TEL AVIV UNIVERSITY אוניברסיטת תל אביב

SACKLER FACULTY OF MEDICINE
DR. MIRIAM AND SHELDON G. ADELSON
GRADUATE SCHOOL OF MEDICINE

הפקולטה לרפואה ע״ש סאקלר
המדרשה לתארים מתקדמים
ע״ש ד״ר מרים ושלדון ג׳ אדלסון

Department of Physiology and Pharmacology

החוג לפיזיולוגיה ופרמקולוגיה

Nucleotide variation of regulatory motifs may lead to distinct expression patterns

Thesis submitted as part of the requirements for the degree of

Master of Science (M.Sc.)

in
Tel Aviv University
Sackler Faculty of Medicine

by

Liat Segal

The research work for this thesis has been carried out under the supervision of

Prof. David Horn
School of Physics and Astronomy

Prof. Eytan Ruppin
Sackler Faculty of Medicine
School of Computer Science

March 2007

ACKNOWLEDGMENTS

First and foremost, I wish to thank my academic supervisors, Prof. David Horn and Prof. Eytan Ruppin, for providing me the opportunity of doing this research, for their guidance and helpful insights. I would especially like to express my deep appreciation to David for his constant care and patience and for the numerous ideas he has introduces to me.

I would like to thank Zach Solan for his help in jumpstarting this project, his good advises and friendship. I'd also like to thank my lab mates that have made this journey a pleasant and funny one: Roy Varshavsky, Uri Barkan, Yasmin Meroz, Vered Kunik, Assaf Gottlieb, Yaron Levy, Uri Weingart and Ben Sandbank.

I thank Prof. Yitzchak Pilpel for providing the EC analysis data for this research.

I wish to thank the Adi Lautman interdisciplinary program for outstanding students and the Lautman family, for providing a rare academic freedom and support.

I am especially grateful to my family, for their love, encouragement and support. You have given me so much confidence and love that cannot be put to words!

And… to my beloved Jonathan. You are a fountainhead of wisdom and encouragement to me!

ABSTRACT

Motivation

Current methodologies for selection of putative transcription factor binding sites (TFBS) rely on various assumptions such as over-representation of motifs occurring on gene promoters, and the use of motif descriptions such as consensus or PSSM. As regulatory motifs are not necessarily over represented in the entire genome, the first demand brings the need for pre-processing of the data and to initially group genes that appear to be co-regulated, using additional data sources to the sequential ones. In order to avoid bias introduced by such assumptions we apply an unsupervised motif extraction (MEX) method, originally designed for extracting words from corpora of natural languages, to sequences of promoters. This allows us to seek biological insights that have previously been overlooked due to such assumptions. The work presented in this dissertation is based on an article that has recently been submitted for publishing.

Methods

We have applied MEX on the promoter regions of *S. cerevisiae*, aiming to identify putative cis-regulatory motifs through a genome-wide analysis. MEX does not depend on over-representation of the motifs in the genome, nor does it rely on clustering or other pre-processing of its input. Instead it uncovers motifs that are significant within the relatively local context of the promoters on which they occur.

The putative cis-regulatory motifs have been further screened, in terms of their regulatory significance, via the expression coherence (EC) of their genes in 40 experiments.

We have then clustered the regulatory motifs based both on their DNA sequence and on the biological conditions in which they govern coherent gene expression. Such grouping reveals biological insights that are easily missed by conservative clustering methods, which rely either on sequence or on numerical data alone.

Results

The MEX methodology is applied to all *S. cerevisiae* genes and is found to be very successful when tested on results of 40 gene expression experiments, via the EC analysis. Clustering regulatory motifs that have highly significant scores of EC, we

describe 20 clusters, some of which regroup known TFBS. The clusters display different EC profiles, correlated with typical changes in the nucleotide composition of their relevant motifs. In several cases, a variation of a single nucleotide is shown to lead to distinct differences in expression patterns. These results are confronted with other available information, such as *in-vivo* binding of transcription factors to groups of genes. Detailed analysis is presented for clusters related to MCB/SCB, STRE and PAC. In the first two cases we provide evidence for different binding mechanisms of different clusters of motifs. For PAC related motifs we uncover a new cluster that has so far been overshadowed by the stronger effects of known PAC motifs.

