BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 12 2007, pages 1486–1494
doi:10.1093/bioinformatics/btm125

Gene expression

Group SCAD regression analysis for microarray time course gene expression data

Lifeng Wang[1], Guang Chen[2] and Hongzhe Li[1,3,]*

[1]Department of Biostatistics and Epidemiology, [2]Department of Bioengineering and [3]Genomics and Computational Biology Graduate Group, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

Received on January 4, 2007; revised on March 1, 2007; accepted on March 23, 2007

Advance Access publication April 26, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Since many important biological systems or processes are dynamic systems, it is important to study the gene expression patterns over time in a genomic scale in order to capture the dynamic behavior of gene expression. Microarray technologies have made it possible to measure the gene expression levels of essentially all the genes during a given biological process. In order to determine the transcriptional factors (TFs) involved in gene regulation during a given biological process, we propose to develop a functional response model with varying coefficients in order to model the transcriptional effects on gene expression levels and to develop a group smoothly clipped absolute deviation (SCAD) regression procedure for selecting the TFs with varying coefficients that are involved in gene regulation during a biological process.

Results: Simulation studies indicated that such a procedure is quite effective in selecting the relevant variables with time-varying coefficients and in estimating the coefficients. Application to the yeast cell cycle microarray time course gene expression data set identified 19 of the 21 known TFs related to the cell cycle process. In addition, we have identified another 52 TFs that also have periodic transcriptional effects on gene expression during the cell cycle process. Compared to simple linear regression (SLR) analysis at each time point, our procedure identified more known cell cycle related TFs.

Conclusions: The proposed group SCAD regression procedure is very effective for identifying variables with time-varying coefficients, in particular, for identifying the TFs that are related to gene expression over time. By identifying the TFs that are related to gene expression variations over time, the procedure can potentially provide more insight into the gene regulatory networks.

Contact: hli@cceb.upenn.edu

Supplementary information: http://www.cceb.med.upenn.edu/~hli/gSCAD-Appendix.pdf

1 INTRODUCTION

Since many important biological systems or processes are dynamic systems, it is important to study the gene expression patterns over time in a genomic scale in order to capture the dynamic behavior of gene expression. Microarray technologies have made it possible to measure the gene expression levels of essentially all the genes during a given biological process. Research in analysis of such microarray time course (MTC) gene expression data has focused on two areas: clustering of MTC expression data (Luan and Li, 2003; Ma *et al*., 2006) and identifying genes that are temporally differentially expressed (Hong and Li, 2006; Tai and Speed, 2006; Yuan and Kendziorski, 2006). While both problems are important and biologically relevant, they provide little information about our understanding of gene regulations.

One approach of studying gene regulation is to associate gene expression values with oligomer motif abundance by using a simple linear regression (SLR) for each oligomer of a given length. Those oligomers with significant coefficients in regression analysis are inferred as potential transcriptional factor binding motifs (TFBMs) (Bussemaker *et al.,* 2001; Gao *et al*., 2004; Keles *et al*., 2002). Assuming that in response to a given biological condition, the effect of a TFBM is strongest among genes with the most dramatic increase or decrease in mRNA expression, Conlon *et al*. (2003) proposed to use SLR to relate the motif abundance to gene expression by first selecting genes with large changes in expression levels. While these approaches work reasonably well in discovery of regulatory motifs in lower organisms, they often fail to identify mammalian transcriptional factor binding sites (Das *et al*., 2006). Das *et al*. (2006) proposed to correlate the binding strength of motifs with expression levels using multivariate adaptive smoothing splines (MARS) of Friedman (2001). In addition, all these methods consider gene expression level at single time point as the response in regression analysis, rather than the full time course, which can lead to loss of efficiency in identifying the relevant transcriptional factors (TFs).

In this article, we consider the problem of identifying the TFs from a large set of candidates (e.g. from TRANSFAC database) that may explain the variations of gene expression over time. Identification of such TFs can provide biological insights into the active transcriptional subnetworks anchored on the proximal promotor DNA from genome-wide mRNA profiles during a biological process (Das *et al*., 2006). One approach to analyzing such MTC data is to use the SLR analysis to relate the TFBM score to the expression level of

*To whom correspondence should be addressed.

genes at each time point (Conlon *et al.*, 2003). Since the effects of a relevant TF are expected to change over time during a given biological process, one should expect some gains in power in detecting the TFs involved in gene expression changes over a time course when the expression levels over all the time points are considered simultaneously in a regression framework. We propose to consider functional response regression analysis with varying coefficients in order to identify the relevant TFs. In such models, the ith response is a real function $y_i(t), i = 1, \dots, n, t \in T$ with associated covariate vector $x_i = \{x_{i1}, \dots, x_{ik}\}$, which is constant in time. Of course, it is only possible to observe the function $y_i(t)$ at a finite number of points, possibly with errors. For the problem of modeling the MTC gene expression data, $y_i(t)$ is the measure expression data for the ith gene at time point t during a given biological process, x_{ik} is the binding strength of the kth motif corresponding to the kth TF (Das *et al.*, 2006). The statistical question to be addressed in this article is to select a set of TFs from a large set of K candidate TFs that can explain partially the variation of gene expression levels over time, where the effects of the TFs on gene expression levels are time varying.

Partially motivated by analysis of high-dimensional microarray gene expression data, the problem of variable selection in high-dimensional regression settings has attracted much research attention in recent years. Among those, the most popular approach is based on penalized estimation, including Lasso (Tibishrani, 1996), the smothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and the least angle regressions (LARS) (Efron, 2005) and various extensions. However, all these methods are developed for regression models with parametric scalar parameters. We propose to develop methods for variable selection for varying coefficient models by combining smoothing spline method with the SCAD procedure, where we represent the time-varying coefficients in terms of B-spline basis functions and propose a penalized estimation procedure to select the sets of basis functions. Our approach is similar in spirit to the group LARS or group Lasso of Yuan and Lin (2006). Although the L_1 penalty gives sparse solutions, the estimates can be biased for large coefficients since large penalties are imposed on larger coefficients. In this article, we propose to use the SCAD penalty on sets of basis functions. Such a penalty produces sparse solutions by thresholding small estimates to zero, providing unbiased estimates for large coefficients. In addition, the resulting estimates based on the SCAD penalty have desirable theoretical properties (Fan and Li, 2001).

