BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 9 2008, pages 1175–1182
doi:10.1093/bioinformatics/btn081

Systems biology

Network-constrained regularization and variable selection for analysis of genomic data

Caiyan Li and Hongzhe Li*

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

Received on December 20, 2007; revised on February 17, 2008; accepted on February 27, 2008

Advance Access publication March 1, 2008

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Graphs or networks are common ways of depicting information. In biology in particular, many different biological processes are represented by graphs, such as regulatory networks or metabolic pathways. This kind of a priori information gathered over many years of biomedical research is a useful supplement to the standard numerical genomic data such as microarray gene-expression data. How to incorporate information encoded by the known biological networks or graphs into analysis of numerical data raises interesting statistical challenges. In this article, we introduce a network-constrained regularization procedure for linear regression analysis in order to incorporate the information from these graphs into an analysis of the numerical data, where the network is represented as a graph and its corresponding Laplacian matrix. We define a network-constrained penalty function that penalizes the L_1-norm of the coefficients but encourages smoothness of the coefficients on the network.

Results: Simulation studies indicated that the method is quite effective in identifying genes and subnetworks that are related to disease and has higher sensitivity than the commonly used procedures that do not use the pathway structure information. Application to one glioblastoma microarray gene-expression dataset identified several subnetworks on several of the Kyoto Encyclopedia of Genes and Genomes (KEGG) transcriptional pathways that are related to survival from glioblastoma, many of which were supported by published literatures.

Conclusions: The proposed network-constrained regularization procedure efficiently utilizes the known pathway structures in identifying the relevant genes and the subnetworks that might be related to phenotype in a general regression framework. 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@mail.med.upenn.edu

1 INTRODUCTION

A central problem in genomic research is to identify genes and pathways involved in diseases and other biological processes and to build a prediction model for future outcomes by linking high-dimensional genomic data, such as microarray gene-expression data, to various clinical outcomes. The problem can in general be formulated as a prediction problem with n observations having outcomes $y_1, y_2, \ldots, y_n$ and p predictors x_{ij}, $i=1,\ldots,n$, $j=1,\ldots,p$. The outcome can be quantitative or binary, representing two cases such as 'diseased' and 'healthy'. Consider the usual linear-regression model where the response $\mathbf{y}$ is predicted by

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1 + \cdots + \mathbf{x}_p\hat{\beta}_p, \quad (1)$$

where a model-fitting procedure produces the vector of coefficients $\hat{\beta} = (\hat{\beta}_0, \cdots, \hat{\beta}_p)$. To deal with the problem of high-dimensionality of the genomic data, many new regularized methods have been developed for identifying the genes that are related to clinical phenotypes in regression frameworks, including lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), fused lasso (Tibshirani *et al.*, 2005) and LARS (Efron *et al.*, 2005), and various extensions such as adaptive lasso (Zou, 2006) and group lasso (Yuan and Lin, 2006). Among these procedures, the elastic net regularization and the fused-lasso are particularly appropriate for analysis of genomic data, where the former encourages a grouping effect and the latter often leads to smoothes of the coefficient profiles for ordered covariates.

One limitation of all these popular approaches is that the methods are developed purely from computational or algorithmic points without utilizing any prior biological knowledge or information. For many complex diseases, especially for cancers, much biological knowledge or pathway information is available from many years of intensive biomedical research. The large body of information is now available primarily through databases on different aspects of biological systems. Such databases are often called metadata, which means data about data. Some well-known pathway databases include KEGG, Reactome (www.reactome.org), BioCarta (www.biocarta.com) and BioCyc (www.biocyc.org). Of particular interest are gene-regulatory pathways that provide regulatory relationships between genes or gene products. These pathways are often interconnected and form a network, which can be represented as graphs, where the vertices of the graphs are genes or gene products and the edges of the graphs indicate some regulatory relationship between the genes. This kind of a priori

*To whom correspondence should be addressed.

information is a useful supplement to the standard numerical data coming from an experiment. Incorporating the information from these graphs into an analysis of the numerical data is a non-trivial task that is generating increasing interest. Several statistical methods have been developed to utilize the pathways or network information, including the hidden Markov-random field approaches to utilize the network structures in identifying the differentially expressed genes (Wei and Li, 2007, 2008; Wei and Pan, 2008). Rahnenführer *et al*. (2004) demonstrated that the sensitivity of detecting relevant pathways can be improved by integrating information about pathway topology. However, none of these methods were developed in the framework of regression analysis.

In this article, we propose to develop a network-constrained regularization procedure for fitting linear-regression models and for variable selection, where the predictors in the regression model are genomic data with graphical structures. The goal of such a procedure is to identify genes and subnetworks that are related to diseases or disease outcomes. In order to achieve automatic variable selection and to account for the network structures, we define a network-constrained penalty that is a combination of the lasso penalty and a penalty induced by the Laplace matrix of the graph. Such a procedure can select subgroups of correlated features in the network, thus enjoying global smoothness over the network. Our proposed procedure, which includes the elastic net regulation procedure as a special case, is similar in spirit to the fused-lasso (Tibshirani *et al*., 2005). It induces smoothed coefficient profiles, which can result in more interpretable identification of genes and subnetworks that are related to the responses in the context of known biology. However, it is different from fused-lasso in that our procedure does not requires that the neighboring genes to have the same coefficients and the network-structure is explicitly modeled using the Laplacian matrix of the graph.

