gal−), where the main source of carbon is raffinose. Our primary goal is to determine the pathways that are *involved* (either induced or suppressed) in cell growth in gal+ compared to gal− environments. In other words, we would like to test whether each of 15 gene sets defined by yeast pathways in the network of Ideker et al. (2001) is differentially expressed in gal+ compared to gal− medium.

In this section, we analyze the Ideker data using methods of gene set analysis. More specifically, we apply the *Gene Set Enrichment Analysis* (GSEA) method of Subramanian et al. (2005). This method uses a permutation-based test (permuting the class labels) to determine whether genes in *a priori* defined gene sets have non-random associations with the phenotype. To that end, we first normalize the data so that the expression levels only represent the effect of the growth environment.[1] The results of the analysis are displayed in Table 1.

The first line of the table presents an expected outcome; the expression levels of genes in the Galactose Utilization pathway is expected to change in response to perturbations of GAL genes in the gal+ environment. On the other hand, although some of the pathways seem to have differential expression when cells lack galactose (e.g., Stress and Vesicular Transport), no other pathway appears significant after adjusting for multiple testing using the False Discovery Rate (FDR) controlling procedure of Benjamini and Hochberg (1995) with a $q$-value of 0.05. In Section 5, we revisit the analysis of the Ideker data based on the method proposed in this paper, which directly incorporates the network information represented by the gene network in Figure 1.

[1]The mean expression levels of the two samples corresponding to each perturbation is subtracted from the two columns of data.

### *1.2. Background*

Recent research on gene set analysis can be broadly classified into permutation-based methods motivated by the GSEA paper and model-based approaches that make specific distributional assumptions about the gene expression data. The literature can be further categorized on whether direct or indirect external information on the gene network is employed. Tian et al. (2005) considered the problem of gene set analysis and described two hypotheses that should be considered when studying the significance of sets of genes. One of these hypotheses, which is the same as the hypothesis considered in GSEA, focuses on non-random association of genes in the gene set with the phenotype. The other hypothesis, considers non-random correlations between genes in a gene set. The test method proposed for the first hypothesis is based on permuting the class labels (column permutation) and the second hypothesis is tested by permuting genes (row permutation). Efron and Tibshirani (2007) formalized the idea of gene set analysis in a coherent statistical framework and examined the hypotheses presented in Tian et al. (2005). They also proposed an alternative test statistic with superior power properties and analyzed the effects of row and column permutations. Goeman and Bühlmann (2007) reviewed different methods proposed for testing significance of gene sets and highlighted important issues in selecting appropriate methods.

Although the above permutation-based methods are computationally intensive, they include minimum assumptions about the underlying biological model and are therefore more robust to model misspecification. An alternative approach is based on model-based tests procedures, where specific distributions for the expression data are assumed. In one such approach, Jiang and Gentleman (2007) extended the idea of gene set analysis by adapting a linear model approach and adjusting for other covariates. They presented the gene sets in the form of an index matrix and offered a heuristic argument for using a normal approximation for testing per gene set sums. One major difficulty regarding model-based methods is the large number of variables (genes) compared to the small number of samples—the large $p$, small $n$ problem (West, 2000). In such situations, estimation of model parameters becomes a challenging task and may result in unstable outcomes. However, additional sources of information besides the expression levels of genes could be used to make the estimation more accurate. One such source of external information is the underlying relationship between genes which itself is of independent interest. It is known that genes interact with each other through their protein products and form gene regulatory networks. Also, the protein products of groups of genes are involved in controlling specific functions in cells through genetic pathways. Increasing amount of information about these relationships is becoming available in public repositories, like the KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000) and the Gene Ontology (GO) (Ashburner et al., 2000), and can be used to improve the estimates of model parameters.

A number of researchers have recently used external information about gene networks to improve the analysis of gene sets. Rahnenführer et al. (2004) demonstrated that the sensitivity of detecting relevant pathways can be improved by integrating information about pathway topology. Barry et al. (2005) presented a permutation based procedure, called SAFE, that considers the underlying network structure. More recently, Wei and Li (2007) have proposed a Markov random field model to incorporate the information on the gene network in the analysis. In a related approach, Wei and Pan (2008) have modeled the network information via latent variables into a spatially correlated mixture model. Both of these methods, consider the problem of analysis of *single genes* on the network.

The above methods either assume that the underlying network does not change as the experimental conditions change or they do not incorporate this change directly into the model. However, changes in the underlying network structure can amplify the change in expression patterns and should be included in the analysis. For instance, Li (2002) demonstrated that the correlation patterns among ARG2 and other members of the urea-cycle pathway can change drastically as the expression level of ARG2 changes. Another concern in analyzing network data is to decorrelate subnetworks from the effects of other nodes in the network and to deal with nodes that belong to multiple networks. Alexa et al. (2006) present one such method which is an attempt to decorrelate GO graph structures. Their method focuses on decorrelating nodes at lower levels (children) from upper level nodes (parents).

In this paper, we propose a latent variable model to directly incorporate the underlying gene network and present test statistics for testing the significance of arbitrary sub-networks based on the theory of mixed linear models. One major advantage of the method proposed in this paper is that not only does it consider

the change in the expression levels of the genes in different conditions, but also reflects the change in network structures and correlations among genes. We also present a systematic approach that decorelates each subnetwork from the other nodes while maintaining the interactions among genes in the subnetwork.

The rest of the paper is organized as follows. In the next section, the proposed latent variable model is introduced and some basic graph theoretical properties related to this model are discussed. In Section 3, we represent the latent variable model using the framework of mixed linear models and propose a general testing scheme based on the theory of mixed linear models. Section 3 ends with a result that is used to test the *pure* effect of each subnetwork. This result prevents tests of significance of subnetworks to be confounded with the effects of other subnetworks and also allows testing the effect of genes that belong to multiple networks. Section 4 includes three simulation studies for evaluating the performance of the new model under different testing conditions as well as studying the effect of noise in the network information on the proposed inference procedure. In Section 5, we revisit the Ideker data, introduced in Section 1.1, and test the significance of pathways using the proposed model. Section 6 includes a discussion on limitations of the proposed model and future extensions.

## 2. THE LATENT VARIABLE MODEL

Consider gene expression data $\mathcal{D}$ organized as a $p \times n$ matrix comprised of the expression levels of $p$ genes for $n$ samples, and let $Y$ be the $k$th sample in the expression data ($k$th column of $\mathcal{D}$).

To model the correlation structure caused by the gene network, we represent the network as a directed graph $G = (V, E)$ with vertex set $V$, and edge set $E$, where $E$ is represented by the $p \times p$ adjacency matrix $A$. Each nonzero element of the adjacency matrix, $A_{ij}$, represents a directed edge in the network. Elements of the adjacency matrix correspond to the strength of association among genes in the graph and are real values in $(-1, 1)$.

Consider the simple network of Figure 2: Suppose $Y = X + \varepsilon$, where $X$ represents the *signal* and $\varepsilon \sim N_p(0, \sigma_\varepsilon^2 I_p)$ the *noise*. Consider two adjacent genes $i$ and $j$, where $i$ affects $j$. One can represent the relationship between $i$ and $j$ using a simple linear model $X_j = \rho_{ij} X_i$. However, to account for unknown associations among genes and/or errors in the association weights, $\rho_{ij}$, we also add *latent variables* $\gamma_j \sim N_p(\mu_j, \sigma_\gamma^2)$ to represent the baseline expression level of gene $j$. For instance, $\gamma_2$ represents the expression level of gene 2 without the effect from gene 1. Thus, for the simple gene network of Figure 2, we obtain

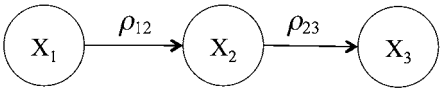


**FIG. 2.** A simple gene network.

$$
\begin{aligned}
X_1 &= \gamma_1 \\
X_2 &= \rho_{12} X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2 \\
X_3 &= \rho_{23} X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3
\end{aligned}
$$

These equations can be summarized in vector notation as:

$$
Y = \Lambda\gamma + \varepsilon, \qquad \gamma \sim N_p(\mu, \sigma_\gamma^2 I_p), \qquad \varepsilon \sim N_p(0, \sigma_\varepsilon^2 I_p) \tag{1}
$$

where $\Lambda$ is called the *influence matrix* of the graph. In the simple example above, we have

$$
\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & 1 & 0 \\ \rho_{12}\rho_{23} & \rho_{23} & 1 \end{pmatrix}
$$

Under such a model, $Y$ is a normal random variable with mean $\mathbb{E}[Y] = \Lambda\mu$ and variance $\text{Var}(Y_i) = \sigma_\gamma^2 \Lambda\Lambda' + \sigma_\varepsilon^2 I_p$, where $\Lambda'$ denotes the transpose of matrix $\Lambda$.

