*University of Pennsylvania*
UPenn Biostatistics Working Papers

*Year* 2008 *Paper* 24

# Incorporation of Genetic Pathway Information into Analysis of Multivariate Gene Expression Data

Zhi Wei[*] Jane E. Minturn[†] Eric Rappaport[‡]
Garrett Brodeur[**] Hongzhe Li[††]

[*] zhiwei@mail.med.upenn.edu
[†] Children's Hospital of Philadelphia, minturn@email.chop.edu
[‡] Children's Hospital of Philadelphia, rappaport@email.chop.edu
[**] Children's Hospital of Philadelphia, brodeur@email.chop.edu
[††] University of Pennsylvania, hongzhe@mail.med.upenn.edu

http://biostats.bepress.com/upennbiostat/art24

# Incorporation of Genetic Pathway Information into Analysis of Multivariate Gene Expression Data

Zhi Wei, Jane E. Minturn, Eric Rappaport, Garrett Brodeur, and Hongzhe Li

## Abstract

Abstract: Multivariate microarray gene expression data are commonly collected to study the genomic responses under ordered conditions such as over increasing/decreasing dose levels or over time during biological processes. One important question from such multivariate gene expression experiments is to identify genes that show different expression patterns over treatment dosages or over time and pathways that are perturbed during a given biological process. In this paper, we develop a hidden Markov random field model for multivariate expression data in order to identify genes and subnetworks that are related to biological processes, where the dependency of the differential expression patterns of genes on the networks are modeled by a Markov random field. Simulation studies indicated that the method is quite effective in identifying genes and the modified subnetworks and has higher sensitivity than the commonly used procedures that do not use the pathway information, with similar observed false discovery rates. We applied the proposed methods for analysis of a microarray time course gene expression study of TrkA- and TrkB-transfected neuroblastoma cell lines and identified genes and subnetworks on MAPK, focal adhesion and prion disease pathways that may explain cell differentiation in TrkA-transfected cell lines.

# Incorporation of Genetic Pathway Information into Analysis of Multivariate Gene Expression Data

Wei Zhi, Jane Minturn, Eric Rappaport, Garrett Brodeur, and Hongzhe Li
Genomics and Computational Biology Graduate Program,
University of Pennsylvania School of Medicine, PA 19104, USA.
Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

April 10, 2008

**Abstract**

Multivariate microarray gene expression data are commonly collected to study the genomic responses under ordered conditions such as over increasing/decreasing dose levels or over time during biological processes. One important question from such multivariate gene expression experiments is to identify genes that show different expression patterns over treatment dosages or over time and pathways that are perturbed during a given biological process. In this paper, we develop a hidden Markov random field model for multivariate expression data in order to identify genes and subnetworks that are related to biological processes, where the dependency of the differential expression patterns of genes on the networks are modeled by a Markov random field. Simulation studies indicated that the method is quite effective in identifying genes and the modified subnetworks and has higher sensitivity than the commonly used procedures that do not use the pathway information, with similar observed false discovery rates. We applied the proposed methods for analysis of a microarray time course gene expression study of TrkA- and TrkB-transfected neuroblastoma cell lines and identified genes and subnetworks on MAPK, focal adhesion and prion disease pathways that may explain cell differentiation in TrkA-transfected cell lines.

# 1 Introduction

Multivariate microarray gene expression data are commonly collected to investigate dose-dependent alterations in gene expression or time-dependent gene expression during a biological process. For example, dose-dependent gene expression data are often measured in the area of toxicology (Lehmann *et al.*, 2004; Seidel *et al.*, 2006) and time-course gene expression data are often collected during a dynamic biological process. For both the dose-dependent and time-course gene expression experiments, the data can be summarized as multivariate vectors, and one goal of such multivariate gene expression studies is to identify genes that have different overall expression patterns between two experiments and the pathways or subnetworks that are perturbed or activated during a given dose-dependent experiment or a dynamic biological process. We call these genes the multivariate differentially expressed (MDE) genes. Compared to gene expression studies of one single experimental condition, such multivariate gene expression data can potentially identify more genes that are differentially expressed (Yuan and Kendzioski, 2006; Tai and Speed, 2006; Hong and Li, 2006).

One important feature of the multivariate gene expression data is that the data are expected to be dependent across dosages or time points. Efficiently utilizing such dependency can lead to a gain in efficiency in identifying the MDE genes. Yuan and Kendzioski (2006) and Wei and Li (2008) developed the hidden Markov model and hidden Markov random field model to identify the differentially expressed genes at each time point for analysis of microarray time-course gene expression data. Instead of identifying genes that are differentially expressed at each time point during a biological process or at a given dosage level, the investigators sometimes are only interested in identifying the genes that show different expression patterns over all the experimental points. Tai and Speed (2006) developed an empirical Bayes method treating the observed time-course gene expression data as multivariate vectors. Hong and Li (2006) developed a functional empirical Bayes method using B-splines. Both approaches treat the data as multivariate vectors to account for possible correlations of gene expressions over different dosages or time points. Although the existing methods can be used to identify the MDE genes, they often do not provide direct information on which key molecular mechanisms are involved in the biological process or which biological pathways are being activated or modified during a given biological process. It is therefore important to develop novel statistical methods for identifying these MDE genes in the context of known biological pathways.

Information about gene regulatory dependence has been accumulated from many years of biomedical experiments and is summarized in the form of pathways and assembled into pathway databases. Some well-known pathway databases include KEGG, Reactome (www.reactome.org), BioCarta (www.biocarta.com) and BioCyc (www.biocyc.org). Several methods have recently been developed to incorporate the pathway structures into analysis of microarray gene expression data. Rahnenführer *et al.* (2004) demonstrated that the sensitivity of detecting relevant pathways can be improved by integrating information about pathway topology. In Sivachenko *et al.* (2005), a network topology extracted from the literature was used jointly with microarray data to find significantly affected pathway regulators. Nacu *et al.* (2006) proposed an interesting permutation-based test for identifying subnetworks from a known network of genes that are related to phenotypes. Wei and Li (2007) have recently developed a hidden Markov random field (HMRF) model for identifying the subnetworks that show differential expression patterns between two conditions, and have demonstrated using both simulations and applications to real data sets that the procedure is more sensitive in identifying the differentially expressed genes than those procedures that do not utilize the pathway structure information. However, none of these explicitly models the multivariate expression data. Wei and Li (2008) further extended the HMRF model of Wei and Li

(2007) and the HMM model of Yuan and Kendzioski (2006) to analyze the microarray time course gene expression in the framework of a hidden spatial-temporal MRF model. However, this approach assumes the same network-dependency of the gene differential expression states at all the time points.

In this paper, to efficiently identify the MDE genes in the multivariate gene expression experiments, we develop the HMRF model of Wei and Li (2007) further into a hidden MRF model for multivariate gene expression data in order to take into account the known biological pathway information. We treat the multivariate gene expression data as multivariate data, allowing for dependency of the data across the dosage levels or over time points. The key to our approach is that the information of a known network of pathways is efficiently utilized in the analysis of multivariate expression data in order to identify more biologically interpretable results. We introduce the HMRF models for both longitudinal and cross-sectional designs and present an iterative conditional modes (ICM) algorithm (Besag, 1986) for estimating the model parameters and for calculating the posterior probabilities of being an MDE gene.

We introduce the hidden MRF model for multivariate expression data in Section 2 for both longitudinal and cross-sectional experiments and present an efficient algorithm for parameter estimation by the ICM algorithm in Section 3. We present results from simulation studies in Section 4 to demonstrate the application of the hidden MRF model, to compare with existing methods, and to evaluate the sensitivity of the method to misspecification of the network structure. In Section 5, for a case study, we apply the hidden MRF model to analyze the time-course gene expression data of TrkA- and TrkB-transfected neuroblastoma cell lines in order to identify the pathways that are related to cell differentiation in TrkA-transfected cell lines. Finally, we present a brief discussion in Section 6

# 2 A Hidden MRF Model for Multivariate Gene Expression Data

We first introduce the HMRF models for multivariate gene expression data for both the longitudinal design and cross-sectional design, where the network structure is represented as an undirected graph. The models are an extension of the HMRF model of Wei and Li (2007) for multivariate gene expression data, where the distribution of latent MDE states of the genes is modeled as a discrete MRF based on the network structure, and the empirical Bayes models of Tai and Speed (2006) are used for modeling the emission density for the observed multivariate gene expression data.

