BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 22 2005, pages 4169–4175
doi:10.1093/bioinformatics/bti680

Gene expression

RSIR: regularized sliced inverse regression for motif discovery

Wenxuan Zhong[1,†], Peng Zeng[2,†], Ping Ma[3,†], Jun S. Liu[1,*] and Yu Zhu[4,*]

[1]Department of Statistics, Harvard University, Cambridge, MA 02138, USA, [2]Department of Mathematics and Statistics, Auburn University, Auburn, AL 36849, USA, [3]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA and [4]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

Received on May 17, 2005; revised on September 3, 2005; accepted on September 13, 2005
Advance Access publication September 15, 2005

ABSTRACT

Motivation: Identification of transcription factor binding motifs (TFBMs) is a crucial first step towards the understanding of regulatory circuitries controlling the expression of genes. In this paper, we propose a novel procedure called regularized sliced inverse regression (RSIR) for identifying TFBMs. RSIR follows a recent trend to combine information contained in both gene expression measurements and genes' promoter sequences. Compared with existing methods, RSIR is efficient in computation, very stable for data with high dimensionality and high collinearity, and improves motif detection sensitivities and specificities by avoiding inappropriate model specification.

Results: We compare RSIR with SIR and stepwise regression based on simulated data and find that RSIR has a lower false positive rate. We also demonstrate an excellent performance of RSIR by applying it to the yeast amino acid starvation data and cell cycle data.

Availability: Matlab programs are available upon request from the authors.

Contact: jliu@stat.harvard.edu; yuzhu@stat.purdue.edu

INTRODUCTION

Gene transcription is regulated by proteins called transcription factors (TFs) binding to their recognition sites located mostly in upstreams of the genes (promoter regions), but also not infrequently in downstreams or intronic regions. The common pattern of the recognition sites of a specific TF is referred to as the transcription factor binding motif (TFBM). Discovering binding sites and motifs of specific TFs of an organism is an important first step towards the understanding of gene regulation circuitry. This problem has attracted much attention from both experimentalists and quantitative researchers. Experimental techniques such as electrophoretic mobility shift assays and DNase footprinting have been used to locate TFBMs on a gene-by-gene and site-by-site basis, but these methods are laborious and time-consuming, and thus unsuitable for genome-wide studies. Recent years have seen a rapid adoption of the ChIP-on-chip technology (Ren *et al*., 2000; Lieb *et al*., 2001; Lee *et al*., 2002), where chromatin immunoprecipitation (ChIP) is carried out in conjunction with mRNA microarray analysis (chip) to identify genome-wide interaction sites of a DNA-binding protein. However, this method only yields a resolution of hundreds to thousands of base pairs, whereas the actual binding sites are 10–15 bp long. Computational methods developed over the past 10–15 years have proven extremely helpful for pinning down the exact binding site locations (see Stormo, 2000 and Jensen *et al*., 2004 for a review).

The latest attempt to improve upon computational TFBM finding methods is to incorporate information from gene expression data. An innovative method is REDUCE, proposed in Bussemaker *et al*. (2001), which utilizes the correlation between gene expression values and the occurrences of certain 'words' in the promoter regions of genes. The method has been extended in Keles *et al*. (2002, 2004). Conlon *et al*. (2003) proposed another method, motif regressor, which selects 'functional' TFBMs from a large pool of motif candidates by regressing the genes' mRNA expression levels against their promoter regions' matching scores to each of the candidate motifs. Along a similar line of thought, Beer and Tavazoie (2004) built Bayesian models to predict a gene's cluster membership based on the motif features in its promoter region.

A fundamental assumption to facilitate motif discovery using gene expression information is that a gene's mRNA copy number is associated with the gene's upstream matching score (or more intuitively, number of TFBM copies) corresponding to a functional TFBM. The simplest association is the linear relationship as considered by both Bussemaker *et al*. (2001) and Conlon *et al*. (2003). Hence, to find the TFBMs, they look for motif candidates that have the strongest linear association with the gene expression values. To identify all the functional TFBMs, Conlon *et al*. (2003) used the classic stepwise regression technique in their motif regressor. Though the idea of motif regressor is promising, the linear model combined with stepwise regression may not be completely satisfactory due to the following drawbacks. First, the relationship between gene expression values and motif scores is likely to be more complex than the simple linear relationship; second, the number of motif candidates in consideration is usually in hundreds, however most stepwise regression algorithms can only explore a small portion of all the possible models and third, motif scores are often highly correlated, which makes regression fitting unstable and may lead to falsely inflated regression coefficients that contribute to high false positives and negatives. In order to avoid the first drawback, more sophisticated methods using non-linear basis functions, such as polynomial basis and spline basis, were recently proposed

[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

for motif discovery (Sinisi and van der Laan, 2004; Das *et al*., 2004). But these methods also face the other two drawbacks.

In this paper, we propose a more flexible model that consists of an unspecified (non-linear) link function and linear combinations of motif scores for the relationship between gene expression values and motif matching scores. Based on the model, we develop a novel procedure, called regularized sliced inverse regression (RSIR), to directly identify the linear combinations and further the functional TFBMs while avoiding the estimation of the link function. RSIR is an extension of sliced inverse regression (SIR) proposed by Li (1991) for data of high dimensionality, high collinearity and relatively small sample size.

After demonstrating the performance of RSIR in a non-linear model setting via simulation, we applied it to two real datasets: the amino-acid starvation data (Gasch *et al*., 2000) and the cell cycle data (Spellman *et al*., 1998). For the former data, we found 16 motif patterns that are functionally active for amino acid starvation, 11 of which are known in the literature. We further explored the results and found two biologically interpretable modules of motif patterns. For the cell cycle data, we successfully identified all the known cell cycle regulating TFBMs. The correlations between the gene expression values and the motif scores of the identified TFBMs reveal periodical patterns, which are consistent with the underlying biological mechanism.

