Bioinformatics Advance Access published May 5, 2007

BIOINFORMATICS

A Markov Random Field Model for Network-based Analysis of Genomic Data

Zhi Wei [a] and Hongzhe Li [a,b]

[a]Genomics and Computational Biology Graduate Group and [b]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104

Associate Editor: Dr. Olga Troyanskaya

ABSTRACT

Motivation: A central problem in genomic research is the identification of genes and pathways involved in diseases and other biological processes. The genes identified or the univariate test statistics are often linked to known biological pathways through gene set enrichment analysis in order to identify the pathways involved. However, most of the procedures for identifying differentially expressed genes do not utilize the known pathway information in the phase of identifying such genes. In this paper, we develop a Markov random field (MRF)-based method for identifying genes and subnetworks that are related to diseases. Such a procedure models the dependency of the differential expression patterns of genes on the networks using a local discrete MRF model.

Results: Simulation studies indicated that the method is quite effective in identifying genes and subnetworks that are related to disease and has higher sensitivity and lower false discovery rates than the commonly used procedures that do not use the pathway structure information. Applications to two breast cancer microarray gene expression datasets identified several subnetworks on several of the KEGG transcriptional pathways that are related to breast cancer recurrence or survival due to breast cancer.

Conclusions: The proposed MRF-based model efficiently utilizes the known pathway structures in identifying the differentially expressed genes and the subnetworks that might be related to phenotype. As more biological networks are identified and documented in databases, the proposed method should find more applications in identifying the subnetworks that are related to diseases and other biological processes.

Contact: Hongzhe Li, email: hongzhe@mail.med.upenn.edu.

INTRODUCTION

Identification of genes and pathways involved in diseases and other biological processes is one of the important problems in genomic research. Microarray technology makes it possible to measure the expression levels of almost all human genes and therefore facilitate the identification of genes and pathways that are related to disease initiation and development. In a typical experiment, several phenotypes are compared, with a certain number of biological replicates for each phenotype. The goal is to identify the differentially expressed (DE) genes among the different phenotype groups.

There are many novel statistical methods that have been developed for identifying the DE genes. A general approach is to conduct a hypothesis test at each gene and then correct for multiple testing. Most of the statistics used are t statistics and differ primarily in the estimation of the variance (Dudoit *et al.*, 2002; Tusher *et al.*, 2001). Other methods include the empirical Bayes methods that can effectively pool data from different genes (Efron *et al.*, 2001; Lonnstedt and Speed, 2002; Newton *et al.*, 2003; Kendziorski *et al.*, 2003). These DE genes identified are often linked to a pre-defined list of groups of genes such as known pathways in order to identify which groups include more DE genes than expected by chance using a hypergeometric distribution. Alternatively, the gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005; Tian *et al.*, 2005) can be used. For such a GSEA, one starts with a pre-defined list of groups of genes and assigns every such group a score that is essentially the average of the univariate test statistics of its member genes. The groups with high scores are more likely to be DE and p-values can be obtained by permutation methods.

One limitation of the commonly used methods of identifying the DE genes or the GSEA is that network structures are not utilized in the analysis. However, the interaction network is a more precise way to represent information than lists of genes or pathways, as it describes which genes are closely connected within a given pathway. Hence it has the potential to detect more subtle changes of gene expressions, such as local disturbances within known pathways. 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 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. The method is essentially based on a spatial scan statistic treating genes collected on the networks as neighbors. However, their method does not explicitly utilize the dependency of gene differential expression patterns on the network. Rapaport *et al.* (2007) proposed to first smooth the gene expression data on the network based on the spectral graph theory and then to use the smoothed data for classification. The method explicitly assumes that the true gene expression levels should be similar for genes that are neighbors on the networks. However, this assumption may be questionable due to both activating and inhibiting effects of gene regulations.

Markov random field (MRF) models have been widely used in image analysis in order to account for the local dependency of the observed pixel intensities (Besag, 1986) and have also been applied for functional prediction of proteins in order to account for the local dependency of protein functions in the protein-protein interaction networks (Deng *et al.*, 2002; Deng *et al.*, 2004; Letovsky and Kasif, 2003). The MRF model has also be applied for discovering molecular pathways from protein interaction and gene expression data

(Segal *et al.*, 2003). In this paper, we propose to develop a Markov random field (MRF)-based method for identifying the DEs and the subnetworks that are related to the phenotypes, where the MRF model is used to capture the dependency of the differential expression patterns for genes on the networks. Our method combines the two-group empirical Bayes method of Newton *et al.* (2003) and Kendziorski *et al.* (2003) with a MRF model to model the dependency of the differential expression patterns. In our model and those of Newton *et al.* (2003) and Kendziorski *et al.* (2003), each gene is either DE or equally expressed. Those genes which are EE present data according to some background Gamma distribution, and those which are DE present data according to a different distribution. The specific forms of these distributions arise by another layer of mixing over the latent mean expression level for each gene and the latent means follow another Gamma distribution. Such empirical Bayes approaches allow a level of information sharing amongst genes.

The rest of the paper is organized as follows. We first introduce the model assumptions and a MRF model. We then provide an iterative conditional mode algorithm (ICM) of Besag (1986) for parameter estimation and for identifying the DE genes. We present simulation studies to demonstrate the methods and to compare the results with other commonly used methods for DE gene identification. Finally, we present results of applying the proposed method to two breast cancer gene expression datasets in order to identify the genes and the subnetworks that are related to breast cancer recurrence or death due to breast cancer. We conclude the paper with a brief discussion of the results and methods.