## 2.1 Data observed and representation of genetic networks as undirected graphs

Consider the multivariate gene expression data measured under two different conditions over $k$ dosage levels or time points, with $n$ independent samples measured under one condition and $m$ independent samples measured under another condition. For each experiment, we assume that the expression levels of $p$ genes are measured. For a given $g$, we denote these data as *i.i.d.* $k \times 1$ random vectors $\mathbf{Y}_{g1}, \cdots, \mathbf{Y}_{gn}$ for condition 1 and $\mathbf{Z}_{g1}, \cdots, \mathbf{Z}_{gm}$ for condition 2. We further assume that $\mathbf{Y}_{gi} \sim N_k(\mu_{\mathbf{gy}}, \mathbf{\Sigma_g})$ and $\mathbf{Z}_{gi} \sim N_k(\mu_{\mathbf{gz}}, \mathbf{\Sigma_g})$. For a given gene $g$, the null hypothesis of interest is

$$H_{g0} : \mu_{\mathbf{gy}} = \mu_{\mathbf{gz}}. \tag{1}$$

Define $\mu_{\mathbf{g}} = \mu_{\mathbf{gy}} - \mu_{\mathbf{gz}}$. For a given gene $g$, let $I_g$ take the value of 1 if $\mu_{\mathbf{g}} \neq 0$ and 0 if $\mu_{\mathbf{g}} = 0$. We call the genes with $I_g = 1$ the MDE genes. Our goal is to identify these MDE genes among the $p$ genes.

Besides the gene expression data, suppose that we have a network of known pathways that can be represented as an undirected graph $G = (V, E)$, where $V$ is the set of nodes that represent genes or proteins coded by genes and $E$ is the set of edges linking two genes with a regulatory relationship. Let $p = |V|$ be the number of genes that this network contains. Note the gene set $V$ is often a subset of all the genes that are probed on the gene expression arrays. If we want to include all the genes that are probed on the expression arrays, we can expand the network graph $G$ to include isolated nodes, which are those genes that are probed on the arrays but are not part of the known biological network. For two genes $g$ and $g'$, if there is a known regulatory relationship, we write $g \sim g'$. For a given gene $g$, let $N_g = \{g' : g \sim g' \in E\}$ be the set of genes that have a regulatory relationship with gene $g$ and $d_g = |N_g|$ be the degree for gene $g$.

### 2.2 A discrete Markov random field model for differential expression states for genes on the network

Our goal is to identify the genes on the network $G$ that are multivariate differentially expressed between the two experimental conditions. Since two neighboring genes $g$ and $g'$ have regulatory relationship on the network, we should expect that the MDE states $I_g$ and $I_{g'}$ are dependent. In order to model the dependency of $I_g$ over the network, following Wei and Li (2007), we introduce a simple MRF model. Particularly, we assume the following auto-logistic model for the conditional distribution of $I_g$,

$$Pr(I_g|I_{g'}, g' \neq g) \quad = \quad \frac{\exp\{I_g F(I_g)\}}{1+\exp\{F(I_g)\}}, \tag{2}$$

where

$$F(I_g) = \gamma + \beta \frac{\sum_{g' \in N_g}(2I_{g'}-1)}{d_g},$$

and $\gamma$ and $\beta \geq 0$ are arbitrary real numbers. Here the parameter $\beta$ measures the dependency of the differential expression states of the neighboring genes. We assume that the true MDE states $(I_g^*) = \{I_g^*, g = 1, \cdots, p\}$ is a particular realization of this locally dependent MRF.

### 2.3 Emission probabilities for multivariate gene expression data from longitudinal designs

To relate the differential expression state $I_g$ to the observed gene expression data $\mathbf{D}_g = (\mathbf{Y}_{g1}, \cdots, \mathbf{Y}_{gn}; \mathbf{Z}_{g1}, \cdots, \mathbf{Z}_{gm})$, we follow the empirical Bayes approach of Tai and Speed (2006) for multivariate gene expression data and use conjugate priors for $\mu_g$ and $\mathbf{\Sigma}_g$, that is, an inverse Wishart prior for $\mathbf{\Sigma}_g$ and a dependent multivariate normal prior for $\mu_g$. To make notation simple, we drop the gene subscript $g$ when introducing the Bayesian model. Let

$$\begin{aligned}
\bar{\mathbf{Y}} &= (\mathbf{Y}_1 + \cdots + \mathbf{Y}_n)/n, \\
\bar{\mathbf{Z}} &= (\mathbf{Z}_1 + \cdots + \mathbf{Z}_m)/m, \\
\bar{\mathbf{X}} &= \bar{\mathbf{Y}} - \bar{\mathbf{Z}}, \\
\mathbf{S_y} &= (n-1)^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})', \\
\mathbf{S_z} &= (m-1)^{-1}\sum_{i=1}^{m}(\mathbf{Z}_i - \bar{\mathbf{Z}})(\mathbf{Z}_i - \bar{\mathbf{Z}})', \\
\mathbf{S} &= (n+m-2)^{-1}((n-1)\mathbf{S_y} + (m-1)\mathbf{S_z}).
\end{aligned}$$

Following Tai and Speed (2006), we assign independent and identical inverse Wishart priors to $\mathbf{\Sigma}$, $\mathbf{\Sigma} \sim W^{-1}((\nu\mathbf{\Lambda})^{-1}, \nu)$. Given $\mathbf{\Sigma}$, we assign multivariate normal priors for the gene-specific mean difference $\mu$ for the two cases $(I = 1)$ and $(I = 0)$:

$$\begin{aligned}
\mu|\mathbf{\Sigma}, I = 1 &\sim N_k(\mathbf{0}, \eta^{-1}\mathbf{\Sigma}), \\
\mu|\mathbf{\Sigma}, I = 0 &\equiv \mathbf{0}.
\end{aligned}$$

Since the statistics $(\bar{\mathbf{X}}, \mathbf{S})$ are the sufficient statistics for the parameters $(\mu, \mathbf{\Sigma})$ (Tai and Speed, 2006), the conditional distribution of the data $\mathbf{D} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_n, \mathbf{Z}_1, \cdots, \mathbf{Z}_m)$ can be written as

$$P(\mathbf{D}|I) = P(\mathbf{Y}_1, \cdots, \mathbf{Y}_n, \mathbf{Z}_1, \cdots, \mathbf{Z}_m|I) \quad = \quad P(\bar{\mathbf{X}}, \mathbf{S}|I).$$

Tai and Speed (2006) further derived

$$\begin{aligned}
P(\mathbf{D}|I=1) \quad = \quad & \frac{\Gamma_k((N+\nu)/2)}{\Gamma_k((N-1)/2)\Gamma_k(\nu/2)} \\
& \times (N-1)^{\frac{k(N-1)}{2}} \nu^{-\frac{kN}{2}} (\pi(n^{-1}+m^{-1}+\eta^{-1}))^{-\frac{k}{2}} \\
& \times \frac{|\mathbf{\Lambda}|^{-\frac{N}{2}} |\mathbf{S}|^{\frac{N-k-2}{2}}}{|\mathbf{I}_k + ((n^{-1}+m^{-1}+\eta^{-1})\nu\mathbf{\Lambda})^{-1}\bar{\mathbf{X}}\bar{\mathbf{X}}' + (\nu\mathbf{\Lambda}/(N-1))^{-1}\mathbf{S}|^{\frac{N+\nu}{2}}},
\end{aligned} \tag{3}$$

where $N = n+m-1$. Thus, given $I = 1$, the probability density function of the data is a function of $\bar{\mathbf{X}}$ and $\bar{\mathbf{S}}$ only, which follows a Student-Siegel distribution (Aitchison and Dunsmore, 1975). Following Aitchison and Dunsmore's and Tai and Speed's notation, this distribution is denoted by $StSi_k(\nu, \mathbf{0}, (n^{-1}+m^{-1}+\eta^{-1})\mathbf{\Lambda}, N-1, (N-1)^{-1}\nu\mathbf{\Lambda})$. Similarly, the distribution of $P(\mathbf{D}|I = 0)$ follows $StSi_k(\nu, \mathbf{0}, (n^{-1}+m^{-1})\mathbf{\Lambda}, N-1, (N-1)^{-1}\nu\mathbf{\Lambda})$, with the following density function

$$\begin{aligned} P(\mathbf{D}|I = 0) \quad &= \quad \frac{\Gamma_k((N+\nu)/2)}{\Gamma_k((N-1)/2)\Gamma_k(\nu/2)} \\ &\quad \times (N-1)^{\frac{k(N-1)}{2}} \nu^{-\frac{kN}{2}} (\pi(n^{-1}+m^{-1}))^{-\frac{k}{2}} \\ &\quad \times \frac{|\mathbf{\Lambda}|^{-\frac{N}{2}} |\mathbf{S}|^{\frac{N-k-2}{2}}}{|\mathbf{I}_k + ((n^{-1}+m^{-1})\nu\mathbf{\Lambda})^{-1}\bar{\mathbf{X}}\bar{\mathbf{X}}' + (\nu\mathbf{\Lambda}/(N-1))^{-1}\mathbf{S}|^{\frac{N+\nu}{2}}}. \end{aligned} \tag{4}$$

Equations (3) and (4) are used to defined the emission probabilities for the longitudinal design in the hidden MRF formulation presented in Section 2.5.

