BIOINFORMATICS

Vol. 20 no. 16 2004, pages 2553–2561
doi:10.1093/bioinformatics/bth282

Identification of DNA regulatory motifs using Bayesian variable selection

Mahlet G. Tadesse[1], Marina Vannucci[2,] and Pietro Liò[3]*

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania, PA 19104, USA, [2]Department of Statistics, Texas A&M University, College Station, TX 77843, USA and [3]Computer Laboratory, University of Cambridge, Cambridge CB3 OFD, UK

Received on January 15, 2004; revised on April 4, 2004; accepted on April 19, 2004
Advance Access publication April 29, 2004

ABSTRACT

Motivation: Understanding the mechanisms that determine gene expression regulation is an important and challenging problem. A common approach consists of identifying DNA-binding sites from a collection of co-regulated genes and their nearby non-coding DNA sequences. Here, we consider a regression model that linearly relates gene expression levels to a sequence matching score of nucleotide patterns. We use Bayesian models and stochastic search techniques to select transcription factor binding site candidates, as an alternative to stepwise regression procedures used by other investigators.

Results: We demonstrate through simulated data the improved performance of the Bayesian variable selection method compared to the stepwise procedure. We then analyze and discuss the results from experiments involving well-studied pathways of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. We identify regulatory motifs known to be related to the experimental conditions considered. Some of our selected motifs are also in agreement with recent findings by other researchers. In addition, our results include novel motifs that constitute promising sets for further assessment.

Availability: The Matlab code for implementing the Bayesian variable selection method may be obtained from the corresponding author.

Contact: mvannucci@stat.tamu.edu

1 INTRODUCTION

Identifying the repertoire of regulatory elements in a genome is one of the major challenges in modern biology. Gene transcription is determined by the interaction between transcription factors and their binding sites, called motifs or *cis*-regulatory elements. In eukaryotes the regulation of gene expression is highly complex and often occurs through the coordinated action of multiple transcription factors. This combinatorial regulation has several advantages; it controls gene expression in response to a variety of signals from the environment and allows the use of a limited number of transcription factors to create many combinations of regulators. Identification of the regulatory elements is necessary for understanding mechanisms of cellular processes. In eukaryotes these sites comprise short DNA stretches often found within non-coding upstream regions.

The detection of regulatory *cis*-acting elements has turned out to be challenging for various reasons; non-coding regions show little similarity, even between co-regulated genes; they lack sharp delimitations, such as start and termination signals; and they seem to have larger plasticity than genes. Thus, the regulatory information, if conserved, is hidden in short fragments embedded into large regions of non-coding DNA.

DNA microarrays provide a simple and natural vehicle for exploring the regulation of thousands of genes and their interactions. Genes with similar expression profile are likely to have similar regulatory mechanisms. A close inspection of their promoter sequences may therefore reveal nucleotide patterns that are relevant to their regulation. This motivates the following strategy: (1) candidate motifs can be obtained from the upstream regions of the most induced or most repressed genes; (2) a score can be assigned to reflect how well each motif matches the upstream sequence of a particular gene; and (3) regression analysis and variable selection methods can be used to detect sets of motifs acting together to affect the expression of genes.

In this paper, we implement steps 1 and 2 using existing procedures and available software. For step 3, we propose the use of Bayesian variable selection methods as an alternative to stepwise selection procedures used by other investigators. Bayesian variable selection methods use a latent binary vector to index all possible sets of variables (nucleotide patterns). Stochastic search techniques are then used to explore the high-dimensional variable space and identify sets that best predict the response variable (gene expression). The method provides joint posterior probabilities of sets of patterns, as well as marginal posterior probabilities for the inclusion of single nucleotide patterns. Stepwise methods, on the other hand, perform greedy deterministic searches and can be stuck at local minima. Another limitation of the stepwise procedure is that it presumes the existence of a single 'best' subset

*To whom correspondence should be addressed.

of variables and seeks to identify it. In practice, however, there may be several equally good models. In this paper we first use simulated data with similar structure to the real datasets considered for analysis and show how Bayesian variable selection can outperform stepwise procedures. We then exemplify our method using *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* genomes and microarray data from environmental stress experiments.

In what follows, we first conclude this section with a short description of the state of the art on transcription site detection. In Section 2, we briefly describe the data and provide details on the statistical procedures used. Section 3 uses simulated data to assess the performance of the method and compare it to the stepwise procedure. Section 4 describes the analyses and related findings using genome sequence and microarray data. We conclude the paper with a brief discussion.