The rest of the article is organized as follows. We first define the network-constrained regularization procedure for linear-regression models and present an efficient algorithm for estimating the parameters. We then provide the grouping property and the asymptotic theorem for the parameter estimates and simulation results. We then present an application of the proposed methods to an analysis of a microarray gene-expression dataset of glioblastoma. Finally, we present a brief discussion of the results.

2 NETWORK-CONSTRAINED REGULARIZATION FOR LINEAR MODELS

Suppose that the dataset contains n observations and p predictors, with response vector $\mathbf{y}=(y_1,\ldots,y_n)^T$ and design matrix $\mathbf{X}=(\mathbf{x}_1|\ldots|\mathbf{x}_p)$, where $\mathbf{x}_j=(x_{1j},\ldots,x_{nj})^T, j=1,\ldots,p$. We also assume that the predictors are standardized and the response is centered so that

$$\sum_{i=1}^{n} y_i = 0, \sum_{i=1}^{n} x_{ij} = 0 \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1 \text{ for } j = 1,\ldots,p.$$

Consider a network that is represented by a weighted graph $G=(V, E, W)$, where V is the set of vertices that correspond to the p predictors, $E=\{u\sim v\}$ is the set of edges indicating that the predictors u and v are linked on the network and there is an edge between u and v and W is the weights of the edges, where $w(u,v)$ denotes the weight of edge $e=(u\sim v)$. In applications, the edge weight can be used to measure uncertainty of the edge between two vertices. Define the degree of the vertex v as $d_v=\sum_{u\sim v} w(u,v)$. We say u is an isolated vertex if $d_u=0$. Following Chung (1997), we define the normalized Laplacian matrix L for G with the uvth element defined by

$$L(u,v)=\begin{cases} 1-w(u,v)/d_u & \text{if } u=v \text{ and } d_u\neq 0,\\ -w(u,v)/\sqrt{d_u d_v} & \text{if } u \text{ and v are adjacent,}\\ 0 & \text{otherwise.}\end{cases}$$

This matrix L is always non-negative definite and its corresponding set of the eigenvalues or spectrum reflects many properties of the graph (Chung, 1997).

For any fixed non-negative λ_1 and λ_2, we define the network-constrained regularization criterion

$$L(\lambda_1,\lambda_2,\beta)=(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda_1|\beta|_1+\lambda_2\beta^T L\beta, \quad (2)$$

where $|\beta_1|=\sum_{j=1}^{p}|\beta_j|$ is the L_1-norm, which induces a sparse solution (Tibshirani, 1996), and the second term $\beta^T L\beta$ induces a smooth solution of β on the network. Note that L is non-negative definite and can be written as $L=SS^T$, where $S_{p\times m}$ is the matrix in which rows are indexed by the vertices and in which columns are indexed by the edges of G such that each column corresponding to an edge $e=\{u,v\}$ has an entry $\sqrt{w(u,v)}/\sqrt{d_u}$ in the row corresponding to u, an entry $-\sqrt{w(u,v)}/\sqrt{d_v}$ in the row corresponding to v and has zero entries elsewhere. Based on simple algebra, we can see that $\beta^T L\beta$ can be written as

$$\beta^T L\beta=\sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}}-\frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v),$$

where $\sum_{u\sim v}$ denotes the sum over all unordered pairs $\{u,v\}$ for which u and v are adjacent on the network. Equation (2) can then be rewritten as

$$L(\lambda_1,\lambda_2,\beta)=(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}}-\frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v), \quad (3)$$

and we define the network-constrained regularized estimator $\hat{\beta}$ as the minimizer of Equation (3), i.e.

$$\hat{\beta}=\text{argmin}_\beta\{L(\lambda_1,\lambda_2,\beta)\}. \quad (4)$$

Let $\alpha=\lambda_2/(\lambda_1+\lambda_2)$, then $\hat{\beta}$ in Equation (4) is equivalent to the solution to the optimization problem

$$\hat{\beta}=\text{argmin}_\beta\{|\mathbf{y}-\mathbf{X}\beta|^2\},$$
$$\text{subject to } (1-\alpha)\sum_{j=1}^{p}|\beta_j|+\alpha\sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}}-\frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v)\le t$$

for some t. We call the function

$$(1-\alpha)\sum_{j=1}^{p}|\beta_j|+\alpha\sum_{u\sim v}\{(\beta_u/\sqrt{d_u}-\beta_v/\sqrt{d_v})^2 w(u,v)\}$$