## 2.4 Emission probabilities for multivariate gene expression data from cross-sectional designs

The longitudinal model in Section 2.3 treats the entire data across multiple conditionals as a vector and makes use of the multivariate normality assumption. We consider in this Section the cross-sectional experiments, where there are no true biological correlations among the gene expression values across multiple conditions (Tai and Speed, 2007). However, our goal is still to identify the genes with different expression profiles over multiple experimental conditions or time points while the null hypothesis for gene $g$ remains the same as (1).

Instead of a general covariance matrix $\mathbf{\Sigma}$, we assume $\mathbf{\Sigma} = \sigma^2 I$. We then assume an inverse-gamma prior for $\sigma^2$, denoted by

$$\sigma^2 \sim Inv - gamma\left(\frac{\nu}{2}, \frac{\nu\lambda^2}{2}\right).$$

Given $\sigma^{\mathbf{2}}$, we assign multivariate normal priors for the gene-specific mean difference $\mu$ for the two cases $(I = 1)$ and $(I = 0)$:

$$\begin{aligned} \mu|\sigma, I = 1 \quad &\sim \quad N_k(\mathbf{0}, \eta^{-1}\sigma^{\mathbf{2}} I), \\ \mu|\sigma, I = 0 \quad &\equiv \quad \mathbf{0}. \end{aligned}$$

Define $s_j^2$ to be the $j$th diagonal element of $\mathbf{S}$. Under this hierarchical model, $P(\mathbf{D}|I = 0)$ and $P(\mathbf{D}|I = 1)$ are simply the special case of Equations (3) and (4) and can be written as

$$\begin{aligned} P(\mathbf{D}|I = 1) \quad &= \quad P(\overline{\mathbf{X}}, \mathbf{S}|I = 1) = \prod_{j=1}^{k} P(\overline{\mathbf{X}}_j, s_j^2|I = 1) \\ &= \quad \left(\frac{\Gamma((N+\nu)/2)}{\Gamma((N-1)/2)\Gamma(\nu/2)}\right)^k \\ &\quad \times (N-1)^{\frac{k(N-1)}{2}} \nu^{-\frac{kN}{2}} (\pi(n^{-1}+m^{-1}+\eta^{-1}))^{-\frac{k}{2}} (\lambda^2)^{-\frac{kN}{2}} \\ &\quad \times \prod_{j=1}^{k} \frac{(s_j^2)^{\frac{N-3}{2}}}{\left(1 + ((n^{-1}+m^{-1}+\eta^{-1})\nu\lambda^2)^{-1}\bar{\mathbf{X}}_j^2 + (\nu\lambda^2/(N-1))^{-1}s_j^2\right)^{\frac{N+\nu}{2}}}, \end{aligned} \tag{5}$$

and

$$\begin{aligned} P(\mathbf{D}|I = 0) \quad &= \quad \left(\frac{\Gamma((N+\nu)/2)}{\Gamma((N-1)/2)\Gamma(\nu/2)}\right)^k \\ &\quad \times (N-1)^{\frac{k(N-1)}{2}} \nu^{-\frac{kN}{2}} (\pi(n^{-1}+m^{-1}))^{-\frac{k}{2}} (\lambda^2)^{-\frac{kN}{2}} \\ &\quad \times \prod_{j=1}^{k} \frac{(s_j^2)^{\frac{N-3}{2}}}{\left(1 + ((n^{-1}+m^{-1})\nu\lambda^2)^{-1}\bar{\mathbf{X}}_j^2 + (\nu\lambda^2/(N-1))^{-1}s_j^2\right)^{\frac{N+\nu}{2}}}, \end{aligned} \tag{6}$$

respectively (Tai and Speed, 2007).

Equations (5) and (6) are used to defined the emission probabilities for the cross-sectional designs in the hidden MRF formulation presented in Section 2.5.

### 2.5 Hidden MRF models for multivariate gene expression data

Together the transition probability (2) and the emission probabilities (3) and (4) define the hidden MRF model for multivariate gene expression data observed from longitudinal designs (HMRF-L), with parameters in the emission probabilities $\theta = (\eta, \nu, \mathbf{\Lambda})$. Similarly, the transition probability (2) and the emission probabilities (3) and (4) define the hidden MRF model for multivariate gene expression data observed from across sectional designs (HMRF-C), with parameters in the emission probabilities $\theta = (\eta, \nu, \lambda^2)$. Define $(I_g) = \{I_1, \cdots, I_p\}$ to be a vector of the differential expression states of the $p$ genes on the network. By Bayes rule, $Pr((I_g)|\mathbf{D}) \propto Pr(\mathbf{D}|(I_g)) \times Pr((I_g))$. The estimate $(\hat{I_g})$ that maximizes $Pr((I_g)|\mathbf{D})$ is a MAP estimate under 0-1 loss. In order to estimate the parameters and $(I_g)$, we make the following conditional independence assumption,

*Assumption*: Given any particular realization $(I_g)$, the random variables $(\mathbf{D}) = (\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_g)$ are conditionally independent and each $D_g$ has the same unknown conditional density function $P(\mathbf{D}_g|I_g)$, dependent only on $I_g$. The conditional density of the observed gene expression data $\mathbf{D}$, given $\mathbf{G}$ and parameter $\theta = (\eta, \nu, \mathbf{\Lambda})$, is simply,

$$L_\theta((\mathbf{D}_g)|(I_g)) = \prod_{g=1}^{p} P(\mathbf{D}_g|I_g) \tag{7}$$

where $P(\mathbf{D}_g|I_g)$ is defined as (3) or (4) for the longitudinal designs, and (5) or (6) for the cross-sectional designs.

## 3 Estimation of the Model Parameters and the MDE States

When inferring $(I_g)^*$, parameter estimation must be carried out simultaneously. We propose the following ICM algorithm of Besag (1986) to simultaneously estimate the parameter $\theta$ in the emission probability model and the parameter $\Phi = (\gamma, \beta)$ in the auto-logistic model. For the longitudinal designs and the HMRF-L model, $\theta = (\eta, \nu, \mathbf{\Lambda})$ with a positive definite constraint on the covariance matrix $\mathbf{\Lambda}$. Simultaneously estimating this covariance matrix is difficult due to the fact that its estimate has to be positive definite. We propose to first estimate $\mathbf{\Lambda}'$ using the moment estimator of Tai and Speed (2006). Specifically, by the weak law of large numbers, $\overline{\mathbf{S}}$ converges in probability to $(\nu - k - 1)^{-1}\nu\mathbf{\Lambda}$. Therefore, $\mathbf{\Lambda}$ can be estimated by $\mathbf{\Lambda}' = \hat{\nu}^{-1}(\hat{\nu} - k - 1)\bar{\mathbf{S}}$, where $\hat{\nu} = \max(mean(\hat{\nu}_j), k+6), j = 1, \cdots, k$ and $\hat{\nu}_j$ is the estimated prior degrees of freedom based on the $j$th diagonal elements of the gene-specific sample variance-covariance matrices using the method proposed in Section 6.2 in Smyth (2004). We then fix $\mathbf{\Lambda}$ at its estimate and estimate the other model parameters within the following ICM algorithm (Besag, 1986), which involves the following iterative steps:

**S1.** Obtain an initial estimate $(\hat{I_g})$ of the true state $(I_g)^*$, using simple two sample Hotelling's $T^2$ test.

**S2.** Estimate $\theta$ by the value $\hat{\theta}$ which maximizes the likelihood $L_\theta(\mathbf{D}|(\hat{I_g}))$ (Equation 7).