The rest of the article is organized as follows. We first introduce the functional response model with time-varying coefficients for relating the TFs to the MTC gene expression data. We then present the SCAD procedure for fitting the models and for selecting the variables (i.e. the TFs). We present simulation studies to evaluate the methods. We also present results from analysis of the yeast cell cycle data set of Spellman *et al.* (1998). Finally, we present a brief discussion of the results and methods.

2 FUNCTIONAL RESPONSE MODEL WITH TIME-VARYING COEFFICIENTS FOR MTC GENE EXPRESSION DATA

Let $Y_i(t)$ be the expression level of the ith gene at time t, for $i = 1, \dots, n$. We assume the following regression model with functional response,

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} \beta_k(t) X_{ik} + \epsilon_i(t), \tag{1}$$

where $\mu(t)$ is the overall mean effect, $\beta_k(t)$ is the regulation effect associated with the kth TF, X_{ik} is the matching score or the binding probability of the kth TF on the promoter region of the ith gene and $\epsilon_i(t)$ is a realization of a zero-mean stochastic process. Several different ways and data sources can be used to derive the matching score X_{ik}. One approach is to derive the score using the position-specific weight matrix (PSWM). Specifically, for each candidate TF k, let P_k be the positive-specific weight matrix of length L, b with element $P_{kj}(b)$ being the probability of observing the base b at position j. Then each L-mer l in the promoter sequence of the ith gene was assigned a score S_{ikl} as:

$$S_{ikl} = \sum_{j=1}^{L} \log \frac{P_{kj}(b_{ilj})}{B(b_{ilj})},$$

where, b_{ilj} is the nucleotide at position j on the lth sequence for gene i, and $B(b)$ is the probability of observing b in the background sequence. This score always assumes a value between 0 and 1. We then define $X_{ik} = \max_l S_{ikl}$, which is the maximum of the matching scores over all the L-mer in the promoter region of the ith gene. The maximum scores can then be converted into the binding probabilities using the method described in Chen *et al.* (2007).

Alternatively, we can define the binding probability based on the chromatin immunoprecipitation (ChIP-chip) data. We present some details in the next section.

2.1 Calculation of binding probabilities based on ChIP data

The results produced by a typical ChIP binding experiment for TF k is a set of measures Z_{ik} for the enrichment of each gene i for that TF k. These measures are then standardized, $U_{iK} = (Z_{ik} - \overline{Z_k})/s_{Z_k}$, to have a common mean and SD. For each U_{ik}, a significance test is performed against a null hypothesis of no enrichment, giving a P-value P_{ik} for each gene that is calculated using a standard normal distribution. However, as these P-values cannot be directly interpreted as the probability $X_{ik} = P(\text{TF } k \text{ binds gene } i)$, we adopted the method proposed by Chen *et al.* (2007) to convert P_{ik} into binding probabilities X_{ik} using mixture modeling. For simplicity of notation, we drop the subscript k in the following. We first convert the P-values P_i to normal score Z_i using the inverse CDF for the standard normal distribution. The distribution of these enrichment measures X_i should be a mixture of two different groups: a large group of unenriched genes that should be centered at $X = 0$ and a smaller group of genes that are truly enriched, with center $\mu > 0$. We can model each gene with a latent variable I_i that indicates whether that gene is in the enriched group ($I_i = 1$) or unenriched group ($I_i = 0$). Then the binding probability for each gene is simply defined as $X_i = P(I_i = 1|\text{Data})$. An expeutation maximization (EM) algorithm can then be applied to estimate these probabilities.

It should be noted that the mixture model used the theoretical standard normal null distribution instead of an empirical null distribution, since the use of an unrestricted mixture model (with an empirically fitted null distribution) led to unreasonable mixtures for several TFs. This procedure was repeated for each TF k to generate our full set of binding probabilities X_{ik}. For the yeast data set we analyzed in later section, the correspondence between the number of genes we predicted as enriched based on P-values ($P_i < 0.005$) and binding probabilities ($X_i > 0.5$) is very good, with a correlation of 0.97 between the number of genes predicted across our 113 TFs. However, we noticed that our conversion procedure tended to be overly conservative for genes with very low P-values. In other words, genes with $P_{ik} < 0.001$ had estimated binding probabilities that is smaller than expected, possibly due to our assumption of a standard normal null distribution. For these highly significant genes, the binding probabilities were increased to $X_{ik} = 0.95$ to reflect our extra confidence that these genes were truly enriched in the ChIP binding experiment for TF k.

3 METHODS OF VARIABLE SELECTION FOR VARYING COEFFICIENT MODELS

We present a penalized estimation procedure for Model (1) using SCAD by representing the varying coefficient $\beta_k(t)$ using smoothing splines. In particular, we propose to use B-splines, which have been shown to provide quite reasonable fits to MTC gene expression data (Hong and Li, 2006; Luan and Li, 2003; Storey *et al.*, 2005).

3.1 Estimation using B-splines

We consider estimation of non-parametric function in Model (1) using the smoothing spline method by approximating $\beta_k(t)$ by using the natural cubic B-spline basis,

$$\beta_k(t) = \sum_{l=1}^{L+4} \beta_{kl} B_l(t) \tag{2}$$

where, $B_l(t)$ is the natural cubic B-spline basis function, for $l = 1, \ldots, L+4$, where L is the number of interior knots. Replacing $\beta_k(t)$ by its B-spline approximation in Equation (2), Model (1) can be approximated as

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} \left\{ \sum_{l=1}^{L+4} \beta_{kl}[B_l(t) X_{ik}] \right\} + \epsilon_i(t), \tag{3}$$

where, we have K groups of parameters with $\beta_k^* = \{\beta_{k1}, \ldots, \beta_{kL+4}\}$ being the parameters associated with the group k, and we want to select the groups with non-zero coefficients. This is the grouped variable selection problem considered in Yuan and Lin (2006).