1.1 State of the art on transcription site detection

Motif detection involves the search for DNA patterns that are over-represented in the upstream region of co-regulated genes. Several computational algorithms have been developed to this end. These methods fall into three broad classes: (1) The word-enumeration approach (van Helden *et al*., 1998) compares the frequency counts of substrings in the upstream region to some reference set. (2) The probability-based models update a position-specific probability matrix using local multiple alignment searches. In this approach, the model parameters are estimated using Expectation–Maximization (EM) (Bailey and Elkan, 1995) or Gibbs sampling methods (Liu *et al*., 1995). Various modifications of the latter have been implemented; these include AlignAce (Roth *et al*., 1998) and MotifSampler (Thijs *et al*., 2001), among others. (3) The dictionary model (Bussemaker *et al*., 2000) is based on a probabilistic segmentation of strings of nucleotides into 'words'. A dictionary is then built by adding words that have high probabilities of occurrence.

These motif detection methods require a set of co-regulated genes, which can be determined experimentally or computationally. A common approach consists of clustering high-throughput gene expression data and searching the upstream regions of each cluster for shared sequence patterns. This, however, leads to a large list of candidate motifs. Bussemaker *et al*. (2001) recently proposed refining the search for biologically meaningful motifs by fitting a linear model that relates the expression data to the counts of each motif. The significant ones were then determined using an extreme value statistic. A similar approach was presented by Keleş *et al*. (2002), who scored the motifs according to their frequency of occurrence and their positions with respect to the gene's translation start site. A linear regression model with stepwise selection was used to identify relevant motifs. Conlon *et al*. (2003) also applied linear regression with stepwise selection after getting a list of candidate motifs using MDScan (Liu *et al*., 2002), an algorithm that makes use of word-enumeration and position-specific probability matrix updating techniques. The candidate motifs were scored in terms of number of sites and degree of matching with each gene. Here, we propose using Bayesian variable selection techniques instead of stepwise methods. In general, Bayesian model selection methods perform a more thorough search of the model space and hence may pick up motifs that can be missed by stepwise methods. In Section 3 we illustrate this with simulated data.

2 METHODS

2.1 Data

We consider two cDNA microarray experiments that explore the transcriptional responses of *S.cerevisiae* (Gasch *et al*., 2000) and *S.pombe* (Chen *et al*., 2003) to environmental stresses. We focus on two stress conditions in wild-type cultures: oxidative stress caused by hydrogen peroxide and heat shock caused by temperature increase.

Our other data consist of the organisms genome sequences and related information, such as the start/stop position and orientation of each open reading frame (ORF). These were obtained from the NCBI's FTP site (ftp://ftp.ncbi.nih.gov/genomes/).

2.2 Generating motif candidates

The motif finding algorithms are sensitive to noise, which increases with the size of upstream sequences examined. As reported by van Helden *et al*. (1998), the vast majority of the yeast regulatory sites in the TRANSFAC database are located within 800 bp from the transcription start site. We therefore extract sequences up to 800 bp upstream, shortening them, if necessary, to avoid overlap with adjacent ORF's. For genes with negative orientation, this is done taking the reverse complement of the sequences.

Another source of noise is the presence in the dataset of upstream sequences that do not contain the motif. We therefore restrict our search to the top 20 up-regulated and top 20 down-regulated genes. We then use Motif Regressor (Conlon *et al*., 2003) to generate a large list of candidate motifs and calculate the matching scores for each gene. The software uses MDScan (Liu *et al*., 2002) to search for nucleotide patterns. The algorithm starts by enumerating each segment of width w (seed) in the top t sequences. For each seed, it looks for w-mers with at least n base pair matches in the t sequences. These are used to form a motif matrix and the highest scoring seeds are retained, based on a semi-Bayesian scoring function

$$\frac{\log(x_n)}{w}\left[\sum_{i=1}^{w}\sum_{j=A}^{T} p_{ij}\log(p_{ij}) - \frac{1}{x_n}\sum_{\text{all segments}}\log(p_0(s))\right],$$

where x_n is the number of n-matches aligned in the motif, p_{ij} is the frequency of nucleotide j at position i of the motif matrix, and $p_0(s)$ is the probability of generating the n-match s

from the background model. The updating step is done iteratively by scanning all w-mers in the set of sequences used for refinement and adding in or removing from the weight matrix segments that increase the score. This is repeated until the alignment stabilizes. We used the top 50 up-regulated and top 50 down-regulated genes for refinement. For each organism, the intergenic regions were extracted and used as background models. We searched for nucleotide patterns of length 5 to 12 bp and considered up to 30 distinct candidates for each width.

