*Year* 2001 *Paper* 98

# Identification of Regulatory Elements Using A Feature Selection Method

Sunduz Keles[*] Mark J. van der Laan[†]
Michael B. Eisen[‡]

[*]Department of Statistics, University of Wisconsin, Madison, keles@stat.wisc.edu

[†]Division of Biostatistics, School of Public Health, University of California, Berkeley

[‡]Dept. of Molecular and Cell Biology, University of California, Berkeley

# Identification of Regulatory Elements Using A Feature Selection Method

Sunduz Keles, Mark J. van der Laan, and Michael B. Eisen

## Abstract

Many methods have been described to identify regulatory motifs in the transcription control regions of genes that exhibit similar patterns of gene expression across a variety of experimental conditions. Here we focus on a single experimental condition, and utilize gene expression data to identify sequence motifs associated with genes that are activated under this experimental condition. We use a linear model with two way interactions to model gene expression as a function of sequence features (words) present in presumptive transcription control regions. The most relevant features are selected by a feature selection method called stepwise selection with monte carlo cross validation. We apply this method to a publicly available dataset of the yeast Saccharomyces cerevisiae, focussing on the 800 basepairs immediately upstream of each gene's translation start site (the upstream control region (UCR)). We successfully identify regulatory motifs that are known to be active under the experimental conditions analyzed, and find additional significant sequences that may represent novel regulatory motifs. We also discuss a complementary method that utilizes gene expression data from a single microarray experiment and allows averaging over variety of experimental conditions as an alternative to motif finding methods that act on clusters of co-expressed genes.

# 1 Introduction

Transcriptional regulation in eukaryotes depends, to a large extent, on the activities of hundreds of sequence specific DNA binding proteins - transcription factors (TFs). Each transcription factor, or group of closely related factors, recognizes a unique family of short sequence elements, usually between 5 and 15 basepairs in length. The rate at which a given gene is transcribed is, to first approximation, determined by the amount and activities of TFs bound to DNA in the immediate vicinity of the gene, which, in turn, is determined by concentrations and activities of TFs in the nucleus and, importantly, by the repertoire of transcription factor binding sites found adjacent to the gene. Thus, the non-coding DNA sequence surrounding a gene determines when and where the gene will be expressed, just as the coding sequence determines the gene's structure and molecular function. A major challenge in contemporary biology is understanding precisely how this regulatory information is encoded in the genome. In this paper, we focus on an early step in this process, the identification of biologically significant transcription factor binding sites in the genome of the yeast *Saccharomyces cerevisiae.*

The simultaneous availability of complete DNA sequences and of DNA microarray-based data on the expression levels of all of an organisms' genes across a wide range of experimental conditions (Eisen et al., 1998; Lockhart et al., 1996; Schena et al., 1995) has spurred the development of numerous methods to understand how transcriptional information is encoded in DNA sequences. A common strategy relies on the observation that there are multiple sets of genes that exhibit similar patterns of expression across experimental conditions (Eisen et al., 1998). This co-expression is taken as evidence of co-regulation, and sequences upstream of these co-expressed genes are searched for shared sequence motifs (for microbes, these searches are usually restricted to regions of a few hundred to a thousand basepairs upstream of the gene's translation start sites - from hereon referred to as the upstream control region (UCR)) . Prior work in identifying shared sequences can be divided into two main categories: 1) Multiple alignment methods; 2) Deterministic combinatorial algorithms based on word frequency counts. Methods in the first category (Lawrence and Reilly, 1990; Lawrence et al., 1993; T.L. Bailey, 1995; A.F. Neuwald, 1995; Hertz and Stormo, 1999; Tavazoie et al., 1999; Hughes et al., 2000), usually take as input a set of unaligned UCRs and a range of possible motif lengths, and return probabilistic models of shared motifs in the form of position weight matrices. These methods rely on local search techniques such as Gibbs sampling (Lawrence et al., 1993) or expectation maximization (T.L. Bailey, 1995). Methods in the second category (van Helden et al., 1998; Tompa, 1999; Jensen and Knudsen, 1999; Sinha and Tompa, 2000) search the UCRs for various sized nucleotide sequences exhaustively, and evaluate the significance of the obtained sequences by a statistical method.

In this paper, we follow a different approach and formulate the problem in the feature selection framework. First, we present some important empirical observations that motivate the idea of identifying known and novel motifs based on gene expression data from a single microarray experiment. Following these observations, we model the gene expression from a single experimental condition as a linear function of scores computed for sequence motifs in

UCRs. These sequence motif scores incorporate the number of occurrences of the motifs and their positions with respect to the gene's translation start site. Our model also allows for two way interactions of the sequence motifs. We treat the sequence motifs as explanatory variables (features) and suggest a feature selection method to extract those that are the most relevant. The method, which is described in details in Section 3, is a forward and backward stepwise selection method embedded into monte carlo cross validation. Finally, we briefly discuss a transformation of the gene expression data in terms of sequence motifs. This transformation enables us to find motifs that are either important in a single experiment or have important roles across a number of experiments. We apply our methodology to a publicly available dataset of the yeast *Saccharomyces cerevisiae*.

