5246–5254 *Nucleic Acids Research, 2009, Vol. 37, No. 16* *Published online 3 July 2009*
doi:10.1093/nar/gkp554

Discovery of protein–DNA interactions by penalized multivariate regression

Leonid Zamdborg[1,2] and Ping Ma[2,3,*]

[1]Center for Biophysics and Computational Biology, [2]Institute for Genomic Biology and [3]Department of Statistics, University of Illinois at Urbana-Champaign, IL, USA

Received May 15, 2009; Revised June 11, 2009; Accepted June 14, 2009

ABSTRACT

Discovering which regulatory proteins, especially transcription factors (TFs), are active under certain experimental conditions and identifying the corresponding binding motifs is essential for understanding the regulatory circuits that control cellular programs. The experimental methods used for this purpose are laborious. Computational methods have been proven extremely effective in identifying TF-binding motifs (TFBMs). In this article, we propose a novel computational method called MotifExpress for discovering active TFBMs. Unlike existing methods, which either use only DNA sequence information or integrate sequence information with a single-sample measurement of gene expression, MotifExpress integrates DNA sequence information with gene expression measured in multiple samples. By selecting TFBMs that are significantly associated with gene expression, we can identify active TFBMs under specific experimental conditions and thus provide clues for the construction of regulatory networks. Compared with existing methods, MotifExpress substantially reduces the number of spurious results. Statistically, MotifExpress uses a penalized multivariate regression approach with a composite absolute penalty, which is highly stable and can effectively find the globally optimal set of active motifs. We demonstrate the excellent performance of MotifExpress by applying it to synthetic data and real examples of *Saccharomyces cerevisiae*. Motif Express is available at http://www.stat.illinois.edu/~pingma/MotifExpress.htm.

INTRODUCTION

Transcription factors (TFs) regulate the expression of target genes by binding in a DNA sequence-specific manner to their recognition sites in the promoter regions of these genes. The common pattern of the binding sites for a particular TF is called a TF-binding motif (TFBM), usually modeled by a position-specific weight matrix (PWM). Discovery of TF-binding sites (TFBSs) and TFBMs in TF–DNA interaction is essential for understanding the regulatory circuits that control cellular programs. In recent years, considerable progress has been made in developing both experimental and computational methods for elucidating TFBSs, and the mapping of their locations in a number of model organisms. Experimental techniques such as ChIP-chip (1) on promoter microarrays and whole genome tiling arrays and ChIP-seq (2) have been used to discover genome-wide TF–DNA-binding sites for organisms ranging from *Saccharomyces cerevisiae* (3) to *Homo sapiens* (4). Nonetheless, these experimental techniques are laborious and expensive, and require specialized antibodies which may be difficult to obtain (5). As a given binding site is not necessarily occupied under all conditions (6), a single set of ChIP-chip/ChIP-seq experiments conducted under one experimental condition is insufficient to fully detect all sites to which a TF may bind to in another experimental condition. Because of these limitations, computational methods provide appealing alternatives for pinning down TFBSs. *De novo* motif discovery algorithms based on probability models using PWMs, such as AlignACE (7), MDscan (8), MEME (9) and Weeder (10) have been accepted as important components of the computational biologist's toolkit. Moreover, rapid progress has been made to detect *cis*-regulatory modules which consist of highly coordinated TFBMs (11,12).

A recent trend to improve the aforementioned computational methods is to integrate information from relatively inexpensive and easily obtained gene expression data. The key idea to facilitate motif discovery using gene expression is that a gene's mRNA copy number is associated with active TFBMs' matching scores (or more intuitively, number of TFBM copies) in the promoter region of this gene. A number of attempts have been made along this line of thinking. For example, REDUCE (13) uses an exhaustive oligo-enumeration strategy to identify a potential set of candidate motifs,

*To whom correspondence should be addressed. Tel: +1 217 244 7095; Fax: +1 217 244 7190; Email: pingma@illinois.edu

then 'reduces' this candidate set to a set of active motifs whose binding was best correlated with gene expression. As an improvement over REDUCE, Motif Regressor (14), incorporates gene expression in the initial identification of candidate motifs; the top-ranking genes in a single sample microarray experiment are used to identify an initial set of candidate motifs by MDscan (8). This candidate list is then winnowed using stepwise regression with genome-wide gene expression serving as the response in a multiple regression model, resulting in a subset of motifs that are best correlated with gene expression across the genome.

These two approaches have stimulated many further studies in the past several years (15), and have generated a number of interesting results (16). However, several issues remain that hinder their effective application in real practice. Since they rely on a single sample microarray measurement to carry out ranking and regression, they are sensitive to experimental and biological noise, especially in regards to low copy-number genes. Consequently, they may select different sets of motifs depending which single sample of microarray experiment is used in regression, requiring time-consuming manual merging and validation of the selected sets of motifs. Additionally, the stepwise motif selection in Motif Regressor relies on adding/deleting one motif at a time, a technique which is highly unstable and can only explore a small portion of all the possible models as the number of candidate motifs is usually hundreds (17).