2.3 Bayesian motif selection model

Our goal is to identify regulatory motifs among the over-represented nucleotide patterns obtained as described above. This is accomplished by fitting a linear regression model relating gene expression levels (Y) to pattern scores (X), and using a Bayesian variable selection method to select motifs that best predict the expression. The pattern score of motif m for gene g is given by:

$$S_{mg} = \log_2 \left[\sum_{x \in X_{wg}} \frac{P(x \text{ from } \theta_m)}{P(x \text{ from } \theta_0)} \right],$$

where θ_m is the probability matrix of motif m of width w, θ_0 is the transition probability matrix for the background model, and X_{wg} is the set of all w-mers in the upstream sequence of gene g (Conlon *et al.*, 2003).

The variable selection method proceeds as follows. A latent vector, γ, with binary entries is introduced to identify variables included in the model; γ_j takes on value 1 if the j-th variable (motif) is included and 0 otherwise. The regression model is then given by:

$$Y = X_\gamma \beta_\gamma + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \tag{1}$$

with $N(\mu, \Sigma)$ indicating the multivariate Gaussian distribution with vector mean μ and variance-covariance matrix Σ. Here the columns of X and Y are mean centered. The vector γ indexes the variables included in the model (George and McCulloch, 1993; Brown *et al.*, 1998). We specify independent Bernoulli priors for the elements of γ with common probability $\theta = p_{\text{prior}}/p$, where p_{prior} is the number of covariates expected a priori to be included in the model. For the other model parameters, we take

$$\begin{aligned} \beta_\gamma &\sim N(0, c\sigma^2 \{X'_\gamma X_\gamma\}^{-1}) \\ \sigma^2 &\sim \text{Inv} - \chi^2(a, b), \end{aligned} \tag{2}$$

where Inv-$\chi^2(a, b)$ is the scaled-inverse-χ^2 distribution. The hyperparameters a, b and c need to be specified by the investigator.

2.4 Stochastic search of regulatory motifs

Having set the prior distributions, a Bayesian analysis proceeds by updating the prior beliefs with information that comes from the data. Our interest is in the posterior distribution of the vector γ given the data, $f(\gamma \mid X, Y)$. Vector values with high probability identify the most promising sets of candidate motifs. Given the large number of possible vector values (2^p possibilities with p covariates), we make use of stochastic search Markov chain Monte Carlo (MCMC) techniques to look for sets with high posterior probabilities.

Our method visits a sequence of models that differ successively in one or two variables. At each iteration, a candidate model, γ^{new}, is generated by randomly choosing one of the following two transition moves:

(i) Add or delete one variable from γ^{old}.

(ii) Swap the inclusion status of two variables in γ^{old}.

The proposed γ^{new} is accepted with a probability that depends on the ratio of the relative posterior probabilities of the new versus the previously visited models:

$$\min \left\{ \frac{f(\gamma^{\text{new}} \mid X, Y)}{f(\gamma^{\text{old}} \mid X, Y)}, 1 \right\}, \tag{3}$$

which leads to the retention of the more probable set of patterns.

Our stochastic search results in a list of visited sets and corresponding relative posterior probabilities. The marginal posterior probability of inclusion for a single motif j, $P(\gamma_j = 1 | X, Y)$, can be computed from the posterior probabilities of the visited models:

$$p(\gamma_j = 1 \mid X, Y) \approx \sum_{\gamma : \gamma_j = 1} p(Y | X, \gamma^{(t)}) \cdot p(\gamma^t), \tag{4}$$

where $\gamma^{(t)}$ is the vector γ at the t-th iteration.

Our methodology is summarized in Figure 1. The analysis starts with gene expression data from microarray experiments and a large list of candidate motifs. We fit a linear regression model and make use of the Bayesian variable selection method described above to identify sets of motifs that best explain and predict changes in expression level.

3 SIMULATION STUDY

Before analyzing the yeast genomes, we motivate the use of Bayesian variable selection methods over stepwise procedures by comparing their performance on simulated data designed to imitate the structure and features of the yeast expression data. In DNA microarray data, the gene expression levels are not independent. One reason is that genes that share common roles in cellular processes tend to be co-expressed (Eisen *et al.*, 1998). The form of this dependence, however, is not known. We define a covariance matrix $\Sigma = WW'$ where W is the vector of gene expression levels for *S.cerevisiae* under heat stress and use the matrix of nucleotide pattern scores as the covariate matrix X. We draw a vector of regression coefficients, β, such that all but 30 of its elements are equal to 0 then simulate the