There are at least two other previous works that also pursue a single gene-expression experiment based approach (Bussemaker et al. (2001), Jensen and Knudsen (1999)). In both of these works, association with gene expression is used to identify the most relevant sequence motifs among a set of different length oligomers. Our approach is closer to of Bussemaker et al. (2001). In their work, authors use a linear model of raw counts of sequence motifs for gene expression and determine significance of the sequence motifs by extreme value statistics. Aside from laying out the statistical motivation for focusing on a single condition, the general scheme of our work differs from this approach in a number of aspects. The scores we use for sequence motifs incorporate location with respect to the translation start sites in an *ad hoc* way as well as the raw counts. Our model allows interactions of sequence motifs to capture combinatorial effects. We suggest a robust feature selection method that doesn't employ the assumption of normality on gene expression. Finally, our method starts of with pentamers as sequence motifs and extends them in an appropriate way, rather than considering the huge set of different length oligomers. The nature of the feature selection method allows implementation of this in a computationally tractable way.

## 2 Motivation for focusing on a single microarray experiment

As discussed above, a common regulatory motif finding approach favored in the literature is to cluster genes based on similarity in their gene expression over a set of experimental conditions and to search the UCRs of clustered genes for the presence of shared or overrepresented motifs. This method has been reasonably successful, recovering many known motifs, discovering a few new motifs that have since been experimentally validated, and suggesting a role for dozens of additional motifs that have not yet been experimentally tested. However, the method remains somewhat limited by the degree of association between similarities in gene expression and similarities in the UCRs of the genes. When investigating this association, we came upon some important empirical observations that led us to focus on a single experiment as an alternative.

Our analysis was based on the yeast cell-cycle experiments of Cho et al. (1998), and on known cell-cycle related regulatory elements from the literature (SWI5E, ACE2, ECB, STE12, MCB, SCB, RAP1P, NEG, CBF1, MCM1, and SFF). Details of these experiments are given in Section 4. We compared gene expression distance (dissimilarities in gene expression) among two gene groups where genes in the first group share six of the above cell cycle regulatory motifs (including MCB and SCB) and the genes in the second group do not have any of the cell cycle motifs. Gene expression distances were computed as in the context of clustering (i.e. using cosine angle metric (Eisen et al. (1998)), correlation distance, euclidean distance, etc...). We found that these two groups have very similar gene expression distance distribution. The histograms of gene expression distance of these two groups are given in part (a) of Figure 1. Mean gene expression distances were 0.966 and 0.978 in the first and second group respectively, using a cosine angle metric. Similar results were obtained with different metrics. This suggests that the presence of these motifs does not strongly select for genes with similar patterns of expression across the dataset.

Following this observation, we directly investigated the association between similarities in gene expression across different experimental conditions and similarities in UCRs. The distance in gene expression profiles and in regulatory motif profiles were computed for a group of genes and the correlation between them was found to be 0.00125. Regulatory motif profiles were obtained for each gene by calculating a score for each cell cycle regulatory motif from each gene's UCR. Examples of these scores are given in the next section. This result indicates that genes that are close to each other based on gene expression distance across experiments are no more likely than random to be close in sequence distance among the set of distances that we have considered. These results also confirm the findings of Bilu and Linial (2001), where they report that the similarity in UCRs of the genes based on BLAST E-score is quite low between genes of a cluster.

Although not entirely unexpected - our metric on sequence distance is fairly crude - the observation that there is no simple association between similarities in gene expression across experiments and overall similarities in UCRs led us to turn our attention to single microarray experiments. We directly investigated gene expression among two groups of genes with and without relevant regulatory motifs for a single experimental condition. As a typical illustration, the histograms of gene expression in two groups at time point 20 minutes of the cell cycle are given in part (b) of 1. This time point corresponds to the late G1 phase with the active regulatory motifs MCB and SCB. These histograms demonstrate the strong difference in gene expression of the two groups. The mean gene expression level in the group of genes with MCB and SCB sites is 0.8638638 where as the mean expression level of the other group is -0.3099315.

We conclude that although there is a significant relationship between expression of the genes and their UCRs in a single experiment, averaging across experiments destroys this signal. We therefore focus here on identifying regulatory motifs for each individual experiments separately.

# 3 Method for identifying regulatory motifs

In the following subsections we define the regulatory motifs as features of the UCRs and explain the method that selects the most relevant ones.