In the remainder of this section, we study the relationship between the influence matrix, $\Lambda$, and the adjacency matrix of the graph, $A$. We provide a general result for the relationship between $\Lambda$ and $A$ as well as a compact expression that can be used to efficiently evaluate $\Lambda$ for specific classes of graphs. We also discuss conditions under which the matrix $\Lambda$ has full rank, which will be used in the analysis of the proposed inference procedure in Section 3.

**Lemma 2.1.** *For any graph $G = (V, A)$ we have $\Lambda = A^0 + A^1 + A^2 + \cdots = \sum_{r=0}^{\infty} A^r$ (here $A^0$ is defined to be the identity matrix).*

**Proof.** From the matrix representation of the latent variable model in (1)

$$Y_i = \sum_{j=0}^{p} \Lambda_{ij}\gamma_j + \varepsilon_i, \qquad i = 1, \ldots, p$$

where $\Lambda_{ii} = 1$ and $\Lambda_{ij} \neq 0$ only if there is a path (of some length) on the graph from node $i$ to node $j$. But for any graph $G$, the number of paths of length $r$ ($r \in \mathbb{N}$) from $v_i$ to $v_j$ is given by the $(i, j)$ element of $A^r$ (Diestel, 2006). Therefore, $\Lambda_{ij} \neq 0$ whenever there exists $r$ such that $[A^r]_{ij} > 0$. Hence, all possible paths from $i$ to $j$ are given by $[\sum_{r=0}^{\infty} A^r]_{ij}$. This implies that $\Lambda = \sum_{r=0}^{\infty} A^r$. ■

**Corollary 2.2.** *For any Directed Acyclic Graph (DAG), $\Lambda = A^0 + A^1 + A^2 + \cdots + A^p$.*

**Proof.** This follows immediately from Lemma 2.1 by noting that since there are no loops in DAGs, the maximum length of paths equals $p$. ■

The following results provide sufficient conditions for the matrix $\Lambda$ to be of full rank. Although this guarantees validity of the model for at least some classes of directed graphs, it does not provide a necessary condition. Based on experiments with randomly generated adjacency matrices, there are in fact larger classes of graphs satisfying this property.

**Lemma 2.3.** *For any Directed Acyclic Graph (DAG), the matrix $\Lambda$ has full rank.*

**Proof.** The full rankness of $\Lambda$ is proved by showing that $\Lambda$ can be re-arranged into a lower triangular matrix with 1's on the diagonal.

First observe that $\Lambda_{ij} \times \Lambda_{ji} = 0$, since otherwise there will be a cycle in the graph. Also, from 2.1 we have $\Lambda_{ii} = 1$.

Consider a reordering of rows (and correspondingly of columns) of the matrix in decreasing number of zeros. Every DAG has at least one root (a node that is not affected by any other node). This means that there is at least one row with $\Lambda_{kk} = 1$ and $\Lambda_{kj} = 0$ for all $j$. Permute $\Lambda$ so that row $k$ is the first row of the matrix and continue in the same way. Denote the number of zero elements of row $i$ by $\phi_i$ and number of zeros in column $j$ as $\phi_{Cj}$. Then by the above observation, $\phi_{Ri} \geq p - \phi_{Ci}$ (here $p - \phi_{Ci}$ is the number of nonzero elements in column $i$).

To complete the proof, we need to show that the rearranged matrix $\Lambda$ can be further permuted to result in a lower diagonal matrix. Suppose there exists $j > i$ such that $\Lambda_{ij} > 0$ and therefore $\Lambda_{ji} = 0$. If $\phi_{Rj} = \phi_{Ri}$ switch $i$ and $j$ to get a lower triangular matrix. However, if $\phi_{Rj} < \phi_{Ri}$ (i.e., if $i$ is affected by a row with less number of zeros) there exists $l$ such that $\Lambda_{jl} > 0$ but $\Lambda_{il} = 0$. However, $\Lambda_{jl} > 0$ means there exists a path from $l$ to $j$ and $\Lambda_{ij} > 0$ means that there exists a path from $j$ to $i$. Thus there exists a path from $l$ to $i$, i.e. $\Lambda_{il} > 0$, a contradiction. Therefore $\Lambda$ must be a lower triangular matrix with $\Lambda_{ii} = 1$. ■

**Lemma 2.4.** *Consider a graph G = (V,A) with influence matrix* $\Lambda$
*a) If G is a Directed Acyclic Graph (DAG), then* $A = I - \Lambda^{-1}$.
*b) If the sum of absolute values of weights of edges ending at every node of the graph G is less than 1 (i.e. A is sub-stochastic), then* $A = I - \Lambda^{-1}$ *and* $\Lambda$ *has full rank.*

**Proof.** a) From Corollary 2.2, $\Lambda = \sum_{r=0}^{p} A^r$ and hence

$$A\Lambda = \sum_{r=0}^{p} A^{r+1} = \Lambda + A^{p+1} - I$$

But when $G$ is a DAG, $A^{p+1} = 0$ hence $A\Lambda = \Lambda - I$. By full rankness of $\Lambda$, $A = I - \Lambda^{-1}$.
b) The condition in (b) implies that the sum of the absolute values of off-diagonal elements of $A$ is less than 1. Let $s_i$ be the sum of absolute values of off-diagonal elements of the $i$th row of $A$. Since the diagonal elements of $A$ are 0, by the Gershgorin's Ring Theorem (Friedberg et al., 1996) if $\lambda$ is an eigenvalue of $A$, we have $|\lambda| \leq s_i \leq 1$. Now let $\Lambda_m = \sum_{r=0}^{m} A^r$. Then $\Lambda = \lim_{m\to\infty} \Lambda_m$ and using an argument similar to part (a),

$$A\Lambda_m = \Lambda_m - I + A^{m+1}$$

Since eigenvalues of $A$ are less than 1 in magnitude, $\lim_{m\to\infty} \Lambda_m$ exists (Friedberg et al., 1996) and by the eigen-decomposition of $A$, $A^{m+1} \to 0$ as $m \to \infty$. Hence, taking the limit, we get $A\Lambda = \Lambda - I + A$. On the other hand, the established bound on the eigenvalues of $A$ implies that all eigenvalues of $I - A$ are nonzero, which means that $I - A$ and therefore, $\Lambda$ are full rank. Thus $A = I - \Lambda^{-1}$. ■

Lemma 2.4 establishes an alternative relationship between $\Lambda$ and $A$ and determines two classes of graphs for which such a relationship is valid. As noted before, conditions presented in this result are only sufficient. For the general graph $G = (V, A)$, if the spectral radius of $A$ is less than 1, $\Lambda$ has full rank and the relationship between $A$ and $\Lambda$ established in Lemma 2.4 holds. On the other hand, in special cases where $\Lambda$ is not of full rank, it may be possible to modify the graph and therefore apply the model presented here. For instance, one large class of graphs where $\Lambda$ is not full rank consists of *cyclic* graphs. The cycles in biological networks are often representatives of feedback loops which are common features of cell cycle related networks. However, the feedback is usually effective after a time delay and therefore, when time series data is used to study these networks, the cycles can be broken down by distinguishing between nodes at the beginning and end of each cycle. Undirected edges (e.g., protein-protein interactions) can also be transformed into two directed edges using a common latent variable affecting both nodes. More generally, it is often possible to transform the graph by introducing dummy nodes and can hence apply the model presented here.