STATISTICAL MODELS AND METHODS

Notation and Assumptions

Given microarray gene expression profiling data under two conditions, we want to determine which genes are differentially expressed. Each gene can have two states, labeled 0 and 1, representing equally expression (EE) and DE, respectively. An arbitrary state assignment of gene set S will be denoted by $x = (x_1, x_2, \cdots, x_p)$, where x_i is the corresponding state of gene i and is 1 if gene i is differentially expressed (i.e., DE) and 0 otherwise. We write x^* for the true but unknown gene state and interpret this as a particular realization of a random vector $X = (X_1, X_2, \cdots, X_p)$, where X_i assigns state to gene i. We let the y_i denote the observed mRNA expression level of gene i and y the corresponding vector, interpreted as a realization of a random vector, $Y = (Y_1, Y_2, \cdots, Y_n)$, where Y_i itself is a vector $y_i = (y_{i1}, y_{i2}, \cdots, y_{im}; y_{i(m+1)}, \cdots, y_{i(m+n)})$, composed of the m replicates under one condition and n replicates for the other. We further introduce the notation $y_{i.m} = \sum_{j=1}^{m} y_{ij}$ and $y_{i.n} = \sum_{j=m+1}^{m+n} y_{ij}$.

In order to specify the joint distribution of Y, we make the following assumptions:

Assumption 1. Given any particular realization x, the random variables $Y = (Y_1, Y_2, \cdots, Y_p)$ are conditionally independent and each Y_i has the same unknown conditional density function $f(y_i|x_i)$, dependent only on x_i. The conditional density of the observed gene expression y, given, x, is simply,

$$l(y|x) = \prod_{i=1}^{p} f(y_i|x_i).$$

Assumption 2. The true state x^* is a realization of a locally dependent discrete MRF with a specified distribution $\{p(x)\}$, which is defined in the next Section.

Gamma-Gamma model for gene expression data

In this section, we first briefly review the Gamma-Gamma model for gene expression data introduced in Newton *et al.* (2003) and Kendziorski *et al.* (2003). We define $f(.|\mu_i)$, which characterizes fluctuations in repeated measurements under the same condition from a gene i having latent mean expression level μ_i, and $\pi(\mu_i)$, which describes fluctuations in these means among genes. We assume that the observation y_i is a sample from a gamma distribution having shape parameter $\alpha > 0$ and a mean value μ_i; thus, with scale parameter $\lambda_i = \alpha/\mu_i$. The corresponding density function can be written as

$$f(y|\mu_i) = \frac{\lambda_i^{\alpha} y^{\alpha-1} exp\{-\lambda_i y\}}{\Gamma(\alpha)}$$

for measurement $y > 0$. Note that the coefficient of variation in this distribution is $1/\sqrt{\alpha}$, taken to be constant across genes i. Following Newton *et al.* (2003), we take $\pi(\mu_i)$ to be an inverse gamma. More specifically, fixing α, the quantity $\lambda_i = \alpha/\mu_i$ has a gamma distribution with shape parameter α_0 and scale parameter v. Let $\theta = (\alpha, \alpha_0, v)$ be the parameters used to specify these two distributions. The joint predictive density for the replicates $\mathbf{y_i}$ of gene i under the same condition is

$$f(\mathbf{y_i}) = \int \left(\prod_{y \in \mathbf{y_i}} f(y|\mu_i) \right) \pi(\mu_i) d\mu_i.$$

Under this general model, we have for the first condition

$$f(y_{i1}, \cdots, y_{im}) = K_1 \frac{(\prod_{j=1}^{m} y_{ij})^{\alpha-1}}{(v + y_{i.m})^{m\alpha+\alpha_0}},$$

where

$$K_1 = \frac{v^{\alpha_0}\Gamma(m\alpha + \alpha_0)}{\Gamma^m(\alpha)\Gamma(\alpha_0)},$$

and for the second condition

$$f(y_{i(m+1)}, \cdots, y_{i(m+n)}) = K_2 \frac{(\prod_{j=m+1}^{m+n} y_{ij})^{\alpha-1}}{(v + y_{i.n})^{n\alpha+\alpha_0}},$$

where

$$K_2 = \frac{v^{\alpha_0}\Gamma(n\alpha + \alpha_0)}{\Gamma^n(\alpha)\Gamma(\alpha_0)}.$$

Therefore, given the differential expression state x_i, we have

$$\begin{aligned} f(y_i|x_i;\theta) &= [f(y_{i1}, \cdots, y_{im}) * f(y_{i(m+1)}, \cdots, y_{in})]^{x_i} \\ &\quad \times [f(y_{i1}, \cdots, y_{im}, y_{i(m+1)}, \cdots, y_{in})]^{(1-x_i)} \\ &= \left[K_1 K_2 \frac{\left(\prod_{j=1}^{m+n} y_{ij}\right)^{\alpha-1}}{(v + y_{i.m})^{m\alpha+\alpha_0} (v + y_{i.n})^{n\alpha+\alpha_0}} \right]^{x_i} \\ &\quad \times \left[K \frac{\left(\prod_{j=1}^{m+n} y_{ij}\right)^{\alpha-1}}{(v + y_{i.m} + y_{i.n})^{(m+n)\alpha+\alpha_0}} \right]^{1-x_i}, \end{aligned}$$

where

$$K = \frac{v^{\alpha_0}\Gamma((m+n)\alpha+\alpha_0)}{\Gamma^{m+n}(\alpha)\Gamma(\alpha_0)}.$$