## 3.1 Feature Extraction and the Statistical Model for Gene Expression

The first requirement of the methodology is to extract useful, relevant explanatory variables from the UCRs of the given organism. In van Helden et al. (1998), it was shown that more than 98% of the 308 yeast regulatory motifs lie within 800bp upstream of the translation start site. Regulatory motifs in yeast are of variable length (5-15bp) and have degenerate as well as highly conserved components. However, in most cases the DNA binding domain of the transcription factor contacts relatively short sequences with low internal variation. It is also well known that in identification of the regulatory motifs, not only the length but also the location of the site is an important factor. We obtained an empirical distribution of the regulatory motif locations in yeast using data from SCPD (Zhu and Zhang (1999)) by counting the number of predicted regulatory motifs occurring in each of the 50bp intervals of all UCRs and normalizing with the total number of occurrences. We thus obtained an estimate of the probability that each 50bp interval of the 800bp UCRs contains a regulatory motif.

We now describe how the explanatory variables are extracted from 800bp UCRs. We restrict our analysis to pentamers for computational reasons. We define the random variable $L_{ig}$ to be location of $i$th occurrence of a pentamer $w$ in UCR of gene $g$ and denote its realizations by $l_{ig} \in \{-1, \cdots, -800\}$. Let $N_{wg}$ be the total number of occurrence of word $w$ in the UCR of gene $g$. We compute the following score for each word, gene combination (van Zwet, 2001):

$$S_{wg} = \sum_{i=1}^{N_{wg}} P(L_{ig} = l_{ig}), \tag{1}$$

where $P(L_{ig} = l_{ig})$ is the probability of location $l_{ig}$ being a regulatory motif location. We obtain a score profile $\overrightarrow{S}_g$ for each gene $g, g = 1, \cdots, P$. We use linear models with two way interactions to model the gene expression data as a function of scores from UCRs. Denoting the gene expression by $Y$ and the total number of pentamers by $M$, the expected gene expression given the score profile is modeled as

$$E(Y|\overrightarrow{S}) = \beta_0 + \sum_{w=1}^{M} \beta_w S_w + \sum_{w_1}^{M} \sum_{w_2, w_1 \neq w_2}^{M} \beta_{w_1,w_2} S_{w_1} S_{w_2} \equiv m(\overrightarrow{S}|\beta). \tag{2}$$

In Bussemaker et al. (2001), authors show that using the number of occurrences of words as explanatory variables and adopting a linear model succeeds in identifying regulatory motifs. Our preliminary analysis (scatter plots of gene expression versus scores) also suggests that a linear model is a reasonable model to work with. The interaction term serves two purposes:

to identify cooperative regulatory motifs and to gather different parts of the regulatory motifs that are longer than 5 base pairs.

## 3.2 Statistical Method for Feature Selection

Having defined the features and the statistical model, our next goal is to select the features that explain the observed gene expression best. We propose a feature selection method that uses forward/backward selection adaptively with monte carlo cross-validation. We will refer the method as stepwise selection with cross-validation (SCV). Both forward/backward selection and cross-validation are among the well known methods in statistics literature. We will first outline the method and then present the differences with the general feature selection methods of the literature. We now present SCV in the context of the model given in eq. 2.

1. Randomly split the total number of genes to obtain a training sample and test sample of sizes $n_{tr}$ and $n_{te}$.

2. Perform forward model selection:

   (a) Initial step: Fit a univariate linear regression model for each variable on training sample, get the parameter $\hat{\beta}_{tr,j}$ and compute the residual sum of squares, $rss(\hat{\beta}_{tr,j})$, on the test sample where

   $$rss(\hat{\beta}_{tr,j}) = \sum_{i=1}^{n_{te}} \{Y_i - m(S_{ij}|\hat{\beta}_{tr,j})\}^2.$$

   (b) Add $\min_j^{-1}\{rss(\hat{\beta}_{tr,j})\}$ to the model.

   (c) Keeping the already added variable(s) in the model, continue adding other variables. For each variable added consider its interaction with the rest of the variables that are already in the model.

   (d) Steps (b) and (c) are repeated until no improvement is observed in the rss of the test sample and $d_k$ is reported as the set of all variables added to the model.

   (e) For all variables in $d_k$, rss on the test sample is computed for the model that excludes that one variable. The variable that improves the rss most is now deleted from $d_k$ and remaining set is denoted as $d_k^{'}$. If none of the variables provide improvement, stop, otherwise continue backward deletion from the set $d_k^{'}$. The final set obtained is denoted as $d_k^f$.

3. Repeat steps 1 and 2 K times and obtain $(d_1^f, d_2^f, \cdots, d_K^f)$

4. Let $d_{FINAL}$ be the union of $d_k^f$, $k = 1, \cdots, K$.

The final output from the procedure consists of following items for each of the variables in the set $d_{FINAL}$:

- Number of splits it is selected,

- Rank profile which represents its order of being selected in each split,

- Average improvement over splits it provides in the rss of the test sample relative to the model where it hasn't been added yet.

