Gene expression

RSIR: regularized sliced inverse regression for motif discovery
Wenxuan Zhong1,†, Peng Zeng2,†, Ping Ma3,†, Jun S. Liu1,∗ and Yu Zhu4,∗

1Department of Statistics, Harvard University, Cambridge, MA 02138, USA, 2Department of Mathematics and Statistics, Auburn University, Auburn, AL 36849, USA, 3Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA and 4Department 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 expression 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 expression-score data, or in short, the expression-score data, are mined as follows. First, a subset (10–100) of the genes with the highest expression-score data is obtained. This task can be accomplished by SIR (Li, 1991), which was originally developed for dimension reduction and data visualization. After having obtained \( \beta_1^*x, \ldots, \beta_k^*x \), we can identify \( x_j \)'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 \( \beta_1^*x, \ldots, \beta_k^*x \), 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 \( S \). Model (3) and the subspace \( S \) 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 \( x \) 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,

\[
\arg \max_{\beta \in \mathbb{R}^m} \beta' \beta \|M\beta,
\]

where \( \beta \) is a \( p \)-dimensional vector, \( M = \text{cov}[E(x|y)] \) and \( \Sigma \) is the variance covariance matrix of \( x \). Once \( \beta_1^* \) is obtained, the constraint in (4) is updated to \( \beta_1^* \Sigma \beta_1 = 0 \) and \( \beta_1^* \beta_1 = 1 \), solving the updated (4) gives \( \beta_2^* \). This procedure continues till \( \beta_1, \ldots, \beta_k \) are obtained. In application, \( \Sigma \) is estimated by the sample covariance matrix \( \hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N-1} (x_i - \bar{x})(x_i - \bar{x})' \), where \( \bar{x} \) is the sample mean. A slicing procedure is proposed in Li (1991) for estimating \( M \). First, divide the range of \( y \)'s into a number of disjoint intervals, say \( H \) intervals, which are denoted by \( S_1, S_2, \ldots, S_H \); second, for \( 1 \leq h \leq H \), calculate \( \bar{x}_h = \frac{1}{n_h} \sum_{i \in S_h} x_i \), where \( n_h \) is the number of \( y_i \) in \( S_h \); third,

\[
M = N^{-1} \sum_{h=1}^{H} n_h (x_h - \bar{x})(x_h - \bar{x})'
\]

is used as an estimate of \( M \). One then proceeds to solve (4) with \( \Sigma \) and \( M \) replaced by their estimates. And the obtained directions are denoted by \( \beta_1, \ldots, \beta_k \).

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 \( \Sigma \) nearly degenerate in a number of directions. Rewrite the sample version of (4) in an equivalent expression,

\[
\arg \max_{\beta \in \mathbb{R}^m} \|\beta\|^2 u'Mu, \quad \text{where } u = \beta' \|\beta\|.
\]

It is easy to see that the target function in (6) depends on both the norm and the orientation of \( \beta \), and the maximization is over \( \beta \) on the ellipsoid \( \mathcal{E} = \{ \beta : \beta' \Sigma \beta = 1 \} \). Along the directions in which \( \Sigma \) is degenerate, \( \beta \)'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 \( \beta_1, \ldots, \beta_k \) very unstable. In order to mitigate the variability caused by the near-degeneracy of \( \Sigma \), we add a
positive definite matrix $sI$ to $\Sigma$, where $I$ is the $p \times p$ identity matrix and $s$ is a prescribed non-negative constant. Thus the ellipsoid $E_s$ is changed to $E_s = \{ \beta : \beta' (G + sI) \beta = 1 \}$, and (6) becomes

$$\text{arg} \max_{\beta \in E_s} \beta' M \beta,$$  

(7)

which can be solved sequentially as (4) to generate $k$ directions denoted by $\beta_1(s), \ldots, \beta_k(s)$. We refer to (7) as RSIR and $\beta_1(s), \ldots, \beta_k(s)$ as the RSIR directions. Note that when $s = 0$, (7) is the sample version of (4) and $\beta_j(0) = \beta_j$ for $j = 1, \ldots, k$.