Then based on the assumption 1, the conditional density of all p genes can be written as

$$l(y|x;\theta) = \prod_{i=1}^{p} f(y_i|x_i;\theta). \qquad (1)$$

Discrete local MRF model for joint differential expression states

The gene differential expression states $x_i's$ are not independent. For example, if a gene is DE, it is more likely that its upstream regulators are also DE and that these regulators in turn affect their downstream-regulated genes. In order to explicitly account for such dependency of differential expression patterns over genes on the networks, we propose to use the known biological network information compiled in the form of pathways. Examples of such pathways include the KEGG pathway (Kanehisa and Goto, 2002) and BioCyc pathways (http://biocyc.com/). In our model, the network is expressed as an undirected graph with the nodes for genes and edges for connections on the network. Consider the p genes on the network, let $x = (x_1, x_2, \cdots, x_p)$ be the vector of unobserved differential expression states for the p genes. We propose to model the dependency of $x = (x_1, x_2, \cdots, x_p)$ using a MRF with parameter $\Phi = (\gamma_0, \gamma_1, \beta)$. Specifically, we assume

$$p(x;\Phi) \propto \exp(\gamma_0 n_0 + \gamma_1 n_1 - \beta n_{01}),$$

where $n_0 = \sum_i^p (1-x_i)$ is the number of genes at state 0, $n_1 = \sum_i^p x_i$ is the number of genes at state 1 and n_{01} is the number of edges linking two genes with different states. The γ_0 and γ_1 are arbitrary parameters and we require that $\beta > 0$, which discourages neighboring genes to have different differential expression states. By considering any two realizations which differ only at gene i, it follows that the conditional probability of state k occurring for gene i, given the states of all other genes is

$$p_i(k|.) \propto \exp(\gamma_k - \beta u_i(1-k)), \qquad (2)$$

where $u_i(1-k)$ denotes the number of neighbors of gene i having state $(1-k)$, $k = 0, 1$ (Besag, 1986). Maximum likelihood estimation of Φ, however, is computationally intractable. In general, it is the constant of proportionality in $p(x;\Phi)$ which cannot be evaluated. A simple alternative to maximum likelihood estimation for a local Markov random field is provided by the “coding method” (Besag, 1986), where the estimate $\hat{\Phi}$ is chosen to maximize the conditional likelihood,

$$\begin{aligned} l(x;\Phi) &= \prod_i^p p_i(x_i|x_{\partial i};\Phi) \qquad (3) \\ &= \prod_i^p \frac{\exp[(1-x_i)(\gamma_0-\beta u_i(1)) + x_i(\gamma_1-\beta u_i(0))]}{\exp[\gamma_0-\beta u_i(1)] + \exp[\gamma_1-\beta u_i(0)]}, \end{aligned}$$

where $x_{\partial i}$ represents the neighbors of gene i.

In order to account for different numbers of neighbors for different genes on the network, we propose to modify the conditional probability (2) as

$$p_i(k|.) \propto \exp(\gamma_k - \beta u_i^*(1-k)), \qquad (4)$$

where $u_i^*(1-k) = u_i(1-k)/d_i$ for $k = 0, 1$ and d_i is the number of neighbors for the ith gene. The conditional likelihood function (3) can be modified accordingly by replacing $u_i(1-k)$ with $u_i^*(1-k)$.

Parameter estimation using ICM and identification of subnetworks

When inferring the true differential expression state x^* for the p genes, the parameter estimation must be carried out simultaneously. We propose the following algorithm based on the ICM algorithm of Besag (1986) to estimate the parameter θ in the Gamma-Gamma model for gene expression data and the parameter Φ in the MRF model. The algorithm involves the following iterative steps:

1. Obtain an initial estimate $\hat{x}$ of the true state x^*, using a simple two sample t-test.

2. Estimate θ by the value $\hat{\theta}$ which maximizes the likelihood $l(y|\hat{x};\theta)$ (see Equation 1).

3. Estimate Φ by the value $\hat{\Phi}$ which maximizes the conditional likelihood $l(\hat{x};\Phi)$ (see Equation 3) based on current $\hat{x}$.

4. Carry out a single cycle of ICM based on the current $\hat{x}, \hat{\theta}$ and $\hat{\Phi}$, to obtain a new $\hat{x}$. Specifically, for $i = 1$ to p, update x_i which maximizes

$$P(x_i|y,\hat{x}_{S/i}) \propto f(y_i|x_i;\hat{\theta})p_i(x_i|\hat{x}_{\partial i};\hat{\Phi}),$$

subject to $x_i = 1$ or $x_i = 0$.

5. Go to step 2 for a fixed number of cycles or until approximate convergence of $\hat{x}$.

The converged $\hat{x}$ are then taken to be the estimate of the true differential expression states. These estimates can then be mapped back to the network to identify the subnetworks, which are defined as those connected genes that show differential expressions between the two experimental conditions.

SIMULATION STUDIES

We first present simulation results to demonstrate our proposed methods. To simulate the data, we first obtained the network structure of 33 human regulatory pathways from the KEGG database (December 2006 version). We are only interested in gene-gene regulatory relations and any non-gene-gene interactions, e.g., compound-gene relations, compound-compound relations, were excluded from our analysis. The remaining gene-gene regulatory data are represented 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 them. Loops (nodes connected to themselves) were eliminated. This results in a graph with 1668 nodes and 8011 edges.