With the SCV method we don't need any model assumptions other than the linearity of the regression model. Our criteria of selecting explanatory variables solely depend on the minimization of the least squares residual sum of squares and cross-validation provides an unbiased estimate of this criteria (Breiman et al. (1984)). Repeating the splitting procedure many times overcomes the instability problem which is common to most of the feature selection methods. In our application, cross-validation serves two purposes: to decide on the variable to be added to or to be deleted from the model and to provide prediction error(rss) on the test sample. The most common usage of cross validation in the literature is to select among a series of a priori nested models by minimizing average prediction error on the test samples. This is typically illustrated by Breiman and Spector (1992) in the context of linear regression. In this work, the authors obtain a nested sequence of models by simply backward deletion from a full model (using all data and all covariates) and then they use cross-validation to select among these models. The main difference of our approach is that we don't construct sequence of models a priori but use the cross-validation in an embedded way into stepwise selection to choose the variables to enter the model. A similar idea was also briefly mentioned in a further discussions section by Shao (1993) though to the best of our knowledge, it wasn't further analyzed or implemented. Shao (1993) suggests the addition to or deletion from the model to be based on average rss on different test samples. The final output of this method is a single model. Our method provides importance measures of features, which we discuss in the next subsection, that are used to transform the UCRs. The source codes for both of the methods are available through our complementary website. Some other typical model selection methods include AIC, BIC criterion and the practical out performance of monte carlo cross validation against these was shown by Pavlic and van der Laan (2001) in the mixture of normals context.

### 3.3 Summary Measures

**Experiment Specific Importance Measure:** Let $R_{w,k}(n)$ denote the rank of motif $w$ and $I_{w,k}(n)$ denote the indicator whether motif $w$ is selected in split $k$ of the experiment $n$. We then define a particular overall experiment specific importance measure of a motif $w$ as

$$R_w(n) = \frac{1}{K}\sum_{k=1}^{K}\frac{I_{w,k}(n)}{J}(J - R_{w,k}(n) + 1),$$

where $J$ represents the total number of terms added to the model. $R_w(n)$ represents a rank weighted proportion of times motif $w$ is selected within $K$ splits. Thus, it provides an overall importance measure that is between 0 and 1 for each motif. Hence, a motif that gets into the model as the first variable in all of the $K$ splits gets an $R_w(n)$ of 1 whereas the ones that are never selected get an importance measure of 0. We restricted the total number of terms in the model to be 10 based on the empirical observation that for situations where the prediction power of the model is low, the gain through additional terms decreases exponentially.

**Clustering of Experiment Specific Motifs:** A way of summarizing the regulatory motifs for each experiment is clustering them based on sequence similarity. Any clustering method that accepts user defined distance metric can be used. We used PAM (Kaufman and Rousseeuw (1990)) with the similarity measure of Vilo et al. (2000) which is defined as the length of maximum overlap between two sequences divided by the length of the shorter sequence. Clustering of the motifs is particularly important when there are variants of the regulatory motifs that are highly correlated.

# 4 Application and Results: Mitotic Cell Cycle in Yeast

We successfully tested the SCV with a simulation study where the response and the features were generated to represent the structures in a real data set. This included models of low $R^2$(e.g. 0.1) values (the proportion of variability due to linear relationship of response with the explanatory variables(features)). Low $R^2$ values raised even if one used as features the scores corresponding to exact consensus of the regulatory motifs. To speed up the computations we used a cutoff for percentage improvement on rss of the test sample. More detailed results of our simulation study can be reached through our complementary website.

We applied the SCV to the cell cycle data of yeast by Cho et al. (1998). The eukaryotic cell cycle consists of four phases: M (mitosis), S (synthesis, DNA replication), G1 (preparation for S phase and growth), G2 (preparation for mitosis). Some of the well known regulatory motifs in cell cycle dependent transcription are early G1 element ECB and late G1 elements MCB and SCB. Transcription factor MCM1 plays an important role in G2 phase and possibly with transcription factor SFF in M phase. Moreover, SWI5E, ACE2, RAP1P, NEG, CBF1, MET31p/MET32p are other potential regulatory motifs of different phases of the cell cycle.

In the experiments of Cho et al. (1998), cells were collected at 17 time points with 10 minute intervals to cover two full cell cycles. In our analysis, we discarded two time points (90 min, 100 min ) due to less efficient labelling of their mRNA prior to hybridization (Tavazoie et al. (1999)). We used normalized expression profiles of the most variable $\sim$ 2900 ORFs.

To demonstrate the strength of the method in identifying motifs relevant for a given experiment among a set of known regulatory motifs, we first focused on 50 regulatory motifs for which SCPD (Zhu and Zhang (1999)) provided a consensus sequence. We computed the scores as in eq. 1 using the consensus sequence of these regulatory motifs. Detailed results for time point 20 minutes (late G1 phase) are given in Table 1. The analyses were performed both by only using the scores for the consensus sequence and by combining the scores for the consensus sequence and its reverse complement. Since the results are not dramatically different, we report here only the latter case. The two regulatory motifs MCB and SCB are the most frequently selected ones with $R_w(n)$ values of 1 and 0.78. There is also an interaction term of SCB and MCB indicating a combinatorial effect of these regulatory motifs. This result is consistent with the biological findings of (Cho et al. (1998)). We conclude that SCV picks up most of the relevant explanatory variables with high $R_w(n)$ values. The detailed results can be found in our complementary web site.