## 3. INFERENCE

### *3.1. Preliminaries*

In this section, we study the inference procedure for the proposed model. Although this method can be used to test a variety of hypotheses, in order to simplify the presentation, we focus on testing the equality of means of two experimental conditions. The extension to more complicated settings is discussed at the end of the section. As before, let $Y$ be a given sample in the expression data ($k$th column of data matrix $\mathcal{D}$) and let $Y^C$ and $Y^T$ represent *control* and *treatment* conditions, with $n_1$ columns of $\mathcal{D}$ corresponding to control samples and $n_2 = (n - n_1)$ columns to treatment samples. Also let two sets of parameters $(\mu^C, \Lambda^C)$ and $(\mu^T, \Lambda^T)$ represent mean vectors and influence matrices under control and treatment conditions, respectively.

Let $\mathbf{b}$ be an indicator vector determining genes that belong to a specific gene set (pathway). In other words, $\mathbf{b}_j = 1$ if gene $j$ is in gene set and 0 otherwise. We can test the significance of the gene sets by defining the test statistic $\mathbf{V} = \mathbf{b}Y^T - \mathbf{b}Y^C$ and testing:

$$H_0 : \mathbb{E}[\mathbf{V}] = 0 \quad \text{vs.} \quad H_1 : \mathbb{E}[\mathbf{V}] \neq 0 \tag{2}$$

Then under $H_0$:

$$\mathbb{E}_0[\mathbf{V}] = 0$$

and

$$\mathrm{Var}_0(\mathbf{V}) = (1/n^2)[n_2(\mathbf{b}\Lambda^T)(\mathbf{b}\Lambda^T)' + n_1(\mathbf{b}\Lambda^C)(\mathbf{b}\Lambda^C)']$$

Although the hypothesis in (2) can be tested using a generalized likelihood ratio test, it turns out that the latent variable model of Section 2 can be represented as a *Mixed Linear Model* (MLM). Using this framework, we can study a variety of spatio-temporal models and consider more general hypothesis testing problems.

### 3.2. Mixed linear model representation

Let $\mathbf{Y}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ represent the rearrangement of vectors $Y$, $\gamma$, and $\varepsilon$ into $np \times 1$ column vectors. Then $\mathbf{Y} = \boldsymbol{\Psi}\beta + \boldsymbol{\Pi}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ where:

$$\beta = (\mu_1^C, \ldots, \mu_p^C, \mu_1^T, \ldots, \mu_p^T)'$$

$$\boldsymbol{\Psi} = \begin{pmatrix} \Lambda^C & \cdots & \Lambda^C & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \Lambda^T & \cdots & \Lambda^T \end{pmatrix}'$$

$$\boldsymbol{\Pi} = \mathrm{diag}(\Lambda^C, \ldots, \Lambda^C, \Lambda^T, \ldots, \Lambda^T)'$$

In this model, $\boldsymbol{\gamma}$ is the vector of (unknown) *random effects* and $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are normally distributed random vectors with:

$$\mathbb{E}\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and

$$\mathrm{Var}\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \Sigma_\gamma & 0 \\ 0 & \Sigma_\varepsilon \end{bmatrix}$$

For the latent variable model presented in the previous section, $\Sigma_\gamma = \sigma_\gamma^2 I$ and $\Sigma_\varepsilon = \sigma_\varepsilon^2 I$ and the variance of $Y^j$, $j \in \{C, T\}$ is given by $\sigma_\gamma^2 \Lambda^j (\Lambda^j)' + \sigma_\varepsilon^2 I$.

The estimate of $\beta$ in the mixed linear model is given by (Searle, 1971):

$$\hat{\beta} = (\boldsymbol{\Psi}'\hat{\mathbf{W}}^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\hat{\mathbf{W}}^{-1}\mathbf{Y}$$

$\mathbf{W} = (\sigma_\gamma^2 \boldsymbol{\Pi}\boldsymbol{\Pi}' + \sigma_\varepsilon^2 \mathbf{I}_{np})$. The estimate of $\beta$ depends on estimates of $\sigma_\gamma^2$ and $\sigma_\varepsilon^2$ which can be estimated via *Restricted Maximum Likelihood* procedure (REML).

The framework of mixed linear models allows us to test a variety of hypotheses about $\beta$ by considering tests of the form:

$$H_0 : l\beta = 0 \quad \text{vs.} \quad H_1 : l\beta \neq 0 \tag{3}$$

Here $l$ is in general any *estimable* linear combination of $\beta$'s (Searle, 1971). An example of such a vector is a *contrast vector*, which satisfies the constraint $\mathbf{1}'l = 0$. In the ensuing discussion, any linear combination of $\beta$'s satisfying the estimability requirement is referred to as a *contrast vector*.

Based on the theory of mixed linear models, we can test (3) using the test statistic:

$$T = \frac{l\hat{\beta}}{\sqrt{l\hat{C}l'}} \tag{4}$$

where $C = (\boldsymbol{\Psi}'\mathbf{W}^{-1}\boldsymbol{\Psi})^{-1}$.

Under the null hypothesis in (3), $T$ has approximately a $t$ distribution with $\nu$ degrees of freedom, where the degrees of freedom is estimated using the Satterthwaite approximation method (McLean and Sanders, 1988):

$$\nu = \frac{2(l\hat{C}l')^2}{\tau' K \tau}$$

with $\tau = (\frac{\partial}{\partial\sigma^2_\gamma} lCl', \frac{\partial}{\partial\sigma^2_\varepsilon} lCl')'$ and $K$ is the empirical covariance matrix of $(\sigma^2_\gamma, \sigma^2_\varepsilon)'$.

### 3.3. Computational issues and the use of the mixed linear model

The mixed linear model facilitates the representation of the latent variable introduced in Section 2. However, estimation and inference in this framework involves forming the matrices $\mathbf{\Psi}$ and $\mathbf{\Pi}$, and performing operations involving products and inverses of these matrices. In the context of analysis of genetic data, the dimensions of these matrices ($np \times 2p$ and $np \times np$) can cause serious difficulties in terms of computation time, memory requirement and numerical stability of the estimation algorithms. It is therefore necessary to derive alternative methods for estimation of parameters in the model. It turns out that due to the special structure of the model presented in Section 2, and the sparsity pattern of matrices $\mathbf{\Psi}$ and $\mathbf{\Pi}$, the formulas presented in the previous section can be substantially simplified. More specifically, for the problem stated in Section 3.2 we have:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}^C \\ \hat{\beta}^T \end{pmatrix} = \begin{pmatrix} \bar{Y}^C \\ \bar{Y}^T \end{pmatrix}$$

and

$$C = \begin{bmatrix} \frac{1}{n_1}(\sigma^2_\gamma I_p + \sigma^2_\varepsilon(\Lambda^{C\prime}\Lambda^C)) & 0 \\ 0 & \frac{1}{n_2}(\sigma^2_\gamma I_p + \sigma^2_\varepsilon(\Lambda^{T\prime}\Lambda^T)) \end{bmatrix}$$

In the particular case considered here, the REML estimates of the variance components can be directly computed as the maximizers of the REML equation without any need for iterative methods. However, profiling out one of the variance components may result in more stable solutions.

### 3.4. Role of the contrast vector

The estimates of $\beta$ based on the mixed linear model represent the individual expression level of each gene in the network. Thus, in order to evaluate the combined effect of each gene set using the test statistic $T$, the choice of contrast vector $l$ proves fairly crucial. More specifically, the choice of $l$ determines the null and alternative hypotheses of the test in (3), which in turn affects its significance level and power. In this section, we present different choices of contrast vectors and study their properties and effects on the power of tests.

A simple choice for the contrast vector $l$ is to use the indicator vector of the gene set. In other words,

$$l^{(1)} = (-\mathbf{b}, \mathbf{b}) \tag{5}$$

This simple choice of $l$ corresponds to testing the following hypothesis:

$$H_0^{(1)} : \mathbf{b}(\mu^T - \mu^C) = 0 \quad \text{vs.} \quad H_1^{(1)} : \mathbf{b}(\mu^T - \mu^C) \neq 0 \tag{6}$$

which for each gene set $g$ is equivalent to

$$H_0^{(1)} : \sum_{i \in g} \mu_i^T - \mu_i^C = 0 \quad \text{vs.} \quad H_1^{(1)} : \sum_{i \in g} \mu_i^T - \mu_i^C \neq 0 \tag{7}$$

Such a contrast vector however, only considers the mean expression levels of genes and does not reflect the combined effect of the set of genes in **b**, which is affected by interactions among genes in the network.