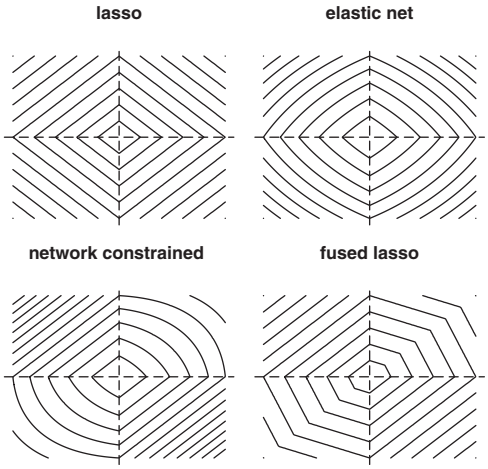

Fig. 1. Contours for four penalty functions for a bivariate argument $\beta=(\beta_1,\beta_2)$. The upper left shows contours of the lasso penalty. The upper right shows contours of the elastic net penalty. The lower left shows the contours of the network-constrained penalty and the lower right shows the contours of the fused lasso penalty, both for $\alpha=0.3$.

the network-constrained penalty, in which the second term imposes smoothness of the parameters β over the network via penalizing the weighted sum of squares of the scaled difference of the coefficients between neighbor vertices in the network. We re-scale the β coefficients in order to account for different degrees of the vertices on the network, allowing the genes with more connections (e.g. the hub genes) to have larger coefficients so that small changes of expressions of such genes can lead to large changes in the response. The biological motivation of this penalty is that we expect the genes that are linked on the networks to have similar functions and therefore smoothed-regression coefficients. Note that we do not require these coefficients to be the same or have the same signs. If the weight $w(u,v)$ represents the probability that vertices u and v are connected, we impose smoothness over these two vertices with probability $w(u,v)$. This provides one way of accounting for uncertainty of the network.

Note that when $\alpha=0$, the network-constrained penalty reduces to the lasso, a singular penalty function at zero and for all $\alpha\in(0,1)$, it is strictly convex, and hence retains the good properties of both sparsity and smoothness. When $L=I$, the network-constrained penalty becomes the elastic net penalty of Zou and Hastie (2005). Figure 1 shows contours for four penalty functions for a bivariate argument $\beta=(\beta_1,\beta_2)$, where $\alpha=0.3$ for the elastic net, fused lasso and the network-constrained penalties. Like the fused lasso penalty, one important feature of the network-constrained penalty is that it is not symmetric over the x-axis or y-axis; therefore, β parameters of different signs will have different penalties.

2.1 Solution and algorithm

Following Zou and Hastie (2005), we develop a similar efficient computation procedure to solve the network-constrained regularization problem. As shown in the following lemma, minimizing Equation (3) is equivalent to solving a lasso-type optimization problem, thus enjoying the computational advantage of the lasso.

LEMMA 1. *Given dataset* $(\boldsymbol{y}, \boldsymbol{X})$ *and two fixed scalars* (λ_1, λ_2), *define an artificial dataset* $(\boldsymbol{y}^*, \boldsymbol{X}^*)$ *by*

$$\boldsymbol{X}^*_{(n+p)\times p}=(1+\lambda_2)^{-1/2}\begin{pmatrix}\boldsymbol{X}\\ \sqrt{\lambda_2}\boldsymbol{S}^T\end{pmatrix},\ \boldsymbol{Y}^*_{(n+p)}=\begin{pmatrix}\boldsymbol{Y}\\ \boldsymbol{0}\end{pmatrix},$$

where $L=U\Gamma U^T$ *and* $S=U\Gamma^{1/2}$. *Let* $\gamma=\lambda_1/\sqrt{1+\lambda_2}$ *and* $\beta^*=\sqrt{1+\lambda_2}\beta$. *Then the network-constrained criterion can be written as*

$$L(\lambda_1,\lambda_2,\beta)=L(\gamma,\beta^*)=(\boldsymbol{y}^*-\boldsymbol{X}^*\beta^*)^T(\boldsymbol{y}^*-\boldsymbol{X}^*\beta^*)+\gamma\sum_{j=1}^{p}|\beta^*_j|$$

Let $\hat{\beta}^*$ *be the solution to the above lasso problem, i.e,*

$$\hat{\beta}^*=\mathrm{argmin}_{\beta^*}\{L(\gamma,\beta^*)\};$$

then the solution to (3) becomes

$$\hat{\beta}=\frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*. \tag{5}$$

Following Zou and Hastie (2005), to correct for potential bias due to double shrinkage, we adjust the network-constrained estimate $\hat{\beta}$ by a factor $1+\lambda_2$. Lemma 1 indicates that the network-constrained penalty problem can be reformulated as an equivalent lasso-type problem by creating an augmented dataset, thus enjoying the automatic variable selection property. Note that this augmented dataset increases the sample size from n to $(n+p)$, which means that this model can potentially select all p variables even when $n\ll p$. Similar to the elastic net, this feature overcomes the limitation that lasso can select at most n (when $n<p$) variables before it saturates. In the next section we will show that the network-constrained criterion can perform the grouped variables selection procedure in a fashion similar to the elastic net.

Finally, if only training samples are available, 10-fold cross-validation (CV) can be used for estimating the prediction error and for comparing models. For each fixed λ_2, we can use the number of steps for the lasso solution of the optimization problem (1) as the second tuning parameter besides λ_2, which is selected by 10-fold CV. The chosen λ_2 is the one giving the smallest CV error.