We now extend the set of explanatory variables to include all possible pentamers. The scores for each of these are computed as in eq. 1. In table 2 the results for time interval 0-30 minutes are summarized including the corresponding $R_w(n)$ values. We only report pentamers with $R_w(n)$ greater than 0.1 since in a model with only 10 terms this corresponds to the $R_w(n)$ of a motif that is selected always tenth in all of the splits. If a term enters the model at lower ranks, it is important that this happens in a consistent way along many number of splits.

We observe that the selected pentamers have partial or exact match to most of the relevant regulatory elements (shown in parenthesis in table 2). Regulatory motifs MCB and SCB have quite strong effects in determining the gene expression throughout different phases of the cell cycle. The method identifies these regulatory motifs for all the phases in which they are known to be active. The pentamer AGGGG, which is the STRE regulatory element, is also identified with high importance measure. This is likely a result of the manner in which the cell culture was synchronized, supported by the observation that the highest importance measure of STRE is obtained at time 0. Transcription factor MCM1 is also known to be active in late G2/M phases. However, since MCM1 binds to a very degenerate binding site we were unable to conclude whether any of the matches we have do correspond to MCM1 sites.

We clustered the motifs of the time point 20 minutes to illustrate motif clustering as a summary measure. All the pentamers that were selected in at least two of the splits were clustered. Since the number of motifs in this set was relatively small (only 28 motifs) it was possible to try out all possible choices of number of clusters. Average silhouette, which is a measure of goodness of clusters (Rousseeuw (1987)), was maximized when number of clusters was set to 7. The seven clusters obtained are given in Table 3. The two subsequences of SCB (CGCGA and CGAAA) are in cluster 3 forming the exact motif when aligned. In cluster 4, 4 of the 7 motifs are one base pair variants of STRE (AGGGG) and one of them is the STRE itself.

Wolfsberg et al. (1999) identified pentamers and hexamers as potential regulatory motifs by analyzing UCRs of the genes that might have been involved in the cell-cycle dependent regulation of transcription. We compared the significant pentamers they found ($p \leq 0.05$) with our findings. Pentamers like ACGCG, CGCGA, AACAA are among the highly scored findings of our and their method. In general, most of the late G1 pentamers identified by Wolfsberg et al. (1999) were picked up by our method as well, however the pentamers selected for other phases were quite different.

Finally, it is worth commenting briefly on the construction of models based on candidate motifs to explain gene expression data. The motifs that enter the model are in general the biologically most relevant ones. The $R^2$ value of the fitted models of the test sample increases by $\sim 50$ % compared to a mean fit model indicating that selected motifs are truly helping to explain the gene expression. Moreover, the aggregated prediction (mean value of predicted gene expression over splits) has a lower prediction error (rss) than any of the individual predictions obtained at each split. However, the overall ability to predict gene expression solely on the basis of UCRs remains poor.

### 4.1 Extending the Pentamers

We had restricted our analysis to pentamers due to computational constraints. This enabled us to find either the relevant regulatory motif (eg. ACGCG for MCB) if indeed it is a pentamer or a core for it (CGCGA for SCB). Obviously, it is of interest to consider sequences of longer lengths and investigate whether these provide an improved prediction of gene expression. One possible extension that we implement on SCB core motif is as follows: During the SCV, whenever the core CGCGA is selected, we consider its length 6 extensions that are obtained by adding one base pair to left and right of the core, respectively. We then repeat the selection procedure for the set including the core word and its length 6 extensions. If a length 6 extension is selected as the best then it is immediately extended to length 7 in the same way and comparisons are performed. This process continues until there is no improvement from extending a length l word to length $l+1$ word. We illustrated this extension for time point 20 minutes of the cell cycle. Core CGCGA was extended to CGCGAAA which is the full predicted consensus for SCB regulatory motif. This extension method provides a quick way of extending the pentamer cores to longer sequences since at each extension step there are at most 8 motifs to compare with the selected motif. We have also applied the extension method to the MCB regulatory motif for which the latest reported consensus is ACGCGN (Iyer et al. (2001)). The extension method didn't extend ACGCG at time point 20 minutes. Considering that the empirical correlation between the pentamer ACGCG and hexamer ACGCGT was 0.78, this is a promising result. However, it is quite possible that when the correlation between the core motif and its extension is very high, extension from the core might not succeed. One is still left with a core that is a good representation of the real regulatory motif, and the clustering on the final output will bring the selected variants of the regulatory motif together.

## 5 Discussion

These results demonstrate that utility of motif detection methods that consider individual genome-wide expression experiment separately. Our and related methods nicely complement those based on the pattern of gene expression variation across numerous conditions.