To overcome these obstacles, we propose MotifExpress, a novel method that selects a set of motifs that best correlate with multiple samples of gene expression measured by microarrays simultaneously. We utilize multivariate regression to link gene expression (as responses) and candidate motifs (as predictors) together. In the multivariate regression framework, the selection of active motifs is very challenging as the number of parameters is much larger than the number of motifs. Thus we have a huge space to search for the globally optimal model which gives rise to the set of active motifs. To surmount this challenge, we fit a model using a composite absolute penalty (CAP). Unlike the stepwise regression procedure, the CAP procedure selects motifs via a convex optimization and can effectively find the globally optimal set of active motifs. We use the Bayesian information criterion (BIC) to select the regularization parameter. We demonstrate the excellent performance of MotifExpress by applying it to synthetic data as well as GCN2 constitutive activation and heat shock experiments in *Saccharomyces cerevisiae*. It is evident from these results that incorporating multiple samples of gene expression substantially reduced the number of spurious results.

MATERIALS AND METHODS

Microarray and sequence data

Microarray data was retrieved from Gene Expression Omnibus (GEO) database and $\log_2$ base transformed. Missing values were estimated using k-nearest-neighbor imputation (18), implemented in the R (19) *impute* package. The upstream 800-bp sequence for each gene was used, with repetitive sequences masked using RepeatMasker (20).

MotifExpress—framework

First, significance analysis of microarrays (SAM) (21), a microarray analysis algorithm, is used to identify genes that are differentially expressed between the treatment and control conditions. The upstream promoter sequences of significantly differentially expressed genes are then used as input in a *de novo* motif discovery algorithm (MDscan) to search for candidate motifs. The upstream promoter sequences of all genes for which expression was measured are then scored for matches to each candidate motif; our MotifExpress algorithm then uses multivariate regression to link gene expression in all samples with the motif matching scores of all candidate motifs. The active motifs that are significantly associated with gene expression are identified using a composite absolute penalty approach with the regularization parameter selected through minimizing BIC (22) (Figure 1).

MotifExpress—motif discovery

SAM (21) analysis is run on the gene expression profiles to determine which genes are differentially expressed between treatment and control conditions at a pre-specified false discovery rate. MDscan (8) is then run on the upstream promoter sequences of the significantly differentially expressed genes to discover motifs ranging from 7 to 15 bp in width. The 30 most significant motifs for each motif width are combined to form a set of candidate motifs. We then calculate the motif matching score x_{ik} which indicates how likely motif k binds upstream of gene i in terms of both goodness of matching and number of sites as defined in Conlon *et al.* (14).

$$x_{ik} = \log_2\left[\sum_{s\in S_{iw}} \frac{\Pr(s \text{ from } \theta_k)}{\Pr(s \text{ from } \theta_o)}\right] \quad \mathbf{1}$$

where $k = 1, \ldots, p$, and p is the total number of candidate motifs, θ_k is the probability matrix of motif k of width w, θ_0 is the third-order Markov model learned from intergenic sequences, and S_{iw} is the set of all w-mers in the upstream of gene i.

MotifExpress—integration of gene expression with motif selection

The gene expression profile of gene i is denoted as $y_i = (y_{i1}, y_{i2}, \ldots, y_{im})$ where y_{ij} is the expression ratio in sample j for gene i, and m is the number of samples. Motif discovery further associates expression of gene i with all candidate motifs' matching scores x_{ik}, $k = 1, \ldots, p$. We assume a multivariate regression model between expression y and motif matching scores x

$$y_{ij} = x_{i1}\beta_{1j} + x_{i2}\beta_{2j} + \cdots + x_{ip}\beta_{pj} + \varepsilon_{ij} \quad \mathbf{2}$$

where random errors ε_{ij} are independently and identically distributed with mean zero and standard deviation