METHODS

Model assumptions

Let $g_1, g_2, \ldots, g_N$ denote the genes and let $y = (y_1, \ldots, y_N)$ be their corresponding mRNA expression levels measured by a microarray experiment under some specific condition. The initial set of TFBM candidates is determined as follows. First, a subset (10–100) of the genes with the highest absolute expression values is obtained. Second, a motif finding algorithm, such as MDscan (Liu *et al*., 2002) in our case, is used to search the promoter regions of the selected genes to give rise to the initial motif set. We denote these TFBM candidates by $m_1, m_2, \ldots, m_p$ and their consensus matrices by $\theta_1, \theta_2, \ldots, \theta_p$, respectively. For gene g_i $(1 \le i \le N)$, the matching score of the motif candidate m_j in g_i's promoter region is calculated as

$$x_{ij} = \log_2 \sum_{k=1}^{n_i - w_j} \frac{P(s_{i,k}|\theta_j)}{P(s_{i,k}|\theta_0)}, \tag{1}$$

where n_i is the size (length) of promoter region of g_i, w_j is the width of the motif candidate m_j, θ_0 is the third-order Markov model parameter estimated from intergenic sequences and $s_{i,k}$ is the sequence segment of width w_j starting at the k-th position in the promoter region of g_i. The matching score x_{ij} describes the abundance and intensity of m_j in the entire promoter region of g_i. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$. Then $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$ form the gene expression and motif matching score data, or in short, the expression-score data.

We use a toy example to motivate our general model for identifying TFBMs. Assume that the initial set contains $p = 9$ motif candidates, and that the gene expression value is related to the motif scores as

$$y = f(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \beta_3'\mathbf{x}, \varepsilon), \tag{2}$$

where ε is a random error independent of $\mathbf{x}$, and the vectors of coefficients are $\beta_1 = (1, 0, 0, 0, 0, 0, 0, 0, 0)'$, $\beta_2 = (0, 0, 0.3, 0.4, 0, 0, 0, 0, 0)'$ and $\beta_3 = (0, 0, 0, 0, 0, 0, 0, 0.8, 0.9)'$.

Model (2) states that the gene expression value depends on the motif matching scores through three of their linear combinations. Since only x_1, x_3, x_4, x_8 and x_9 have nonzero coefficients in these linear combinations, motifs m_1, m_3, m_4, m_8 and m_9 are the true functional TFBMs whereas the rest are not. Model (2) further suggests that, to identify the true TFBMs using the expression-score data, we need to estimate the linear combinations first and then identify the motif scores with nonzero contributions. Note that the link function f is left unspecified in (2).

In general, we assume that the expression value of a gene depends on the motif scores through k (unknown) linear combinations,

$$y = f(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \ldots, \beta_k'\mathbf{x}, \varepsilon). \tag{3}$$

Since f is not specified, (3) can accommodate a wide variety of models including the linear model. The expression-score data are usually high-dimensional and noisy, so a direct fitting of f using non-parametric methods is impractical. It is thus desirable to estimate the linear combinations without fitting f. This task can be accomplished by SIR (Li, 1991), which was originally developed for dimension reduction and data visualization. After having obtained $\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x}$, we can identify x_i's with nonzero contributions to the linear combinations and their corresponding functional TFBMs. Because k is much smaller than p, if further desirable, non-parametric methods can be used to fit f only using $\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x}$, and the fitted model can be used to predict the gene expression value y. In this paper, we only focus on the identification of functional TFBMs and their linear combinations with biological interpretation. The subspace spanned by $\beta_1, \ldots, \beta_k$ is defined to be the sufficient dimension reduction subspace, which is denoted by $\mathcal{S}$. Model (3) and the subspace $\mathcal{S}$ can also be defined from the perspective of conditional independence, and other methods exist for estimating $\beta_1, \ldots, \beta_k$; see Cook (1998) for details.

Regularized sliced inverse regression

When $\mathbf{x}$ satisfies a linearity condition, Li (1991) showed that $\beta_1, \ldots, \beta_k$ in (3) can be estimated by solving the following optimization problem sequentially,

$$\arg \max_{\beta'\Sigma\beta=1} \beta' M \beta, \tag{4}$$

where β is a p-dimensional vector, $M = \text{cov}[E(\mathbf{x}|y)]$ and Σ is the variance covariance matrix of $\mathbf{x}$. Once β_1 is obtained, the constraint in (4) is updated to $\beta'\Sigma\beta_1 = 0$ and $\beta'\Sigma\beta = 1$, solving the updated (4) gives β_2. This procedure continues till $\beta_1, \ldots, \beta_\text{k}$ are obtained. In application, Σ is estimated by the sample covariance matrix $\hat{\Sigma} = [1/(N-1)] \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$, where $\bar{\mathbf{x}}$ is the sample mean. A slicing procedure is proposed in Li (1991) for estimating M. First, divide the range of $\{y_i\}_{i=1}^N$ into a number of disjoint intervals, say H intervals, which are denoted by $S_1, S_2, \ldots, S_H$; second, for $1 \le h \le H$, calculate $\bar{\mathbf{x}}_h = n_h^{-1} \sum_{y_i \in S_h} \mathbf{x}_i$, where n_h is the number of y_i in S_h; third,

$$\hat{M} = N^{-1} \sum_{h=1}^{H} n_h (\bar{\mathbf{x}}_h - \bar{\mathbf{x}})(\bar{\mathbf{x}}_h - \bar{\mathbf{x}})' \tag{5}$$

is used as an estimate of M. One then proceeds to solve (4) with Σ and M replaced by their estimates. And the obtained directions are denoted by $\hat{\beta}_1, \ldots, \hat{\beta}_k$.