It is natural to consider how our method can be applied to the very large multi-condition gene expression experiments that are now available. Using the importance measures, $R_w(n)$, one can build an $M$ by $N$ matrix where rows are sequence motifs and columns are the experiments in consideration. This matrix is a transformation of the $P$ by $N$ gene expression data matrix to a potential regulatory motif by experiment matrix in which each cell represents the significance of a given motif in a given experimental condition (a similar transformation matrix based on genome-mean expression profiles was also suggested by Chiang et al. (2001)). From such a matrix one can readily identify motifs that have importance across many experiments - we report motifs with highest mean importance measure, $\bar{R}_w$, across experiments in table 4. More significantly, it allows for the identification of motifs that are significant in limited numbers of conditions and those whose significance is obscured by stronger motifs active in the same conditions.

In essence, this transformation matrix brings us back into the statistical framework where interesting subsetting rules can be defined and clustering analysis can be performed to identify motifs that have similar regulatory effects across different experiments. Moreover, bootstrap analysis will be useful to study the reproducibility of the constructed clusters and subsets as in van der Laan and Bryan (2001). We are in the process of applying such tools and developing complementary methods.

To summarize, the method we have suggested both serves experiment specific regulatory motif identification and generates groups of motifs that have similar roles across different experimental conditions. One future research direction will be the extensive comparison of this method with other regulatory motif finding methods based on clustering of the genes using expression data across variety of conditions.

## 6 Acknowledgments

We thank Peter Bickel, Katherina Kechris, Erik van Zwet, Jaimyoung Kwon and the Eisen Lab for many helpful discussions. In particular, we are also thankful to Jaimyoung Kwon for providing the empirical distribution of the known regulatory motifs using SCPD. This work has been supported by NIAID grant 1R01 AI46182-01.

## References

C.E. Lawrence A.F. Neuwald, J.S. Liu. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science*, 4:1618–1632, 1995.

Y. Bilu and M. Linial. On the predictive power of sequence similarity in yeast. In *Proceedings of RECOMB*, Montreal,Canada, 2001.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks/Cole, Monterey, 1984.

L. Breiman and P. Spector. Submodel selection and evaluation in regression. the x-random case. *International Statistical Review*, 60:291–319, 1992.

H.J. Bussemaker, H. Li, and E.D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27:167–171, 2001.

D.Y. Chiang, P.O. Brown, and M.B. Eisen. Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics*, 17:49–55, 2001.

R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J.Lockhart, and R.W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol.Cell*, 2:65–73, 1998.

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

G.Z. Hertz and G.D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–577, 1999.

J.D. Hughes, P.W. Estep, S. Tavazoie, and G.M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *Journal of Molecular Biology*, 296(5):1205–1214, 2000.

V.R. Iyer, C.E. Horal, M. Synders D. Botstein, and P.O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409:533–538, 2001.

L.J. Jensen and S. Knudsen. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, 16(4): 326–333, 1999.

L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990.

C.E. Lawrence, S.F. Altschul, M.S. Boguski, A.F. Neuwald J.S. Liu, and J.C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

C.E. Lawrence and A.A. Reilly. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function and Genetics*, 7:41–51, 1990.

D. J. Lockhart, H. Dong, M. C. Byrne, M. V. Gallo M. T. Follettie, M. S. Chee, M. Mittmann, C. Want, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high density oligo nucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

M. Pavlic and M.J. van der Laan. Fitting of mixtures with unspecified number of components using cross-validation distance estimate. *Technical Report #89, Division of Biostatistics, UC Berkeley*, 2001.

P.J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.

J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88:486–494, 1993.

S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. In *Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 344–354, San Diego, CA, 2000. AAAI.

S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genet.*, 22:281–285, 1999.

C. Elkan T.L. Bailey. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21(1-2):51–80, 1995.

M. Tompa. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 262–271, Heidelberg, Germany, 1999. AAAI.

M.J. van der Laan and J.F. Bryan. Gene expression analysis with the parametric bootstrap. *Biostatistics*, 2:1–17, 2001.

J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Bilogy*, 281:827–842, 1998.

E. van Zwet, 2001. Personal Communication.

J. Vilo, A. Brazma I. Jonassen, A. Robinson, and E.Ukkonen. Mining for putative regulatory elements in the yeast genome using gene expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 384–394. AAAI, 2000.

T.G. Wolfsberg, A.E. Gabrielian, M.J. Cambpbell, R. J. Cho, and D. Landsman J. L. Spouge. Candidate regulatory sequence elements for cell cycle-dependent transcription in saccharomyces cerevisiae. *Genome Research*, 9:775–792, 1999.