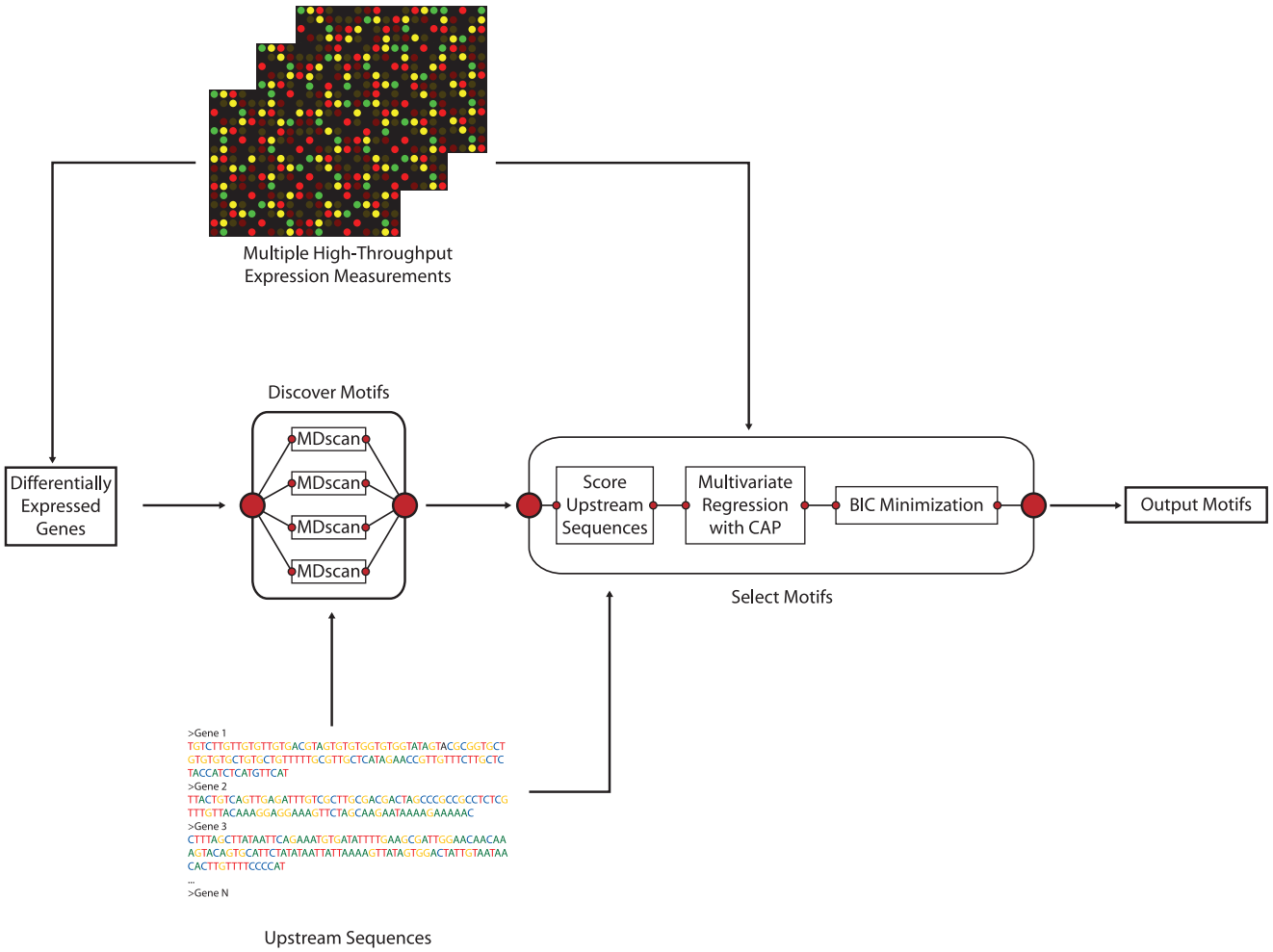

Figure 1. MotifExpress system diagram.

σ_ε, and β_{kj} is the unknown coefficient which relates the expression of gene i to the motifs that putatively regulate it. We may write Equation (2) as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} (\beta_{11} \ldots \beta_{1m}) + \begin{pmatrix} x_{12} \\ \vdots \\ x_{n2} \end{pmatrix} (\beta_{21} \ldots \beta_{2m}) + \cdots + \begin{pmatrix} x_{1p} \\ \vdots \\ x_{np} \end{pmatrix} (\beta_{p1} \ldots \beta_{pm}) + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \quad \mathbf{3}$$

Note that for any motif k, if $\beta_{k1} = \ldots = \beta_{km} = 0$, then motif k is not associated with gene expression. We thus infer that motif k is not active in the biological conditions under which gene expression was measured. Otherwise, motif k is inferred to be active. Identifying active motifs thus becomes a variable selection problem in Equation (3).

MotifExpress—motif selection using penalized multivariate regression with CAP

By combining regression with variable selection, it is possible to select a set of active motifs which is significantly associated with gene expression. Classically, stepwise regression is used for variable selection; however, it is sensitive to perturbation of the data and can only explore a small portion of all the possible models as the number of candidate motifs is usually hundreds.

Lasso (23) has recently received significant attention as an efficient variable selection method.

Lasso estimates the coefficients of predictors through minimizes the following expression

$$\text{RSS} + \lambda \times \text{Sum of Absolute values of coefficients} \quad \mathbf{4}$$

where the residual sum of squares (RSS) and the sum of Absolute values of coefficients are two conflicting measures: the model with a smaller residual sum of squares tends to have more nonzero coefficients, which in turn results in a higher sum of absolute values of coefficients. λ is a regularization parameter controlling the trade-off between these two goals. Compared to a solution that minimizes only the residual sum of squares, i.e. the least squares estimate, the estimated coefficients in Lasso are closer to zero, which is referred to as 'shrinking'. It has been shown that the Lasso can select predictors consistently, i.e. selecting correct predictors

with probability one asymptotically, by shrinking the coefficients of the insignificant predictors to zero. However, Lasso was developed for regression with a single response rather than that with multivariate responses as in Equation (2). Moreover, we are interested in eliminating any inactive motif in Equation (2), which requires simultaneously shrinking all m coefficients corresponding to that motif to zero, rather than shrinking an individual coefficient. Recently, the simultaneous variable selection (23) and group Lasso (24,25) methods have been developed for selecting groups of variables. These methods have been nicely summarized in a unified shrinkage method, the composite absolute penalty (CAP) approach (25). We apply the idea of a composite absolute penalty to a multivariate regression model in Equation (2).