SIR is not as successful for our TFBM identification task as in some other applications due to the high dimensionality and high multicollinearity of the expression-score data, which makes $\hat{\Sigma}$ nearly degenerate in a number of directions. Rewrite the sample version of (4) in an equivalent expression,

$$\arg \max_{\beta'\hat{\Sigma}\beta=1} \|\beta\|^2 u'\hat{M}u, \quad \text{where } u = \beta/\|\beta\|. \tag{6}$$

It is easy to see that the target function in (6) depends on both the norm and the orientation of β, and the maximization is over β on the ellipsoid $\mathcal{E} = \{\beta : \beta'\hat{\Sigma}\beta = 1\}$. Along the directions in which $\hat{\Sigma}$ is degenerate, β's on $\mathcal{E}$ have extremely large norms, so that they may be falsely selected as the estimates of $\beta_1, \ldots, \beta_k$, which makes $\hat{\beta}_1, \ldots, \hat{\beta}_k$ very unstable. In order to mitigate the variability caused by the near-degeneracy of $\hat{\Sigma}$, we add a positive definite matrix sI to $\hat{\Sigma}$, where I is the $p \times p$ identity matrix and s is a prescribed non-negative constant. Thus the ellipsoid $\mathcal{E}$ is changed to $\mathcal{E}_s = \{\beta : \beta'(\hat{\Sigma} + sI)\beta = 1\}$, and (6) becomes

$$\arg \max_{\beta'(\hat{\Sigma}+sI)\beta=1} \beta'\hat{M}\beta, \tag{7}$$

which can be solved sequentially as (4) to generate k directions denoted by $\hat{\beta}_1(s), \ldots, \hat{\beta}_k(s)$. We refer to (7) as RSIR and $\hat{\beta}_1(s), \ldots, \hat{\beta}_k(s)$ as the RSIR directions. Note that when $s = 0$, (7) is the sample version of (4) and $\hat{\beta}_j(0) = \hat{\beta}_j$ for $j = 1, \ldots, k$.

In fact, $\hat{\beta}_1(s), \ldots, \hat{\beta}_k(s)$ can be obtained by solving the following system of equations and constraints,

$$\begin{gathered} \hat{M}\beta_i = \lambda_i(sI + \hat{\Sigma})\beta_i, \\ \beta_i'(sI + \hat{\Sigma})\beta_j = 1 \quad \text{if } i = j; = 0 \text{ if } i \neq j, \\ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0, \end{gathered} \tag{8}$$

where $i, j = 1, \ldots, k$.

Similar to other regularized methods, RSIR reduces estimation variability at the cost of inducing estimation bias. The success of RSIR depends on its proper choice of s especially for data of high dimensionality, high multicollinearity and relative small sample size such as the expression-score data. The same argument and method were originally used by Hastie *et al.* (1995) and Yu *et al.* (1999) when proposing penalized linear discriminant analysis.

IMPLEMENTATION

The flow chart of our procedure is given in Figure 1. Note that we choose MDscan as the tool to search for the motif candidates. Since MDscan is publicly available, we do not discuss it in detail here; see Liu *et al.* (2002) for more information. Readers can choose their favorite motif searching algorithms when using our procedure. Some computational issues related to RSIR are discussed in this section. The entire Matlab code can be requested from the authors.

Computing scheme for RSIR directions

Suppose k is known and the regularization parameter s is given. First, we calculate $\hat{M}$ using (5) with fixed H. Second, we solve (8) to obtain $\hat{\beta}_1(s), \ldots, \hat{\beta}_k(s)$. The linear combinations $\hat{\beta}_1'(s)\mathbf{x}, \ldots, \hat{\beta}_k'(s)$ are referred to as the RSIR variates in the rest of the paper. For visualization, we further generate k plots of y against $\hat{\beta}_\ell'(s)$, $1 \leq \ell \leq k$, respectively, which reveal the relationships between the gene expression values and the k RSIR variates. Similar to SIR (Li, 1991), RSIR is insensitive to the choice of H.

Motif selection based on RSIR variates

For motif candidate m_j, its coefficients in the RSIR variates are $\hat{\beta}_{1j}(s)$, $\hat{\beta}_{2j}(s), \ldots, \hat{\beta}_{kj}(s)$. According to model (3), intuitively, if all the $\hat{\beta}_{1j}(s)$, $\hat{\beta}_{2j}(s), \ldots, \hat{\beta}_{kj}(s)$ are close to zero, then m_j is not a functional TFBM. Next we propose a scoring procedure to determine functional TFBMs.

Let $\hat{\beta}_{*j}(s) = (\hat{\beta}_{1j}(s), \ldots, \hat{\beta}_{kj}(s))'$, and let $\Sigma_{*j}(s)$ be the covariance matrix of $\hat{\beta}_{*j}(s)$. We use the squared Mahalanobis distance between $\hat{\beta}_{*j}(s)$ and the origin as a significance score for m_j, which is

$$\Gamma_j = \hat{\beta}_{*j}'(s) \sum\nolimits_{*j}^{-1}(s)\hat{\beta}_{*j}(s).$$

The exact sampling distribution of $\hat{\beta}_{*j}(s)$ is difficult to derive. When $s = 0$, Li (1991) stated in his Remark 5.1 that $\hat{\beta}_{*j}(0)$ is asymptotically normal with mean $\beta_{*j}(0)$ and covariance $\hat{\Sigma}_{*j}(0)$. For $s > 0$, the same result still holds, i.e. $\hat{\beta}_{*j}(s)$ asymptotically follows a normal distribution with mean $\beta_{*j}(s)$ and covariance $\Sigma_{*j}(s)$. Therefore, $(\hat{\beta}_{*j}'(s) - \beta_{*j}'(s))\Sigma_{*j}^{-1}(s)(\hat{\beta}_{*j}(s) - \beta_{*j}(s))$ follows a chi-squared distribution χ_k^2 where k is the degree of freedom. When $\beta_{*j}(s) = 0$, Γ_j follows χ_k^2 asymptotically. Note that the advantage of regularization is to trade bias for significant reduction in variation, and usually the introduced bias is very small when s is properly chosen, especially in our setting. This was also observed in penalized discriminant analysis by Yu *et al.* (1998). Although $\beta_{*j}(s)$ is different from $\beta_{*j}(0)$, their difference is expected to be small. Hence, we can use Γ_j as a scoring function for checking whether m_j is a functional motif or not.