In fact, $\beta_1(s), \ldots, \beta_k(s)$ can be obtained by solving the following system of equations and constraints,

$$M \beta_j = \lambda_j (I + sI) \beta_j, \quad \beta_j' (I + sI) \beta_j = 1 \quad \text{if} \quad i = j = 0 \text{if} \quad i \neq j,$$

(8)

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0,$$

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 $M$ using (5) with fixed $H$. Second, we solve (8) to obtain $\beta_1(s), \ldots, \beta_k(s)$. The linear combinations $\beta_i(s)x$, $i = 1, \ldots, k$ are referred to as the RSIR variates in the rest of the paper. For visualization, we further generate $k$ plots of $y$ against $\beta_i(s)$, $1 \leq i \leq k$, 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 $\beta_{1j}(s), \beta_{2j}(s), \ldots, \beta_{kj}(s)$. According to model (3), intuitively, if all the $\beta_{1j}(s), \beta_{2j}(s), \ldots, \beta_{kj}(s)$ are close to zero, then $m_j$ is not a functional TFBM. Next we propose a scoring procedure to determine functional TFBMs.

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

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

The exact sampling distribution of $\hat{\beta}_{1j}(s)$ is difficult to derive. When $s = 0$, Li (1991) stated in his Remark 5.1 that $\hat{\beta}_{1j}(0)$ is asymptotically normal with mean $\beta_{1j}(0)$ and covariance $\Sigma_j(0)$. For $s > 0$, the same result still holds, i.e. $\hat{\beta}_{1j}(s)$ asymptotically follows a normal distribution with mean $\beta_{1j}(s)$ and covariance $\Sigma_j(s)$.

**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

$$L(s) = \sum_{i=1}^{k} \text{tr}(\text{cov}(\hat{\beta}_{1j}(s))) + \sum_{i=1}^{k} \|E(\hat{\beta}_{1j}(s)) - \beta_{1j}\|^2,$$

(9)

where $\text{tr}$ 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 $V(s) = \sum_{i=1}^{k} \text{tr}(\text{cov}(\hat{\beta}_{1j}(s)))$ is the sum of the variances of $\hat{\beta}_{1j}(s)$, and the second part $B(s) = \sum_{i=1}^{k} \|E(\hat{\beta}_{1j}(s)) - \beta_{1j}\|^2$ is the sum of the squared biases. When $s$ increases from 0, $L(s)$ first decreases. After $L(s)$ reaches its minimum at $s_0$, it begins to increase as $s$ further increases. We choose $s_0$ to be the amount of regularization in RSIR.
Using $s_\alpha$, 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 $\beta_{(s)}^{(1)}(s), \ldots, \beta_{(s)}^{(m)}(s)$ generated by the bootstrap procedure discussed in the previous subsection to approximate $V(s)$ and $E(\beta_{(s)}(s)).$ The true directions $\beta$ are approximated by the bootstrap mean of $E(\beta_{(s)}(s))$ with $s_0$ being a suitably small positive value. Although the proposed procedure for determining $s$ is approximative, 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 $M = \text{cov}(E(x|y))$. 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 $\beta_{(s)}(i|x)$ 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 $\lambda_1(s) \geq \lambda_2(s) \geq \cdots \geq \lambda_p(s)$ be the eigenvalues calculated from (8). Define $\Lambda_d(s) = \sum_{i=d+1}^p \lambda_i(s)$. When $s = 0$, Li (1991) proved that $\Lambda_{d}(0)$ follows a $\chi^2$ 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 $20 \times 20$ orthogonal matrix, and let $\Lambda_0$ 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 $x_\iota$ from $N(0, \Lambda_0/F)$, and use them as the ‘motif matching scores’; (2) Generate 500 iid random errors $e_i$ from $N(0,1)$; (3) The 500 gene expression values are computed as $y_j = \text{sign}(\beta^T x) \times \log(|\beta^T x| + 5) + 0.3 e_i$, where $\beta_1 = (1, 1, 1, 1, 0, \ldots, 0)^T$ and $\beta_2 = (0, \ldots, 0, 1, 1, 1, 1)^T$. Thus, the functional TFBMs in this case are $m_1, m_2, m_3, m_4, m_5, m_6, m_7, m_8, m_9, m_{10}$, and the covariance matrix of $x_\iota$ 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.

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 $s_\alpha = 1.0$. 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 shows 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.
The groups. The third column is motif logos of the 28 selected TFBMs. The correlation between any pair of them is identified in Conlon et al regulators known to respond to amino acid starvation and were also (STRE, GCN4, M3A, M3B, MET4, PHO4, RAP1 and URS1) are directly to the reduced expression-score data that only include the unknown and invite further biological study. The biological functions of the remaining five patterns are currently identified in Conlon et al regulators known to respond to amino acid starvation and were also (STRE, GCN4, M3A, M3B, MET4, PHO4, RAP1 and URS1) are directly to the reduced expression-score data that only include the unknown and invite further biological study.

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.

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.
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.

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 preceding 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.
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