To simulate X, the gene expression states, we initialized the genes in the K pathways to be DE and the rest of genes to be EE, which gives us the initial X_0. Then we performed sampling five times based on X_0, according to Equation (3), with $\gamma_0 = 1, \gamma_1 = 1, \beta = 2$. We chose $K = 5, 9, 13, 17$, so that we have different percentages of genes in DE states. Next, given X, we simulated the gene expression level Y according to GG model (Equation (1)) for

Table 1. Comparison of performance for the proposed MRF approach (MRFGG), the Gamma-Gamma model (GG) of Kendziorski *et al.* and standard two-sample t-test applied to the simulated data. Summaries are averaged over 100 simulations; standard errors are shown in parentheses. tTest1: two-sample t-test using p-value of 0.05 as cutoff point; tTEST2: two-sample t-test for FDR=0.05 using the procedure of Benjamini and Hochberg.

% of DE in simulated data	Model	Sensitivity	Specificity	FDR
p=0.115(0.005)	MRFGG	0.682(0.064)	0.999(0.001)	0.013(0.011)
	GG	0.640(0.035)	0.998(0.001)	0.023(0.015)
	tTEST1	0.495(0.033)	0.966(0.005)	0.347(0.037)
	tTEST2	0.007(0.009)	1.000(0.000)	0.014(0.075)
p=0.189(0.008)	MRFGG	0.743(0.067)	0.997(0.003)	0.018(0.014)
	GG	0.664(0.027)	0.996(0.002)	0.023(0.012)
	tTEST1	0.495(0.029)	0.966(0.005)	0.229(0.029)
	tTEST2	0.010(0.010)	1.000(0.000)	0.009(0.041)
p=0.357(0.009)	MRFGG	0.793(0.037)	0.991(0.006)	0.020(0.011)
	GG	0.698(0.020)	0.990(0.004)	0.024(0.008)
	tTEST1	0.497(0.020)	0.966(0.005)	0.110(0.017)
	tTEST2	0.019(0.012)	1.000(0.000)	0.008(0.023)
p=0.486(0.008)	MRFGG	0.835(0.036)	0.975(0.011)	0.030(0.012)
	GG	0.718(0.018)	0.982(0.006)	0.025(0.008)
	tTEST1	0.496(0.017)	0.966(0.006)	0.068(0.012)
	tTEST2	0.026(0.014)	1.000(0.001)	0.011(0.022)

1668 genes in two conditions, having three replicates in each condition. We took model parameters similar to those in Newton *et al.* ($\alpha = 10, \alpha_0 = 0.9$ and $v = 0.5$). Simulations were repeated 100 times to assess the sensitivity, specificity, and false discovery rates of the proposed MRF GG model (MRFGG). We used the conditional probability (2) and the conditional likelihood (3) in our analysis. As a comparison, we also performed analysis on the simulated data sets using the standard two sample t-test which doesn't consider any prior information at all, and the empirical Bayesian GG model of Kendziorski *et al.* (2003).

The results over 100 replications are presented in Table 1, where the sensitivity is calculated as the average over the 100 replications of the fraction of DE genes correctly identified by the method; specificity is the average of the EE genes correctly identified; and the false discovery rate (FDR) is the average of the ratio of the number of false positives to the number of the genes identified as DE. For t-tests, a cut-value of 0.05 was used for declaring a gene to be the DE. We observed that overall specificity is high for all three procedures and the MRFGG model resulted in higher sensitivity than the GG model while the FDRs are similar. As expected, using a p-value of 0.05 can result in substantially higher FDRs. On the other hand, if the FDR controlling procedure of Benjamini and Hochberg (1995) was used, the two sample t-test resulted in very low sensitivity. The gain in sensitivity over the GG model is greater when there are more DE genes. These results demonstrated that by incorporating the network structure information, we can indeed gain sensitivity in identifying the DE genes.

APPLICATION TO REAL DATASETS

We present results from application of the proposed methods to two breast cancer microarray gene expression studies in order to identify the subnetworks that are related to breast cancer metastasis or survival from breast cancer. We used the modified conditional probability (4) and the corresponding conditional likelihood function in the following analyses.

Application to the breast cancer gene expression dataset of Wang *et al.*

Wang *et al.* (2005) reported a large Affymetrix-based gene expression profiling for 286 patients with lymph-node-negative primary breast cancer. These patients were treated between 1980-1995 with age at surgery ranging 26-86 years and a median age at surgery of 52 yrs. No patient received any adjuvant therapy. During the follow-up period, 179 of these patients were relapse-free at 5 yrs, and 107 of them developed distant metastasis. Gene expression profiling using Affymetrix HG-133A was performed on all these patients, including 17,819 transcripts that were present in two or more samples. We merge the gene expression data with the 33 KEGG regulatory pathways and identified 1533 genes on the U133A array that can be found in the 1668-node KEGG network with 8011 edges. Our goal is to identify which genes and which subnetworks of the KEGG network of 33 pathways are related to breast cancer metastasis.