When the underlying network structure and therefore the correlation among genes is known, a natural alternative to $l^{(1)}$ is to also include the influence matrices $\Lambda^C$ and $\Lambda^T$. This leads to the following choice of contrast vector:

$$l^{(2)} = (-\mathbf{b}\Lambda^C, \mathbf{b}\Lambda^T) \tag{8}$$

which corresponds to testing the following hypotheses:

$$H_0^{(2)} : \mathbf{b}(\Lambda^T \mu^T - \Lambda^C \mu^C) = 0 \quad \text{vs.} \quad H_1^{(2)} : \mathbf{b}(\Lambda^T \mu^T - \Lambda^C \mu^C) \neq 0 \tag{9}$$

The null hypothesis presented in (9) may first seem less intuitive and the choice of $l^{(2)}$ rather arbitrary. However, the rationale behind the latter choice of contrast vector becomes clearer when we examine the test statistics corresponding to each one of the two null hypotheses in (6) and (9). In the case of the two-population test considered here, the above choices of contrast vectors lead to (after some algebra) the following test statistics:

$$T_1 = \frac{\mathbf{b}((\Lambda^T)^{-1}\bar{Y}^T - (\Lambda^C)^{-1}\bar{Y}^C)}{\sqrt{\hat{\sigma}_\gamma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{b}\mathbf{b}' + \hat{\sigma}_\varepsilon^2 \left[\mathbf{b}\left(\frac{1}{n_2}(\Lambda^{T\prime}\Lambda^T)^{-1} + \frac{1}{n_1}(\Lambda^{C\prime}\Lambda^C)^{-1}\right)\mathbf{b}'\right]}} \tag{10}$$

and

$$T_2 = \frac{\mathbf{b}(\bar{Y}^T - \bar{Y}^C)}{\sqrt{\hat{\sigma}_\gamma^2 \left[\mathbf{b}\left(\frac{1}{n_2}\Lambda^T\Lambda^{T\prime} + \frac{1}{n_1}\Lambda^C\Lambda^{C\prime}\right)\mathbf{b}'\right] + \hat{\sigma}_\varepsilon^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{b}\mathbf{b}'}} \tag{11}$$

From the above two equations it becomes clear than choosing $l^{(2)}$ as the contrast vector leads to a very familiar test statistic. The numerator of test statistic $T_2$ considers the difference in average observed values of expression levels and its denominator represents the variance of $\bar{Y}^T - \bar{Y}^C$ based on the mixed linear model.

It is also important to study the effect of the contrast vector on the power of tests. The two null hypotheses presented in (6) and (9) are different and therefore the usual power analysis cannot be applied to choose the right test. However, when $\Lambda^C = \Lambda^T = \Lambda$, the hypothesis presented in (6) is a special case of (9) (assuming that $\Lambda$ has full rank) and it is possible to compare the powers of the two tests in this special case. When $\Lambda^C = \Lambda^T = \Lambda$, the null and alternative hypotheses are given in (6) and the test statistics $T_1$ and $T_2$ have the following simplified forms:

$$T_1 = \frac{\mathbf{b}(\Lambda^{-1}(\bar{Y}^T - \bar{Y}^C))}{\sqrt{\mathbf{b}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)(\hat{\sigma}_\gamma^2 I + \hat{\sigma}_\varepsilon^2(\Lambda'\Lambda)^{-1})\mathbf{b}'}} \tag{12}$$

$$T_2 = \frac{\mathbf{b}(\bar{Y}^T - \bar{Y}^C)}{\sqrt{\mathbf{b}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)(\hat{\sigma}_\gamma^2(\Lambda\Lambda') + \hat{\sigma}_\varepsilon^2 I)\mathbf{b}'}} \tag{13}$$

From these equations we can see that when no underlying network structure is taken into account, ($\Lambda = I$) the two test statistics are the same. However, if there is an underlying network structure ($\Lambda \neq I$), the test statistic in (13) represents the likelihood ratio test for testing the null hypothesis in (6), which is asymptotically most powerful. On the other hand, as $\|\Lambda^T - \Lambda^C\|$ increases, the test presented by $l^{(1)}$ will no longer be appropriate and we could expect $l^{(2)}$ to have a better performance.

In the more general case, where $\Lambda^C \neq \Lambda^T$, it is desirable for the test statistic to account for all of the interactions between genes in the specific subnetwork and to not include any effects from genes outside the

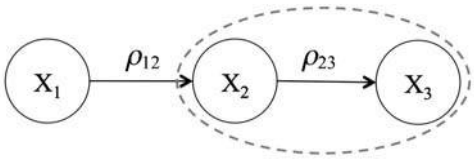


**FIG. 3.** Illustration of the Network contrast vector on a simple network; dashed line indicates the interactions that are included in the contrast vector.

subnetwork. Consider again the simple gene network in Figure 2 and let $\mathbf{b} = (0, 1, 1)$. It is then desirable for the test statistic to include the interaction between genes 2 and 3, while excluding the effect of gene 1 (Fig. 3). The following result describes a choice of a contrast vector that achieves this goal.

**Lemma 3.1.** *Consider a* $1 \times p$ *indicator vector* $\mathbf{b}$ *and let* $x \cdot y$ *represent the element-wise product of* $x$ *and* $y$.
*Then* $(\mathbf{b}\Lambda \cdot \mathbf{b})\gamma$ *includes the effects of all the nodes in* $\mathbf{b}$ *on each other, but it is not affected by any node outside of the set of nodes indexed by* $\mathbf{b}$.

**Proof.** Let $I_{\mathbf{b}} = \{i : \mathbf{b}_i = 1\}$. Based on the latent variable model, the $j$th column of $\Lambda$ includes the influences of node $j$ on all other nodes in the network. Therefore, $(\mathbf{b}\Lambda)_j$ is the influence of the $j$th node on all nodes in $\mathbf{b}$. Also, note that $\Lambda_{ii} = 1$ for all $i$ and $\Lambda_{ji}$ is non-zero only if there is a path from $j$ to $i$. Thus,

$$(\mathbf{b}\Lambda)_j = \begin{cases} \sum_{i \in I_{\mathbf{b}}} \Lambda_{ji} & j \notin I_{\mathbf{b}} \\ 1 + \sum_{i \in I_{\mathbf{b}}, i \neq j} \Lambda_{ji} & j \in I_{\mathbf{b}} \end{cases}$$

But $(\mathbf{b}\Lambda \cdot \mathbf{b})_j$ is non-zero only if $j \in I_{\mathbf{b}}$ and therefore

$$(\mathbf{b}\Lambda \cdot \mathbf{b})\gamma = \sum_{j \in I_{\mathbf{b}}} \gamma_j + \sum_{j \in I_{\mathbf{b}}} \sum_{i \in I_{\mathbf{b}}, i \neq j} \Lambda_{ji}\gamma_j$$

which means that $(\mathbf{b}\Lambda \cdot \mathbf{b})\gamma$ only includes the effects of elements of $\mathbf{b}$ on each other. ■

The estimated $\beta$'s in the latent variable model reflect the individual effect of each gene and therefore, can be thought of as the "pure signals." Based on Lemma 3.1, in order to include interactions among genes in each subnetwork and prevent any confounding effects, we define the *network contrast vector* by

$$l^{(N)} = (-\mathbf{b} \cdot \mathbf{b}\Lambda^C, \mathbf{b} \cdot \mathbf{b}\Lambda^T)$$

### *3.5. Comparison with other gene set analysis techniques*

In this section, we discuss the main differences between the approach proposed in this paper and the idea of gene set enrichment analysis (GSEA) presented in Subramanian et al. (2005) and generalized by Efron and Tibshirani (2007).

Permutation based methods of gene set analysis, including GSEA, first compute an association measure relating the expression levels of each gene in the list to the phenotype (e.g., the $p$-value from the two sample $t$-test). The individual association measures are then combined into an *enrichment score* for each gene set (GSEA uses a version of Kolmogorov-Smirnov test statistic, while a maxmin function is used in GSA). The main strength of the GSEA method, that is also inherited by its extensions, is that the correlation structure of genes in the gene set is preserved, and the permutation based distribution of the enrichment score also represents the correlation among genes. However, these methods compute the individual association measures of each gene separately and do not directly include the correlation among genes when calculating the enrichment score.