**S3.** Estimate $\Phi$ by the value $\hat{\Phi}$ which maximizes the following pseudo-likelihood

$$L_\Phi((\hat{I_g})) = \prod_{g=1}^{G} \frac{\exp\{I_g F((\hat{I_g}))\}}{1 + \exp\{F((\hat{I_g}))\}}.$$

**S4.** Carry out a single cycle of ICM based on the current $(\hat{I_g}), \hat{\theta}$ and $\hat{\Phi}$, to obtain a new $(\hat{I_g})$: for $g = 1$ to $p$, update $I_g$ which maximizes

$$P(I_g|\mathbf{D}, \hat{I_{g'}}, g' \neq g) \propto P(\mathbf{D}_g|I_g; \hat{\theta}) P(I_g|\hat{I_g}', g' \neq g; \hat{\Phi}).$$

**S5.** Go to step 2 for a fixed number of cycles or until there is convergence in the estimates.

In Step 2, $\theta = (\eta, \nu, \lambda^2)$ for the cross-sectional designs and the HMRF-C model and $\theta = (\eta, \nu)$ in the HMRF-L model and they can be estimated using any numerical optimization procedure.

## 4 Simulation Study

We performed simulation studies to evaluate the proposed methods and to compare results with other methods for identifying the MDE genes. Following Wei and Li (2007), we first obtained 33 human regulatory pathways

Table 1: Comparison of parameter estimates of three different procedures for four sets of simulations with different percentages of MDE genes ($p$). HMRF-L: the proposed HMFR model and the ICM algorithm with longitudinal emission probabilities using the network structures; HMRF-I: the proposed HMFR model and the ICM algorithm with longitudinal emission probabilities without using the network structures; MB: algorithm of Tai and Speed (2006). Parameter estimates are averages over 100 simulations; standard error is shown in parentheses. The true parameters are $(\eta, \nu)$=(0.5,13).

| | | Percentage of MDE genes ($p$) | | | |
|---|---|---|---|---|---|
| Method | Parameter | 0.115(0.005) | 0.189(0.008) | 0.357(0.009) | 0.486(0.008) |
| HMRF-L | $\hat{\eta}$ | 0.375(0.026) | 0.395(0.028) | 0.414(0.018) | 0.436(0.020) |
| | $\hat{\nu}$ | 13.010(0.061) | 13.059(0.061) | 13.124(0.059) | 13.178(0.057) |
| HMRF-I | $\hat{\eta}$ | 0.314(0.019) | 0.338(0.017) | 0.368(0.013) | 0.386(0.012) |
| | $\hat{\nu}$ | 13.025(0.060) | 13.065(0.060) | 13.151(0.057) | 13.221(0.056) |
| MB | $\hat{\eta}$ | 0.067(0.004) | 0.053(0.003) | 0.042(0.002) | 0.039(0.001) |
| | $\hat{\nu}$ | 7.265(0.207) | 7.434(0.210) | 7.858(0.230) | 8.212(0.254) |

from the KEGG database (December 2006), where we retained only gene-gene regulatory relations. These 33 regulatory pathways are inter-connected and formed a network of pathways. We represent such a network as an undirected graph where each node is a gene and two nodes are connected by an edge if there is a regulatory relation between corresponding genes. Loops (nodes connected to themselves) were eliminated. This results in a graph with 1668 nodes and 8011 edges.

To simulate the differential expression states of the genes on this network, we initialized the genes in the $K$ pathways to be DE and the rest genes to be EE, which gives us the initial $\mathbf{G_0}$. We then performed sampling five times based on the current gene differential expression states, according to the Markov random field model with $\gamma_0 = \gamma_1 = 1$ and $\beta = 2$ (Wei and Li, 2007). We chose $K = 5, 9, 13, 17$ to obtain different percentages of genes in MDE states. After obtaining the differential expression states for the genes, we simulated the multivariate gene expression levels based on the empirical Bayes models, using the same parameters as Tai and Speed (2006): $\eta = 0.5, \nu = 13$ and

$$\mathbf{\Lambda} = \begin{pmatrix} 14.69 & 0.57 & 0.99 & 0.40 & 0.55 & 0.51 & -0.23 \\ 0.57 & 15.36 & 1.22 & 0.84 & 1.19 & 0.91 & 0.86 \\ 0.99 & 1.22 & 14.41 & 2.47 & 1.81 & 1.51 & 1.07 \\ 0.40 & 0.84 & 2.47 & 17.05 & 2.40 & 2.32 & 1.33 \\ 0.55 & 1.19 & 1.81 & 2.40 & 15.63 & 3.31 & 2.75 \\ 0.51 & 0.91 & 1.51 & 2.32 & 3.31 & 13.38 & 3.15 \\ -0.23 & 0.86 & 1.07 & 1.33 & 2.75 & 3.15 & 12.90 \end{pmatrix} \times 10^{-3}.$$

For each condition, we chose the number of independent replications to be 3 for each group and repeated the simulation 100 times.

## 4.1 Comparison with the method of Tai and Speed

We first examined the parameter estimates of $\theta = (\eta, \nu)$ using three different methods: the MB method of Tai and Speed (2006), the ICM algorithm incorporating the network structures and the ICM algorithm assuming that all the nodes are singletons (i.e., no dependency of the differential expression states), both assuming the longitudinal emission probabilities. The performance results are shown in Table 1. We observed that both ICM algorithms provide better estimates of both $\eta$ and $\nu$ than the MB algorithm.

We then compare the sensitivity, specificity and FDR in identifying the MDE genes with the MB method of Tai and Speed (2006). Since the MB method only provides ranks of the genes and does not infer gene states, for the purpose of comparison, we chose a cutoff value to declare genes to be MDE using their method so that their approach would have the closest observed FDR levels to our proposed method. We applied both the HMRF-L model and the HMRF-C model to the simulated data sets. The results are summarized in Table 2, clearly showing that our approach obtained significant improvement in sensitivity compared to the other approaches making an independence assumption of genes. The smaller $p$ was, the more improvements we obtained. At the same time, our approach also achieved lower FDRs and comparable specificity. Our proposed algorithm assuming that the

Table 2: Comparison of performance in terms of sensitivity (SEN), specificity (SPE) and false discovery rate (FDR) of three different procedures based on 100 replications for four different scenarios with different percentages of MDE genes ($p$). HMRF-L: the proposed HMFR model with longitudinal emission probabilities using the network structures; HMRF-C: the proposed HMFR model with cross-sectional emission probabilities using the network structures; HMRF-I: the proposed HMFR model with longitudinal emission probabilities without using the network structures; MB-L: algorithm of Tai and Speed (2006) with FDRs matched to the HMRF-L algorithm; MB-C: algorithm of Tai and Speed (2006) with FDRs matched to the HMRF-C algorithm. Summaries are averaged over 100 simulations; standard deviation is shown in parentheses.

| $p$ | Method | Sensitivity | Specificity | FDR |
|---|---|---|---|---|
| 0.115(0.005) | HMRF-L | 0.80(0.029) | 1.00(0.0023) | 0.045(0.019) |
| | HMRF-C | 0.80(0.04) | 0.98(0.0053) | 0.16(0.037) |
| | HMRF-I | 0.70(0.042) | 0.99(0.0027) | 0.079(0.025) |
| | MB-C | 0.78(0.04) | 0.98(0.0046) | 0.16(0.032) |
| | MB-L | 0.69(0.054) | 0.99(0.0027) | 0.079(0.05) |
| 0.189(0.008) | HMRF-L | 0.87(0.033) | 0.99(0.0049) | 0.058(0.020) |
| | HMRF-C | 0.84(0.038) | 0.97(0.0058) | 0.12(0.024) |
| | HMRF-I | 0.76(0.03) | 0.99(0.004) | 0.074(0.018) |
| | MB-C | 0.81(0.029) | 0.97(0.0059) | 0.12(0.023) |
| | MB-L | 0.75(0.032) | 0.99(0.0041) | 0.075(0.018) |
| 0.357(0.009) | HMRF-L | 0.91(0.016) | 0.97(0.0065) | 0.054(0.010) |
| | HMRF-C | 0.87(0.025) | 0.96(0.0070) | 0.077(0.011) |
| | HMRF-I | 0.84(0.020) | 0.97(0.0063) | 0.066(0.011) |
| | MB-C | 0.85(0.025) | 0.96(0.0073) | 0.077(0.011) |
| | MB-L | 0.83(0.022) | 0.97(0.0064) | 0.066(0.011) |
| 0.486(0.008) | HMRF-L | 0.95(0.012) | 0.94(0.012) | 0.061(0.012) |
| | HMRF-C | 0.90(0.025) | 0.94(0.011) | 0.065(0.011) |
| | HMRF-I | 0.88(0.015) | 0.95(0.0086) | 0.060(0.0093) |
| | MB-C | 0.88(0.020) | 0.94(0.011) | 0.065(0.011) |
| | MB-L | 0.88(0.015) | 0.95(0.0087) | 0.060(0.0094) |

genes are independent give very similar results to the MB method of Tai and Speed (2006). We also observed that when the data are simulated for longitudinal design, analysis of the data using the HMRF-C can result in higher FDRs. However, given the same FDR level, the HMFR-C method performed similarly to the MB method.