Two-sample t-tests identified only 8 DE genes for FDR of 0.05 using the Benjamini and Hochberg's procedure. As a comparison, our proposed procedure identified 72 DE genes. The parameter estimates were $\gamma_0 = 2.64, \gamma_1 = 0.71$ and $\beta = 1.24$ in the MRF model. Figure 1 shows 17 of these genes that are mapped to the KEGG pathways, where the largest connected subnetwork includes six genes on the Cytokine-cytokine receptor interaction pathway. This

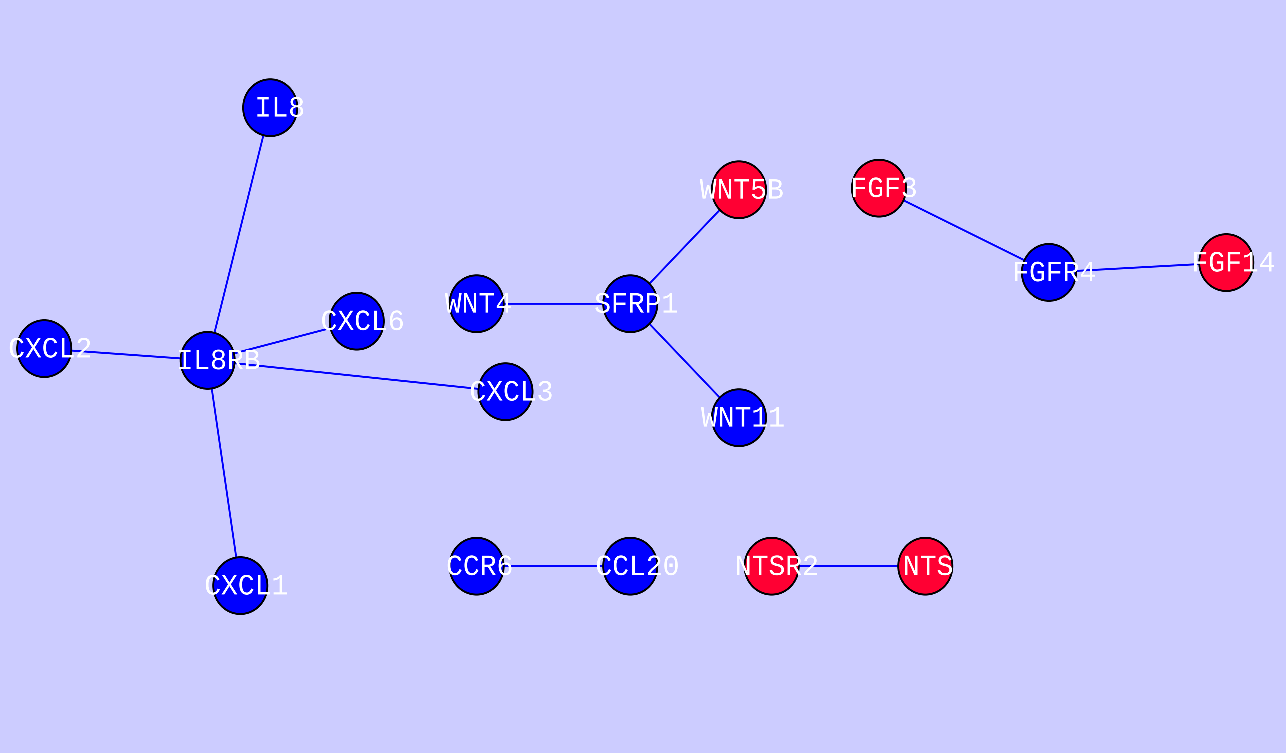

Fig. 1. Results from analysis of gene expression dataset of Wang *et al.* (2005). DE genes identified by the MRFGG method linked to the KEGG pathways, where genes in red are over-expressed and those in blue are under-expressed in cancer cells with breast cancer metastasis.

subnetwork, centered around interleukin 8 receptor Beta (IL8RB), is down-regulated in cancer with relapse. In addition, the chemokine receptor (CCR6) and chemokine (C-C motif) ligand 20 (CCL20) are also down-regulated in cancers with relapse, indicating that a chemokine pathway is down-regulated in cancers with relapse. In addition, CCL20/CCR6 involvement in the neoplastic progression and metastatic spread was reported in several tumor types (Rubie *et al.*, 2006), including breast cancer metastasis (Muller *et al.*, 2001).

Another subnetwork includes 4 genes on the Wnt signaling pathway, including WNT4, WNT11 and secreted frizzled-related protein 1 (SFRP1), which were down-regulated in cancer with metastasis. SFRPs are secreted Wnt antagonists that directly interact with the Wnt ligand to inhibit signaling and members of the SFRP class bind directly to Wnts, thereby altering their ability to bind to the Wnt receptor complex. In particular, the SFRP1 gene is found at chromosome 8p21, a site of frequent loss of heterozygosity in human tumors and is down-regulated in cervical carcinoma, breast carcinoma and ovary and kidney carcinomas (Shulewitz *et al.*, 2006). Wnt5b partially inhibits the canonical Wnt/beta-catenin signaling pathway. These findings agree with current knowledge of the involvement of the Wnt signaling pathway in breast cancer progression (Barker and Clevers, 2006).

We also found that the fibroblast growth factor 3 (FGF-3) and fibroblast growth factor 14 (FGF-14) are up-regulated in breast cancers with metastasis, while the fibroblast growth factor receptor-4 (FGFR4) is down-regulated. These three genes are connected on the MAPK signaling pathway. The four closely related human FGFRs and their more than 20 known ligands control a multitude of cellular processes, including cell growth, differentiation, and migration, and it has been shown that the FGF/FGFR system plays a critical role in cancer development due to its angiogenic potential or direct enhancement of tumor growth (Burke *et al.*, 1998).