3 PROPERTIES OF THE PROPOSED PROCEDURE

We present several properties related to the proposed network-constrained regularization procedure, including the grouping effect and the asymptotic property in the case when p is fixed and $n\to\infty$.

3.1 The grouping effect

We show in this section that the estimates of network-constrained regularization can lead to desirable grouping effects for predictors that are correlated or linked on the network. The following Lemma, which is the direct result from the Lemma 2 of Zou and Hastie (2005) since the network-constrained loss function is a convex function, guarantees the

grouping effect for network-constrained penalization regression in the situation with identical predictors.

LEMMA 2. *Assume that* $\hat{\beta}$ *is determined by equation (5), also assume that* $\boldsymbol{x}_i = \boldsymbol{x}_j$, *then* $\hat{\beta}_i = \hat{\beta}_j$, *for any* $\lambda_2 > 0$.

If we consider the simple case when the two genes are linked only to each other on the network, the following theorem provides an upper bound on the difference of the estimates from the network-regularization procedure.

THEOREM 1. *Given dataset* $(\boldsymbol{y}, \boldsymbol{X})$ *and two fixed scalars* (λ_1, λ_2), *the response* $\boldsymbol{y}$ *is centered and predictors* $\boldsymbol{X}$ *are standardized. Let* $\hat{\beta}(\lambda_1, \lambda_2)$ *be the solution to Equation (4). Suppose that* $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$, *and the two vertices u and v are only linked to each other on the network,* $d_u = d_v = w(u,v)$. *Define*

$$D_{\lambda_1, \lambda_2}(u,v) = \frac{1}{|\mathbf{y}|_1}|\hat{\beta}_u(\lambda_1, \lambda_2) - \hat{\beta}_v(\lambda_1, \lambda_2)|,$$

then

$$D_{\lambda_1, \lambda_2}(u,v) \le \frac{1}{2\lambda_2}\sqrt{2(1-\rho)} \tag{6}$$

where $|\boldsymbol{y}|_1 = \sum_{i=1}^{n} |y_i|$ *and* $\rho = \boldsymbol{x}_u^T \boldsymbol{x}_v$ *is the sample correlation.*

The proof of this theorem is similar to that in Zou and Hastie (2005) and can be found in Li and Li (2007). The upper bound in (6) gives a quantitative description for the grouping effect of the network-constrained regularization, which is half of the upper bound in the elastic net model. In a pathway, for two adjacent vertices i and j satisfying $d_i = d_j = w(i,j)$, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly correlated, i.e., $\rho \doteq 1$, then the difference between the coefficient paths of features i and j is almost 0.

3.2 Asymptotic property

In this section, we derive asymptotic results for the estimates from network-constrained penalization under the assumption that p is fixed and the sample size $n \to \infty$. The result and proof is similar in spirit to the estimates based on the fused lasso (Tibshirani *et al.*, 2005). Consider the following linear-regression model,

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \cdots + \mathbf{x}_p\beta_p + \epsilon,$$

where ϵ is the error term of mean 0 and variance σ^2. For a given n *i.i.d.* observations, recall that the network-constrained penalized least squares criterion is

$$\sum_{i=1}^{n}(y_i - \mathbf{x_i}^T\beta)^2 + \lambda_n^{(1)}\sum_{j=1}^{p}|\beta_j| + \lambda_n^{(2)}\sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v),$$

where the Lagrange multipliers $\lambda_n^{(1)}$ and $\lambda_n^{(2)}$ are functions of the sample size n. We have the following asymptotic theorem for the estimates:

THEOREM 2. *If* $\lambda_n^{(l)}/\sqrt{n} \to \lambda_0^{(l)} \ge 0$ *for* $l = 1, 2$ *and*

$$C = \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right)$$

is non-singular, then

$$\sqrt{n}(\hat{\beta}_n - \beta) \to^d \operatorname{argmin}(V)$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0^{(1)}\sum_{j=1}^{p}\{u_j sgn(\beta_j)I(\beta_j \ne 0) + |u_j|I(\beta_j = 0)\} + 2\lambda_0^{(2)}\sum_{i\sim j}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_j}{\sqrt{d_j}}\right)\left(\frac{u_i}{\sqrt{d_i}} - \frac{u_j}{\sqrt{d_j}}\right)w(i,j),$$

and

$$\mathbf{W} \sim N(0, \sigma^2 C).$$

The proof of this theorem is can be found in Li and Li (2007). For the special case when $p = 2$ and $w(i,j) = 1$, it is easy to check that the estimates follow a bivariate normal distribution.

4 SIMULATION STUDIES

To demonstrate the performance of the proposed network-constrained regularization procedure, we first simulated the following simple regulatory network: suppose that we have 200 transcription factors (TFs) and each regulates 10 genes. The resulting network includes 2200 genes and edges between each of the TFs and the 10 genes that they regulate. We assume that four TFs and the genes that they regulated are related to response Y. For the first model, we assume that the data are simulated from the following models:

- $y = X\beta + \varepsilon$ and

$$\beta = (5, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{10}, -5, \underbrace{\frac{-5}{\sqrt{10}}, \ldots, \frac{-5}{\sqrt{10}}}_{10}, 3, \underbrace{\frac{3}{\sqrt{10}}, \ldots, \frac{3}{\sqrt{10}}}_{10}, -3, \underbrace{\frac{-3}{\sqrt{10}}, \ldots, \frac{-3}{\sqrt{10}}}_{10}, 0, \ldots, 0),$$

where $\varepsilon \sim N(0, \sigma_e^2)$.