Our estimate is defined as the minimizer of:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-x_{i1}\beta_{1j}-x_{i2}\beta_{2j}-\cdots-x_{ip}\beta_{pj})^2 + \lambda\times \mathrm{pen}(\beta_{11},\beta_{12},\ldots,\beta_{\mathrm{pm}}) \quad (5)$$

where λ is a regularization parameter. The composite absolute penalty uses a combination of various metrics to achieve the objective of group predictor selection. One possible choice of penalty to achieve this goal is $\mathrm{pen}(\beta_{11},\beta_{12},\ldots,\beta_{pm})=\sum_{k=1}^{p}\sqrt{\beta_{k1}^2+\beta_{k2}^2+\cdots+\beta_{km}^2}$, which reduces to the group Lasso penalty. However, the computation of group Lasso relies on a shooting algorithm, and the cost of computation is very high (24). We instead elected to use the following CAP:

$$\mathrm{pen}(\beta_{11},\beta_{12},\ldots,\beta_{pm})=\sum_{k=1}^{p}\max(|\beta_{k1}|,|\beta_{k2}|,\ldots,|\beta_{km}|) \quad (6)$$

which is the penalty used in (23). For each motif, the corresponding m coefficients are grouped through their maximum absolute values. As λ is increased, the group of motif coefficients β shrink simultaneously; a motif whose group of coefficients have shrunk to zero falls out of the model.

Since minimizing Equation (5) is a convex optimization problem, a solution satisfying the Karush-Kuhn–Tucker conditions is a global minimum (26,27). A beneficial feature of the proposed method Equation (5) with penalty Equation (6) is that the solution has a piecewise linear solution path for all values of λ. We adopt the homotopy algorithm (28,29), also known as the LARS/Lasso algorithm (30) to find the solutions for all values of λ. Even though the solution path for all values of λ can be effectively computed, it is still highly desirable that one solution is given for a fine-tuned value of λ. To choose a value of λ with a good balance of goodness-of-fit of the model and model parsimony, we minimize the BIC (22),

$$nm\ln\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(y_{ij}-x_{i1}\hat{\beta}_{1j}(\lambda)-\cdots-x_{ip}\hat{\beta}_{pj}(\lambda))^2}{nm} + p_A(\lambda)\times\ln(nm) \quad (7)$$

where $\hat{\beta}_{1j}(\lambda)$, $\hat{\beta}_{2j}(\lambda),\ldots,\hat{\beta}_{pj}(\lambda)$ are CAP estimates of β_{1j}, $\beta_{2j},\ldots,\beta_{pj}$, $p_A(\lambda)$ is the number of estimated active motifs, i.e. nonzero coefficients, and they all depend on λ. Since the solution path is piecewise linear, the smallest BIC can be found by comparing Equation (7) for a number of λ values.

To test the significance of the selected motif, we calculate a pooled P-value for each selected motif by combining all P-values of corresponding m coefficients using Stouffer's method (31,32).

Functional annotation of discovered motifs

To verify identifications and further elucidate biological relationships, manual analysis and functional annotation was carried out on discovered motifs. Results were validated where possible by comparison to known TF-binding sites by ChIPCodis (33). Discovered motifs were further putatively identified by Tomtom (34) against the MacIsaac *et al*. yeast TFBM dataset (3) as well as additional STAMP (35) comparison to the common ribosome-associated RRPE and PAC motifs (36).

RESULTS

Simulation results

Extensive simulations were carried out to examine the effectiveness of MotifExpress in identifying active motifs. We set number of genes, $n = 5000$, and number of the motif candidates as 100, among which 10 motifs were active. We generated motif scores x_{ik} from $N(0, 1)$ and random error from $N(0,\sigma_\varepsilon)$. We gradually increased the standard deviation of random error from $\sigma_\varepsilon = 1$ to 5. Gene expression was generated as the summation of linear combinations of the active motif scores and random error as in Equation (2). We let $m = 3, 4, 5$. We generated 100 datasets for each setting, i.e. each combination of σ_ε and m. The Matthews correlation coefficient (MCC) (37), was calculated for each dataset, where a value of 1 indicates perfect selection of active motifs and rejection of spurious motifs, while a value of 0 is average random selection.

The summary statistics of the resulting MCCs are given in Table 1. It can be seen that our proposed method consistently performs very well across all settings. Since AIC with second order correction (AICc) was also suggested as an alternative criterion to choose the regularization parameter, we tested the performance of our proposed method using AICc. MotifExpress with BIC-minimization consistently outperformed that with AICc-minimization; average MCCs were significantly higher across all simulation regimes. MotifExpress with AICc was observed to identify more motifs spuriously; its performance was robust as σ_ε increased, but had a lower MCC throughout the settings (Figure 2). In comparison, the performance of MotifExpress with BIC did degrade as σ_ε increased, but still had much higher MCC as we see in Figure 2. We also ran Motif Regressor one-response-at-a-time and combined the identified motifs.