Conclusions

While conventional representations of motifs by consensus or PSSM are common and simple, such representations involve the loss of information and may lead to wrong predictions. As MEX does not use such representations, we can analyze each motif independently, and only then generate clusters of regulatory motifs, gaining a better understanding of the regulation without reducing the sequence information or biasing the results. We have learned from our analysis that single changes of a nucleotide within a motif can go a long way in affecting the regulation of genes. The strength of regulation may depend on various mechanisms. We have tested the repetition rates of motifs on the promoters and the localization of motifs upstream to genes to decide whether any of them should carry the burden for higher or lower regulation strength, or whether it is the binding mechanism of the TF to specific motifs that does it. In both the MCB/SCB and STRE clusters we have concluded that the latter is the case. Both examples demonstrate that small variations in regulatory motifs lead to high magnitude effects on regulation. Even a single nucleotide substitution at the motifs of these clusters is sufficient for such effects, acting as a tuner of regulation.

Key words

Saccharomyces Cerevisiae, transcription factor, binding site, regulatory motif, gene expression, clustering, motif extraction

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Transcription factor binding sites

Regulation of gene expression is mainly mediated through specific interactions of transcription factors (TF) with DNA promoter elements. The TF binding sites (TFBS) are short (typically of length 6-20 bases) and comprise a minority of the nucleotides within a promoter region. The binding sites are embedded within a sequence that is assumed to be nonfunctional with respect to transcription. Furthermore, a single transcription factor protein may interact with a variety of sequences. Identifying genuine binding sites is a challenging task as the physical extent of a promoter is rarely well defined, and within this ill-defined region we are seeking sparsely distributed, short and imprecise sequence motifs.

1.2 Related work on motif extraction from promoter regions

Advances in genome research, including whole genome sequencing and mRNA expression monitoring have allowed the development of computational methods for binding site prediction. Among the most popular and powerful methods for *ab initio* detection of regulatory motif is Gibbs-sampling [23, 20]. In this method motifs that are over represented in the data may be found. However since regulatory motifs are very short, while in contrast, the regulatory portion of the genome is very long (e.g., 6,000,000 base-pairs in yeast, and much longer in mammals), and since the size of gene regulatory networks is relatively small (typically tens of genes), most regulatory motifs are not expected to be over-represented on a genome-wide scale. The task of motif identification is thus often first tackled by grouping together relatively small sets of genes (tens or hundreds) that are likely to be co-regulated, followed by motif searching within such groups [9, 19, 43, 32].

Other methods employ phylogenetic footprinting for the task of motif finding. Such methods compare upstream regions of orthologous genes from related species, assuming that TFBS are relatively conserved. The choice of species is crucial for obtaining reliable results; Comparing species with a short divergence time may yield false positives, as conservation is likely to reflect evolutionary proximity rather than functional constraints. A choice of too distant species will fail to recover species-specific sites. For instance, about 40% of human functional TFBS are expected to be non functional in rodents [13]. Furthermore, the alignment of orthologous intergenic

sequences is non-trivial. Well-conserved sequence blocks of different lengths are interspersed with sequences that show little conservation. It is common practice to restrict the binding site search to genomic regions that are relatively conserved among all selected species. However, regulatory sites are not necessarily restricted to such conserved genomic segments, as has been shown in yeast and flies [14, 16, 38].

For most TFs, there appears to be no unique sequence of bases that is shared by all recognized binding sites. However there are typically clear biases in the distribution of bases that occur at each binding site position. These biases are commonly represented mathematically by position specific scoring matrices (PSSM), whose components give the probabilities of finding each nucleotide at each binding site position [7, 41].