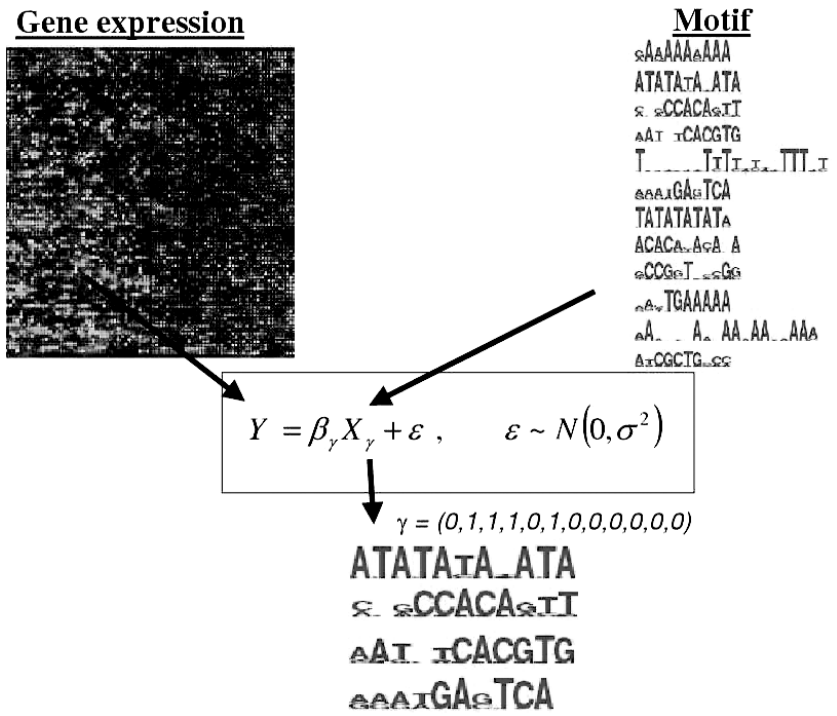

Fig. 1. Graphical representation of methodology—a linear regression model that relates gene expression data to pattern scores is fitted and a Bayesian variable selection method is used to identify variables. This is accomplished through a latent binary vector γ, updated via stochastic search MCMC techniques. Motifs with high posterior probability indicate promising sets.

response vector $Y \sim N(X\beta, \Sigma)$. In order to assess the effect of the magnitude of the regression coefficients on the results, we generate two datasets where the non-zero elements of β fall in the range $[-2, 2]$ and $[-10, 10]$.

Let us first briefly describe the results from the stepwise method. The procedure involves (1) identifying an initial model, (2) iteratively adding or removing a predictor variable from the model previously visited according to a 'stepping criteria', and (3) terminating the search when adding/removing variables is no longer possible given the specified criteria. Detailed descriptions of the procedure can be found in many linear models books (Draper and Smith, 1981). We ran the stepwise regression using three different criteria (P-values of 0.05, 0.01 and 0.001) for adding or deleting variables from the subsets considered at each iteration. Table 1 summarizes the results. For the simulated data where the non-zero β_js are in the range $[-2, 2]$, the stepwise procedure with a P-value of 0.05 selected 57 variables, 38 of which were false positives. As expected, the procedure becomes more conservative with a more stringent criterion but the proportion of false positives remains high. With a P-value of 0.01, 42 variables were selected with 26 false positives and there were 14 false negatives. With a P-value of 0.001, 11 out of the 24 variables included in the model are false positives and there were 17 false negatives. The stepwise method performed equally poorly with the second simulated data where larger regression coefficients in the range $[-10, 10]$ were considered. The false negatives, i.e. the non-selected predictive variables were not necessarily those with smallest regression coefficients.

For the Bayesian variable selection method, we ran two parallel MCMC chains with 100 000 iterations each and started the searches from widely different points to avoid possible dependence of the results on the initial model. For the first chain, all the γ_js were set to 0 except for 10 randomly selected, and for the second chain, γ_j was set to 1 for 100 randomly selected js. We chose the number of variables expected a priori to be included in the model to $p_{\text{prior}} = 20$. For the prior on σ^2, we took $a = 3$, which corresponds to the smallest integer such that the expectation of σ^2, $\mathrm{E}[\sigma^2] = \frac{ab}{(a-2)}$, exists. The scaling value b was taken to be comparable in size to the expected error variance of the standardized data. We assessed the sensitivity of the results to the choice of the hyperparameter c by running three separate analyses with c set to 20, 50 and 100.