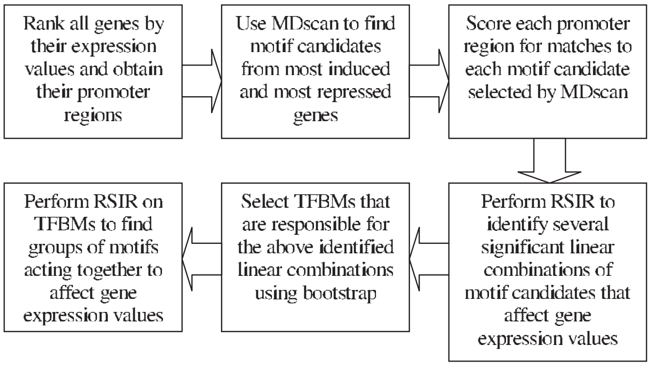
Rank all genes by their expression values and obtain their promoter regions
Use MDscan to find motif candidates from most induced and most repressed genes
Score each promoter region for matches to each motif candidate selected by MDscan
Perform RSIR to identify several significant linear combinations of motif candidates that affect gene expression values
Select TFBMs that are responsible for the above identified linear combinations using bootstrap
Perform RSIR on TFBMs to find groups of motifs acting together to affect gene expression values

Fig. 1. Flow chart for the procedure of motif discovery.

There exists another difficulty in using Γ_j directly, because we do not have an analytic expression for $\Sigma_{*j}(s)$. We use the following bootstrap procedure to derive an estimate of $\Sigma_{*j}(s)$. For $m = 1, \ldots, B$, we draw with replacement N pairs $(y_i^{(m)}, \mathbf{x}_i^{(m)})$ from the original sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$ completely at random. Each such formed new sample is called a bootstrap sample. We apply RSIR to each bootstrap sample to obtain the estimates $\hat{\beta}_1^{(m)}(s), \ldots, \hat{\beta}_k^{(m)}(s)$ of $\beta_1(s), \ldots, \beta_k(s)$, and estimate $\Sigma_{*j}(s)$ by the sample covariance matrix of $\{(\hat{\beta}_{1j}^{(m)}, \ldots, \hat{\beta}_{kj}^{(m)})'\}_{m=1}^B$, denoted by $\hat{\Sigma}_{*j}(s)$. Finally, we approximate Γ_j by $\hat{\Gamma}_j = \hat{\beta}_{*j}'(s)\hat{\Sigma}_{*j}^{-1}(s)\hat{\beta}_{*j}(s)$. If $\hat{\Gamma}_j > \chi_k^2(\alpha)$, where α is a significance level determined by the user, m_j is declared to be a functional TFBM. When the number of motif candidates is large, multiple comparison procedures can be incorporated.

Choice of regularization parameter

Because regularization controls the tradeoff between the bias and the variability of the estimate, it is important to determine the amount of regularization s. For a given s, the total mean squared error (MSE) of the RSIR directions is

$$L(s) = \sum_{i=1}^{k} \text{tr}(\text{cov}(\hat{\beta}_i(s))) + \sum_{i=1}^{k} \|E(\hat{\beta}_i(s)) - \beta_i\|^2, \tag{9}$$

where tr is the trace of a matrix. $L(s)$ can be interpreted as the average distance between the RSIR directions and the true directions $\beta_1, \ldots, \beta_k$, and it consists of two parts. The first part $V(s) = \sum_{i=1}^k \text{tr}(\text{cov}(\hat{\beta}_i(s)))$ is the sum of the variances of $\hat{\beta}_{ij}(s)$, and the second part $B(s) = \sum_{i=1}^k \|E(\hat{\beta}_i(s)) - \beta_i\|^2$ is the sum of the squared biases. When s increases from 0, $L(s)$ first decreases. After $L(s)$ reaches its minimum at s_*, it begins to increase as s further increases. We choose s_* to be the amount of regularization in RSIR.

Using s_*, RSIR can benefit from the significant reduction in variance while the increase of bias is limited to a minimum.

Note that both $V(s)$ and $B(s)$ in $L(s)$ are not directly computable. We use $\hat{\boldsymbol{\beta}}_1^{(m)}(s), \ldots, \hat{\boldsymbol{\beta}}_k^{(m)}(s)$ generated by the bootstrap procedure discussed in the previous subsection to approximate $V(s)$ and $E(\hat{\boldsymbol{\beta}}_i(s))$. The true directions $\boldsymbol{\beta}_i$ are approximated by the bootstrap mean of $E(\hat{\boldsymbol{\beta}}_i(s_0))$ with s_0 being a suitably small positive value. Although the proposed procedure for determining s is approximate, our experience from simulation and real data suggests that it works well.