## 4.2 Sensitivity to misspecification of the network structure

Due to the fact that our current knowledge of biological networks is not complete, in practice, it is possible that the network structures that we use for network-based analysis are misspecified. The misspecification can be due to either the true edges of the networks being missed or the wrong edges being included in the network, or both of these scenarios. We performed simulation studies to evaluate how sensitive the results of the HMRF-L approach are to these three types of misspecifications of the network structures. We used the same data sets of 100 replicates as in the previous section but used different misspecified network structures when we fitted the hidden MRF model.

For the first scenario, we randomly removed 801 (10%), 2403 (30%) and 4005 (50%), respectively, from the 8011 true edges from the true KEGG networks when we fit the hidden MRF model. For the second scenario, we randomly added approximately 801, 2403 and 4005 new edges to the KEGG network, respectively. Finally, for the third scenario, we randomly selected 90%, 70% and 50% of the 8011 true edges and also randomly added approximately 801, 2403 and 4005 new edges to the network, respectively, so that the total number of edges remains approximately 8011. The results of the simulations over 100 replications are summarized as Figure 1. First, as expected, since the true number of MDE genes is small, the specificities of the HMRF-L procedure remain very high and are similar when the true network structure is used. Second, we also observed that the FDR rates also remain almost the same as when the true structure is used. However, we observed some decreases in sensitivity in identifying the true DE genes. It is worth pointing out even when the network structure is largely

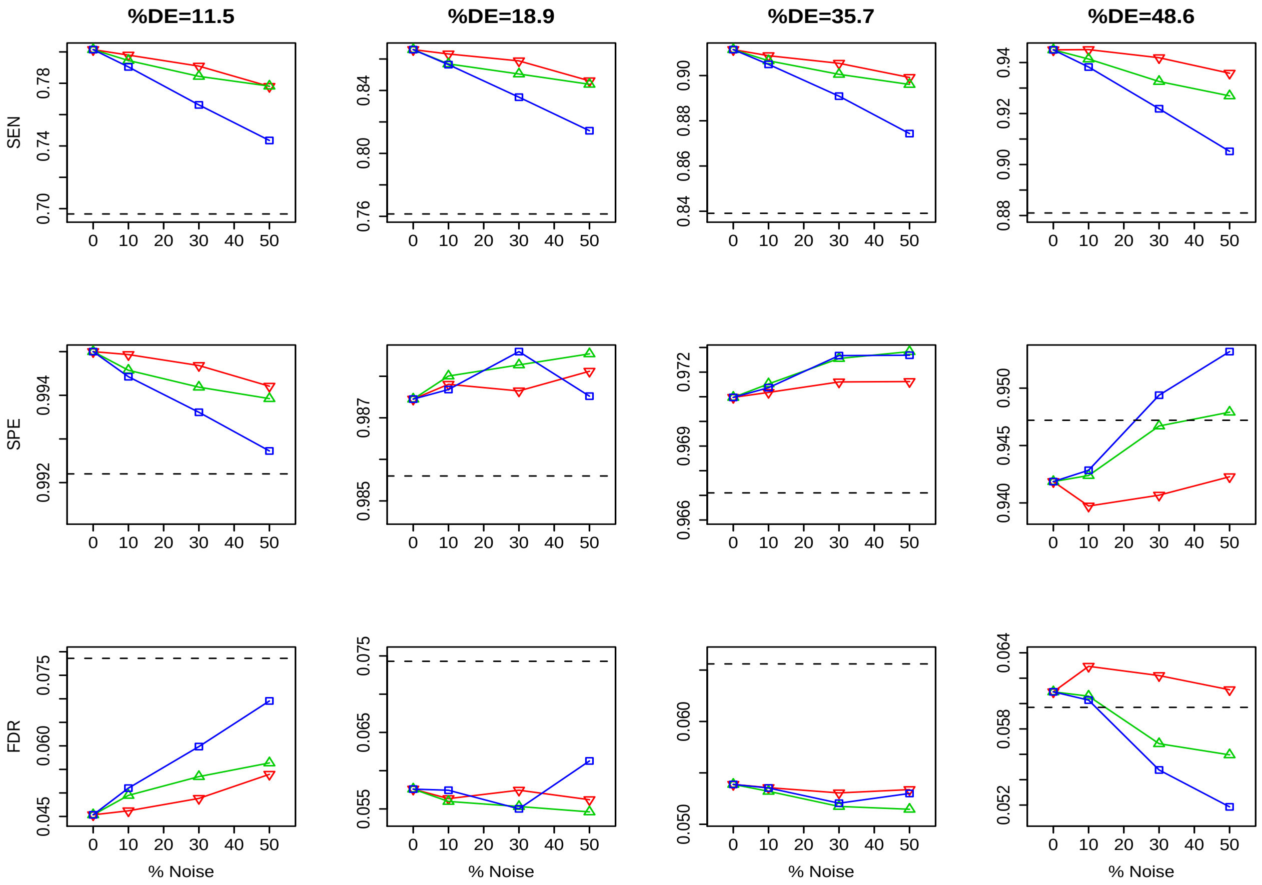

Figure 1: *Results in sensitivity, specificity and false discovery rate when the network structure is misspecified for four different sets of simulations corresponding to different proportions of MDE genes.* ▽*: randomly deleting 10%, 30% and 50% of the true edges of the network;* △*: randomly adding approximately 801 (10%), 2403 (30%) and 4005 (50%) new edges to the network;* □*: randomly choosing 90%, 70% and 50% of the true edges and randomly adding 10%, 30% and 50% new edges to the network.*

misspecified as in scenario 3, the results from the HMRF-L model are still comparable to those obtained from the HMRF-I approach where the network structure is not utilized.

Finally, we also applied these simulated data with a randomly created network structure with the same number of nodes and edges. As expected, in this case, the estimate of the $\beta$ parameter was always zero or very close to zero, and therefore, the results in sensitivity, specificity and FDR are essentially the same as the method that does not utilize the network structure. These simulations seem to indicate that the results of the HMRF-L model are not too sensitive to the misspecification of the network structure unless the structure is greatly misspecified.

## 5 Application to Time-Course Gene Expression Study of TrkA- and TrkB-transfected Neuroblastoma Cell Lines

Neuroblastoma is the most common and deadly solid tumor in children, but this tumor also has a very high propensity to undergo spontaneous differentiation or regression. Evidence suggests that the Trk family of neurotrophin receptors plays a critical role in tumor behavior (Broduer, 2003). Neuroblastomas expressing TrkA are biologically favorable and prone to spontaneous differentiation or regression. In contrast, neuroblastomas expressing TrkB usually have MYCN amplification and are among the most aggressive and deadly tumors known. These tumors also express the TrkB ligand, resulting in an autocrine survival pathway. Unlike the TrkA-expressing tumors, exposure to ligand promotes survival under adverse conditions, but does not cause differentiation. In order to explore the biological basis for the very different behavior of neuroblastomas expressing these highly homologous neurotrophin receptors, a microarray time-course gene expression study was conducted by transfecting

TrkA and TrkB into SH-SY5Y cells, a neuronal subclone from the NB cell line SK-N-SH. In particular, full length TrkA and TrkB were cloned into the retroviral expression vector pLNCX and transfected into SH-SY5Y cells. Cells were then serum starved overnight and treated with either nerve growth factor (NGF) and brain-derived neurotrophic factor (BDNF) at $37^o$ for 0 to 12 hours. Fifteen micrograms of total RNA were then collected from TrkA- and TrkB-SY5Y cells exposed to 0, 1.5, 4 and 12 hrs of NGF or BDNF and the gene expressions were profiled using the Affymetrix GeneChip 133A. Four and three replicates were performed for the TrkA and TrkB cells, respectively. The robust multi-array (RMA) procedure (Irizarry *et al.*, 2003) was used to obtain the gene expression measures.