Let us first consider the simulated data with β in the range $[-2, 2]$. The MCMC samplers mostly visited models with 30–35 variables. For each value of c, we pooled together the output of the parallel chains and computed the normalized posterior probabilities of each distinct visited model. We also derived the marginal posterior probabilities of inclusion for each variable. In general, we found good agreement between variables with high marginal posterior probabilities

Table 1. The results of simulated data using the stepwise procedure with different P-value criteria for adding/deleting variables and the Bayesian variable selection method with different values for the hyperparameter c

	No. of selected variables	% False positives	No. of false negatives	R^2
$\beta_j \in [-2, 2]$				
P-value cutoff				
0.05	57	66.7	11	0.782
0.01	42	61.9	14	0.763
0.001	24	45.8	17	0.712
c				
20	29	0	1	0.977
50	29	0	1	0.977
100	29	0	1	0.977
$\beta_j \in [-10, 10]$				
P-value cutoff				
0.05	52	65.4	12	0.865
0.01	39	56.4	13	0.854
0.001	30	43.3	13	0.842
c				
20	26	3.8	5	0.997
50	26	0	4	0.998
100	25	4.0	6	0.976

and those selected in the 'best' models, defined as the sets with the highest joint posterior probability among all distinct models visited by the MCMC sampler. Table 1 reports the variables selected in the 'best' model based on the pooled output from the parallel chains for each value of c considered. These variables correspond also to the ones with highest marginal posterior probabilities. For all values of c, the 'best' model contained 29 variables with no false positives and a single false negative. The Bayesian stochastic search method therefore does an excellent job at identifying the predictive variables. Consequently, this approach leads to larger coefficients of determination. It explained around 97% of the variability in the response compared to 70–80% with the stepwise procedure. The analysis of the second simulated data with non-zero β_js in the range $[-10, 10]$ also gave good results. For $c = 20$ and 50, 26 variables were selected with 1 and 0 false positives respectively. For $c = 100$, 25 variables were included in the model with a single false positive and six false negatives. The R^2 values for this case were close to 100% with the Bayesian approach compared to around 85% with the stepwise procedure.

This simulation study is meant to show that the Bayesian variable selection approach can provide better results compared to the stepwise procedure, especially in situations where the space of variable subsets is prohibitively large. There are many well-documented problems with stepwise regression. Our results are in line with findings of other authors that have compared the performance of the two procedures on experimental data [see, e.g. the critique of Viallefont *et al.* (2001) on the use of stepwise methods and P-values in the context of case-control studies].

4 RESULTS

We now return to the analyses of the yeast data. For each organism and stress condition, we regressed the expression levels on the pattern scores using separate models. In all cases, the analyses were started with a set of around 400 patterns. We chose the same prior settings as in the simulation study. For every regression model, we ran two parallel MCMC chains with 100 000 iterations each. One of the chains was started with 10 and the other with 100 randomly selected γ_js set to 1. We pooled together the sets of patterns visited by the two MCMC chains and computed the normalized posterior probabilities of each distinct visited set. We also computed the marginal posterior probabilities, $p(\gamma_j = 1 \mid X, Y)$, for the inclusion of single nucleotide patterns.

For comparison, we repeated the analysis with Motif Regressor (Conlon *et al.*, 2003), which uses stepwise regression to select motifs. Both the Bayesian variable selection method and the stepwise procedure started with the same set of around 400 motifs found by MDscan and used the same sequence scores. As described in the next section, the two methods differed in the selected final models.

4.1 Findings

Interesting sets of motifs can be found by exploring the visited sets with highest joint posterior probabilities. An alternative would be to select nucleotide patterns with high marginal posterior probabilities. It is important to stress that the sets of motifs with high posterior probability represent groups of motifs that work in combination to explain gene expression. Marginal probabilities are instead indications of how well single motifs work. Here we found good agreement between motifs with high marginal posterior probabilities and those in the 'best' models. Indeed, all motifs in the model with highest joint posterior probability are also the ones with highest marginal posterior probabilities. These indicate promising motifs for further investigation.

Tables 2–5 report the motifs selected in the 'best' model in each of the regression analyses, ordered according to their marginal probabilities. Selections that are robust to the choice of c, i.e. motifs that show up in the best model of all MCMC analyses run with different values of c, are represented with two asterisks, and those that appear in two of the three MCMC analyses have a single asterisk. The tables also present the motifs selected by the stepwise procedure implemented in Motif Regressor (Conlon *et al.*, 2003) with a P-value cutoff of 0.01 for inclusion/removal of variables.