Determining the number of directions

Recall that k is the number of true directions and it is equal to the rank of $M = \mathrm{cov}[E(\mathbf{x}|y)]$. So, in RSIR, the choice of k shall not depend on the choice of s. A graphical method for determining k is to plot y against the derived RSIR variates $\hat{\boldsymbol{\beta}}_i'(s)\mathbf{x}$ for $1 \leq i \leq p$, and pick up the variates that generate plots with visible patterns. A more formal yet conservative approach as proposed by Li (1991) is to sequentially test a series of hypotheses: $H_0 : k = d$ versus $H_1 : k > d$, for $d = 0, 1, 2, \ldots, p-1$. Let $\hat{\boldsymbol{\lambda}}_1(s) \geq \hat{\boldsymbol{\lambda}}_2(s) \cdots \geq \hat{\boldsymbol{\lambda}}_p(s)$ be the eigenvalues calculated from (8). Define $\Lambda_d(s) = n\sum_{i=d+1}^{p} \hat{\boldsymbol{\lambda}}_i(s)$. When $s = 0$, Li (1991) proved that $\Lambda_d(0)$ follows a χ^2 distribution with $(p-k)(H-k-1)$ degrees of freedom asymptotically. Using the test statistic and its asymptotic distribution, we start with $d = 0$ and sequentially test the subsequent hypotheses until H_0 is accepted. As indicated by Li (1991), the sequential test is conservative and sometimes underestimates k. In this paper, we combine both the graphical method and the sequential test to choose k.

RESULTS

Simulation

Suppose that there are 20 motif candidates and 500 genes. Let U be a fixed 20×20 orthogonal matrix, and let Λ_0 be a diagonal matrix with diagonal entries (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 25, 50, 100). We generate the expression-score data using the following scheme: (1) Generate 500 iid $\mathbf{x}_i$ from $N(0, U\Lambda_0 U')$, and use them as the 'motif matching scores'; (2) Generate 500 iid random errors ε_i from $N(0,1)$; (3) The 500 gene expression values are computed as $y_i = \mathrm{sign}(\beta_1'\mathbf{x}_i) \times \log(|\beta_2'\mathbf{x}_i + 5|) + 0.3\varepsilon_i$, where $\beta_1 = (1, 1, 1, 1, 0, \ldots, 0)'$ and $\beta_2 = (0, \ldots, 0, 1, 1, 1, 1)'$. Thus, the functional TFBMs in this case are m_1, m_2, m_3, m_4, m_{17}, m_{18}, m_{19} and m_{20}, and the covariance matrix of $\mathbf{x}_i$ is nearly degenerate in some directions.

We generate 1000 expression-score datasets following the above procedure. For each data, we apply RSIR, SIR and the stepwise regression procedure used by Conlon *et al.* (2003) to identify functional TFBMs, and record false negative and false positive rates at various levels, which are used to generate the ROC curves in Figure 2. We observed that RSIR outperformed the SIR and stepwise regression at all sensitivity levels for this example.

Amino acid starvation

DNA microarrays were used to obtain the expression values of 5970 yeast genes before and after 0.5 h of amino acid starvation; see Gasch *et al.* (2000) for details. Conlon *et al.* (2003) extracted the upstream sequence up to 800 bp of each gene, and used MDscan to find 414 motif candidates of width between 5 and 15 bp from the upstream sequences of the 100 most induced and 100 most repressed genes. The motif-matching scores of the selected 414 motif candidates were assigned according to (1). So expression-score data were generated for 5970 genes and 414 motif candidates. The 414 motif matching scores are highly correlated due to their generating mechanism. For example, a large number of pairwise correlations between motif candidates are >0.9.

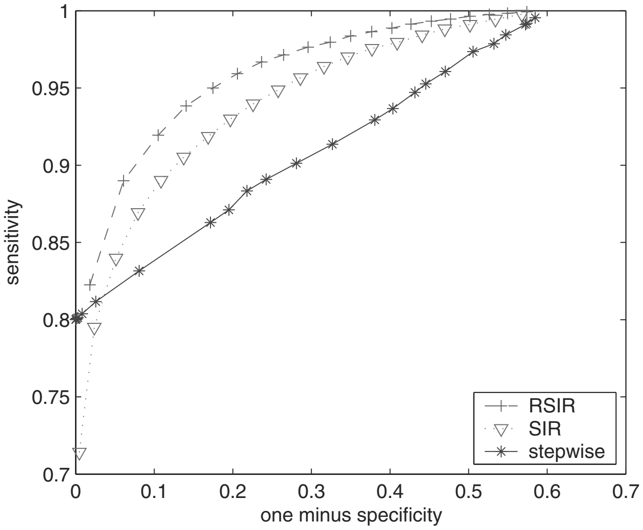

Fig. 2. ROC curves for RSIR, SIR and stepwise regression. The false negative rate is defined as the number of false rejections over the number of motifs, and the false positive rate is defined as the number of false selection over the number of motifs.

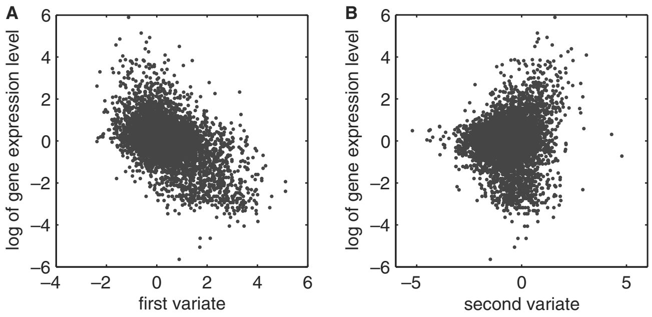

Fig. 3. (**A**) Logarithm-transformed gene expression values versus the first RSIR variate with $s = 1$. (**B**) Logarithm-transformed gene expression values versus the second RSIR variate with $s = 1$.

Using RSIR ($H = 80$) and the sequential test, we identified two significant RSIR directions ($k = 2$) at $\alpha = 5\%$. We further used the bootstrap procedure to approximate $L(s)$ at various s, and found that $L(s)$ is minimized at $s_* = 1.0$. Figure 3 displays the plots of the logarithm-transformed gene expression values against the two RSIR variates separately. Figure 3A shows that the gene expression values decrease when the first RSIR variate increases, while Figure 3B demonstrates that the dispersion of gene expression values increases as the second RSIR variate increases.