- The expression levels for the 200 TFs follow standard normal, $X_{TF_j} \sim N(0,1)$
- The expression levels of the TF and the gene that it regulates are jointly distributed as a bivariate normal with a correlation of 0.7. This implies that conditioning on the expression level of the TF, the expression level of the gene it regulates, follows a $N(0.7 * X_{TF_j}, 0.51)$.

For the second model, the expression levels are simulated in the same way as for Model 1, except that we assume that

$$\beta = (5, \frac{-5}{\sqrt{10}}, \frac{-5}{\sqrt{10}}, \frac{-5}{\sqrt{10}}, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{7}, -5, \frac{5}{\sqrt{10}}, \frac{5}{\sqrt{10}}, \frac{5}{\sqrt{10}}, \underbrace{\frac{-5}{\sqrt{10}}, \ldots, \frac{-5}{\sqrt{10}}}_{7}, 3, \frac{-3}{\sqrt{10}}, \frac{-3}{\sqrt{10}}, \frac{-3}{\sqrt{10}}, \underbrace{\frac{3}{\sqrt{10}}, \ldots, \frac{3}{\sqrt{10}}}_{7}, -3, \frac{3}{\sqrt{10}}, \frac{3}{\sqrt{10}}, \frac{3}{\sqrt{10}}, \underbrace{\frac{-3}{\sqrt{10}}, \ldots, \frac{-3}{\sqrt{10}}}_{7}, 0, \ldots, 0)$$

Table 1. Results of the simulation study, sensitivity, specificity and the prediction mean-squared-errors (PMSE) are calculated based on 50 simulations, where standard errors are given in parentheses

	Sensitivity			Specificity			PMSE		
Model	Lasso	Enet	Net	Lasso	Enet	Net	Lasso	Enet	Net
1	0.482	0.471	1.00	0.996	0.996	0.906	90.2	77.0	46.9
	(0.06)	(0.06)	(0.00)	(0.002)	(0.002)	(0.04)	(17.4)	(14.7)	(7.3)
2	0.351	0.332	0.766	0.993	0.995	0.966	90.1	86.6	81.3
	(0.05)	(0.003)	(0.06)	(0.002)	(0.003)	(0.007)	(14.18)	(13.6)	(12.0)
3	0.504	0.668	1.00	0.996	0.993	0.909	34.4	32.9	27.5
	(0.11)	(0.13)	(0.00)	(0.002)	(0.002)	(0.004)	(6.67)	(6.41)	(4.37)
4	0.455	0.413	0.940	0.996	0.997	0.943	34.9	32.3	33.6
	(0.11)	(0.11)	(0.03)	(0.002)	(0.002)	(0.01)	(6.06)	(5.79)	(5.28)

Enet: the elastic net of Zou and Hastie (2005); Net: the proposed network-constrained regulation procedure.

This model assumes that genes that are regulated by the same TF can have both positive and negative effects on the response Y.

The third model is similar to Model 1, except that we replace the $\sqrt{10}$ in the denominators in β with 10. The fourth model is similar to Model 2, which assumes that genes that are regulated by the same TF can have both positive and negative effects on the response Y. For this model, we replace the $\sqrt{10}$ in the denominators in β with 10.

For each of these four models, the noise variance was chosen to be $\sigma_e^2 = (\sum_j \beta_j^2)/4$ so that the signal-to-noise ratio was 21.68, 7.34, 10.70 and 5.82 for Models 1, 2, 3 and 4, respectively. We simulated a training set and an independent test set with sample sizes of 100 for both sets. A 10-fold CV was conducted on the training dataset to select the tuning parameters and then the parameter estimates were obtained using all of the training dataset. For each model, we repeated the simulations 50 times. We then computed the prediction mean-squared error (PMSE) on the test dataset. In addition, we also calculated both the sensitivity and specificity for each procedure. Table 1 summarizes the simulation results for these four different models. For all four models, our proposed network-constrained procedure gave much smaller or comparable PMSEs than the lasso or elastic net regressions. The network-constrained procedure also resulted in much higher sensitivity in identifying the relevant genes. The specificity is somewhat reduced, but not greatly as compared to the gains in sensitivity.