3.2 A group SCAD penalization procedure

We propose a general group SCAD (gSCAD) procedure for selecting the groups of variables in a linear regression setting. Selecting important variables in Model (1) corresponds to the selection of groups of basis functions in Model (3). Yuan and Lin (2006) proposed several procedures for such group variable selection, including group LARS and group LASSO. Instead of using the L_1 penalty for group selection as in Yuan and Lin (2006), we propose to use the SCAD penalty of Fan and Li (2001). Specifically, to select non-zero $\beta_k(t)$, we can minimize the following penalized loss function

$$\begin{aligned} l(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{T} [y_{ij} - \mu(t_j) - \sum_{k=1}^{K} \sum_{l=1}^{L+4} \beta_{kl} B_l(t_j) X_{ik}]^2 \\ + nT \sum_{k=1}^{K} p_\lambda(||\beta_k^*||_2), \end{aligned} \tag{4}$$

where, y_{ij} is the observed gene expression level for gene i at time t_j, $p_\lambda(.)$ is the SCAD penalty with λ as a tuning parameter, which is defined as

$$p_\lambda(|w|) = \begin{cases} \lambda|w| & \text{if } |w| \le \lambda, \\ -\frac{(|w|^2 - 2a\lambda|w| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |w| < a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |w| > a\lambda \end{cases} \tag{5}$$

and $||\beta_k^*||_2 = \sqrt{\sum_{l=1}^{L+4} \beta_{kl}^2}$. The penalty function (5) is a quadratic spline function with two knots at λ and $a\lambda$, where a is another tuning parameter. Fan and Li (2001) showed that the Bayes risks are not sensitive to the choice of a and suggested to use $a = 3.7$, which was also used in this article.

3.3 Algorithm and selection of tuning parameters

Because of non-differentiability of the penalized loss $l(\beta)$ in Equation (4), the commonly used gradient method is not applicable. Instead, we develop an iterative algorithm based on local quadratic approximation of the non-convex penalty $p_\lambda(\|\beta_k\|_2)$ as in Fan and Li (2001). More specifically, in a neighborhood of a given non-zero $\beta_0 \in \mathbb{R}$, we can approximate the SCAD penalty at value β as the following,

$$p_\lambda(|\beta|) \approx p_\lambda(|\beta_0|) + 1/2\{p'_\lambda(|\beta_0|)/|\beta_0|\}(\beta^2 - \beta_0^2).$$

In our algorithm, a similar quadratic approximation is used by substituting β with $\|\beta_k^*\|_2$, $k = 1, \ldots, K$. Given an initial value of β_k^{*0} with $\|\beta_k^{*0}\|_2 > 0$, $p_\lambda(\|\beta_k^*\|_2)$ can be approximated by a quadratic form

$$p_\lambda(\|\beta_k^{*0}\|_2) + 1/2\{p'_\lambda(\|\beta_k^{*0}\|_2)/\|\beta_k^{*0}\|_2\}((\beta_k^*)^t \beta_k^* - (\beta_k^{*0})^t \beta_k^{*0}),$$

where, the subscript t represents vector or matrix transpose. Using this approximation and letting

$$\beta^* = (\beta_{11}, \ldots, \beta_{1(L+4)}, \ldots, \beta_{K1}, \ldots, \beta_{K(L+4)}),$$

the Equation (4) becomes

$$l(\beta^*) = (Y - C\mu - \tilde{X}\beta^*)^t (Y - C\mu - \tilde{X}\beta^*) + \frac{1}{2} nT \beta^{*t} \Sigma \beta^*,$$

where $Y = (y_{11}, \ldots, y_{1T}, \ldots, y_{n1}, \ldots, y_{nT})^t$, $\mu = (\mu(t_1), \ldots, \mu(t_T))$, $C = \vec{1}_n \bigotimes I_T$, $\tilde{X} = X \bigotimes B$ with $B_{lj} = B_l(t_j)$, $l = 1, \ldots, L+4$, $j = 1, \ldots, T$, $X = \{X_{ik}\}_{i=1,\ldots,n}^{k=1,\ldots,K}$ and

$$\Sigma = \text{diag}\left\{ \frac{p'_\lambda(\|\beta_1^{*0}\|_2)}{|\beta_1^{*0}\|_2}, \ldots, \frac{p'_\lambda(\|\beta_K^{*0}\|_2)}{|\beta_K^{*0}\|_2} \right\} \bigotimes I_{(L+4)}.$$

Here, $\bigotimes$ represents the Kronecker product of two matrices and I is the identify matrix. This is a quadratic form and can be solved by

$$\begin{aligned} (\tilde{X}^t \tilde{X} + \frac{1}{2} nT \Sigma)\beta^* = \tilde{X}^t (Y - C\mu), \\ \mu = C^t (Y - \tilde{X}\beta^*). \end{aligned} \tag{6}$$

We outline the algorithm as follows:

Step 1: Initialize $(\mu^{(1)}, \beta^{*(1)})$.
Step 2: Set $\beta^{*0} = \beta^{*(k)}$, solve $(\mu^{(k+1)}, \beta^{*(k+1)})$ by Equation (6).

Step 3: Iterate Step 2 until convergence of β^* and denote the final estimate of β^* as $\hat{\beta}^*$.

In the initialization step, we obtain an initial estimation of (μ, β) using a ridge regression, which substitutes $p_\lambda(\|\beta_k^*\|_2)$ in (4) with a quadratic function $\|\beta_k^*\|_2^2$. At any iteration of step 2, if some $\|\beta_k^*\|_2$ is smaller than a cutoff value $\epsilon_1 > 0$, we set $\hat{\beta}_{kl} = 0$ for all $l = 1, \ldots, L+4$ and treat X_{ik} as irrelevant. If any matrix is singular when solving Equation (6), a small perturbation ϵ_2 is added to the diagonal entry of the matrix. In our algorithm both ϵ_1 and ϵ_2 are set to 10^{-3}. Note that adding a small perturbation ϵ_2 is equivalent to adding another L_2 penalty to the penalized loss function (4), which also facilitates the selection of highly correlated features (Zou and Hastie, 2005).