Following the motif selection procedure described in the implementation, we have identified 28 active TFBMs from the 414 motif candidates at $\alpha = 1\%$. So the original expression-score data can be reduced to include only 28 TFBMs. We further calculate the sample correlation matrix of the motif scores of the 28 TFBMs.

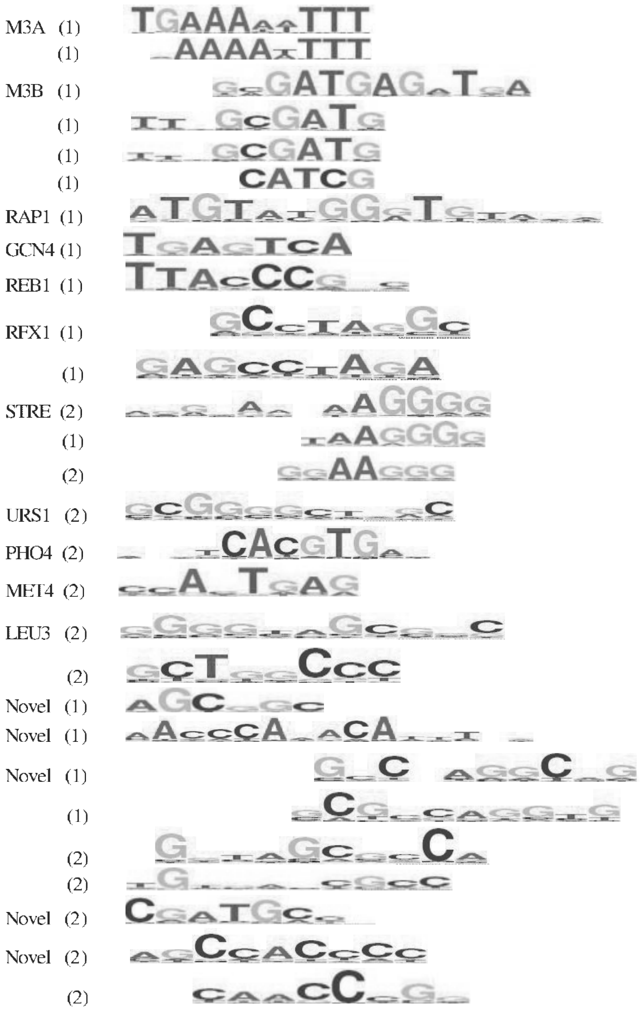

Fig. 4. The first column is motif names. The second column is the label of the groups. The third column is motif logos of the 28 selected TFBMs.

The correlation between any pair of them is <0.4, indicating that the reduced data do not have high multicollinearity as the original data do.

According to their position weight matrices, the 28 TFBMs can be classified into 16 motif patterns (Fig. 4). Eight motif patterns (STRE, GCN4, M3A, M3B, MET4, PHO4, RAP1 and URS1) are regulators known to respond to amino acid starvation and were also identified in Conlon *et al*. (2003). Another three patterns (REB1, LEU3 and RFX1) we have found also respond to amino acid starvation, but they were not identified in Conlon *et al*. (2003). It is known that REB1, which stimulates or inhibits transcription, is essential for cell growth (Morrow *et al*., 1993); LEU3, a zinc-finger transcription factor, regulates genes involved in amino acid biosynthesis; RFX1 plays an important role in the regulation of protein synthesis activities. Our findings are consistent with results from the more recent study by Harbison *et al*. (2004). The biological functions of the remaining five patterns are currently unknown and invite further biological study.

To further investigate the identified TFBMs, we apply RSIR directly to the reduced expression-score data that only include the 28 TFBMs. As we discussed above, the multicollinearity is not serious in the reduced data. Therefore we use RSIR with $s = 0$, i.e. SIR, to avoid bias. Two RSIR directions were identified at $\alpha = 5\%$ with $H = 80$. We further explore the possible linear combinations of the two RSIR directions in order to find the directions that have more interesting biological interpretations. Two directions were identified using an oblique rotation and a promax criterion. The two new directions represent two disjoint modules of TFBMs with the first involving 16 TFBMs marked (1) in Figure 4, and the second involving the remaining 12 TFBMs marked (2) in Figure 4. The plot of the logarithm of gene expression values along the first module shows the same negative linear trend as in Figure 3A, and that along the second module demonstrates the same heteroscedastic pattern as in Figure 3B. Hence, the two modules well preserve the information regarding the relationship between gene expression values and motif matching scores.

The TFBMs in the first module, which are marked (1) in Figure 4, have a significant effect on cell growth and protein synthesis. From the downstream genes of the motifs in the first module, we use GeneMerge (Castillo-Davis and Hartl, 2003) to identify significantly enriched gene function categories defined in Gene Ontology Consortium (http://www.geneontology.org/). The identified genes are closely associated with ribosome functioning, translation factors, selenoamino acid metabolism and aminoacyl-tRNA biosynthesis. In the case of amino acid starvation, processes of transcription, translations and protein synthesis should be slowed down. Thus the genes associated with these processes are expected to be downregulated. The negative trend between gene expression values and motif scores of the first module is consistent with the biological mechanism just described.

The TFBMs in the second module respond to environmental stress and mediate transcription activity. Using GeneMerge again, we identified the significantly enriched functions among the genes downstream of the TFBMs in the second module. We found that many metabolism pathways are enriched, such as arginine and proline metabolism, propanoate metabolism, pyruvate metabolism, and alanine and aspartate metabolism. It is expected that the effects of the TFBMs on these metabolism pathways are quite different during amino acid starvation. Thus, the genes could be up- or downregulated. This phenomenon is clearly demonstrated by the heteroscedastic pattern.