Finally, recent study by Souaze *et al.* (2006) supports the contribution of neurotensin receptor (NTSR) in human breast cancer progression and pointed out the utility to develop therapeutic molecules targeting the neurotensin or NT1 receptor signaling cascade. We found that NTS and NTSR in the neuroactive ligand-receptor interaction pathway are linked and are both up-regulated in breast cancer with metastasis.

Application to the breast cancer gene expression dataset of Miller *et al.*

Miller *et al.* (2005) reported a gene expression profiling study of 251 primary breast cancer tissues resected in Uppsala County, Sweden from January 1, 1987 to December 31, 1989, using Affymetrix Chip HG-133A and HG-133B (GEO Accession No. GSE3494). The authors identified an expression signature for p53 which can be used for predicting the mutation status, transcriptional effects, and patient survival. Among these patients, 236 of them had follow-up information in terms of time and event of disease-specific survival. Different from the previous dataset, these patients included both lymph-negative and lymph-positive patients.

In single gene analysis using t-tests, we obtained only four genes that are DE for FDR=0.05 using the Benjamini and Hochberg's FDR proedure. The GG methods of Kendziorski *et al.* (2003) identified

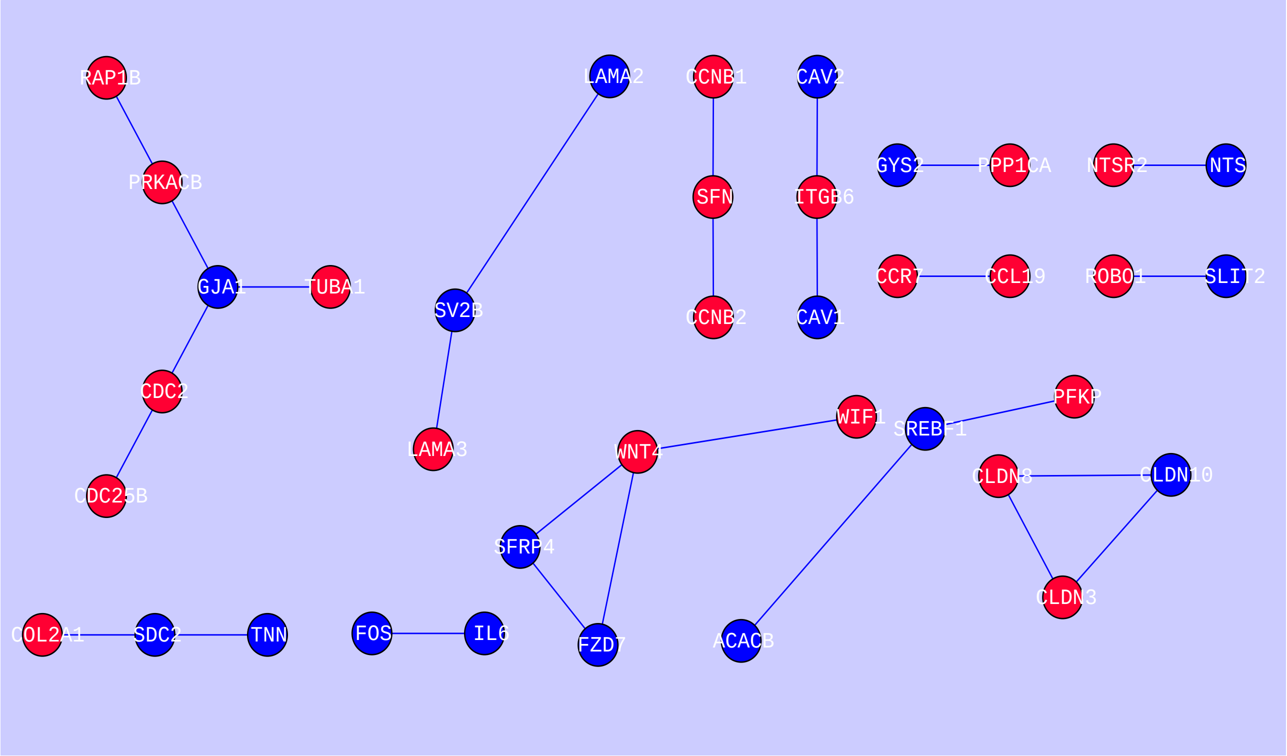

Fig. 2. Results from analysis of gene expression dataset of Miller *et al.* (2005). DE genes identified by the MRFGG method linked to the KEGG pathways, where genes in red are over-expressed and those in blue are under-expressed in cancer cells from the patients who died of breast cancer.

82 genes and our MRFGG method identified 103 genes. The parameter estimates were $\gamma_0 = 2.54, \gamma_1 = 0.38$ and $\beta = 0.55$ in the MRF model. Figure 2 shows several of the connected subnetworks that were identified by the MRFGG method. Similar to the previous example, we found the genes in the WNT pathway (WNT4, SFRP4 and FZD7), genes related to the chemokine pathway (CCR7 and CCL19) and neurotensin and its receptor (NTS and NTSR2) are related to survival from breast cancer. We also found that Caveolin-1, Caveolin-2 (CAV1 and CAV2) are down-regulated in cancers in patients who died of breast cancer. A recent study by Sagara *et al.* (2004) indicated that a reduced CAV1 mRNA level was significantly associated with increasing tumor size and negative estrogen receptor status. There was also a significant association between low CAV2 mRNA level and negative progesterone receptor status. Sagara *et al.* (2004) further indicated that CAV1 suppression correlated closely with that of CAV2 in breast cancer, that CAV1 level was inversely correlated with tumor size, and that CAV1 and CAV2 levels were correlated with hormonal receptor status. Therefore, CAV1 and CAV2 play an important role in tumor progression in breast cancer patients.