5 APPLICATION TO ANALYSIS OF A MICROARRAY GENE-EXPRESSION DATASET GLIOBLASTOMA

We demonstrate the proposed methods by analyzing a microarray gene expression study ofglioblastoma by Horvath *et al.* (2006). Glioblastoma is the most common primary malignant brain tumor of adults and one of the most lethal of all cancers. Patients with this disease have a median survival of 15 months from the time of diagnosis despite surgery, radiation and chemotherapy. Global gene-expression data from two independent sets of clinical tumor samples of $n = 55$ and $n = 65$ were obtained by high-density Affymetrix arrays. The gene-expression datasets were normalized using the RMA methods (Irizarry *et al.*, 2003). Among the first set of 55 patients, five were alive at the last followup and four were alive for the second set. In our analysis, we built a predictive model using the first set of 50 patients with time to death information and tested the predictive performance using the second set of 61 patients with time to death information. We used the logarithm of time to death as the response variable in our analysis.

Table 2. Results from analysis of the glioblastoma dataset, where the test set mean-squared errors are calculated based on an independent set of 61 glioblastoma patients

Method	Test mean-squared error	Number of genes selected
lasso	1.18	23
elastic net	1.02	5
network-constraint	1.06	95

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 can be found in the 1668-node KEGG network of 33 pathways. Instead of considering all the genes on the Hu133A chip, we only focused analysis on these 1533 genes and aimed to identify which genes and which subnetworks of the KEGG network of 33 pathways are related to survival times from brain cancer. Table 2 shows the results from three different procedures in terms of prediction errors in the test datasets and the number of genes selected by these procedures in the training set. Both the elastic network and the network-constrained regularization procedures resulted in similar and slightly smaller prediction errors than lasso. However, the network-constrained procedure selected more genes than the lasso or elastic net, about half of these genes (44 genes) are connected on the KEGG pathways. As a comparison, the lasso identified three pairs of connected genes (ITGB7∼SYNJ2, PCK1∼ PTEN and FOXO1A∼PRKCG), and the elastic net identified only one pair of connected genes (PRKCG∼ ITGB7). These genes do not provide much information on which pathways/subnetworks might be related to survival from glioblastoma. Finally, the genes identified by the network-constrained procedure include all the genes identified by the elastic network and lasso.

Results from our network-constrained analysis indeed suggest that several pathways might be related to time to death from glioblastoma. Figure 2 shows the connected subnetworks of KEGG that were identified by the proposed network-constrained procedure. The largest subnetwork includes genes involving the MAPK signaling pathway (e.g. genes PLCE1, PRKCG, MAP2K7, ZAK, KBKG, TRAF2 and MAPK11) and its connected pathways, such as the PI3K/Akt signaling pathway (e.g. genes GYS1) and its target FOXO1A. Of particular interest is the identification of the FOXO1A that might be related to risk of death from glioblastoma. FOXO1A is

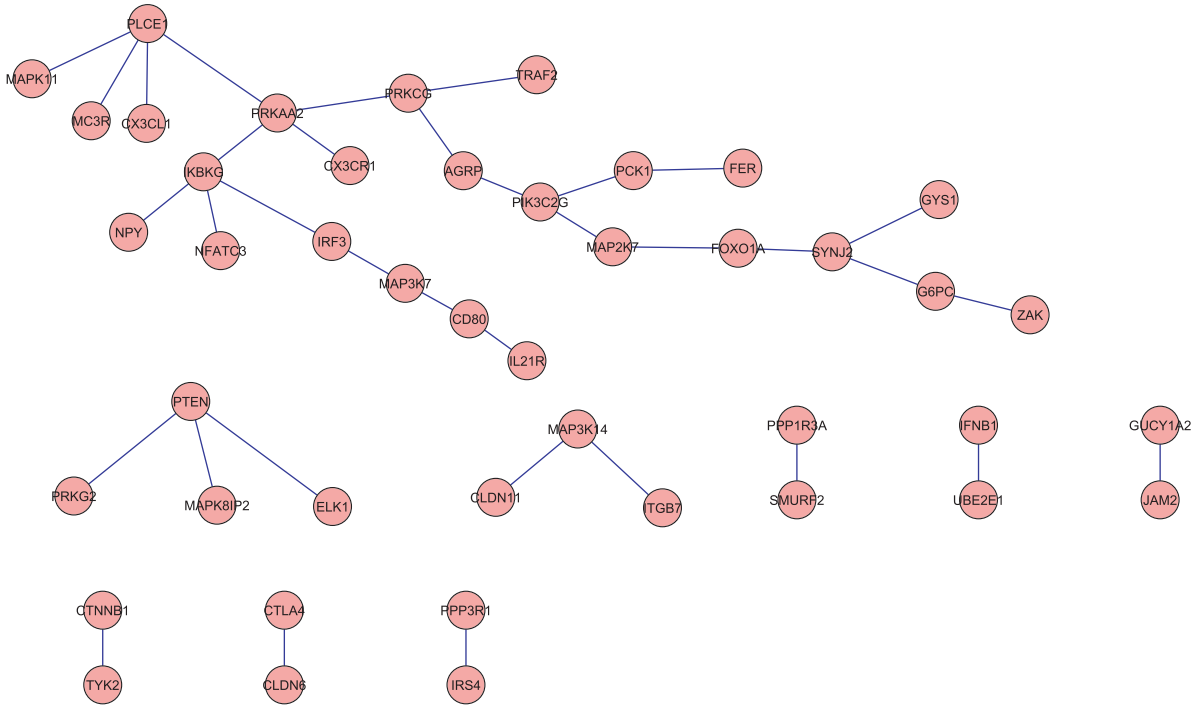

Fig. 2. Subnetworks identified by the network-constrained regulation method that might be related to survival time from glioblastoma based on a sample of 50 patients.

an important TF involved in the regulation of a range of critical processes in mammalian cells, including proliferation, differentiation, apoptosis, metabolism and responses to oxidative stress and DNA damage (Accili and Arden, 2004). The prognostic relevance of MAPK expression in glioblastoma multiforme was reported in Mawrin *et al.* (2003) and Pelloski *et al.* (2006).