There are two tuning parameters that we need to choose in order to implement the proposed procedure: the number of knots L in the B-spline basis expansion (see Equation 2) and the tuning parameter λ in the SCAD penalty function. These two parameters can be selected simultaneously using the generalized cross-validation (GCV). In practice, since the number of time points in typical MTC experiments is usually small, we choose a small number of basis functions in our analysis. Key to the performance of gSCAD is selection of the tuning parameter λ. When λ is too large, it leads to biased estimates of the coefficients, whereas a too small λ often fails to yield a sufficiently sparse solution. Note that in our algorithm, the estimated $\hat{\beta}^* = (\tilde{X}^t\tilde{X} + 1/2nT\Sigma_\lambda(\hat{\beta}^*))^{-1}\tilde{X}^t$, and thus $\hat{y} = \tilde{X}\hat{\beta}^* = M(\lambda)y$ with $M(\lambda) = \tilde{X}(\tilde{X}^t\tilde{X} + 1/2nT\Sigma_\lambda(\hat{\beta}^*))^{-1}\tilde{X}^t$. Therefore, an optimal λ can be obtained by minimizing the following estimated GCV error

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\frac{\|y - M(\lambda)y\|_2^2}{(1 - tr[M(\lambda)]/n)^2}.$$

3.4 Oracle property of gSCAD group selection

In Fan and Li (2001), the oracle property of the SCAD penalized estimates for standard linear models was established, which indicates that the SCAD penalty enables consistent variable selection and parameter estimation simultaneously, as if the subset of relevant variables is already known. To study theoretical properties of gSCAD, we generalize the arguments in Fan and Li (2001) to the group selection settings assuming that the knot locations are held fixed as the sample size increases. Without loss of generality, we consider the model given by equation (3) with $\mu = 0$ and assume that a random design where $\mathbf{z}_{ij} = (\mathbf{x}_{ij}, y_{ij})$, $i = 1, \ldots, n, j = 1, \ldots, T$, are independently and identically distributed with

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}^* + \varepsilon_{ij},$$

$x_{ij}^t = (B_1(t_j)X_{i1}, \ldots, B_{L+4}(t_j)X_{iK})$, $E\varepsilon_{ij} = 0$ and $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$. We further denote $\boldsymbol{X} = (\boldsymbol{x}_{11}^t, \ldots, \boldsymbol{x}_{nT}^t)^t$. Let $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*t}, \ldots, \boldsymbol{\beta}_K^{*\,t})^t$ and assume that $\boldsymbol{\beta}^*(1) = (\boldsymbol{\beta}_1^{*t}, \ldots, \boldsymbol{\beta}_s^{*t})^t$ are the non-zero coefficients, and $\boldsymbol{\beta}^*(2) = (\boldsymbol{\beta}_{s+1}^{*\,t}, \ldots, \boldsymbol{\beta}_K^{*\,T})^t = 0$, i.e. the first s TFs in Model (3) are relevant to gene expression levels over time and the next $K - s$ TFs are not relevant. Let $\Omega = \mathrm{diag}\{(\partial^2 l(\|\boldsymbol{\beta}_1^*\|)/\partial\boldsymbol{\beta}_1^*\partial\boldsymbol{\beta}_1^{*t})|_{\boldsymbol{\beta}_1^* = \hat{\boldsymbol{\beta}}_1^*}, \ldots, (\partial^2 l(\|\boldsymbol{\beta}_s^*\|)/\partial\boldsymbol{\beta}_s^*\boldsymbol{\beta}_s^{*t})|_{\boldsymbol{\beta}_s^* = \hat{\boldsymbol{\beta}}_s^*}\}$ and $m = nT$ and we have the following asymptotic theorem:

THEOREM 1. *Assume* $\Sigma = E(\boldsymbol{XX}^t)$ *is positive definite, and* $\lambda_m \to 0$ *and* $\sqrt{m}\lambda_m \to \infty$ *as* $\mathrm{m} \to \infty$. *Then,*

(1) $\hat{\boldsymbol{\beta}}^*(2) = 0$ with probability approaching 1.

(2) $m^{1/2}(\Sigma(1) + \Omega)(\hat{\boldsymbol{\beta}}^*(1) - \boldsymbol{\beta}^*(1)) \to N(0, \Sigma(1))$ *in distribution, where* $\Sigma(1)$ *is the covariance matrix* Σ *corresponding to* $\boldsymbol{\beta}(1)$.

The proof of this theorem is given in the Supplementary Materials. Such an oracle property of the gSCAD procedure distinguishes it from other group variable selection procedures, such as group LARS or group Lasso of Yuan and Lin (2006). As a consequence, the asymptotic covariance matrix of $\hat{\beta}^*(1)$ is

$$\frac{1}{m}(\Sigma(1) + \Omega)^{-1}\Sigma(1)(\Sigma(1) + \Omega)^{-1},$$

which can be used to derive the confidence intervals for $\hat{\beta}_k(t)$.

4 SIMULATIONS

We conducted simulation studies to evaluate the proposed gSCAD procedure in selecting relevant variables and in estimating the regression coefficients. Specifically, we simulated MTC gene expression data for 500 genes over 11 time points at $0, 0.1, 0.2, \ldots, 0.9$ and 1.0 based on Model (3), where $\beta_k(t)$ was generated using B-splines with 1 interior knot, which corresponds to five basis functions. We assume that there are 10 TFs that affect the MTC expression levels over time. For each TF, the binding probabilities were generated from a uniform (0,1) distribution. The true time-varying coefficients for the TF 2 is shown as the solid lines in Figure 1 (coefficients for other TFs are given in Supplementary Materials). We also assume that the 500 genes can be divided into 25 regulatory modules, each including 20 genes that have similar promoter motif-matching scores. Finally, the noises in Model (3) are generated from $N(0, \sigma^2)$, where $\sigma^2 = 1$ or 3 for low and high noise levels. For each model, we repeated the simulation 100 times and summarized the results.