Alternatively, if efficient estimation of the covariance matrix is possible, parametric test statistics may be used to test the difference between the expression levels of the two treatment groups. This is not usually possible since in most microarray analysis applications the number of parameters needed to be estimated is considerably larger than the number of samples available ($n \ll p$). However, the external information about the underlying gene network can make this estimation problem tractable. For instance, in the mixed linear model proposed in this paper, the covariance matrix is modeled as a function of few parameters which can be efficiently estimated from the data. Thus, it is possible to test the significance of each gene set using tests that include the expression levels of *all* genes in the gene set and also directly incorporates the covariance structure of the genes in each subnetwork. An example of such a test statistic is the $T_2$ test statistic discussed in Section 3.4, which is a version of the two-sample $t$-test. If the model is correctly specified, one could expect such a test statistic to be sensitive to changes in both the expression levels and also in the covariance structure. However, in the absence of external information about the network, estimation of the covariance matrix may be impractical and non-parametric methods like GSEA, may offer better inference properties.

In the next section, we carry out simulation studies to illustrate the difference between the proposed model and the GSEA method. We will also examine the effect of the choice of the contrast vector on the performance of the proposed test statistic.

## 4. PERFORMANCE ANALYSIS

Three sets of simulation studies are considered in this section. In the first simulation, we study different choices of contrast vectors and compare their performance with GSEA in a simple network. The second simulation study is designed to analyze the combined effect of change in mean and covariance between control and treatment conditions. In the last simulation, we evaluate the sensitivity of the proposed inference procedure to the presence of noise in the association weights. Note that in simulation studies of this section, it is assumed that the effect of the gene network is appropriately modeled using the latent variable model of Section 2 and that the the topology of the network is correctly specified.

### *4.1. Simulation 1: Different choices of contrast vector*

In the first setting, a simple network structure consisting of an eight-level binary tree with 255 nodes is used. It is assumed that there are no interactions in the network under the control condition ($\Lambda^C = I$). Under the treatment condition, genes on the network are assumed to be positively correlated with different association strengths: The association for the first three levels of the genes in the network (top seven genes in the tree) is assumed to be 0.8, genes in the next three levels (56 genes) have association equal to 0.5 and the remainder of the genes are weakly associated with $\rho = 0.2$. Under control, the mean vector for mRNA expression levels of genes is set to zero ($\mu^C = 0$). Scenarios for mean expression levels under treatment are presented in Table 2 and Gene sets considered in this simulation are given in Table 3. The gene sets are chosen so that for each mean scenario there exists gene sets with highly expressed genes and also gene sets that represent non-differentially expressed genes.

Table 4 presents the estimated powers of the GSEA method and tests based on the three contrast vectors, $l^{(1)}$, $l^{(2)}$, and $l^{(N)}$, introduced in Section 3.3 based on 1000 simulations. The powers are calculated based on the FDR controlling procedure of Benjamini and Hochberg (1995) with a $q$-value of 0.05.

TABLE 2. MEAN SCENARIOS UNDER TREATMENT FOR THE FIRST SIMULATION STUDY

| *Scenario* | *Mean parameters* |
|---|---|
| 1 | $\mu^T = \mu^C = 0$ |
| 2 | $\mu^T = 2$ for top one-third levels of the tree, $\mu^T = \mu^C = 0$ for rest |
| 3 | $\mu^T = 2$ for top two-third levels of the tree, $\mu^T = \mu^C = 0$ for rest |
| 4 | $\mu^T = 2$ in the left branch of the tree (including the root node), $\mu^T = \mu^C = 0$ in the right branch |

TABLE 3. GENE SETS CONSIDERED IN THE FIRST SIMULATION STUDY[a]

| *Gene set* | *Genes considered* |
|---|---|
| 1 | All genes in the network |
| 2 | Ttop one-third levels of the tree |
| 3 | First two-third levels of the tree |
| 4 | The last level of the tree |
| 5 | Left branch of the tree (including the root) |
| 6 | Right branch of the tree (excluding the root) |
| 7 | 20% of genes in the network selected randomly |

[a]GSEA method tests the significance of genes in the gene set against other genes not included in the gene set. We have excluded the last gene from gene set 1 to make this comparison possible, but this may not be an appropriate gene set for GSEA.

The positive correlation structure of the network affects the significance of the subnetworks selected for this comparison. When a specific gene in the network becomes differentially expressed, the other genes in the network that are influenced by that gene will also have modified expression levels in the same direction and the combined subnetwork becomes strongly significant. This propagation mechanism explains the abundance of powers of 1 in the table. The first mean scenario in this study corresponds to the case that $\Lambda^C \mu^C = \Lambda^T \mu^T$. All the methods have nominal significance level of 0.05 for this test. On the other hand, there are some differences between the tests based on different contrast vectors and the GSEA method. As one expects from the discussion in Section 3.4, the test based on $l^{(2)}$ has higher power than the test based on $l^{(1)}$. It can also be seen that in all but one case, the power resulted from test based on $l^{(2)}$ is higher than the power for the GSEA method verifying the discussion of Section 3.5. There are few cases that deserve special attention. The GSEA method indicates no power for testing all the genes in the network under scenario 2. However, in this case the top 1/3 levels of the tree are significant and therefore it is natural to expect significant differences in overall expression levels. The same pattern can

TABLE 4. RESULTS OF FIRST SIMULATION STUDY

| *Scenario* | *Method* | *All* | *Top 1/3* | *Top 2/3* | *Last level* | *Left branch* | *Right branch* | *Random* |
|---|---|---|---|---|---|---|---|---|
| 1 | GSEA | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | $l^{(1)}$ | 0.024 | 0.015 | 0.014 | 0.014 | 0.019 | 0.022 | 0.018 |
| | $l^{(2)}$ | 0.023 | 0.020 | 0.015 | 0.012 | 0.011 | 0.023 | 0.019 |
| | $l^{(N)}$ | 0.022 | 0.021 | 0.015 | 0.010 | 0.011 | 0.021 | 0.018 |
| 2 | GSEA | *0.000* | 1.000 | 1.000 | 0.000 | 1.000 | *0.000* | **0.000** |
| | $l^{(1)}$ | *0.119* | 1.000 | *0.535* | 0.047 | *0.127* | *0.056* | **0.046** |
| | $l^{(2)}$ | **1.000** | 1.000 | 1.000 | 0.090 | 0.980 | *0.956* | *0.523* |
| | $l^{(N)}$ | **1.000** | 1.000 | 1.000 | 0.070 | 0.979 | **0.562** | **0.067** |
| 3 | GSEA | 1.000 | *0.000* | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 |
| | $l^{(1)}$ | 1.000 | **1.000** | 1.000 | 0.089 | 1.000 | 1.000 | 0.999 |
| | $l^{(2)}$ | 1.000 | **1.000** | 1.000 | 0.568 | 1.000 | 1.000 | 1.000 |
| | $l^{(N)}$ | 1.000 | **1.000** | 1.000 | 0.089 | 1.000 | 1.000 | 1.000 |
| 4 | GSEA | 1.000 | *0.000* | 1.000 | 1.000 | 1.000 | **0.000** | 1.000 |
| | $l^{(1)}$ | 1.000 | **0.997** | 1.000 | 1.000 | 1.000 | **0.089** | 1.000 |
| | $l^{(2)}$ | 1.000 | **1.000** | 1.000 | 1.000 | 1.000 | *0.476* | 1.000 |
| | $l^{(N)}$ | 1.000 | **1.000** | 1.000 | 1.000 | 1.000 | **0.089** | 1.000 |

Powers of tests based on GSEA and three contrast vectors for different mean scenarios. Multiple testing adjustment is based on FDR with $q^* = 0.05$. Entries in italic indicate powers that are lower or higher than expected, and numbers in bold show powers close to values expected from the simulation design.

be observed when comparing the two methods for testing the right branch of the tree under the second scenario and the top 1/3 of genes under the third scenario. On the other hand, the test based on $l^{(2)}$ has a high false positive rate for testing the right branch of the tree in the situation where only the left branch is up-regulated (scenario 4), while the GSEA method correctly shows no deviation from the null hypothesis. The same phenomenon can be seen for the results of testing the last level of the tree in the case where the top 2/3 levels of the tree are significant. The test based on $l^{(2)}$ is not able to isolate the significance of the genes under consideration from the effect of other genes in the network and can therefore result in high false positive rates. As expected based on Lemma 3.1, the test based on $l^{(N)}$ resolves these shortcomings. The power of this test is close to the nominal significance level for testing the above two cases while it offers a high power in cases where the GSEA method fails to distinguish the significance of the subnetworks.