Some of the motifs we discovered are experimentally known to be related to stress. Others are novel and constitute a

Table 2. Selected motifs for *S.cerevisiae* under heat shock, using expression data as response variable

Bayesian variable selection Discovered motif	Known	Marginal probabilities	Motif regressor Discovered motif	Known
WTAAGGGAK**		1.0000	WTAAGGGAK	
T**GAAA****	M3A	1.0000	T**GAAA**	M3A
ACCYT**GAAA****	M3A	0.9999	ACCYT**GAAA**	M3A
TCYAGAATRTT**	Cliften *et al*.	0.9998	**TCYAGAA**TRTT	Cliften *et al*.
GGCAGGAMA**		0.9991	GGGCCWGGM	
HYCCWTMCAT**		0.9991	WTGYAYKGGTK	
W**ARGGG****	STRE	0.9987	AA**ARGGRG**MMG	STRE
M**GATGAG**ATGAR**	M3B	0.9985	M**GATGAG**ATGAR	M3B
GM**GATGAG**MWT**	M3B	0.9839	GM**GATGAG**MWT**	M3B
GAADRAA**AGGGR****	STRE	0.9739	GAADRAA**AGGGR**	STRE
GCCCC*		0.9573	GCCCC	
AGGGRGSGAAD*	STRE	0.9251	C**CCCCT**T	STRE
GCWCATCCACC		0.8441	GRCCCC	
CMAACAAAS		0.8195	AATTT	
GSCCKGSWA		0.5308	TGCGATG	
ARGGGGSGGR*	STRE	0.5136	GAWTM**AGGGG**	STRE
AMRWGCCAGAA		0.4364	AGATC	

Motifs are ordered according to their marginal posterior probabilities. Selections that are robust to the choice of the hyperparameter c are shown with asterisks. Selected patterns by Motif Regressor are also reported.
Characters in bold face correspond to matches between known and discovered motifs.
IUPAC codes are used for degenerate nucleotides: K = G/T; M = A/C; R = A/G; S = C/G; Y = C/T; W = A/T; B = C/G/T; D = A/G/T; H = A/C/T; V = A/C/G; and N = A/C/G/T.

Table 3. Selected motifs for *S.cerevisiae* under oxidative stress, using expression data as response variable

Bayesian variable selection Discovered motif	Known	Marginal probabilities	Motif regressor Discovered motif	Known
GAAWGRCWGTAG**	Cliften *et al*.	1.0000	TTACT	
YGATTAGTAAKS**		1.0000	TTTAT	
AAATT**		0.9999	TAAAA	
AATM**AGGGG****	STRE	0.9987	AGAGGG	
CA**CCCCT**TW**	STRE	0.9959	GATTASTAATS	
ACTAMGTGTAT**		0.9943	WGGCTAGSM	
G**CCCCYT****	STRE	0.9929	G**CCCCYT**	STRE
GATGAATAA**	M3B	0.9751	TGCWTGACTTGM	
RCGGGTARC**		0.9655	GCTKCTAAA	
CGGATCCG**		0.9534	TACATACAC	
ATGMGTCARG**		0.8403	RSCTAGSCTA	
ACCCGCCG		0.7051	CTTTT	
GCTYCKCTCT*		0.6773	YCATYTCTTGA	
AATGGA*		0.6540	GGCTKTTATTT	
GTGATCAG*		0.6620	GATTTTA	
CTGAMRTMGTA*		0.6594	AGAKGAAVCTG	
GGTGACGCAAA		0.3567	RGCTKYYAWTTT	

set of promising candidates for future experimental work. Nucleotide patterns that match known motifs appear in the tables with bold characters, along with the associated binding site or a reference.

The analyses of *S.cerevisiae* led to the selection of around 25 motifs for each stress condition (Tables 2 and 3). They explained respectively 16 and 8% of the expression variability in response to heat and oxidative stress. Among the selected motifs, we identified some that contain matches to three well known stress-related motifs: STRE, M3A and M3B. STRE is known to respond to general environmental stress and to positively regulate transcription (Schmitt and McEntee, 1996). M3A and M3B have previously been found in genes repressed under environmental stress and act by slowing down cell

Table 4. Selected motifs for *S.pombe* under heat shock, using expression data as response variable