J. Zhu and M.Q. Zhang. Scpd: A promoter database of yeast saccharomyces cerevisiae. *Bioinformatics*, 15:607–611, 1999.

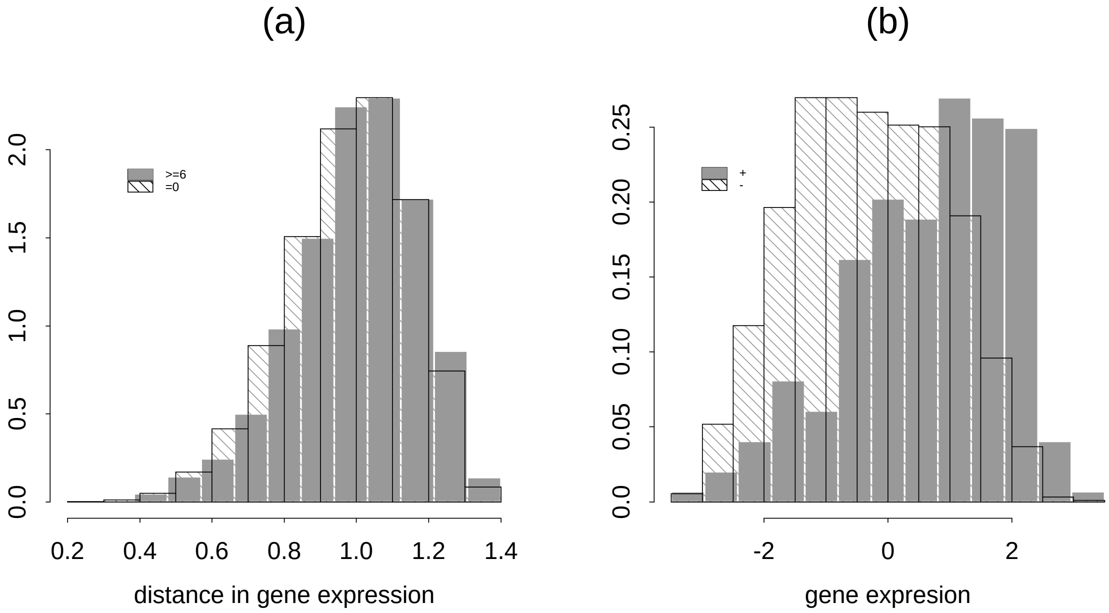

Figure 1: (a) Distribution of gene expression distance in groups of genes with no cell cycle related regulatory elements versus at least 6 (out of 11) cell cycle related regulatory elements. (b) Histogram of gene expression for time point 20 minutes of the cell cycle for the groups with(+) and without(-) the MCB and SCB regulatory motifs.

| MOTIF | # | A.RSS | RANK |
|---|---|---|---|
| **MCB** | 50 | 5.8497 | 1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1 |
| STRE | 46 | 0.6509 | 2.2.2.2.2.6.2.4.2.31.2.2.3.2.2.4.2.2.2.2.4.19.5.2.4.2.4.5.4.18.4.2.2.4.5.2.43.2.2.2.2.2.6.2.4.3 |
| **SCB** | 48 | 0.5799 | 3.4.3.3.3.4.3.2.3.3.3.2.4.2.3.3.3.8.2.6.3.3.2.3.2.3.2.2.3.2.4.2.2.3.2.2.3.2.3.3.3.3.4.2.3.2.2.2 |
| MET31 | 38 | 0.2059 | 8.12.7.10.11.23.8.8.5.3.8.6.19.6.19.9.6.7.9.5.6.9.7.17.9.15.31.21.5.5.3.14.6.7.7.7.7.5 |
| GCR1 | 36 | 0.1939 | 12.9.12.8.9.11.17.34.8.6.12.8.29.7.13.21.6.13.8.9.9.11.8.9.7.16.7.34.48.7.7.9.8.13.9.11 |
| REB1 | 35 | 0.1277 | 16.15.8.53.34.16.32.12.6.39.7.27.9.18.10.38.15.16.10.27.14.7.15.14.17.13.10.5.13.6.6.7.11.35.29 |
| URSPHR | 41 | 0.089 | 18.14.25.22.28.53.13.17.15.12.16.38.15.28.26.34.19.10.13.22.13.27.12.16.12.16.30.18.12.8.16.14.7.16.15.8.17.25.43.9.29 |
| SCB*MCB | 37 | 0.1641 | 4.5.4.4.5.4.3.18.4.4.5.3.4.4.9.3.7.4.4.3.4.3.4.3.3.5.3.4.33.4.4.4.4.5.3.3.3 |

Table 1: Regulatory motifs selected at Late G1 phase (time point 20 minutes) of the cell cycle. Second column gives the total number of splits the motif is selected, and third column is the average percentage decrease it provides in rss of the test sample, and the last column is the rank profile over splits. Total number of splits performed is 50.