Plots (a) and (b) in Figure 1 show the means and +/− 1 SE of the estimated time-varying coefficients for the simulated TF 2 using the proposed gSCAD when the noise variance $\sigma^2 = 1$ and 3, respectively, indicating that the gSCAD procedure estimates the parameters very well (plots for other nine TFs are given in the Supplementary Materials). As a comparison, plots (c) and (d) in Figure 1 show the results based on SLR analysis for each time point. Specifically, for a given TF, an SLR was used to estimate the corresponding coefficient at each time point. The means of these estimates (+/−1 SE) are plotted in Figure 1 at each of the 11 time points. Although these estimates can also capture the trend of the true functions well, it is clear that these estimates have much larger variances than those from the gSCAD procedure. The same results were also observed for all other nine TFs (see plots in Supplementary Materials).

Table 1 shows the frequencies over 100 replications of the relevant TFs and irrelevant TFs that were identified by gSCAD and simple point-wise linear regression models. For linear regression, we used the false discovery rate (FDR) procedure of Benjamini and Hochberg (1995) to select the relevant TFs. For a given FDR value, if a TF is significant at at least one time point, we call this TF significant. When the noise variance $\sigma^2 = 1$, both methods identified almost all the true TFs with gSCAD resulting smaller rate of false positives. When the noise variance is increased to $\sigma^2 = 3$, the frequency of the relevant TFs being identified by the gSCAD is 92% with a false positive rate of 4%. As a comparison, the SLRs identified fewer relevant TFs for similar rate of false positives. When $\sigma^2 = 5$, gSCAD has a true positive rate of 66% and a false positive rate of 4%. In contrast, SLR identified much fewer TFs. In order to

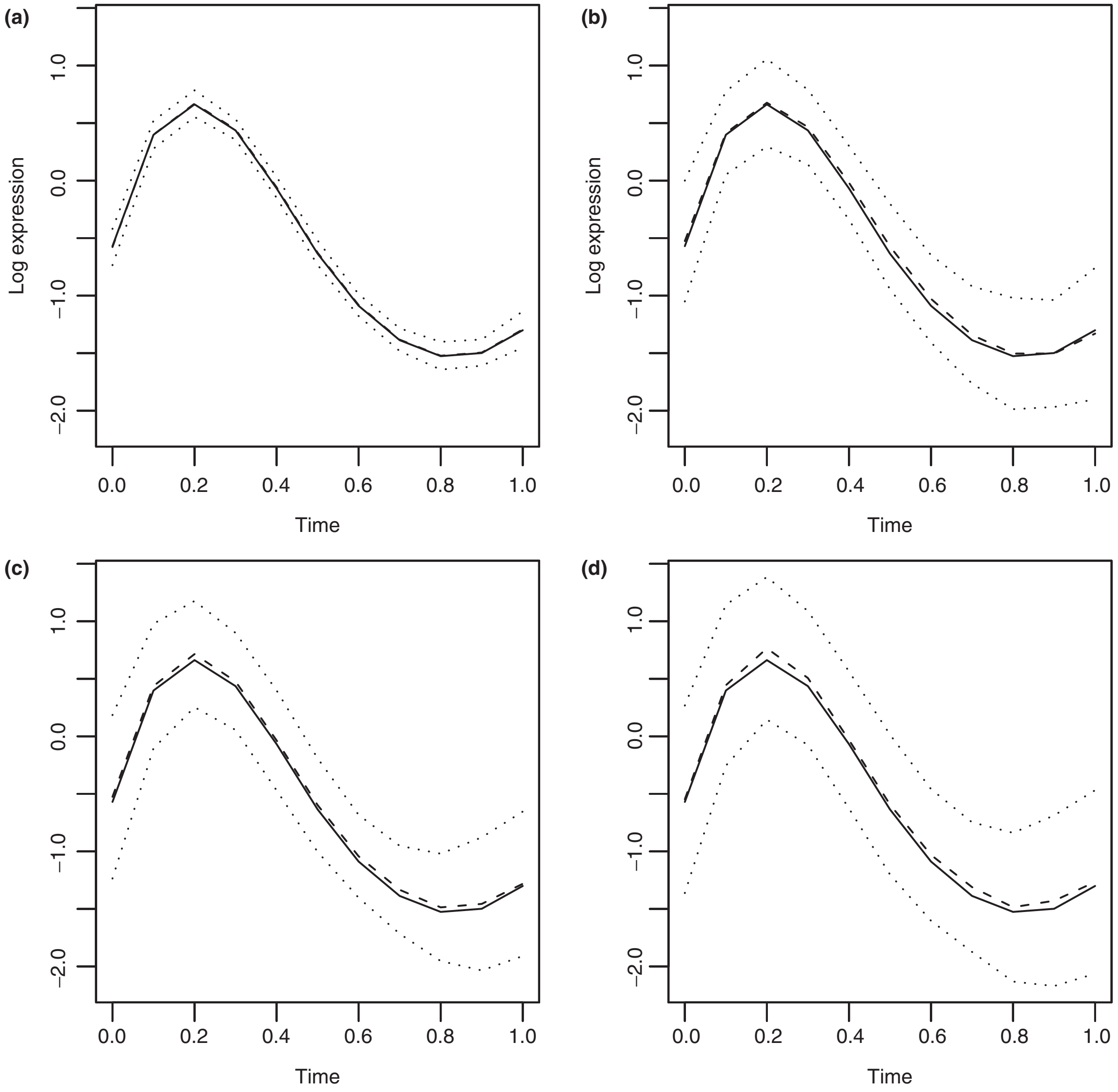

Fig. 1. True (solid line) and mean (+/− 1 SE) of the estimated (dashed and dotted lines) time-dependent transcriptional effects for TF 2 using the gSCAD (plots (**a**) and (**b**)) and SLR (plots (**c**) and (**d**)) for noise variance of 1 (left) and 3 (right).

achieve a true positive rate of 66% using SLR, the false positive rate has to be 9.61%.