### *4.2. Simulation 2: Simultaneous changes in mean & covariance*

The second simulation study is designed to evaluate simultaneous changes in expressions levels as well as associations among genes. The network structure in this simulation consists of three root nodes and seven five-level trees (220 genes total). The network consists of low and high association subnetworks and also includes both positive and negative correlations. Three of the subnetworks are considered to be differentially expressed (the level of expression increases in increments of 0.2) and the other subnetworks have equal values of mean in treatment and control conditions. Figure 4 illustrates the setting of parameters of this simulation study.

Table 5 presents the estimates of powers for the GSEA method and the test based on $l^{(N)}$ for testing different trees with increasing expression levels in a simulation with 1000 repetitions. It can be seen from the results that both of these methods reject the null hypothesis for tests related to trees with high positive correlation (subtrees 1, 2, and 7 in Fig. 4). The GSEA method can only detect the significance of subtree 3 for large values of increase in the expression level while the test based on $l^{(N)}$, can detect this change for smaller values of increase. Subtrees 4 and 5 correspond to cases where the correlation among genes is minimal. Subtree 4 is affected by root genes 1 and 2 that are both up regulated but they have opposite correlations with genes in subtree 4. As one would expect, the powers for subtree 4 are similar to those of subtree 5, which suggests that the combined effect of genes 1 and 2 on subtree 4 is the same as the effect of gene 3 on subtree 5. Subtree 6 illustrates the fact that the test based on $l^{(N)}$ takes advantage of the known correlation structure even if the genes in the network are negatively correlated while the GSEA method cannot detect the change in the correlation structure between control and treatment conditions.

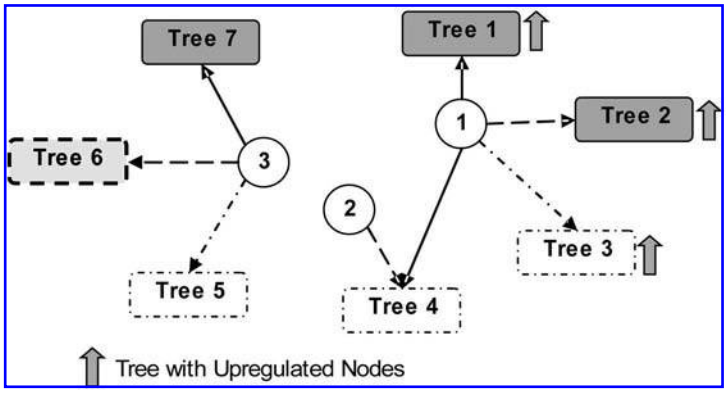


**FIG. 4.** Design of the second simulation study. Solid arrows and boxes represent high positive association (0.6 here), dashed arrows and boxes represent high negative association (−0.6), and dotted arrows and boxes indicate low positive association (0.1). The root genes 1 and 2 are upregulated, while the expression level for gene 3 does not change.

TABLE 5. RESULTS OF THE SECOND SIMULATION STUDY

| *Mean increase* | | *0* | *0.2* | *0.4* | *0.6* | *0.8* | *1.0* |
|---|---|---|---|---|---|---|---|
| Tree 1 | GSEA | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | NetGSA | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Tree 2 | GSEA | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | NetGSA | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Tree 3 | GSEA | 0.000 | *0.000* | *0.000* | *0.000* | *0.000* | 1.000 |
| | NetGSA | 0.2500 | **0.9580** | **1.00** | **1.00** | **1.00** | 1.000 |
| Tree 4 | GSEA | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | NetGSA | 0.263 | 0.277 | 0.298 | 0.278 | 0.298 | 0.295 |
| Tree 5 | GSEA | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | NetGSA | 0.281 | 0.296 | 0.290 | 0.297 | 0.305 | 0.281 |
| Tree 6 | GSEA | *0.000* | *0.000* | *0.000* | *0.000* | *0.000* | *0.000* |
| | NetGSA | **0.982** | **0.984** | **0.986** | **0.980** | **0.978** | **0.976** |
| Tree 7 | GSEA | *0.000* | *0.000* | *0.000* | *0.000* | *0.000* | *0.000* |
| | NetGSA | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |

Estimated powers for the GSEA and the test based on $l^{(N)}$ for different mean scenarios and different subnetworks. In results for each subnetwork, the first row represents the power for the GSEA method, and the second row displays the power for the test based on $l^{(N)}$. Settings of fonts and colors are similar to Table 4.

### *4.3. Simulation 3: Effect of noise in network information*

In the last simulation, we evaluate the sensitivity of the proposed inference procedure to presence of noise in association weights of the gene network. The network consists of four similar subnetworks, each with 40 genes. Under control, genes have mean $\mu^C = 1$ and the weights of the adjacency matrix are set to 0.2. The settings of the parameters under treatment are given in Table 6. The estimated powers of tests of significance of each subnetwork using a test based on $l^{(N)}$ are plotted in Figures 5 and 6. Figure 5 represents the case where the errors are introduced at random, that is, each weight in the adjacency matrix under treatment is perturbed by a uniform noise in the range $[-e, e]$ where $e$ is a value between 0 and 0.4. On the other hand, Figure 6 represents the estimated powers of tests when a systematic bias is included in the weights of the adjacency matrix under treatment. It can be seen that if the underlying model is correctly specified, presence of random noise in weights of adjacency matrix will not significantly affect the power of the test. However, presence of systematic bias in the estimated weights can introduce both type I, as well as type II errors. This is illustrated by the increase of power of the test as the difference between weights under treatment and control becomes more significant (Fig. 6). It is important to note that the simulation considered here does not include errors in the topology of the network. These errors become more critical if the topology of the network, as well as the association weights, are estimated from expression data, which is beyond the scope of this article.

TABLE 6. SIGNIFICANT PARAMETERS OF THE THIRD SIMULATION STUDY UNDER TREATMENT CONDITION

| *Subnetwork* | *Mean* | *Association weight* |
|---|---|---|
| 1 | $\mu^T = 1.5$ | $\rho^T = 0.6$ |
| 2 | — | $\rho^T = 0.6$ |
| 3 | — | — |
| 4 | $\mu^T = 1.5$ | — |

In all other cases, $\mu^T = \mu^C = 1$ and $\rho^T = \rho^C = 0.2$.

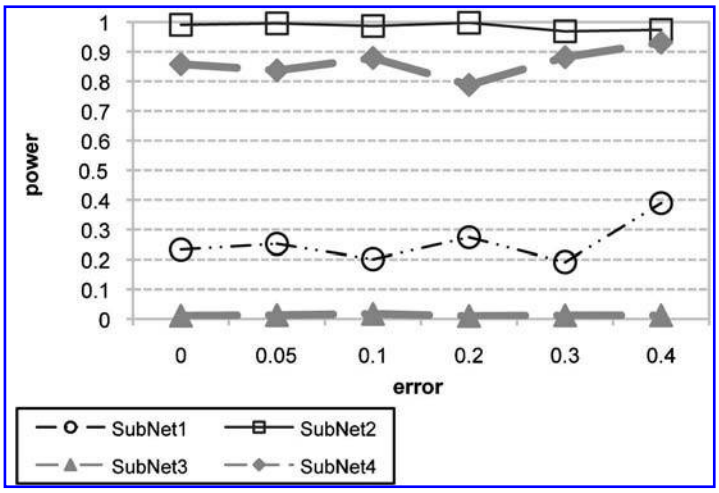


**FIG. 5.** Estimated powers of test of significance of subnetworks of Table 6 with *random noise* in weights of the adjacency matrix. Plots in gray represent the powers of subnetworks whose true adjacency matrices in control and treatment are the same.