The largest subnetwork identified includes 6 genes on the GAP Junction pathway, of which five genes (CDC25B, CDC2, TUBA1, PRKACB and RAP1B) are up-regulated in cancer samples from individuals who died of cancer. Interestingly, the gap junction membrane channel protein alpha 1 (GJA1) is, however, down-regulated. GJA1 has been reported to suppress cell proliferation (Yu *et al.*, 2006) and therefore its down-regulation can lead to more cancer cell proliferation and increase the chance of death from breast cancer. In addition, we also observed over-expression of the tight junction proteins claudin-3 (CLDN3) and claudin-8 (CLDN8). Claudin-3 and claudin-4 are frequently over-expressed in several neoplasias, including ovarian, breast, pancreatic, and prostate cancers (Morin 2005; Hewitt *et al.*, 2006).

We also observed over-expression of genes related to Cyclin B1 (CCNB1) and B2 (CCNB2), together with over-expression of STRATIFIN (SFN) in cancer samples from individuals who died of cancer. These three genes are on the cell cycle pathway. Cyclins are a family of regulatory proteins that play a key role in controlling the cell cycle. Abnormalities of cell cycle regulators, including cyclins and cyclin-dependent kinases, have been reported in various malignant tumors. Zhao *et al.* (2006) observed significantly greater cyclin B1 expression in invasive cervical cancer than in normal cervical tissue and indicated that aberrant expression of cyclin B1 might play an important role in cervical carcinogenesis. In addition, CCNB1 expression was highly correlated with the labeling index for antigen identified by mAb ki067 (Ki067, associated with increased tumor cell proliferation), which suggests a key role for CCNB1 in the regulation of neuroendocrine tumor cell proliferation (Igarashi *et al.*, 2004; Lahad *et al.*, 2005).

Other genes related to breast cancer survival include genes involving ECM-receptor interaction and cell communication pathway (COL2A1,SDC2, TNN, LAMA2, LAMA3 and SV2B), and genes in the insulin signaling pathway (GYS2,PPP1CAACACB, SREBF1 and PFKP).

CONCLUSION AND DISCUSSION

We have proposed a MFR-based procedure that uses information of interaction networks in identification of DE genes. The proposed method utilizes the structure information of the interaction networks in order to capture the dependency of differential expression patterns for genes on the network. By doing so, we can expect to obtain results that are not found by single-gene analysis. Simulation studies and application to several real microarray gene expression data sets demonstrated that our methods are more sensitive in identifying the DE genes than some of the commonly used methods while maintaining low false discovery rates. Results from analysis of two breast cancer microarray gene expression datasets identified several subnetworks that are related to breast metastasis or death from breast cancer. Some of the subnetworks were reported in the literature.

We make several assumptions for the proposed methods. First, our proposed methods depend on the reliability of the structure of the interaction networks. We used KEGG interaction networks made of 33 regulatory pathways in our analysis of the breast cancer gene expression data. The edges of the networks include both protein-protein and DNA-protein interactions. Our methods treat all interactions equally, regardless of type and direction. However, if different types of interactions can be clearly defined, we can modify our method to allow for different dependency parameters for different interaction types. Important future research is to refine network structures and to extend our MFR models to more complex and refined network structures. Second, we used the Gamma-Gamma model of Newton *et al.* (2003) and Kendziorski *et al.* (2003) for modeling the gene expression data. Alternatively, one can assume a log-normal-normal model for the gene expression data. However, as shown in Kendziorski *et al.* (2003), the Gamma-Gamma model is quite robust to model misspecification. Third, in our model formulation, for each gene, we only consider its immediate neighbors on the network as its neighbors (i.e., first degree neighbors). However, if the differential expression patterns are dependent in neighbors centered at this gene with a radius r, we may want to include as its neighbors all the genes in this ball. This can potentially increase the sensitivity of identifying more DE genes.

In this paper, we have focused on the problem of identifying the differentially expressed genes between two experimental conditions. The MRF-methods can however be extended in several ways. First, it can be easily extended for identifying genes that show differential expressions among multiple groups using replicated gene expression profiles following the parametric empirical Bayes setup of Kendziorski *et al.* (2003). Second, the methods can also be extended to deal with other phenotypes such as continuous or censored survival phenotypes by considering whether a gene is related to the phenotype as a latent state and using the MFR for modeling such latent states. Third, the methods can also be extended to microarray time course gene expression data in order to identify subnetworks that change their expression states during a biological time course such as cancer initiation and progression. We are currently working on these extensions. Finally, important future research will include how to represent and assess the uncertainly of the inference of the true differential expression states x^*.

In summary, we have proposed a Markov random field model for identifying differentially expressed genes between two experimental conditions in order to utilize the network structure information. As more and more networks become available, we expect more applications of such methods for identifying genes and pathways that are related to various phenotypes.

ACKNOWLEDGMENTS

This research was supported by NIH grant ES009911 and a grant from the Pennsylvania Department of Health. We thank Mr. Edmund Weisberg, MS at Penn CCEB for editorial assistance and two referees for helpful comments.

REFERENCES

Barker N and Clevers H (2006): Mining the Wnt pathway for cancer therapeutics. *Nature Reviews Drug Discovery*, 5: 997-1014.

Benjamini Y and Horchberg Y (1995): Controling the false discovery rate: a practical and pwoerful approach to multiple testing. *Journal of the Royal Statistical Society*, Ser. B 57: 289-300.