The second subnetwork includes four genes, PTEN, PRKG2, MAPK8IP2 and ELK1. Li *et al.* (1997) describe a phosphatase and tensin homolog deleted on the chromosome 10 (PTEN) protein that is mutated in a number of human cancers including those from breast, brain and prostate. This protein interacts with actin filaments and is a putative protein tyrosine phosphatase, and acts as a tumor suppressor, at least in part, by antagonizing phosphoinositide 3-kinase (PI3K)/Akt signaling. Uht *et al.* (2007) suggested that PKC-eta-mediated glioblastoma proliferation involves MEK/mitogen-activated protein (MAP) kinase phosphorylation, activation of ERK and subsequently of Elk-1. The MAPK8IP2 (mitogen-activated protein kinase 8 interacting protein) is closely related to MAPK8IP1/IB1/JIP-1, a scaffold protein that is involved in the c-Jun amino-terminal kinase signaling pathway. This protein is expressed in brain and pancreatic cells and has been shown to interact with, and regulate the activity of MAPK8/JNK1 and MAP2K7/MKK7 kinases. This protein thus is thought to function as a regulator of signal transduction by protein kinase cascade in brain (Uht *et al.*, 2007). Finally, the gene PRKG2, encoding the cGMP-dependent protein kinase II, was targeted by insertions in brain tumors. Overexpression of PRKG2 in human glioma cell lines led to a reduction in colony formation, cell proliferation and migration (Uht *et al.*, 2007).

Among the small subnetworks of two genes, their involvement in glioblastoma has also been reported in the literature for some of the pairs. Perego *et al.* (2002) showed that the invasive behavior of glioblastoma cell lines is associated with altered organization of the cadherin-catenin adhesion system, where the catenin (cadherin-associated protein), beta 1 (CTNNB1) protein is a major component. Leach *et al.* (1996) suggested that a blockade of the inhibitory effects of CTLA-4 can allow for, and potentiate, effective immune responses against tumor cells. One reason for the poor immunogenicity of many tumors may be that they cannot provide signals for the CD28-mediated costimulation necessary to fully activate T cells. It has recently become apparent that CTLA-4, a second counter receptor for the B7 family of costimulatory molecules, is a negative regulator of T-cell activation. In addition, the family of more than 20 claudin (CLDN) proteins comprises one of the major structural elements within the apical tight junction apparatus, a dynamic cellular nexus for maintenance of a luminal barrier, paracellular transport, and signal transduction. Loss of normal tight junction functions constitutes a hallmark of human carcinomas. CLDN1 may support tumor suppressive functions in tissues such as the brain, where dramatic loss of expression has been demonstrated in glioblastoma multiforme (Swisshelma *et al.*, 2005).

In summary, these results indicate that by considering the KEGG pathways, our proposed methods can identify subnetworks that are potentially relevant to time to death from glioblastoma. Some of these subnetworks are well-supported by previously published work. In contrast, the genes identified by lasso or the elastic network cannot suggest the involvement of any possible pathways that are related to the risk of death from glioblastoma.

6 DISCUSSION

We have introduced a network-constrained regularization procedure for linear models in order to incorporate information coded in known genetic networks. Such a regularization procedure can also be regarded as a penalized least-squared estimation where the penalty is defined as a combination of the L_1 penalty and L_2 penalty on degree-scaled differences of coefficients between variables linked on the networks. Such a penalty induces both sparsity and smoothness with respect to the network structure of the regression coefficients. Our proposed network-constrained regularization procedure is similar in spirit to the fused lasso (Tibshirani *et al.*, 2005), both of which try to smooth the regression coefficients in certain ways. However, the fused lasso does not utilize prior genetic network information; instead, it first clusters genes to provide a gene order for the fusion process. Second, instead of using L_2-norm on the differences of the coefficients of the nearby genes, the fused lasso uses the L_1-norm on the differences, which tends to lead to the same regression coefficients for genes that are nearby. However, when the gene neighbors are defined by the prior network information, we should expect that the corresponding coefficients are similar but not the same. So for the settings that we consider in this article, it makes more sense to use the L_2-norm on the scaled coefficients in our definition of the network penalty. It is important to note that our proposed network-constraint regularization procedure does not require the coefficients of the genes that are linked on the network to have the same values or even the same signs. As shown in our simulations (Models 2 and 4), even when the coefficients of the neighboring genes are different, the proposed procedure still performs well in terms of the sensitivity and the prediction errors.