5 APPLICATION TO YEAST CELL CYCLE DATA SET

The cell cycle is one of life's most important processes, and the identification of cell cycle regulated genes has greatly facilitated the understanding of this important process. Spellman *et al.* (1998) monitored genome-wide mRNA levels for 6178 yeast ORFs simultaneously using several different methods of synchronization including an α-factor-mediated G_1 arrest, which covers approximately two cell cycle periods with measurements at 7 min intervals for 119 min with a total of 18 time points (http://genome-www.stanford.edu/cellcycle/data/rawdata/). Using data based on different synchronization experiments, Spellman *et al.* (1998) identified a total of about 800 cell cycle regulated genes, some showing periodic expression patterns only in a specific experiment. Using a model-based approach, Luan and Li (2003) identified 297 cell cycle regulated genes based on the α-factor synchronization experiments. We applied the mixture model approach described in previous section using the ChIP data of Lee *et al.* (2002) to derive the binding probabilities X_{ik} for these 297 cell cycle regulated genes for a total of 96 TFs with at least one non-zero binding probability in the 297 genes.

We applied the gSCAD procedure with $L=2$ and an additional L_2 penalty in order to identify the TFs that affect the expression changes over time for these 297 cell cycle regulated genes in the α-factor synchronization experiment. The gSCAD procedure identified a total of 71 TFs that are related to yeast cell cycle processes, including 19 of the 21 known and experimentally verified cell cycle related TFs. The estimated transcriptional effects of these 21 TFs are shown in Figure 2, except for the two TFs (CBF1 and GCN4) that were not selected by the gSCAD procedure and the TF LEU3, the other 18 TFs all showed time-dependent effects of these TFs on gene expression levels. In addition, the effects followed similar trends between the two cell cycle periods. It was not clear why CBF1 and GCN4 were not selected by the gSCAD. The minimum P-values over 18 times points from SLRs are 0.06 and 0.14, respectively, also indicating that CBF1 and GCN4 were not related to expression variation over time. Overall, the model can explain 43% of the total variations of the gene expression levels.

Table 1. Simulation comparison of the gSCAD method and the SLR method using Benjamin and Hochberg's FDR (5 and 15%, respectively) for noise variances of $\sigma^2=1$, 3 and 5

Method	$\sigma^2=1$	$\sigma^2=3$	$\sigma^2=5$
gSCAD	1.00/0.016	0.92/0.040	0.66/0.040
SLR,FDR = 5%	0.997/0.033	0.74/0.0094	0.18/0.002
SLR,FDR = 15%	0.998/0.12	0.87/0.049	0.33/0.011

For each entry, the numbers are the frequencies of the relevant TFs (first number) and the irrelevant TFs (second number) that were identified over 100 replications.

The 52 additional TFs (see Table 2) that were selected by the gSCAD procedure almost all showed estimated periodic transcriptional effects. Figure 3 showed the estimated transcriptional effects for eight of these TFs (CIN5, PHD1, NDD1, STP1, YAP6, NRG1, HSP1 and MBP1), all showing periodic transcriptional effects (plots for other 10 randomly selected TFs can be found in the Supplementary Materials). The identified TFs include many pairs of cooperative or synergistic pairs of TFs involved in the yeast cell cycle process reported in the literature (Banerjee and Zhang, 2003; Tsai *et al.*, 2005). Of these 52 TFs, 34 of them belong to the cooperative pairs of the TFs identified by Banerjee and Zhang (2003). The results are not surprising, since by adding a L_2 penalty term to the SCAD penalized loss function, our procedure can effectively identify the TFs that bind to similar genes or the TFs that have similar binding scores. To assess false identifications of the TFs that are related to a dynamic biological procedure,