In summary, our analysis suggests that two primary sources are responsible for the transcription response to amino acid starvation. The first module slows down 'inessential activities' and the second module changes metabolism patterns, e.g. increases the uptake of certain amino acids.

Cell cycle regulation

Transcriptional regulation is one of the crucial regulation mechanisms for the cell cycle clock. We applied RSIR to the yeast cell-cycle data of 18 time points over two complete cell cycles, starting from release from alpha-factor arrest in the M/G1 phase (Spellman *et al*., 1998). We repeated the same procedure as in the previous example to generate the expression-score datasets at each time point. TFBMs were selected using RSIR with $H = 80$ and $s = 0.1$ at $\alpha = 5\%$ with false discovery rate correction (Storey, 2002). A total of 143 TFBMs were obtained by combining the selected TFBMs at each time point.

Table 1. Some well-known cell cycle regulator TFBMs identified using RSIR along with their active phases and *P*-values

Phase	Motif	*P*-value
M/G1	STRE	0.0023
	SWI5	0.0015
	STE12	0.0030
G1	SCB	0.0012
	MCB	0.0016
S	PHO4	0.0088
	STE12	0.0030
S/G2	SFF	0.0060
G2/M	SFF	0.0060
	SWI5	0.0015
	MCM1	0.0011

Among the TFBMs we have identified, SWI5, SCB, MCB, SFF and MCM1 are well-known cell cycle regulator TFBMs. Their active phases and *P*-values are summarized in Table 1. We find that SWI5 is active during the M/G1 phase, and SCB and MCB are active during the G1 phase. Our findings are consistent with the results in the literature. The complex of Mcm1, Fkh2 and Ndd1 proteins plays a key role in activating G2/M genes (Harbison *et al*., 2004). We also find that SFF and MCM1 are active during the G2/M phase, which agrees with recent experimental findings (e.g. Simon *et al*., 2001). In addition to the five TFBMs above, we further discover three other TFBMs, STE12, PHO4 and STRE. Because the Ste12 protein is an important transcriptional activator that has a pheromone inductive effect, STE12 is active during many phases of the cell cycle. PHO4 is active during the S phase, which is consistent with Makhnevych *et al*. (2003). STRE is active at the beginning of the cell cycle after release from alpha-factor arrest, and it responds to the stress resulting from cell cycle release.

The expression values of the cell-cycle-associated genes vary periodically over cell cycles. The correlations between gene expression values and the motif scores are expected to demonstrate the same pattern. Using *K*-means clustering, we clustered the 143 TFBMs into 10 clusters similarly as in Conlon *et al*. (2003). Eight of the ten clusters contain TFBMs that have strong effects on the cell cycle. Figure 5 shows the heatmaps for the correlations between the gene expression values and the matching scores of the TFBMs in each cluster. It is clear that the peaks of the correlations for the first eight clusters shift slowly to the right, which suggests that the TFBMs in different clusters are active one after another during the cell cycle. The correlations for the last two clusters do not fluctuate in the cell cycles, which indicates that their motifs are not necessarily cell cycle related. We notice that most of these motifs were identified in the first cell cycle, and their presence may reflect the limitation of the cell cycle experiment. The yeast cell cycle program was blocked by the alpha-factor at G1 phase in this experiment. After release, this tight synchrony decayed gradually due to the diversity of individual cell growth rates. In general, motifs identified in the second cycle are not always the same motifs identified during the first cycle.

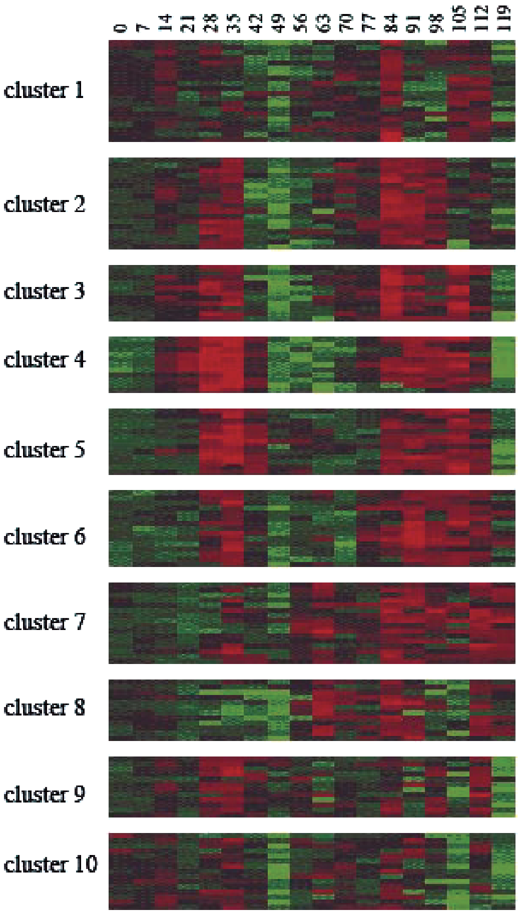

Fig. 5. Motif clusters during cell cycle. The 143 significant TFBMs selected by RSIR are clustered into 10 clusters by the correlation coefficients between the gene expression values and the upstream sequence motif matching scores.

For 4 of the 10 clusters, we plot the correlations between gene expression values and matching scores along time (Fig. 6). All the plots demonstrate cyclic patterns in two periods, which clearly reflect the effects of the motifs, identified by RSIR, on the gene expression values over time. However, the pattern in the second cycle is not as sharp as that in the first cycle, indicating the correlation between gene expression values and their upstream matching scores of TFBMs weakens as time increases. This phenomenon is also caused by the limitation of the experiment as stated in the preceeding paragraph.