To perform network-based analysis of the data, we merged the gene expression data with the 33 KEGG regulatory pathways and identified 1533 genes on the Hu133A chip that could be found in the 1668-node KEGG network of 33 pathways. Instead of considering all the genes on the Hu133A chip, we only focused our analysis on these 1533 genes and aimed to identify which genes and which subnetworks of the KEGG network of 33 pathways are potentially related to the cell differentiation of TrkA-transfected cell lines. We analyzed the data using the HMRF-L model with longitudinal emission probabilities and obtained parameter estimates of $\hat{\alpha} = -1.58$ and $\hat{\beta} = 0.39$, indicating that there are more genes with similar expression patterns than those with different expression patterns. Our method identified 210 MDE genes out of the 1533 KEGG genes, among these 118 are connected on the KEGG pathways and 92 are isolated, not collecting to other MDE genes. The heat map plot of these 210 MDE genes is show in Figure 2, showing clear different expression patterns between TrkA and TrkB time courses. There is a large cluster of genes that are largely up-regulated in the Trk A transfected cells but are down-regulated in the Trk B transfected cells. Similarly, there is a cluster of genes that are up-regulated in the Trk B-transfected cells but are down-regulated in the Trk A transfected cells.

Among the 33 KEGG regulatory pathways, enrichment analysis using DAVID Tools (Dennis *et al.*, 2003) (http://david.abcc.ncifcrf.gov/home.jsp) identified that the mitogen-activated protein kinase (MAPK) signaling pathway, focal adhesion pathway and pathway related to prion diseases are enriched with $p$-values of 0.012, 0.029 and 0.05, respectively, of which the MAPK signaling pathway and the focal adhesion pathway are inter-connected. The MAPK (Erk1/2) signal transduction pathway is expressed and active in both TrkA and TrkB expressing NB cells after specific ligand-mediated Trk receptor phosphorylation. The distinct role that this signaling pathway plays in the biologic heterogeneity of NB is not well known; however, we have shown that the time course of pathway activation by phosphorylation of signal effector proteins is different between TrkA- and TrkB- expressing NB cells, and this may, in part, explain the biological differences between TrkA- vs. TrkB-expressing tumors. To give a detailed comparison of TrkA- and TrkB-mediated genomic responses, we present in Figure 3 and Figure 4 the MDE genes on the MAPK signaling pathway and on the KEGG focal adhesion pathway. On the MAPK pathway, it is not surprising that the TrkA/B shows different expression patterns. We also observed that a cluster of genes (or a subnetwork) in the neighborhood of ERK shows different expression patterns, including MEK2, MP1, PTP, MKP, Tau, cPLA2, MNK1/2 and c-Myc. This subnetwork, leading to cell proliferation and differentiation, may partially explain the difference in cell differentiation between the TrkA- and TrkB-infected NB cells. Another interesting subnetwork in the neighborhood of p38, including MKK3, MKK6, PTP, MKP, MAPKAPK, GADD153 and HSP27, also showed differential expression patterns. This subnetwork also related to cell proliferation and differentiation. Activation of these two subnetworks on the MAPK pathway may explain the different biological behaviors of these two types of NB cells, especially in terms of cell differentiation. MAPK signaling in the nervous system has been shown to promote a broad array of biologic activities including neuronal survival, differentiation, and plasticity. Regulating the duration of MAPK signaling is important in neurogenesis, and likely plays a similar role in the behavior of Trk-expressing neuroblastomas. Prolonged activation of MAPK is correlated with neurotrophin-dependent cell cycle arrest and terminal cellular differentiation in the PC12 pheochromocytoma cell line, whereas short-duration MAPK signaling is correlated with mitogenic and proliferative cell signaling in PC12 cells (Tombes *et al.*, 1998; Kao *et al.*, 2001; Marshall, 1995; Qui and Green, 1992). TrkA-expressing NB cells treated with NGF (which activates MAPK) increase the number and length of extended neurites and decrease cell proliferation resulting in a more mature neuronal appearing cell, while TrkB-expressing NB cells treated with ligand (BDNF) increase cell proliferation without morphologic differentiation.

Increasing evidence suggests an important role for the focal adhesion kinase (FAK) pathway in regulating cancer cell adhesion in response to extracellular forces or mechanical stress. Studies have demonstrated that tumor cells are able to regulate their own adhesion by over-expression or alteration in activity of elements within the FAK signaling pathway, which may have implications in the survival, motility and adhesion of metastatic tumor cells (Basson, 2008). While mechanotransduced stimulation of the FAK signaling pathway appears to be a cell surface receptor independent process, the FAK pathway also acts downstream of receptor tyrosine kinases and has been shown to be phosphorylated in response to external cytokine/ligand stimuli. The insulin-like growth factor-1 receptor (IGF-1R) and FAK physically interact in pancreatic adenocarcinoma cells resulting in activation of a common signal transduction pathway that leads to increased cell proliferation and cell survival (Liu *et al.*, 2008). In neuroblastoma, MYCN regulates FAK expression by directly binding to the FAK promoter, and increasing transcription of FAK mRNA. Beierle *et al.* (2007) have correlated FAK mRNA abundance with MYCN expression in MYCN-amplified and non-amplified NB cell lines by real time quantitative PCR, and their data suggest that MYCN regulation of FAK expression directly impacts cell survival and apoptosis. On the focal adhesive pathway, we observed that a subnetwork of 6 genes, including Actinin, Filamin, Talin, Zyxin, VASP, Vinculin, that show differential expression patterns. In addition, PI3K and its neighboring genes GF, RTK, Shc and Ha-Ras show differential expression patterns. We have not yet explored the regulation of FAK pathway activity by TrkA or TrkB expression and activation in our NB cell lines, but the differential expression states for genes on the KEGG FAK pathway suggest differential mediation by TrkA vs. TrkB, that may have downstream biological relevance.

Finally, on the pathway related to prion disease, we observed that Prion Protein (PrPc) and its neighboring genes HSPA5, APLP1, NRF2 and LAMB1 show differential expression patterns.

## 6 Conclusion and Discussion

In this paper we have proposed a hidden MRF model and an ICM algorithm that utilizes the gene regulatory network information to identify multivariate differentially expressed genes. Different from the approach of Wei and Li (2008) for network-based analysis of microarray time-course gene expression data, this new approach identify the genes that show different expression patterns over time rather than identifies the differentially expressed genes at each time point. Simulation studies show that our methods outperform the methods that do not utilize network structure information. We applied our method to analyze the MTC data of TrkA- and TrkB-transfected neuroblastoma cell lines and identified the MAPK and focal adhesive pathways from the KEGG that are related to cell differentiation in TrkA-transfected cell lines. Note that the proposed methods can also be applied to other types of genomic data such as proteomic data and protein-protein interaction data.

In this paper, we analyzed the neuroblastoma MTC data using KEGG pathways and aimed to identify the KEGG pathways that may explain the differentiation states of the two different NB cell lines. However, the proposed methods can be applied to any other networks of pathways. An important question is to decide which pathways one should use in analyzing the MTC data. This partially depends on the scientific questions to be addressed. If an investigator is only interested in a particular pathway, the proposed method can be applied to that particular pathway. If an investigator is interested in fully exploring his/her data and all available pathways, one should use a large collection of pathways, e.g., the pathways collected by Pathway Commons (http://www.pathwaycommons.org/pc/). It should also be noted that our proposed methods can include all the genes probed on microarray by simply adding isolated nodes to the graphs. Another related issue is that our knowledge of pathways is not complete and can potentially include errors or misspecified edges on the networks. Although our simulations demonstrate that our methods are not too sensitive to the misspecification of the network structures, the effects of misspecification of the network on the results deserve further research. One possible solution to this problem is to first check the consistency of the pathway structure using the data available. For example, if the correlation in gene expression levels between two neighboring genes is very small, we may want to remove the edge from the pathway structure. Alternatively, one can build a set of new pathways using various data sources and compare these pathways with those in the pathway databases in order to identify the most plausible pathways for use in the proposed MRF method. For example, we can construct a large molecular network with the nodes being the gene products and the links extracted from the KEGG database, the Biomolecular Interaction Network Database (BIND) and Human Interactome Map (HIMAP) (Alfarano *et al.*, 2005). This will provide more comprehensive description of known biological pathways and networks than using data from only one source.

In summary, generation of high-throughput genomic data together with intensive biomedical research has generated more and more reliable information about biological pathways and networks. It is very important to incorporate the network information into the analysis of genomic data in order to obtain more interpretable results in the context of known biological pathways. Such integration of genetic network information with high-throughput genomic data can potentially be useful for identifying the key molecular modules and subnetworks that are related to complex biological processes.

## Acknowledgements

This research was supported by NIH grants ES009911, CA127334 and AG025532. We thank Mr. Edmund Weisberg, MS at Penn CCEB for editorial assistance.

## References

[Aitchison1975] Aitchison J and Dunsmore IR (1975): *Statistical prediction analysis*, Cambridge University Press, London.