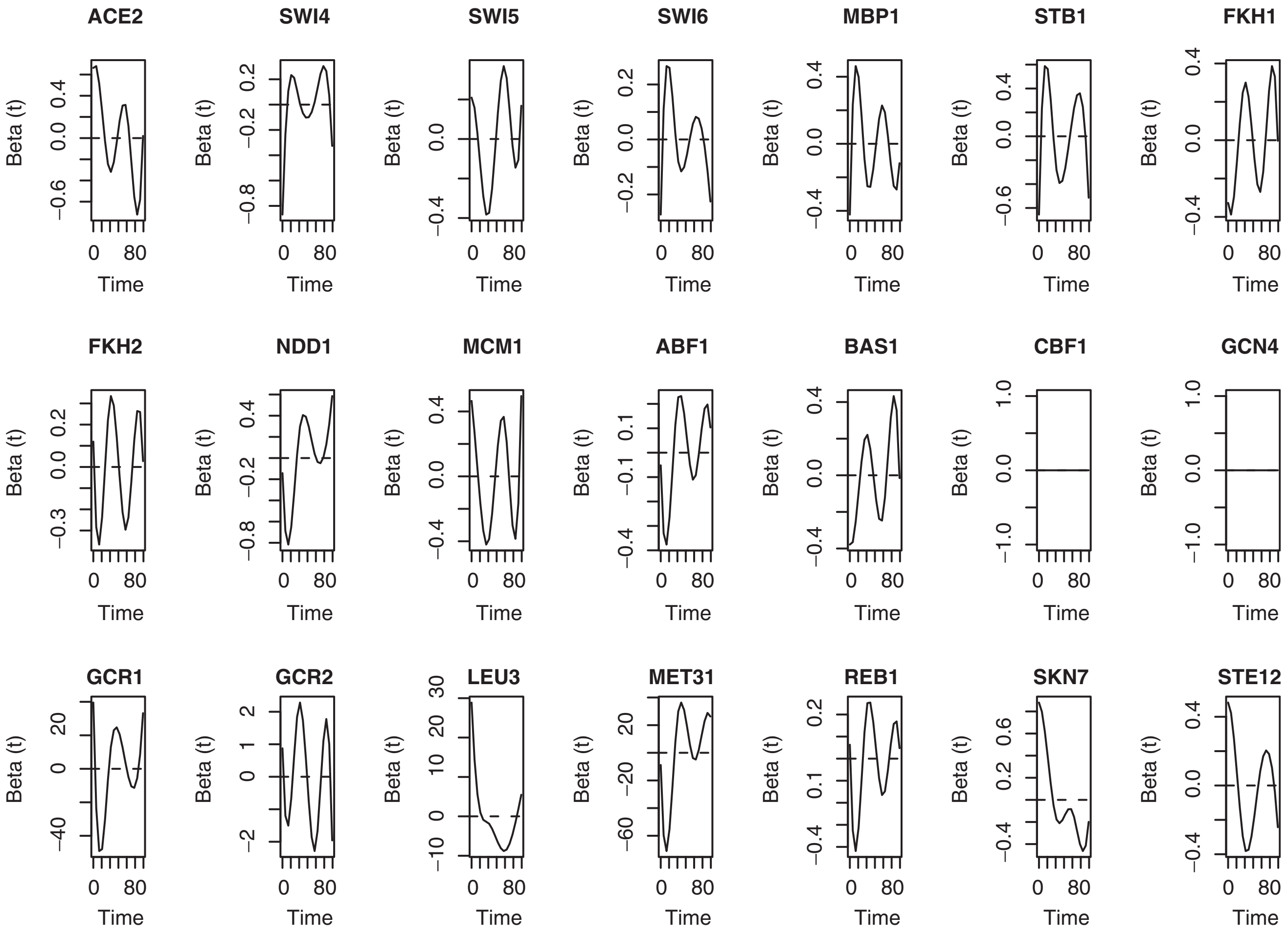

Fig. 2. Estimated time-dependent transcriptional effects for 21 known yeast TFs related to cell cycle process using gSCAD. Note that CBF1 and GCN4 were not selected by gSCAD.

Table 2. Fifty-two additional TFs identified by gSCAD procedure

ARG81	ARO80	ASH1	CIN5	CRZ1	CUP9	DAL81	DOT6	FHL1	FZF1
GAT1	GAT3	GRF10.Pho2.	GTS1	HAL9	HAP2	HAP3	HAP4	HAP5	HIR2
HMS1	HSF1	IME4	INO2	MAC1	MAL13	MATa1	MET4	MIG1	MOT3
MSN4	MTH1	NRG1	PHD1	PUT3	RFX1	RGM1	RLM1	ROX1	RTG1
RTG3	SFP1	SIG1	SIP4	SMP1	SOK2	SRD1	STP1	STP2	YAP5
YAP6	YJL206C								

These include 34 that belong to the cooperative pairs of the TFs identified by Banerjee and Zhang (2003).

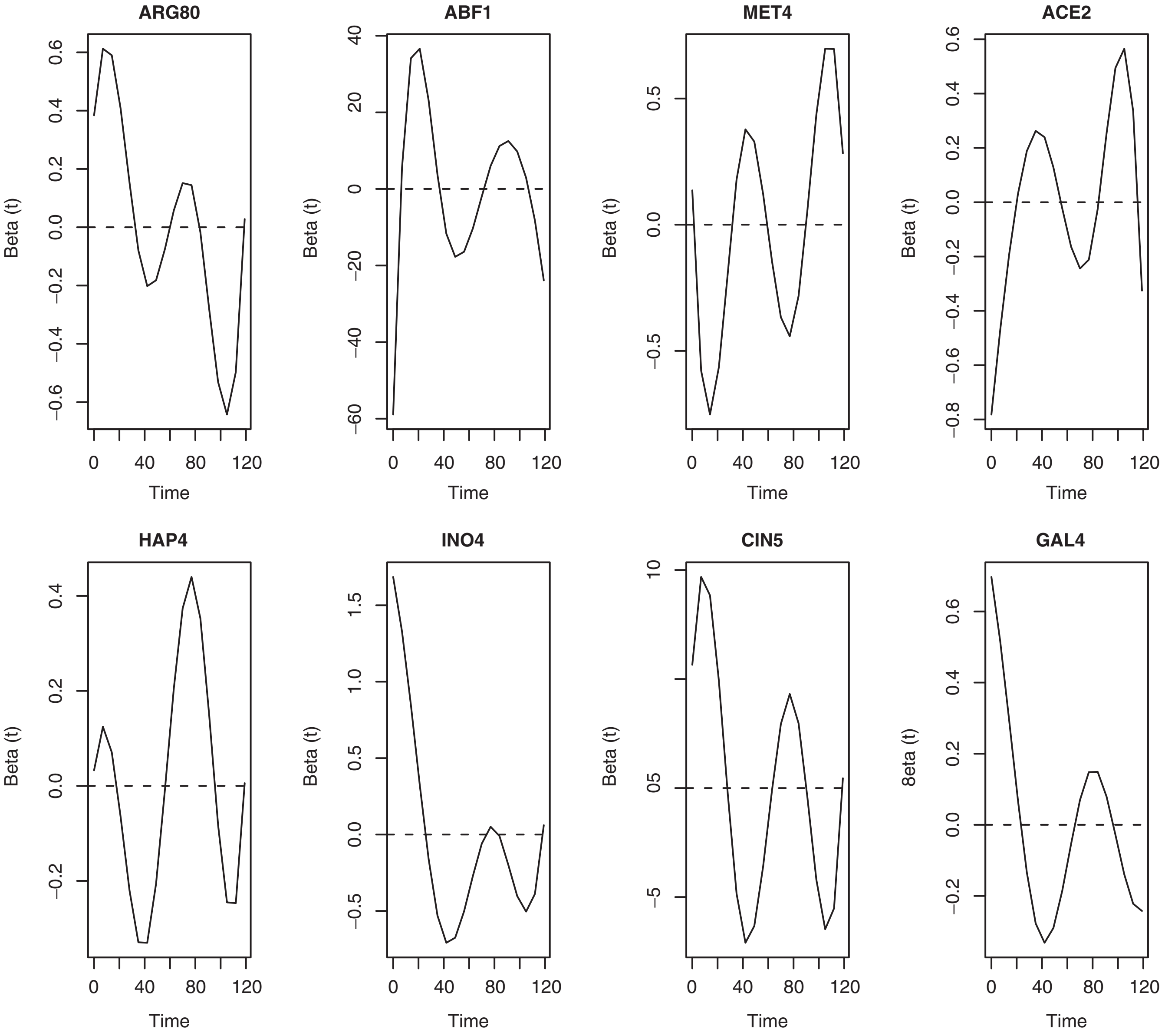

Fig. 3. Estimated time-dependent transcriptional effects for eight out of 52 additional yeast TFs related to the cell cycle process identify gSCAD.

we randomly permuted the gene expression values across genes and time points and applied the gSCAD procedure again to the permuted data sets. We repeated this procedure 50 times. Among the 50 runs, 5 runs selected 4 TFs, 1 run selected 3 TFs, 16 runs selected 2 TFs and the rest of the 28 runs did not select any of the TFs, indicating that our procedure indeed selects the relevant TFs with few false positives.