The mean MCCs of Motif Regressor are consistently higher than those of MotifExpress with

Table 1. Mean (Ave.) and standard deviation (SD) of MCC for MotifExpress with regularization parameters selected through AICc-minimization and BIC-minimization as the random error's standard deviation σ_ε and the number of samples m are varied in the simulation study

σ_ε	$m = 3$				$m = 4$				$m = 5$			
	MCC (AICc)		MCC (BIC)		MCC (AICc)		MCC (BIC)		MCC (AICc)		MCC (BIC)	
	Ave.	SD	Ave.	SD	Ave.	SD	Ave.	SD	Ave.	SD	Ave.	SD
1	0.582	0.143	0.928	0.056	0.620	0.159	0.933	0.057	0.587	0.143	0.940	0.058
2	0.590	0.158	0.923	0.068	0.572	0.147	0.918	0.058	0.639	0.156	0.939	0.057
3	0.580	0.142	0.895	0.069	0.591	0.150	0.904	0.072	0.614	0.148	0.922	0.068
4	0.577	0.149	0.872	0.077	0.556	0.132	0.854	0.073	0.609	0.139	0.864	0.086
5	0.576	0.150	0.839	0.091	0.575	0.135	0.831	0.081	0.598	0.144	0.838	0.078

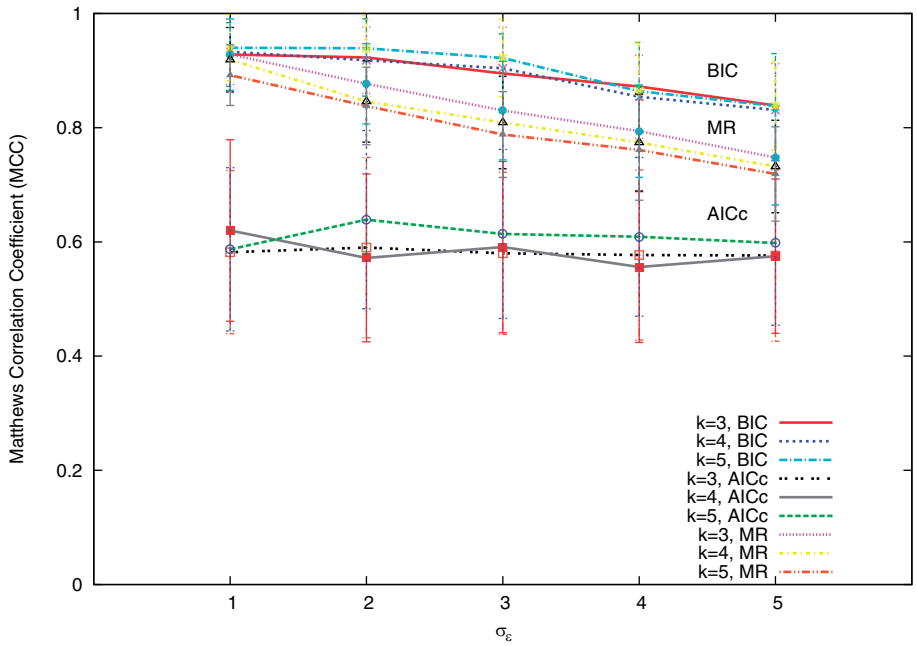

Figure 2. Summary plot of MCC for Motif Regressor (MR) and MotifExpress with regularization parameters selected through AICc-minimization and BIC-minimization as the random error's standard deviation σ_ε and the number of samples m are varied in the simulation study. Motif Regressor performance in the same simulation was computed by pooling results from independent runs.

AICc-minimization, but lower than those of MotifExpress with BIC-minimization. Moreover, as number of response m increase, we notice that the MCC of Motif Regressor drops since the number of false discovery goes up (Figure 2).

GCN2 constitutive activation analysis

The protein kinase *GCN2* has drawn attention in recent years due to its extensive regulatory impact. The Gcn2p homodimer associates with the large ribosome subunit in the cytosol, and when activated, phosphorylates Ser-52 of eIF2α (38). In yeast, this has two immediate effects; firstly, the repression of general translation by sequestration of eIF2β, and secondly the derepression of *GCN4* translation. Gcn4p then acts as a TF that modulates the expression of numerous stress- and starvation-related genes (39). Gcn2p may be activated by numerous signals via multiple pathways, including the drug rapamycin (40).

We elected to use constitutively active Gcn2 as a means of validating our method. As active Gcn2p results in Gcn4p activation, consequentially leading to an activation of downstream genes, it follows that the presence of the GCN4 motif should be strongly correlated with gene expression under the condition of constitutively active Gcn2p. A four-sample dataset of cDNA microarrays, which hybridized four biological replicates of $GCN2^c$ constitutively active mutant samples to a common reference wild-type sample, was retrieved from GEO (GSE8111) (41). The data is the log transformed gene expression ratios between mutant and wild-type, and was used by MotifExpress as the response to fit multivariate regression model. Three motifs were selected by MotifExpress, identified by STAMP as GCN4, PAC and RAP1.