[Affarano et al2005] Alfarano C, Andrade CE, Anthony K, Hahroos N, Bajec M, *et al.* (2005): The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Research*, D418-D424.

[Basson2008] Basson MD (2008): An intracellular signal pathway that regulates cancer cell adhesion in response to extracellular forces. *Cancer Research*, 68(1):2-4.

[Beierle et al1992] Beierle EA, Trujillo A, Nagaram A, Kurenova EV, et al. (2007): N-MYC regulates focal adhesion kinase expression in human neuroblastoma. *Journal of Biological Chemistry*, 282(17):12503-16.

[Besag1972] Besag J (1972): Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society, Series B* 34, 75-83.

[Besag1974] Besag J (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* 36, 192-225.

[Besag1986] Besag J (1986): On the statistical analysis of dirty pictures. *Journal of Royal Statistical Society B*, 48: 259-302.

[Brodeur2003] Brodeur GM (2003): Neuroblastoma: biological insights into a clinical enigma. *Nature Reviews - Cancer*, 3:203-216.

[Dennis et al2003] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003): DAVID: database for annotation, visualization and integrated discovery. *Genome Biology*, 4:P3.

[Eggert et at2000] Eggert A, Ikegaki N, Liu X, Chou TT, Lee VM, Trojanowski JQ and Brodeur GM (2000): Molecular dissection of TrkA signal transduction pathways mediating differentiation in human neuroblastoma cells. *Oncogene*, 19: 2043-2051.

[Hong and Li2006] Hong FX and Li H (2006): Functional hierarchical models for identifying genes with different time-course expression profiles. *Biometrics*, 62: 534-544.

[Irizarry2003] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2003): Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* 4: 249-264.

[Kanehisa and Goto2002] Kanehisa M and Goto S (2002): KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28: 27-30.

[Kao et al2001] Kao S, Jaiswal RK, Kolch W, Landreth GE (2001): Identification of the mechanisms regulating the differential activation of the MAPK cascade by epidermal growth factor and nerve growth factor in PC12 cells. *Journal of Biological Chemistry*, 276(21):18169-77.

[Lehmann et al2004] Lehmann KP, Phillips S, Sar M, Foster PMD and Gaido KW (2004): Dose-dependent alterations in gene expression and testosterone synthesis in the fetal testes of male rates exposed to Di ($n-$butyl) phthalate. *Toxicological Sciences*, 81: 60-68.

[Liu et al2008] Liu W, Bloom DA, Cance WG, Kurenova EV, Golubovskaya VM and Hochwald SN (2008): FAK and IGF-IR interact to provide survival signals in human pancreatic adenocarcinoma cells. *Carcinogenesis*, in press.

[Marshall1995] Marshall CJ (1995): Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell*, 80(2):179-85.

[Nacu et al2007] Nacu S, Critchley-Thorne R, Lee P and Holmes S (2007): Gene expression network analysis, and applications to immunity. *Bioinformatics*, 23(7): 850-858.

[Qui and Green1992] Qui MS, Green SH (1992): PC12 cell neuronal differentiation is associated with prolonged p21ras activity and consequent prolonged ERK activity. *Neuron*, 9(4):705-17.

[Schulte et al2005] Schulte J, Schramm A, Klein-Hitpass L, Klenk M, Wessels H, Hauffa BP, Eils J, Iils R, Brodeur GM, Schweigerer L, Havers W and Eggert A (2005): Microarray analysis reveals differential gene expression patterns and regulation of single target genes contributing to the opposing phenotype of TrkA- and TrkB-expressing neuroblastomas. *Oncogene*, 24: 165-177.

[Seidel et al2006] Seidel S, Stott W, Kan H, Sparrow B, Gollapudi B (2006): Gene expression dose-response of liver with a genotoxic and nongenotoxic carcinogen. *International Journal of Toxicology*, 25, 57-64.

[Smyth2004] Smyth GK (2004): Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), Article 3.

[Tai and Speed2006] Tai YC and Speed T (2006): A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics*, 34: 2387-2412.

[Tai and Speed02007] Tai YC and Speed T (2007): On the gene ranking of replicated microarray time course data. Technical report, Department of Statistics, UC Berkeley.

[Tombes et al1998] Tombes RM, Auer KL, Mikkelsen R, et al. (1998): The mitogen-activated protein (MAP) kinase cascade can either stimulate or inhibit DNA synthesis in primary cultures of rat hepatocytes depending upon whether its activation is acute/phasic or chronic. *Biochemistry Journal*, 330 (Pt 3):1451-60.

[Wei and Li2007] Wei Z and Li H (2007): A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23: 1537-1544.

[Wei and Li2008] Wei Z and Li H (2008): A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *Annals of Applied Statistics*, 2(1), 408-429.

[Yuan and Kendziorski2006] Yuan M and Kendziorski C (2006): Hidden Markov models for microarray time course data under multiple biological conditions (with discussion). *Journal of the American Statistical Association*, 101(476), 1323-1340.

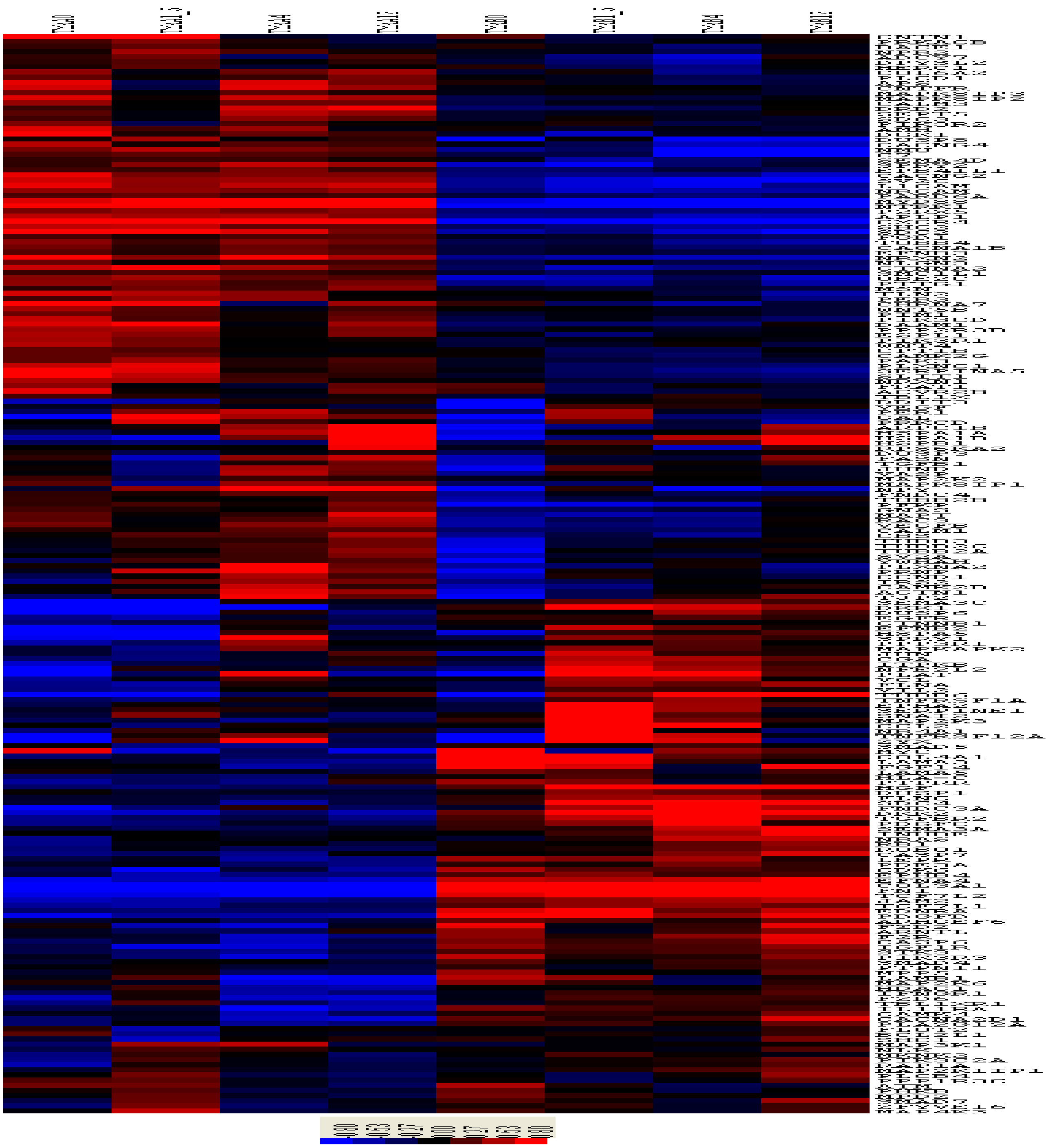

Figure 2: *Heatmap clustering plot of the 210 MDE genes on the KEGG pathways, showing different expression patterns between the TrkA and TrkB time-courses. The first four columns correspond to the TrkA time course experiments at times 0, 1.5, 4 and 12 hr, the second four columns correspond to the TrkB time-course experiments at times 0, 1.5, 4 and 12 hr.*