To compare the gSCAD procedure with SLR, we performed SLR with motif probability as the predictor and the gene expression at each time point as the response. After Bonferroni adjustment for multiple testing, we found that only 7 out of the 21 known cell cycle related TFs that showed statistically significant association with the gene expression levels.

Besides using the binding probabilities derived from ChIP-chip data, we also performed analysis using the binding probabilities derived from the sequence data and the PSWMs. Specifically, we identified 72 TFs with reliable PSWMs for estimating the binding probabilities. The gSCAD identified 56 TFs that can potentially be related regulation of the cell cycle process, including 18 of the 21 known cell cycle related TFs. Overall, the model can explain 33% of the total variation of the gene expression levels, indicating that the sequences and PSWMs-based binding probabilities might not be as reliable as those obtained from ChIP-chip data for explaining the gene expression variations.

6 CONCLUSIONS AND DISCUSSION

Motivated by identifying TFs that can explain (partially) the observed variation of MTC gene expression over time during a given biological process, we introduce a group SCAD penalized estimation procedure for selecting variables with time-varying coefficients in the context of functional response models. Simulation studies indicated that this procedure is very effective in selecting the relevant groups of variables and in estimating the regression coefficients. Results from application to the yeast cell cycle data set indicate that the procedure can be effective in selecting the TFs that potentially play important roles in regulation of gene expressions during the cell cycle process.

In this article, we used B-spline basis functions to approximate the varying coefficients associated with each TF. B-spline basis functions provide flexible models for MTC gene expression data and have been applied for clustering MTC gene expression data (Luan and Li, 2003; Storey *et al.*, 2004) and for identifying temporally regulated genes (Hong and Li, 2006). Our application to real data sets in this article further demonstrated its utility in modeling the MTC gene expression data. However, it should be noted that other basis functions can also be used to approximate the coefficient functions $\beta_k(t)$. For example, one can use linear spline with truncated lines as the basis for regression. Such a linear spline was used in MARS (Friedman, 2001) and in Das *et al.* (2006) for modeling regulatory subnetworks. The proposed gSCAD can equally work for such linear spline approximation.

The proposed methods can be extended in several ways. First, in Model (1), we assume an additive model for the effects of the TFs on the gene expression levels over time. However, genetic regulation often involves interacting *cis*-control motifs. One way to incorporate such interactions is to extend the proposed Model (1) to include interaction effects between two TFs as

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} \beta_k(t) X_{ik} + \sum_{k=1}^{K} \sum_{k' \neq k} \beta_{kk'}(t) X_{ik} X_{ik'} + \epsilon_{it},$$

where, $\beta_{kk'}$ measures the interaction effects between two TFs k and k'. The gSCAD procedure proposed in this paper should be applicable to such models also. Second, although the models and the procedure considered in this article are motivated by analysis of MTC gene expression data, the proposed gSCAD procedure will be easily extended to other regression models such as the generalized linear models and Cox models with varying coefficients. These are the topics that deserve further investigation.

In summary, we have proposed a penalized estimation procedure using SCAD for selection of grouped variable in a linear regression model setting. We particularly considered the application of such a group SCAD procedure to selection of time-varying coefficients in high-dimensional functional response regression model settings. The procedure is useful for identifying the TFs that are related to MTC gene expression data measured during a given biological process. The TFs identified can provide useful information about the transcriptional networks.

ACKNOWLEDGEMENTS

This research was supported by NIH grant ES009911. We thank two reviewers for many helpful comments and Mr. Edmund Weisberg, MS at Penn CCEB for editorial assistance.

Conflict of Interest: none declared.

REFERENCES

Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.

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

Bussemaker,H.J. *et al.* (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Chen,G. *et al.* (2007) Clustering of genes into regulons using integrated moeling(cogrim). *Genome Biol.*, **8**, 1, R4.

Conlon,E.M. *et al.* (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA.*, **100**, 3339–3344.

Das,D. *et al.* (2006) Adaptively inferring human transcriptional subnetworks. *Mol. Syst. Biol.*, msb410067-E1.

Keles,S. *et al.* (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167–1175.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.

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

Friedman,J. (2001) Multivariate adaptive regression splines. *Ann. Stat.*, **19**, 1–141.

Gao,F. *et al.* (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Hong,F. and Li,H. (2006) Functional hierarchical models for identifying genes with different time-course expression profiles. *Biometrics*, **62**, 534–544.

Lee,T.I. *et al*. Transcriptional regulatory networks in *S*. *cerevisiae*. *Science*, **298**, 799–804.

Luan,Y. and Li,H. (2003) Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, **19**, 474–482.

Ma,P. *et al*. (2006) A data-driven clustering method for time course gene expression data. *Nucleic Acids Res*., **34**, 1261–1269.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol*. *Biol*. *Cell*, **9**, 3273–3297.

Storey,J.D. *et al*. (2005) Significance analysis of time course microarray experiments. *Proc*. *Natl Acad*. *Sci*. *USA*., **102**, 12837–12842.

Tai,Y.C. and Speed,T.P. (2006) A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann*. *Stat*., in press.

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

Tsai,H.K. *et al*. (2005) Statistical methods for identifying yeast cell cycle transcription factors. *PNAS*, **102**, 13532–13537.

Yuan,M. and Kendziorski,C. (2006) Hidden Markov models for microarray time course data in multiple biological conditions. *J*. *Am*. *Stat*. *Assoc*., 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.