The results obtained by MotifExpress were compared with those obtained by running Motif Regressor (14) separately on each sample. In contrast to three motifs

Table 2. GCN4 motif discovery on constitutively activated Gcn2 mutant dataset by MotifExpress on all samples simultaneously and by Motif Regressor, on each sample individually

Analysis	Number of Motifs Identified	GCN4 motif with smallest *P*-value	Rank of Result
MotifExpress	3	1.73×10^{-40}	1st in 3
MotifRegressor Sample 1	42	8.86×10^{-3}	9th in 42
MotifRegressor Sample 2	15	4.37×10^{-4}	3rd in 15
MotifRegressor Sample 3	55	1.22×10^{-3}	7th in 55
MotifRegressor Sample 4	18	8.28×10^{-12}	2nd in 18

discovered by MotifExpress, Motif Regressor discovered 130 motifs. For Motif Regressor running on each sample, the smallest *P*-value of the motifs identified by STAMP (35) as GCN4 and the highest rank of the GCN4 motifs sorted by *P*-values in the result is reported in Table 2. MotifExpress analysis resulted in a parsimonious set of results, where the GCN4 motif had the lowest *P*-value and ranked first, while analyzing individual samples in the dataset by Motif Regressor yielded highly heterogeneous results. The *P*-value was much lower in the MotifExpress results than in any of the Motif Regressor results.

The presence of the PAC and RAP1 motifs is likewise unsurprising; the RP regulon (strongly associated with the RAP1 motif) and the RRB regulon (strongly associated with the PAC motif) (42) are both known to be repressed by treatment with rapamycin, which is also known to induce Gcn4p synthesis (40). The GSE8111 dataset shows a transcription profile quite similar to rapamycin treatment, with the RRB and RP genes downregulated, and amino-acid biogenesis genes upregulated. Analysis of genes used for motif discovery via ChIPCODIS revealed a significant overrepresentation of genes to which Gcn4p binds under rapamycin treatment, with a pooled *P*-value of 1.56×10^{-18}.

Heat shock analysis

The heat shock response is a conserved and concerted cellular program in eukaryotes. Temperature changes above the physiological optimum induce the synthesis of heat shock proteins, a diverse class of proteins that have effects on protein folding, metabolism and antioxidant response. Expression of the genes coding for these proteins is regulated by a set of stress-related TFs, most importantly Hsf1p and Msn2/4p.

A three-sample dataset of cDNA microarrays, comparing wild-type cells in midlog-phase grown at 30°C to wild-type cells heat-shocked at 39°C for 15 min, was downloaded from GEO (GSE7665) (43). The data was the log-transformed gene expression ratios between heat shock and control conditions, and was used by MotifExpress as the response to fit a multivariate regression model. A set of 15 active motifs was selected by MotifExpress, among them HSF1, RPH1, MSN2/4, SFP1, FHL1 and PAC; of these, the HSF1 motif was the most significant, with a pooled *P*-value of 7.8×10^{-37}. In contrast, Motif Regressor discovered 113 motifs, many of which were redundant.

It is known that many of the genes regulated by Hsf1p encode chaperones, proteins responsible for inducing and maintaining protein conformation and preventing unwanted protein aggregation, such as HSP82 (44). It would be expected that cells undergoing heat shock would experience an up-regulation of genes regulated by Hsf1p and Msn2/4p; analysis by MotifExpress analysis demonstrates the detection of this response. The PAC motif, as mentioned in the previous example, is a signature of the RRB regulon, which is down-regulated under heat shock conditions (45). The genes in the dataset were confirmed to be significantly overrepresented for Hsf1p, Msn2/4p, Fhlp, and Sfp1p binding under stress conditions by ChIPCODIS (*P*-values ranging from 8.82×10^{-62} to 3.63×10^{-18}).

It is interesting to note that the HSF1 motif is known to consist of repeats of a 5-bp consensus sequence 5′-NGAAN-3′ and its reverse complement 5′-NTTCN-3′ (46). In Figure 3, we plotted the motif logos of the most significant HSF1 motifs discovered by MotifExpress, by Motif Regressor via analyzing each sample independently, and that discovered using ChIP-chip and phylogentic methods in MacIsaac *et al.* (3). The HSF1 motif discovered by MotifExpress is the closest to the consensus sequence reported in (46) among the five motifs (Figure 3).

DISCUSSION

In this article, we developed a novel method, MotifExpress, for identifying TFBMs strongly associated with multiple samples of gene expression. Existing methods for identifying TFBMs correlate sequence information to a single sample of gene expression, one sample at a time, which results in a redundant set of active motifs with many spurious results (47). Additionally, existing methods rely on classical variable selection techniques such as stepwise regression, which is highly unstable and can only explore a small portion of all the possible models as the number of candidate motifs is usually in the hundreds. Our method is designed to integrate multiple samples of gene expression via multivariate regression. Using the CAP approach and selecting the regularization parameter using BIC, we can effectively identify a parsimonious set of active motifs. We examined the performance of MotifExpress using synthetic data under an array of settings with different numbers of samples and various variance magnitudes of random error. MotifExpress performed consistently well throughout all settings. We then analyzed two real experiments using MotifExpress, identifying active motifs correlated with expression. The set of discovered motifs agreed well with current literature.