| T=0min | | T=10min | |
|---|---|---|---|
| Motif | $R_w(n)$ | Motif | $R_w(n)$ |
| AGGGG/CCCCT(STRE) | 0.925 | AAACA/TGTTT(STE12) | 0.6 |
| ACGCG/CGCGT(MCB) | 0.765 | CTTAA/TTAAG | 0.535 |
| GAAAA/TTTTC(ECB) | 0.665 | GTTTA/TAAAC(SFF) | 0.51 |
| CGGAG/CTCCG | 0.49 | GCGAA/TTCGC(SCB) | 0.495 |
| AAGGG/CCCTT | 0.48 | CAGAC/GTCTG | 0.37 |
| CATAA/TTATG | 0.39 | AACAT/ATGTT | 0.355 |
| CATCG/CGATG | 0.22 | ACTTC/GAAGT | 0.3 |
| GGATA/TATCC | 0.185 | TTCAA/TTGAA | 0.27 |
| TCGCA/TGCGA | 0.165 | AGGGG/CCCCT | 0.21 |
| TCCGA/TCGGA | 0.165 | GATGA/TCATC | 0.19 |
| AGATC/GATCT | 0.16 | CCACG/CGTGG | 0.185 |
| AGTTC/GAACT | 0.16 | GGGGA/TCCCC | 0.145 |
| ACCCG/CGGGT | 0.14 | CGCGC/GCGCG | 0.145 |
| AGGGG/CCCCT*ACGCG/CGCGT | 0.125 | CAGGG/CCCTG | 0.135 |
| | | CAGTA/TACTG | 0.1 |
| | | ACGGA/TCCGT | 0.1 |

| T=20min | | T=30min | |
|---|---|---|---|
| Motif | $R_w(n)$ | Motif | $R_w(n)$ |
| ACGCG/CGCGT(MCB) | 1.000 | ACGCG/CGCGT(MCB) | 1 |
| CGCGA/TCGCG(SCB) | 0.745 | CCACA/TGTGG | 0.465 |
| AGGGG/CCCCT(STRE) | 0.55 | CGCGA/TCGCG(SCB) | 0.46 |
| AAACA/TGTTT(STE12) | 0.38 | CTCCA/TGGAG | 0.395 |
| GAAGC/GCTTC | 0.315 | ATAAC/GTTAT | 0.225 |
| CCTGA/TCAGG | 0.305 | AGGAA/TTCCT | 0.225 |
| AAGGG/CCCTT | 0.27 | CACGG/CCGTG | 0.185 |
| TAAAA/TTTTA | 0.19 | GGTAA/TTACC | 0.18 |
| TGAAA/TTTCA | 0.17 | CCCAC/GTGGG | 0.18 |
| AAAGG/CCTTT | 0.17 | GTTGA/TCAAC | 0.175 |
| ATATC/GATAT | 0.155 | CTCAA/TTGAG | 0.16 |
| AAATA/TATTT | 0.15 | CACAA/TTGTG | 0.14 |
| CTATA/TATAG | 0.13 | TTAAA/TTTAA | 0.14 |
| CGAAA/TTTCG | 0.11 | CTAAA/TTTAG | 0.14 |
| GCACC/GGTGC | 0.11 | TAAGA/TCTTA | 0.13 |
| GTTTA/TAAAC | 0.1 | AGGTA/TACCT | 0.11 |
| ACGCG/CGCGT*CGCGA/TCGCG | 0.64 | | |

Table 2: Selected pentamers and their reverse complements for time points {0,10,20,30} minutes of the cell cycle. Selected motifs have exact or partial matches to the regulatory motifs given in parenthesis. Total number of variables in the model is restricted to 10.

| cluster 1(MCB-like) | cluster 2 | cluster 3 (SCB-like) | cluster 4 (STRE-like) |
|---|---|---|---|
| ..ACGCG/CGCGT<br>GCACC../GGTGC<br>ACACG../CGTGT | AAATA./TATTT<br>AAATC./GATTT<br>.GATCA/TGATC<br>CTCTC./GAGAG | CGCGA..../TCGCG<br>..CGAAA../TTTCG<br>....AAACA/TGTTT<br>...TGAAA../TTTCA<br>...TAAAA./TTTTA<br>..CGAGC../GCTCG | ..AGGGG./CCCCT<br>AAAGG.../CCTTT<br>GAAGC.../GCTTC<br>.AAGGG../CCCTT<br>...GGGTT./AACCC<br>.AGAGG../CCTCT<br>...AGGTG/CACCT |
| cluster 5 | cluster 6 | cluster 7 | |
| ATTTC/GAAAT<br>GTTTA/TAAAC | CCTGA./TCAGG<br>.ATGAG/CTCAT | ATATC/GATAT<br>CTATA/TATAG<br>ATAAG/CTTAT<br>ATAGA/TCTAT | |

Table 3: Clusters of motifs for time point 20 minutes.

| motif | $\bar{R}_w$ | match |
|---|---|---|
| ACGCG/CGCGT | 0.541665 | MCB(5/6) |
| AGGGG/CCCCT | 0.247000 | STRE(5/5) |
| CGCGA/TCGCG | 0.193665 | SCB(5/7) |
| CCAGC/GCTGG | 0.114335 | SWI5E(5/8) |
| AAACA/TGTTT | 0.103000 | STE12(5/7) |
| GTTTA/TAAAC | 0.087335 | SFF(5/8) |

Table 4: Motifs with overall importance measure across experiments $\geq 0.08$. Ratio of the length of the pentamer to the length of the exact regulatory motif is given in parenthesis.