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

Burke D, Wilkes D, Blundell TL, Malcolm S (1998): Fibroblast growth factor receptors: lessons from the genes. *Trends in Biochemical Sciences*, 23: 59-62.

Deng MH, Chen T, Sun FZ (2004): Integrated probabilistic model for functional prediction of proteins. *Journal of Computational Biology*, 11:463-476

Deng MH, Zhang K, Mehta S, Chen T and Sun FZ (2002): Prediction of protein function using protein-protein interaction data. *The first IEEE Computer Society Bioinformatics Conference, CSB2002*, 117-126.

Dudoit S, Yang YH, Callow MJ and Speed TP (2002): Statistucal methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12: 111-139.

Efron B, Tibshirani R, Story JD, and Tushe V (2001): Empirical Bayes Analysis of Microarray Experiment *Journal of the American Statistical Association*, 96: 1151-1160.

The Gene Ontology Consortium (2000): Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25: 25-29.

Hewitt KJ, Agarwal R and Morin PJ (2006): The claudin gene family: expression in normal and neoplastic tissues. *BMC Cancer*, 6: 186.

Igarashi T, Jiang S, Kameya T, Asamura H, Sato Y, Nagai K and Okayasu1 I (2004): Divergent cyclin B1 expression and Rb/p16/cyclin D1 pathway aberrations among pulmonary neuroendocrine tumors. *Modern Pathology*, 17: 12591267,

Kendziorski CM, Newton MA, Lan H and Gould MN (2003): On paramertic empirical Bayes methods for comparing multiple groups using replicated gene expressionm profiles. *Statistics in Medicine*, 22: 3899-3914.

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

Lahad JP, Mills JB and Coombes KR (2005):Stem cellness: a ”magic marker” for cancer. *Journal of Clinical Investigation*, 115: 1463-1467.

Letovsky S and Kasif S (2003): Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(Suppl. 1): i197-i204.

Lönnstedt I and Speed T (2002): Replicated microarray data. *Statistica Sinica*, 12:31-46.

Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu E and Bergh J (2005): An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of National Academy of Sciences*, 102: 13550-13555.

Morin PJ (2005): Claudin Proteins in Human Cancer: Promising New Targets for Diagnosis and Therapy. *Cancer Research*, 65: 9603-9606.

Muller A, Homey B, Soto H, *et al.* (2001): Involvement of chemokine receptors in breast cancer metastasis. *Nature*, 410: 50-56.

Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW (2001): On differntial variability of expression ratios: improving statistical inference abou gene expression changes from micorarray data. *Journal of Computational Biology*, 8: 37-52.

Nacu S, Critchley-Thorne R, Lee P and Holmes S (2006): Gene expression network analysis, and applications to immunity. *Technical report, Department of Statistics, Stanford Univerisity*.

Rahnenführer J, Domingues F, Maydt J, Lengauer T (2004): Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 3:Article 16.

Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP.: Classification of microarray data using gene networks. BMC Bioinformatics. 2007 Feb 1;8:35.

Rubie C, Oliveira-Frick V, Rau V and Schilling M (2006): Chemokine receptor CCR6 expression in colorectal liver metastasis. *Journal of Clinical Oncology*, 24: 5173-5174.

Sagara Y, Mimori K, Yoshinaga K, Tanaka F, Nishida K, Ohno S, Inoue H and Mori M (2004): Clinical significance of Caveolin-1, Caveolin-2 and HER2/neu mRNA expression in human breast cancer. *British Journal of Cancer*, 91: 959-965.

Segal E, Wang H and Koller D (2003): Discovering Molecular Pathways from Protein Interaction and Gene Expression Data. *Bioinformatics*, 19(Suppl 1): 264-272.

Sivachenko A, Yuriev A, Daraselia N, Mazo I (2005): Identifying Local Gene Expression Patterns in Biomolecular Networks. *Proceedings of 2005 IEEE Computational Systems Bioinformatics Conference*, Stanford, California.

Shulewitz M, Soloviev I, Wu T, Koeppen H, Polakis P and Sakanaka C (2006): Repressor roles for TCF-4 and Sfrp1 in Wnt signaling in breast cancer. *Oncogene*, 25(31): 4361-9.

Souaze F, Dupouy S, Viardot-Foucault V, Bruyneel E, Attoub S, Gespach C, Gompel A and Forgez P (2006): Expression of neurotensin and NT1 receptor in human breast cancer: a potential role in tumor progression. *Cancer Research*, 66(12): 6243-9.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005): Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of National Academy of Sciences, U S A.*, 102(43):15545-50.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane I and Park P (2005): Discovering statistically significant pathways in expression profiling studies. *Proceedings of National Academy of Sciences*, 103: 13544-13549.

Tusher V, Tibshirani R, Gross V and Chu G (2001): Significance analyusis of miocrarrays applied to ionizing radiation response. *Proceedings of National Academy of Sciences*, 98:5116-5121.

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D and Foekens JA (2005): Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365: 671-9.

Yu K, Ganesan K, Miller LD and Tan P (2006): A modular analysis of breast cancer reveals a novel low-grade molecular signature in estrogen receptorpositive tumors. *Clinical Cancer Research*, Vol. 12, 3288-3296.

Zhao M, Kim YT, Yoon BS, Kim SW, Kang MH, Kim SH, Kim JH, Kim JW and Park YW (2006): Expression profiling of cyclin B1 and D1 in cervical carcinoma. *Experimental Oncology*, 28(1):44-8.