The MotifExpress framework is easily extensible to support other TF–DNA-binding discovery methods, especially *cis*-regulatory module discovery methods. Statistically, penalized multivariate regression with CAP is ready to incorporate additional structural information about motifs. Likewise, as high-throughput transcriptomic studies transition from hybridization-based

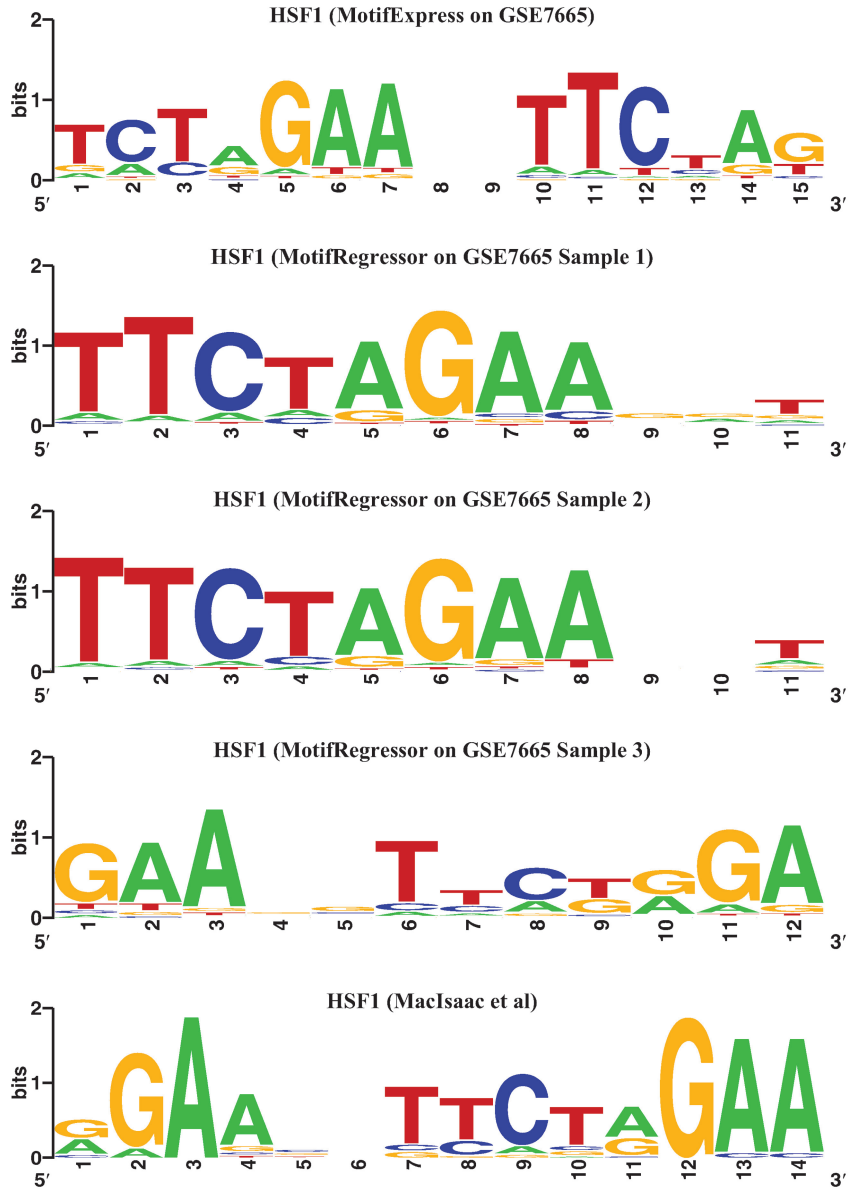

Figure 3. HSF1-binding motif discovered by MotifExpress analyzing all samples in in GSE7665 simultaneously compared to MotifRegressor analyzing each sample individually and current literature. The head-to-head inverted NGAAN motif is prominent in the MotifExpress results.

microarrays to rapid whole-transcriptome sequencing, this new data is easily integrated in MotifExpress.

Aside from motif selection, another challenge is to identify the regulatory targets of a TF. In principle, given the motifs (including promoter sequence) and estimated coefficients, we can predict gene expression. Then the genes with significant high or low expressions could be considered as potential regulatory targets. However, such prediction in practice typically has too large prediction uncertainty to be used for identifying regulatory targets. A possible alternative is to build a prediction model with gene cluster membership as the response, e.g. Beer and Tavazoie (48). The variable selection method that we employed in this article can be adapted to that model.

ACKNOWLEDGEMENTS

The authors are grateful to X. Shirley Liu and Berwin A. Turlach for access to their source code. Majority of the work was done while PM was visiting department of statistics at UC Berkeley. PM is very grateful to Peter Bickel for hosting the visit and having many stimulating discussions.

FUNDING

National Science Foundation [DMS-0800631]. Funding for open access charge: National Science Foundation DMS-0800631.

Conflict of interest statement. None declared.