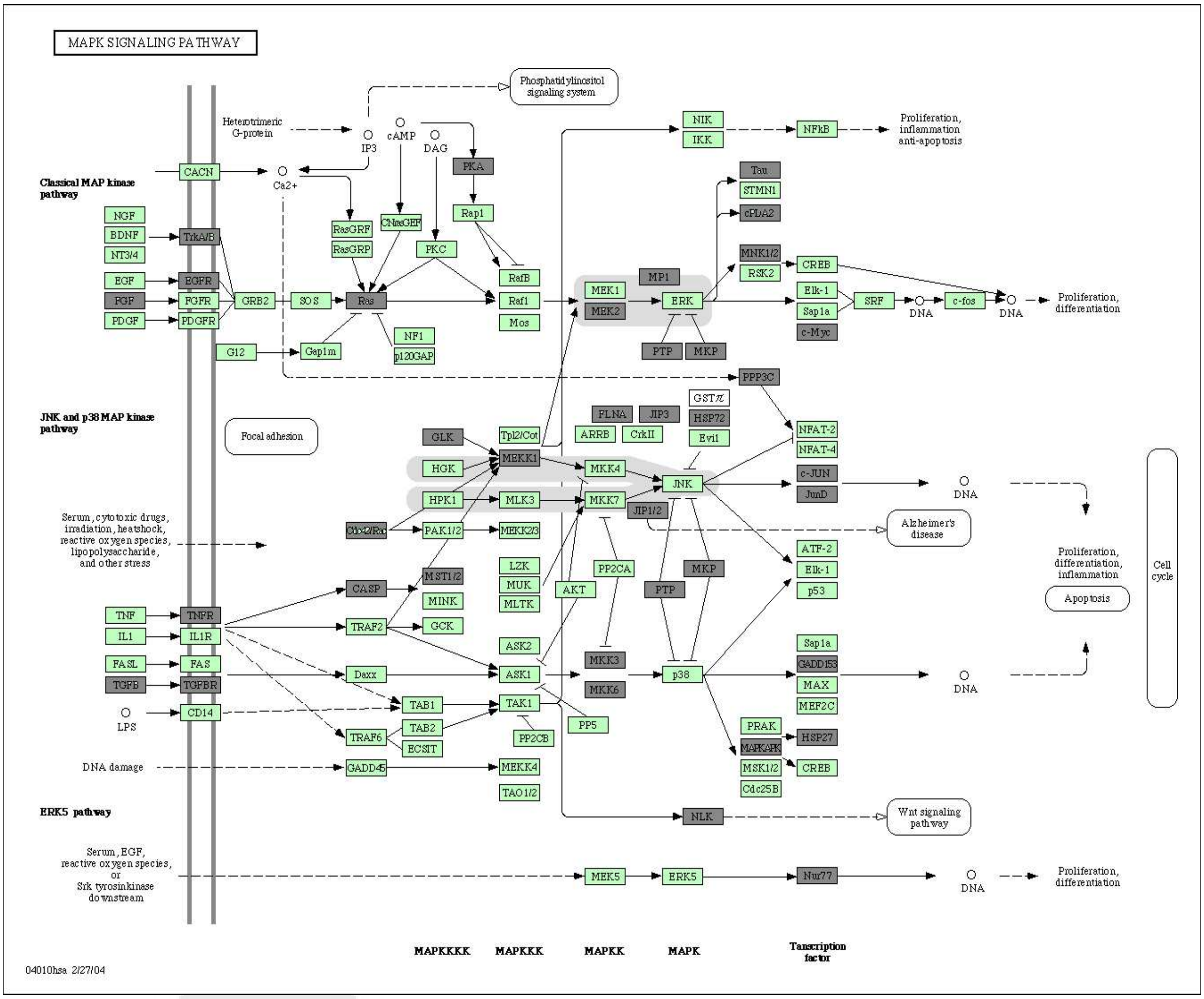

Figure 3: *Differential expression states for genes on the KEGG MAPK pathway, where genes colored in dark gray are multivariate differentially expressed and those colored in light green are equally expressed.*

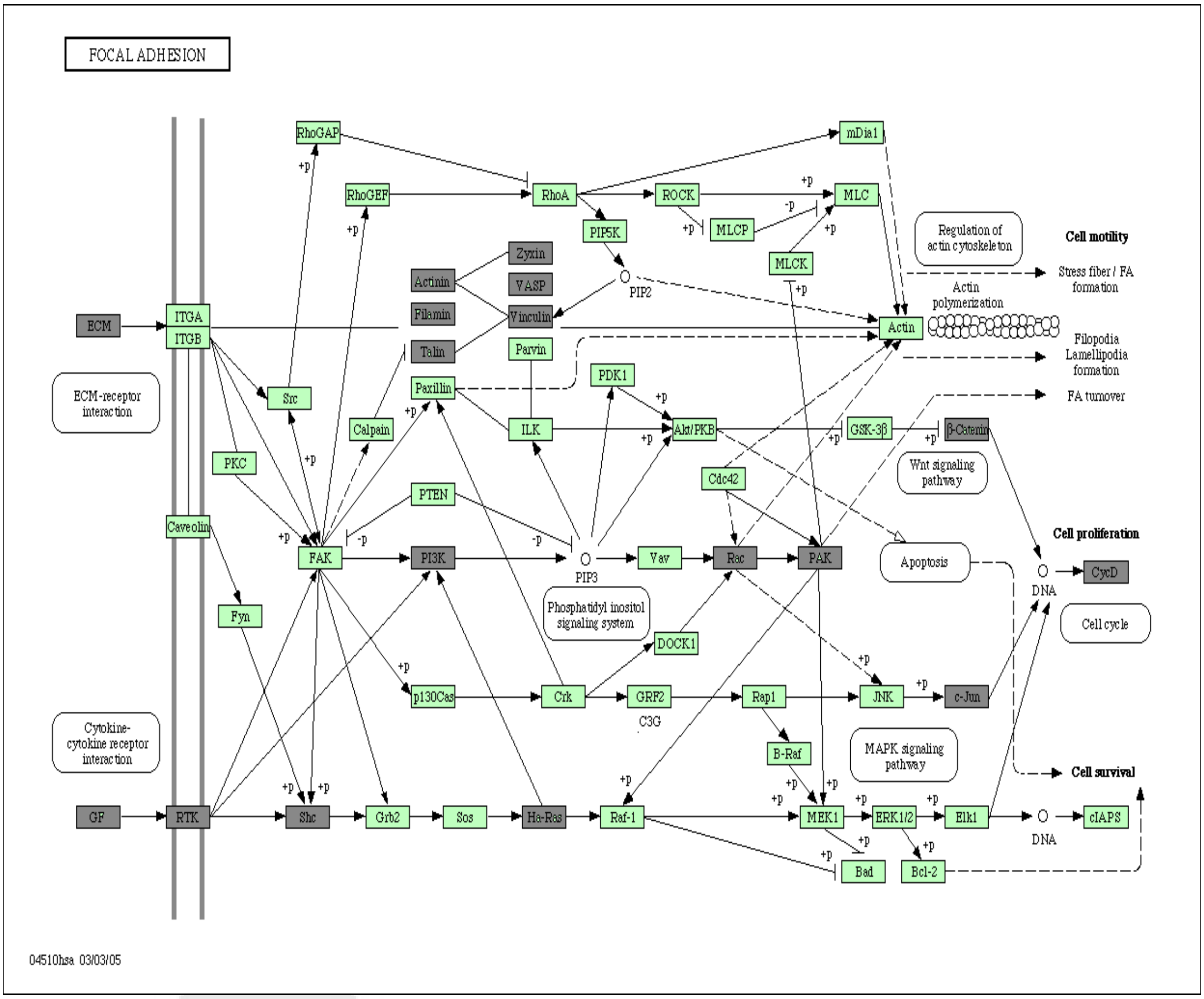

Figure 4: *Differential expression states for genes on the KEGG Focal Adhesion pathway, where genes colored in dark gray are multivariate differentially expressed and those colored in light green are equally expressed.*