## 5. ANALYSIS OF YEAST GALACTOSE UTILIZATION PATHWAY DATA

In Section 1.1, we analyzed the yeast GAL pathway data (Ideker data) using the GSEA method, which revealed that the Galactose Utilization pathway is significantly activated in gal+ condition. In that analysis, the external information provided by the network was only used to determine the gene sets of interest. As discussed in Section 1.1, the Ideker data also includes strength of gene interactions in the network. Therefore, it is possible to directly incorporate the network information and use the proposed network-based inference procedure. It is important to note that the Ideker data only includes one set of association weights for both gal+ and gal− conditions. In other words, in this section we assume $\Lambda^T = \Lambda^C = \Lambda$, and hence the proposed inference procedure cannot test the change in the network structure. Assuming that the latent variable model correctly represents the effect of the underlying network, the increased power of the network based procedure is mainly due to directly incorporating the network information.

Table 7 compares results of analyzing the Ideker data using the GSEA method and the network based method presented in this paper (using $l^{(N)}$). This table also includes results of analyzing this

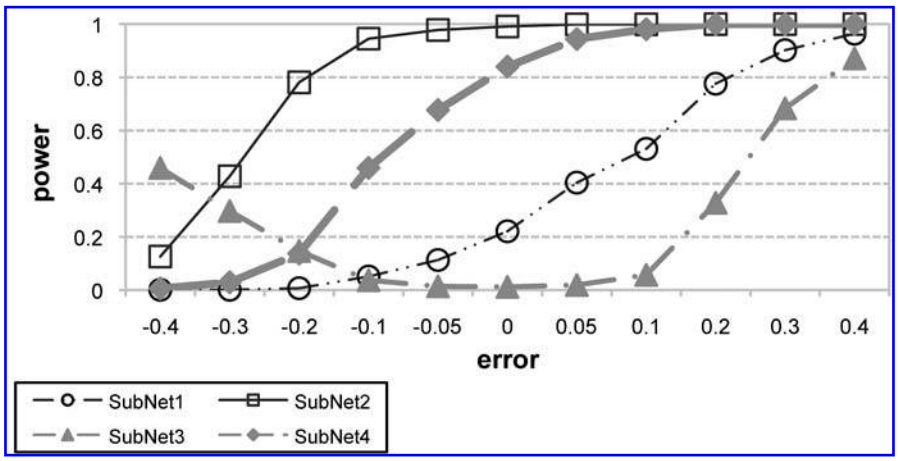


**FIG. 6.** Estimated powers of test of significance of subnetworks of Table 6 with *systematic bias* in weights of the adjacency matrix. Plots in gray represent the powers of subnetworks whose true adjacency matrices in control and treatment are the same.

TABLE 7. ANALYSIS OF IDEKER 2001 DATA USING GSEA, GSA, AND THE PROPOSED METHOD BASED ON THE UNDERLYING GENE NETWORK USING THE $l^{(N)}$ CONTRAST VECTOR

| | | *GSEA* | | *GSA* | | *NetGSA* | |
|---|---|---|---|---|---|---|---|
| *Pathway* | *Size* | *NOM p-val* | *FDR signif* | *NOM p-val* | *FDR signif* | *NOM p-val* | *FDR signif* |
| rProtein synthesis | 28 | 0.5261 | | 0.278 | | 0.0038 | ✓ |
| Glycolytic enzymes | 16 | 0.9683 | | 0.357 | | 0.2825 | |
| RNA processing | 75 | 0.9879 | | 0.386 | | 0.479 | |
| Fatty acid oxidation | 7 | 0.4694 | | 0.299 | | 0.0068 | ✓ |
| $O_2$ stress | 13 | 0.2384 | | 0.285 | | 0.4448 | |
| Mating, cell cycle | 58 | 0.3583 | | 0.417 | | 0.4317 | |
| Vesicular transport | 19 | 0.07243 | | 0.156 | | 0.3693 | |
| Sugar transport | 2 | 0.7358 | | 0.458 | | 0.3319 | |
| Glycogen metabolism | 12 | 0.1321 | | 0.034 | | 0.3057 | |
| Stress | 12 | 0.02004 | | 0.007 | | 0.0000 | ✓ |
| Metal uptake | 4 | 0.8374 | | 0.326 | | 0.0802 | |
| Respiration | 9 | 0.1878 | | 0.091 | | 0.0001 | ✓ |
| Gluconeogenesis | 7 | 0.8455 | | 0.475 | | 0.0383 | |
| Galactose utilization | 12 | 0.002045 | ✓ | 0.001 | ✓ | 0.0000 | ✓ |
| Amino acid synthesis | 30 | 0.1853 | | 0.054 | | 0.0665 | |

For each method, the nominal $p$-value and whether the pathway is significant based on the FDR with $q^* = 0.05$ is reported.

data using the GSA method of Efron and Tibshirani (2007).[2] As one may expect, all three methods find the Galactose Utilization pathway to be statistically significant. Although the GSEA and the GSA methods agree on the significance of other subnetworks, it can be seen from Table 7 and Figure 7, that including the underlying network structure in the analysis, reveals four additional significant pathways. Although additional experiments are needed to verify the result of Table 7, the biology of yeast cells may offer some insight to significance of newly detected pathways. These pathways can be categorized into two groups: Galactose Utilization and rProtein Synthesis pathways are involved in cell growth in gal+ environment, while genes in the Stress, Respiration and Fatty Acid Oxidation pathways are induced in gal− environment. The Stress pathway has a low nominal $p$-value in both GSEA and GSA results; however, these methods do not consider this pathway significant. The significance of the Stress pathway is not surprising and can be explained by the fact that galactose is a more efficient source of carbon than raffinose. Thus, in absence of galactose (gal−), the genes in the Environmental Stress Response (ESR) are induced (Gasch et al., 2000; Gasch and Werner-Washburne, 2002). The Fatty Acid Oxidation and Respiration pathways are also upregulated in gal− environment. The genes in the Respiration pathway are among the genes that are induced in the ESR.

Many of the stress defense mechanisms consume ATP, and therefore, cellular stress could lead to the induced expression of respiration genes (Hohmann and Mager, 2003). Also, many genes involved in importing and exporting fatty acids are induced in ESR and the induction of these genes can increase the local concentration of fatty acids, which in turn may induce the expression of genes in Fatty Acid Oxidation pathway (Hohmann and Mager, 2003). The induction of Fatty Acid Oxidation and Respiration genes can be further explained by the coregulation of genes in these pathways. It should be noted that two of the genes in the Respiration pathway are directly affected by genes in GAL pathway (GAL4 regulates CYC1 and HAP4 is regulated by MIG1), and our proposed model can exploit such relationship in order to gain more statistical power. Finally, the significance of the rProtein Synthesis genes can be explained by growth dependent expression of these genes and the fact that ESR represses the expression of many protein synthesis genes (Hohmann and Mager, 2003).

[2]The minmax criteria is used as the enrichment function in the GSA method.