Bayesian variable selection Discovered motif	Known	Marginal probabilities	Motif regressor Discovered motif	Known
ACGTCAT**	ATF/CRE	1.0000	AGAGGAA	
C**TCTYTYTT**TTT**	Chen *et al*.	1.0000	C**TCTYTYTT**TTT	Chen *et al*.
TTTTACCMAC**		1.0000	TCAATWGCATTG	
TYTYTTKCT**	Chen *et al*.	0.9986	GAATTC	
ACKTAARATCG**		0.9974	TYTATTTAY	
GTTTAC*		0.9967	CTTTTATT	
TAGATAA		0.9261	TTCCATTT	
GAATTGTAG**		0.8778	GAATTGTAG	
CCGGGTTTYKA*		0.7675	TTYTTCTCTYTT	
ATTAA*		0.6709	ATTAA	
RTATATATATA**		0.6545	CTCAATCTC	
AATARAATKGA**		0.6465	TTYTTATAAT	
AAAAT		0.4381	AACGACGTTTM	
TTATYTATYTTC		0.4317	TTATAATWATAA	
TCGTTTTTTG*		0.2619	GTTTA	
			TAMGTAARGWAW	
			TTAAAA	
			AACTWRCGKWAG	
			GGAAA	
			TTTCGTT	
			TGTTTAC	
			TGATTTG	
			TCCTCTT	
			TTGTTTAC	
			CTATTTTT	
			AGKAAACAA	
			TTGCTTWTC	
			TKCCTTTTC	
			WATATATATAY	

growth (Gasch *et al*., 2000). We also selected some patterns that overlap with the stress-related motifs found by Cliften *et al*. (2003) using comparative genomics.

For *S.pombe* (Tables 4 and 5) the selected patterns explained respectively 9 and 13% of the expression variability in response to heat and oxidative stress. Among these, we identified matches to the ATF/CRE motif (Hai and Hartman, 2001). This is a binding site for the Atf1p family of transcription factors, which regulate stress-dependent transcription (Takeda *et al*., 1995). The analyses also confirmed some of the novel motifs described by (Chen *et al*., 2003).

As for the comparison with Motif Regressor, although the R^2 values are comparable, we note some differences in the list of selected motifs. A thorough evaluation of which method performs best in this context is difficult since there are very few stress related motifs that are known for these organisms, in particular for *S.pombe*. With regards to motifs that are experimentally known or cited in the literature, we notice that in some cases both approaches identified matches to the same binding sites (see, e.g. M3A, M3B, STRE for *S.cerevisiae* under heat shock—Table 2). In other cases, the Bayesian variable selection method selected more matches to the known motifs than Motif Regressor; for *S.cerevisiae* under oxidative stress see M3B and motif found in Cliften *et al*., and for *S.pombe* under heat shock see ATF/CRE.

Within each organism similar motifs were selected for the different stress conditions. This can be explained by the general stress response strategy in these organisms. Both have a cross-protection program wherein exposure to a non-lethal dose of one stress can protect against a potentially lethal dose of other stresses (Chen *et al*., 2003).

Instead of the short list presented by Chen *et al*. (2003), we found approximately equal number of motifs for both organisms. Chen *et al*. (2003) hypothesized that there are qualitative differences between stress motifs of budding and fission yeast. The budding yeast genome is organized by stress specific mechanisms while the fission yeast genome is regulated mainly by Sty1p mitogen-activated protein kinase (MAPK) pathway that acts as a main stress switch. These differences reflect the fact that *S.pombe* activates gene expression programs more specialized for a given stress or a subset of stresses than *S.cerevisiae*. On the other hand, fission yeast and budding

Table 5. Selected motifs for *S.pombe* under oxidative stress, using expression data as response variable

Bayesian variable selection Discovered motif	Known	Marginal probabilities	Motif regressor Discovered motif	Known
TGACGTA**	ATF/CRE	1.0000	**TGACGTA**	ATF/CRE
AWGARKAAAATM**		0.9998	AWGARKAAAATM	
TT**ACGTMA**G**	ATF/CRE	0.9701	TT**ACGTMA**G	ATF/CRE
AACTCGTTC**		0.9588	AACTCGTTC	
CCAACAACC**		0.8929	GTAAAATTGG	
ATATAGCAAA*		0.8451	TAGGATTKAAAA	
CATTT*		0.8421	CATTT	
GATGATGATGT		0.8356	GATGATGATGT	
CATCTTCCR**		0.7997	CATCTTCCR	
TGATATCATATY*		0.7521	TGATATCATATY	
TCTTYCTTTTCT*	Chen *et al*.	0.6797	**TCTTYCTT**TTCT	Chen *et al*.
ATGATGTT*		0.5202	ATGATGTT	
GAAGAAG		0.5054	TGATT	
TAAACA*		0.4667	TTTTA	
TATWTATWTATT**		0.4381	ATTYTATTY	
TTAATT		0.3688	TTTCCTTTCTYT	
AATTYTATTYTW		0.3536	TGAAATCA	
TCTTTTCKTATA*		0.3144	ACAAT	
GTAAA*		0.3098	ACGTATA	
			CTTCTTC	
			TTKCATTYC	
			TCTTTAC	
			ATTTCA	
			TTACTCT	
			CTTTCTTT	
			CATCTTTT	
			YCYWCTMCC	
			TTGTGTKTGTGT	
			TATATATATAT	

yeast although distantly related share some similarities. For instance they have comparable genome sizes and similar number of transcription binding sites in different gene networks (Chen *et al*., 2003).