CONCLUSION

The identification of TFBMs is an important research topic in computational biology. In this paper, we have proposed a novel procedure, namely RSIR, to identify TFBMs using microarray gene expression information. Unlike REDUCE, which uses the counts of motif appearances, we follow an idea of motif regressor by using motif matching scores. We do not assume the rigid linear relationship between gene expression values and motif matching scores; instead, a semi-parametric model with an unspecified (and non-linear) link function is used. The RSIR algorithm proposed in this article can identify TFBMs, while avoiding an explicit estimation

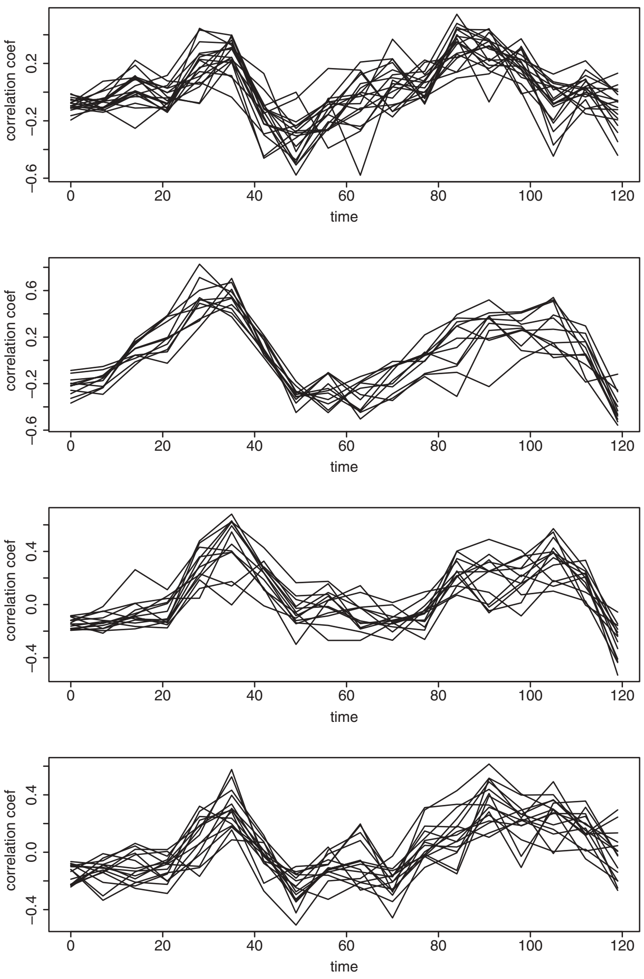

Fig. 6. Correlation coefficients between the gene expression values and the upstream sequence motif matching scores during cell cycle. From top to bottom: the first cluster contains STRE and SWI5 motifs; the second contains the MCB/SCB motifs; the third contains PHO4; the fourth contains the SFF motif.

of the link function. RSIR improves upon the SIR algorithm proposed earlier (Li, 1991) by introducing a regularization term, which allows the user to make a conscious tradeoff between bias and variance. This strategy is especially well suited and important for analyzing data with high dimensionality and high collinearity, which are typical in biological applications. An interesting and potentially useful by-product of RSIR for expression-motif data analysis is also to estimate modules consisting of functionally coherent motifs. The computation cost of RSIR is minimum, and our simulation study and real data applications show that RSIR outperforms other existing procedures. Lastly, RSIR is not limited to motif discovery, and can be applied to many other variable selection and feature identification problems with high dimensional data.

ACKNOWLEDGEMENTS

The authors are indebted to Erin Conlon, Lihua Zou and Xiaole S. Liu for their help with MDscan and motif regressor. We thank Jun-Yi Leu and Cristian Castillo-Davis for clarifying the biological results. We also thank the three anonymous referees for their constructive suggestions and comments that have helped improve an early version of the paper.

Conflict of Interest: none declared.

REFERENCES

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

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

Castillo-Davis,C. and Hartl,D. (2003) Genemerge: post-genomic analysis, data-mining and hypothesis. *Bioinformatics*, **19**, 891–892.

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

Cook,R.D. (1998) *Regression Graphics: Ideas for Studying Regressions Through Graphics*. John Wiley & Sons, New York.

Das,D. *et al.* (2004) Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 16234–16239.

Gasch,A. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Harbison,C. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hastie,T. *et al.* (1995) Penalized discriminant analysis. *Ann. Stat.*, **23**, 73–102.

Jensen,S. *et al.* (2004) Computational discovery of gene regulatory binding motifs: a bayesian perspective. *Stat. Sci.*, **19**, 188–204.

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

Keles,S. *et al.* (2004) Regulatory motif finding using logic regression. *Bioinformatics*, **20**, 2799–2811.

Lee,T. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Li,K.C. (1991) Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.*, **86**, 316–327.

Lieb,J.D. *et al.* (2001) Promoter-specific binding of rap1 revealed by genome-wise maps of protein-DNA association. *Nat. Genet.*, **28**, 327–334.

Liu,X. *et al.* (2002) An algorithm for finding protein-DNA interaction sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **27**, 835–839.

Makhnevych,T. *et al.* (2003) Cell cycle regulated transport controlled by alterations in the nuclear pore complex. *Cell*, **115**, 813–823.

Morrow,B.E. *et al.* (1993) A bipartite DNA-binding domain in yeast reb1p. *Mol. Cell Biol.*, **13**, 1173–1182.

Ren,B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Simon,I. *et al.* (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.

Sinisi,S.E. and van der Laan,M.J. (2004) Deletion/substitution/addition algorithm in learning with applications in genomics. *Stat. Appl. Genet. Mol. Biol.*, **3**, 2799–2811.

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. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B*, **64**, 479–498.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Yu,B. *et al.* (1999) Penalized discriminant analysis of *in situ* hyperspectral data for conifer species recognition. *IEEE Trans. Geosci. Remote Sensing*, **37**, 2569–2577.