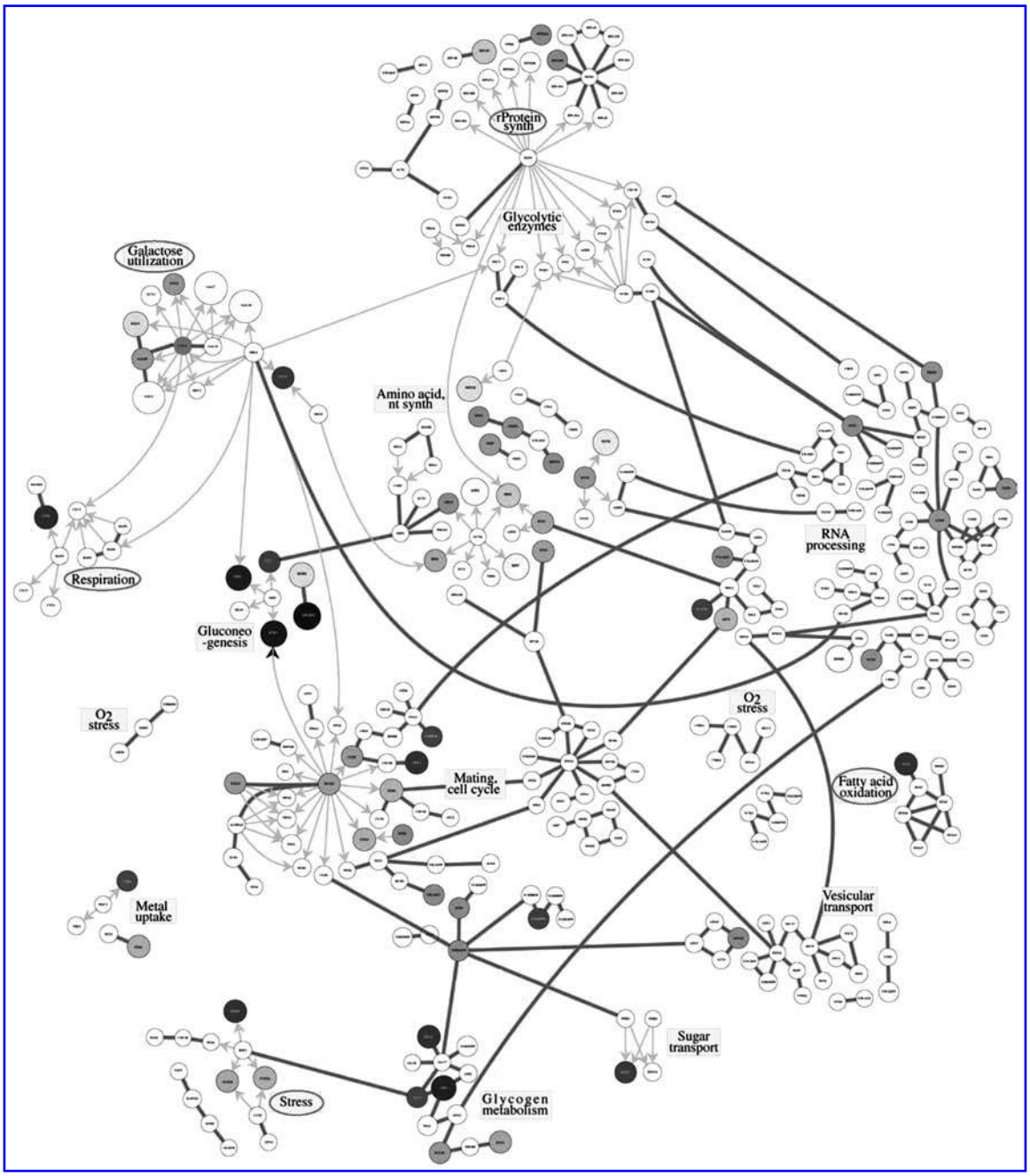


**FIG. 7.** Yeast gene network indicating the significant pathways; significant pathways have been marked with ovals.

## 6. DISCUSSION

Finding significant subnetworks and pathways that are involved in certain biological phenomena has been the focus of many new studies. The main challenge is to formulate the null and alternative hypotheses that consider the change in the expression levels of the genes as well as the change in the network structure in response to environmental factors. In this paper, we proposed a model-based approach for testing the significance of biological pathways using the underlying gene network and studied graph theoretic properties of the model. Our approach uses external information available about the underlying network and it hence depends on availability and quality of such data. The method proposed in this paper,

incorporates the weighted adjacency matrix of the network through a latent variable model and uses a flexible mixed linear representation. We discussed that the inference based on this method depends on the choice of the contrast vector and proposed a choice that offers improvement in power of the test compared to the GSEA method of Subramanian et al. (2005). The simulation studies and the analysis of the yeast galactose utilization pathway reveal the ability of the proposed method in identifying significant pathways that are otherwise difficult to distinguish. Although the focus of this paper was on testing the significance of subnetworks in the two population inference problem, the proposed method provides a general framework for studying a variety of phenotypes including analysis of time series mRNA data and the change in the network over time. More generally, different correlation structures among observations can be implemented in the mixed linear model and therefore, different types of data can be modeled using this framework. Considering parameters for environment factors and gene-gene and gene-environment interactions is also a straight forward extension of the proposed model.

The model presented in this paper relies on two main assumptions: (a) The relationship between the expression levels of genes in the network can be represented linearly using the influence matrix of the network and (b) that the data follows a normal distribution. Although the first assumption is a crucial part of this analysis, the second assumption can be relaxed using the Generalized Mixed Linear Model (GMLM) framework. However, this would make the computational aspects of the problem more challenging.

The growth of information available on the underlying biological networks calls for effective methods that can utilize such information efficiently and requires extensions of statistical methods appropriate for studying of network structures. The model presented in this paper requires external information on the weighted adjacency matrix of the network. Although more data is becoming available on gene and protein networks, many available network data only include the binary association among genes (network topology) and do not include information about the strength or direction of associations among genes. The problem of estimating the weighted adjacency matrix of the network, which is related to estimation of the covariance matrix, is of separate interest and is beyond the scope of this paper. Chaudhuri et al. (2007) propose an efficient algorithm for estimating the association among genes when the topology of the network is known. The method proposed in this paper can also be extended to the cases where only partial information about the network is available.

## ACKNOWLEDGMENTS

We would like to thank the CoEditor-in-Chief, Professor Sorin Istrail, and two anonymous referees for helpful comments and suggestions. We are also thankful to Professor Trey Ideker for providing the yeast Galactose Utilization data and helpful discussions. The work of George Michailidis was partially supported by the NIH (grant 5P 41RR018627) and the MEDC (grant GR-687).

## DISCLOSURE STATEMENT

No competing financial interests exist.

## REFERENCES

Alexa, A., Rahnenfuhrer, J., and Lengauer, T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22, 1600–1607.

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

Bader, J.S., Chaudhuri, A., Rothberg, J.M., et al. 2004. Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* 22, 78–85.

Barry, W.T., Nobel, A.B., and Wright, F.A. 2005. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 21, 1943–1949.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B* 57, 289–300.

Chaudhuri, S., Drton, M., and Richardson, T. 2007. Estimation of a covariance matrix with zeros. *Biometrika* 94, 199–216.

Diestel, R. 2006. *Graph Theory*. Springer-Verlag, New York.

Efron, B., and Tibshirani, R. 2007. On testing the significance of sets of genes. *Ann. Appl. Statist.* 1, 107–129.

Friedberg, S.H., Insel, A.J., and Spence, L.E. 1996. *Linear Algebra*. Prentice Hall, Englewood Cliffs, NJ.

Gasch, A.P., Spellman, P.T., Kao, C.M., et al. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11, 4241–4257.

Gasch, A.P., and Werner-Washburne, M. 2002. The genomics of yeast responses to environmental stress and starvation. *Funct. Integr. Genomics* 2, 181–192.

Goeman, J.J., and Bühlmann, P. 2007. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23, 980–987.

Hohmann, S., and Mager, W. 2003. *Yeast Stress Responses*. Springer, New York.

Ideker, T., Thorsson, V., Ranish, J., et al. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292.

Jiang, Z., and Gentleman, R. 2007. Extensions to gene set enrichment. *Bioinformatics* 23, 306–313.

Kanehisa, M., and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30.

Li, K.-C. 2002. Genome-wide coexpression dynamics: theory and application. *Proc. Natl. Acad. Sci. USA* 99, 16875–16880.

McLean, R.A., and Sanders, W.L. 1988. Approximating degrees of freedom for standard errors in mixed linear models. *Proc. Statist. Comput. Sect. Am. Statist. Assoc.*, pgs. 50–59.

Rahnenführer, J., Domingues, F.S., Maydt, J., et al. 2004. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statist. Appl. Genet. Mol. Biol.* 3, 16.

Searle, S.R. 1971. *Linear Models*. John Wiley & Sons, Inc., New York.

Subramanian, A., Tamayo, P., Mootha, V., et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102, 15545–15550.

Tian, L., Greenberg, S.A., Kong, S.W., et al. 2005. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. USA* 102, 13544–13549.

Wei, P., and Pan, W. 2008. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics* 24, 404–411.

Wei, Z., and Li, H. 2007. A Markov random field model for network-based analysis of genomic data. *Bioinformatics* 23, 1537–1544.

West, M. 2000. Bayesian regression analysis in the large $p$ small $n$ paradigm. Technical report. Institute of Statistics and Decision Sciences.

Address reprint requests to:
*Ali Shojaie*
*Department of Statistics*
*University of Michigan*
*269 West Hall*
*1085 South University Avenue*
*Ann Arbor, MI 48109*

*E-mail:* shojaie@umich.edu