We used the normalized Laplacian L of the graph G in our definition of the smoothness penalty. Alternatively, one may use the combinatorial Laplacian of graph G (Chung, 1997), defined by

$$\mathcal{L}(u,v) = \begin{cases} d_u - w(u,u) & \text{if } u = v \text{ and } d_u \neq 0, \\ -w(u,v) & \text{if } u \text{ and } v \text{ adjacent}, \\ 0 & \text{otherwise}, \end{cases}$$

in the definition of the smoothness penalty. It is easy to verify that $\beta^T \mathcal{L}\beta = \sum_{u\sim v} (\beta_u - \beta_v)^2 w(u,v)$. This penalty may also make biological sense, however, it does not account for the variable degrees of the genes on the network. In addition, the matrix $\mathcal{L}$ is not always non-negative definite and cannot always be decomposed similarly as the L matrix in Lemma 1. The consequence of this fact is that the regularization problem cannot always be converted into an efficient lasso-type solution and some new optimization procedure such as the coordinate descent algorithm (Wu and Lange, 2008) has to be developed. It would be interesting to compare the performance of these two different definitions of the smoothness penalty.

In this article, we analyzed the glioblastoma gene-expression data using KEGG pathways and aimed to identify the KEGG pathways or subnetworks that are related to time to death from the cancer. 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 gene-expression 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/) or build the network of pathways using some existing network construction tools. 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. 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 method. Important future research will be to assess how sensitive the results are to misspecification of the network structures. Note that our proposed smoothness penalty is equivalent to imposing a graph-based Markov-random field prior on the regression coefficients. For the problem of identifying differentially expressed genes, recently studies have indicated that the results are not too sensitive to misspecification of the network structures (Wei and Li, 2007; Wei and Li, 2008; Wei and Pan, 2008). Since majority of the genes on the network are expected not to be associated with the response and therefore to have zero coefficients, we expect that only misspecification of the true response-related subnetworks will have great effects on the results. Finally, we presented the asymptotic property of the network-constrained estimates of the regression parameters for the scenario when p is fixed and $n \to \infty$. Interesting future research will be to derive the asymptotic property of the estimates when $p = p_n \to \infty$ as $n \to \infty$.

The proposed methods can be extended in several ways. First, the methods can be similarly extended to other types of response variables such as binary or survival responses. Second, many genetic networks are given by directed graphs. It is possible to extend our method to directed networks by using the Laplacian matrix for directed graphs (Chung, 1997) in our definition of the network-constraint penalty.

ACKNOWLEDGEMENTS

This research was supported by NIH grants ES009911, CA127334 and AG025532.

Conflict of Interest: none declared.

REFERENCES

Accili,D. and Arden,K.C. (2004) FoxOs at the crossroads of cellular metabolism, differentiation, and transformation. *Cell*, **117**, 421–426.

Chung,F. (1997) *Spectral Graph Theory*, Vol. 92 of *CBMS Reginal Conferences Series*. American Mathematical Society, Providence.

Efron,B. *et al.* (2004) Least angle regression. *Annals of Statistics*, **32**, 407–499.

Fan,J. and Li,R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**, 1348–1360.

Horvath,S. *et al.* (2006) Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a novel molecular target. *Proc. Natl Acad. Sci.*, **103**, 17402–17407.

Irizarry,R.A. *et al.* (2003) Exploration, normalizationsummaries of high-density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

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

Leach,D.R. *et al.* (1996) Enhancement of antitumor immunity by CTLA-4 blockade. *Science*, **271**, 1734–1736.

Li,C. and Li,H. (2007) Network-constrained regularization and variable selection for analysis of genomic data. *UPenn Biostatistics Working Paper*, Working Paper 23. http://biostats.bepress.com/upennbiostat/papers/art23.

Li,J. *et al.* (1997) PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breastprostate cancer. *Science*, **275**, 1943–1946.

Mawrin,C. *et al.* (2003) Prognostic relevance of MAPK expression in glioblastoma multiforme. *Int. J. Oncol.*, **33**, 641–648.

Pelloski,C.E. *et al.* (2006) Prognostic associations of activated mitogen-activated protein kinase and Akt pathways in glioblastoma. *Clin. Cancer Res.*, **12**, 3935–3941.

Perego,C. *et al.* (2002) Invasive behaviour of glioblastoma cell lines is associated with altered organisation of the cadherin-catenin adhesion system. *J. Cell Sci.*, **115**, 3331–3340.

Rahnenführer,J. *et al.* (2004) Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 16.

Swisshelma,K. *et al.* (2005) Role of claudins in tumorigenesis. *Adv. Drug Deliv. Rev.*, **57**, 919–928.

Tibshirani,R.J. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Tibshirani,R. *et al.* (2005) Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B*, **67**, 91–108.

Uht,R.M. *et al.* (2007) The protein kinase C-eta isoform induces proliferation in glioblastoma cell lines through an ERK/Elk-1 pathway. *Oncogene*, **26**, 2885–93.

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

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

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**, 408–429.

Wu,T.T. and Lange,K. (2008) Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, in press.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B*, **68**, 49–67.

Zou,H. (2006) The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, **101**, 1418–1429.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.