5 DISCUSSION

The discovery of regulatory sites is an important field of research. Our work has focused on the detection of regulatory motifs by integrating DNA microarray data with genome sequences. This was accomplished by regressing expression levels on pattern scores and using Bayesian variable selection methods. We examined stress-related transcription sites in *S.cerevisiae* and *S.pombe* and identified some well known stress regulators. We also found novel motifs that may constitute promising candidates for further experimental assessment. Our results were based on an initial set of around 400 motifs. The composition of this initial set depends on the number of genes used in the motif finding procedure described in Section 2.2. We used the top 20 genes to search for candidate motifs and the top 50 for refinement.

We have shown that the Bayesian variable selection method proposed in this paper has lower false positive and false negative rates compared to the greedy stepwise regression used in Motif Regressor. The Bayesian approach is computationally more intensive, both in terms of time and memory requirements, since the method explores a larger set of possible models. This, however, is a cheap price to pay relative to the labor and cost incurred in trying to validate erroneous findings through experimental laboratory techniques.

Alternative approaches for motif discovery are based on sequence alignment among related species at different evolutionary distances (Cliften *et al*., 2003). These analyses, however, are limited by current genome sequence availability. For example, although several *Saccharomyces* species that are closely related to *S.cerevisiae* have become available, none of them are close enough to *S.pombe* to allow transcription site detection by alignment.

The identification of regulatory motifs can provide a better understanding of selection and mutation processes both at the sequence and gene expression levels. It also provides improved ability in building biochemical and signaling

pathways, ultimately leading to the understanding of an entire cell as a vast genetic network. In the future, it would be interesting to look at transcription regulation in human and mouse, an inherently more difficult problem since transcription factor binding sites show more combinatorial complexity and are spread over distances of 10 kb or more from the coding region.

ACKNOWLEDGEMENTS

M.G.T. was supported by NCI grant CA90301 and M.V. by NSF CARREER award DMS-0093208.

REFERENCES

Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learning*, **21**, 51–80.

Brown,P.J., Vannucci,M. and Fearn,T. (1998) Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc., Ser. B*, **60**, 627–641.

Bussemaker,H.J., Li,H. and Siggia,E.D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci., USA*, **97**, 10096–10100.

Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Chen,D., Toone,W.M., Mata,J., Lyne,R., Burns,G., Kivinen,K., Brazma,A., Jones,N. and Bähler,J. (2003). Global transcriptional responses of fission yeast to environmental stress. *J. Mol. Biol. Cell*, **14**, 214–229.

Cliften,P., Sudarsanam,P., Desikan,A., Fulton,F., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

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.

Draper,N.R. and Smith,H. (1981) *Applied Regression Analysis*, 2nd edn. John Wiley and Sons, New York.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.

Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

George,E.I. and McCulloch,R.E. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.

Hai,T. and Hartman,M.G. (2001) The molecular biology and nomenclature of the activating transcription factor/cAMP responsive element binding family of transcription factors: activating transcription factor proteins and homeostasis. *Gene*, **273**, 1–11.

Keleş,H., van der Laan,M. and Eisen,M.B. (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167–1175.

Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.

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. *Nat. Biotechnol.*, **20**, 835–839.

Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Schmitt,A.P. and McEntee,K. (1996) Msn2p, a zinc finger DNA-binding protein, is the transcriptional activator of the multistress response in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci., USA*, **93**, 5777–5782.

Takeda,T., Toda,T., Kominami,K., Kohnosu,A., Yanagida,M. and Jones,N. (1995) *Schizosaccharomyces pombe atf1*$^{+}$ encodes a transcription factor required for sexual development and entry into stationary phase. *EMBO J.*, **14**, 6193–6208.

Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.

van Helden,J., André,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.

Viallefont,V., Raftery,A. and Richardson,S. (2001) Variable selection and Bayesian model averaging in case–control studies. *Stat. Med.*, **20**, 3215–3230.