REFERENCES

1. Buck,M.J. and Lieb,J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.
2. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
3. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, **7**, 113.
4. Wei,C.-L., Wu,Q., Vega,V.B., Chiu,K.P., Ng,P., Zhang,T., Shahab,A., Yong,H.C., Fu,Y., Weng,Z. *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**, 207–219.
5. Wu,J., Smith,L.T., Plass,C. and Huang,T.H. (2006) ChIP-chip comes of age for genome-wide functional analysis. *Cancer Res.*, **66**, 6899–6902.
6. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.-B., Pokholok,D.K., Kellis,M. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
7. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *Cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
8. Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnol.*, **20**, 835–839.
9. Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Sec. Int. Conf. Intell. Sys. Mol. Biol.,* 28–36.
10. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
11. Zhou,Q. and Wong,W.H. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.
12. Gupta,M. and Liu,J.S. (2005) De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **102**, 7079–7084.
13. Roven,C. and Bussemaker,H.J. (2003) REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Res.*, **31**, 3487–3490.
14. Conlon,E.M., Liu,X.S., Lieb,J.D. and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.
15. Das,D., Pellegrini,M. and Gray,J.W. (2009) A primer on regression methods for decoding cis-regulatory logic. *PLoS Comput. Biol.*, **5**, e1000269.
16. Ben-Yehuda,S., Fujita,M., Liu,X.S., Gorbatyuk,B., Skoko,D., Yan,J., Marko,J.F., Liu,J.S., Eichenberger,P., Rudner,D.Z. *et al.* (2005) Defining a centromere-like element in Bacillus subtilis by Identifying the binding sites for the chromosome-anchoring protein RacA. *Mol. Cell*, **17**, 773–782.
17. Breiman,L. (1996) Heuristics of instability and stabilization in model selection. *Ann. Stat.*, **24**, 2350–2383.
18. Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
19. R Development Core Team. (2008). R Foundation for Statistical Computing, Vienna, Austria.
20. Smit,A.F.A., Hubley,R. and Green,P. (1996–2004) *RepeatMaster Open-3.0*.
21. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
22. Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.
23. Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B (Method)*, **58**, 267–288.
24. Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Method)*, **68**, 49–67.
25. Zhao,P., da Rocha,G.V. and Yu,B. (In Press) The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat.*
26. Karush,W. (1939) Minima of functions of several variables with inequalities as side constraints. *Master's Thesis*. University of Chicago, Chicago, Illinois.
27. Kuhn,H.W. and Tucker,A.W. (1951) Nonlinear programming. In *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, California, pp. 481–492.
28. Turlach,B.A., Venables,W.N. and Wright,S.J. (2005) Simultaneous Variable Selection. *Technometrics*, **47**, 349–363.
29. Osborne,M.R., Presnell,B. and Turlach,B.A. (2000) A new approach to variable selection in least squares problems. *IMA J. Numerical Anal.*, **20**, 389–403.
30. Efron,B., Hastie,T., Johnstone,I. and Tibshirani,R. (2004) Least angle regression. *Ann. Stat.*, **32**, 407–451.
31. Burns,P. (2007) In Kontoghiorghes,E.J. and Gatu,C. (eds), *Optimization, Econometric and Financial Analysis*. Springer, Berlin, pp. 239–240.
32. Whitlock,M.C. (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.*, **18**, 1368–1373.
33. Abascal,F., Carmona-Saez,P., Carazo,J.M. and Pascual-Montano,A. (2008) ChIPCodis: mining complex regulatory systems in yeast by concurrent enrichment analysis of chip-on-chip data. *Bioinformatics*, **24**, 1208–1209.
34. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
35. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.
36. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.
37. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
38. Padyana,A.K., Qiu,H., Roll-Mecak,A., Hinnebusch,A.G. and Burley,S.K. (2005) Structural basis for autoinhibition and mutational activation of eukaryotic initiation factor 2alpha protein kinase GCN2. *J. Biol. Chem.*, **280**, 29289–29299.
39. Hinnebusch,A.G. and Natarajan,K. (2002) Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. *Eukaryot. Cell*, **1**, 22–32.
40. Kubota,H., Obata,T., Ota,K., Sasaki,T. and Ito,T. (2003) Rapamycin-induced translational derepression of GCN4 mRNA involves a novel mechanism for activation of the eIF2 alpha kinase GCN2. *J. Biol. Chem.*, **278**, 20457–20460.
41. Menacho-Marquez,M., Perez-Valle,J., Arino,J., Gadea,J. and Murguia,J.R. (2007) Gcn2p regulates a G1/S cell cycle checkpoint in response to DNA damage. *Cell Cycle*, **6**, 2302–2305.
42. Wade,C.H., Umbarger,M.A. and McAlear,M.A. (2006) The budding yeast rRNA and ribosome biosynthesis (RRB) regulon contains over 200 genes. *Yeast*, **23**, 293–306.
43. Shivaswamy,S. and Iyer,V.R. (2008) Stress-dependent dynamics of global chromatin remodeling in yeast: dual role for SWI/SNF in the heat shock stress response. *Mol. Cell Biol.*, **28**, 2221–2234.
44. Lindquist,S. and Craig,E.A. (1988) The heat-shock proteins. *Annu. Rev. Genet.*, **22**, 631–677.

45. Warner,J.R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.*, **24**, 437–440.
46. Bonner,J.J., Ballou,C. and Fackenthal,D.L. (1994) Interactions between DNA-bound trimers of the yeast heat shock factor. *Mol. Cell Biol.*, **14**, 501–508.
47. Foat,B.C., Houshmandi,S.S., Olivas,W.M. and Bussemaker,H.J. (2005) Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc. Natl Acad. Sci. USA*, **102**, 17675–17680.
48. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.