Motif representations by PSSM, however, ignore dependencies between nucleotide positions in regulatory motifs. Such dependencies are known to occur [6, 10]. Statistical models that account for such dependencies include hidden Markov models, and Bayesian networks [15]. Yet, even sophisticated models of this kind have relatively low values of sensitivity and specificity when required to represent the known binding sites [3].

1.3 Motivation and outline

The work presented in this dissertation is based on an article that has recently been submitted for publishing [37]. Here we employ a different approach that attempts to avoid the limitations and inherent assumptions discussed above. We adapt a recently published unsupervised algorithm [39], designed originally to extract patterns from natural-language corpora. This motif extraction algorithm (MEX) is based on a statistical model that identifies consecutive chains of interdependencies between adjacent nucleotide positions. It can thus successfully identify motifs as statistically significant on a genome-wide scale, even without significant over-representation. The algorithm readily detects the motif boundary, as the position where the series of highly probable transitions begins or terminates. MEX both overcomes the requirement to pre-group potentially co-regulated genes, and captures interdependencies between motif positions.

Applying MEX to genome-wide yeast regulatory sequences, we extract sequence motifs. We then validate their biological significance using whole genome mRNA expression data. We use the expression coherence (EC) score [22, 33] in order to

check which of the identified putative motifs exert significant effects on the expression profiles of their down stream genes. The expression analysis shows an enormous enrichment of highly-scoring motifs among MEX's predictions, and it also identifies potential biological conditions in which these motifs act. We further group the high-scoring motifs into subsets based not only on their raw DNA sequence, but also on the biological conditions in which they govern coherent expression. Such grouping reveals biological insights that are easily missed by conservative clustering methods, which rely either on sequence or on numerical data alone. For instance partially overlapping binding sites that are bound by distinct TFs regulating different biological conditions, are indistinguishable by sequence, yet may appear in separate clusters using our method. Another biological phenomenon we can capture is slight variations in binding site sequence which result in different expression outputs. Our analysis shows that the commonly used PSSM description does not capture some very important properties as there exist specific structural relations that correlate with high EC values in particular biological conditions, i.e. they are of functional importance.

2 METHODS

2.1 Motif Extraction algorithm (MEX)

MEX is a motif extraction algorithm [39] that extracts statistically significant motifs from sequential data. MEX is a data driven unsupervised algorithm, hence does not need any preprocessing of the data or additional information apart from the data set itself. Furthermore, MEX finds motifs that are not necessarily over-represented in the data.

MEX was originally developed in a linguistic context, as a distillation tool for extracting words from corpora of natural language. As more intuitive, let us first describe the algorithm in its original context.

Consider a corpus of sentences, whose word delimiters have been removed (such as spaces, capital letters, punctuations, etc.). The problem at hand is to uncover the words that have originally constructed the sentences. MEX receives as an input such corpus, consisting of many sequences of a given finite alphabet of size N (e.g. N=26 letters in the English alphabet, N=20 amino acids in proteins and N=4 nucleic acids in the case of DNA). The algorithm uses a directed graph, whose vertices, V, are composed of the letters of the given alphabet, in addition to a 'begin' and an 'end' vertices. A set of ordered pairs of vertices (directed edges) represent the order in which the letters appear in the corpus. For example, the edge e(t,h), represents a connection from the vertex 't' to the vertex 'h', which means that the letter 'h' appears at some point along the corpus after the letter 't'. MEX loads the given corpus onto a directed graph, one sentence after the other. The edges representing each sentence are built, starting with the 'begin' vertex, followed by the letters composing the sentence, one after the other, and ending with the 'end' vertex. This way, ordered paths are created in the graph, such that each sentence is represented by a path. Each path is saved by MEX and will be used as a search path for patterns. This procedure is demonstrated in figure 2.1.

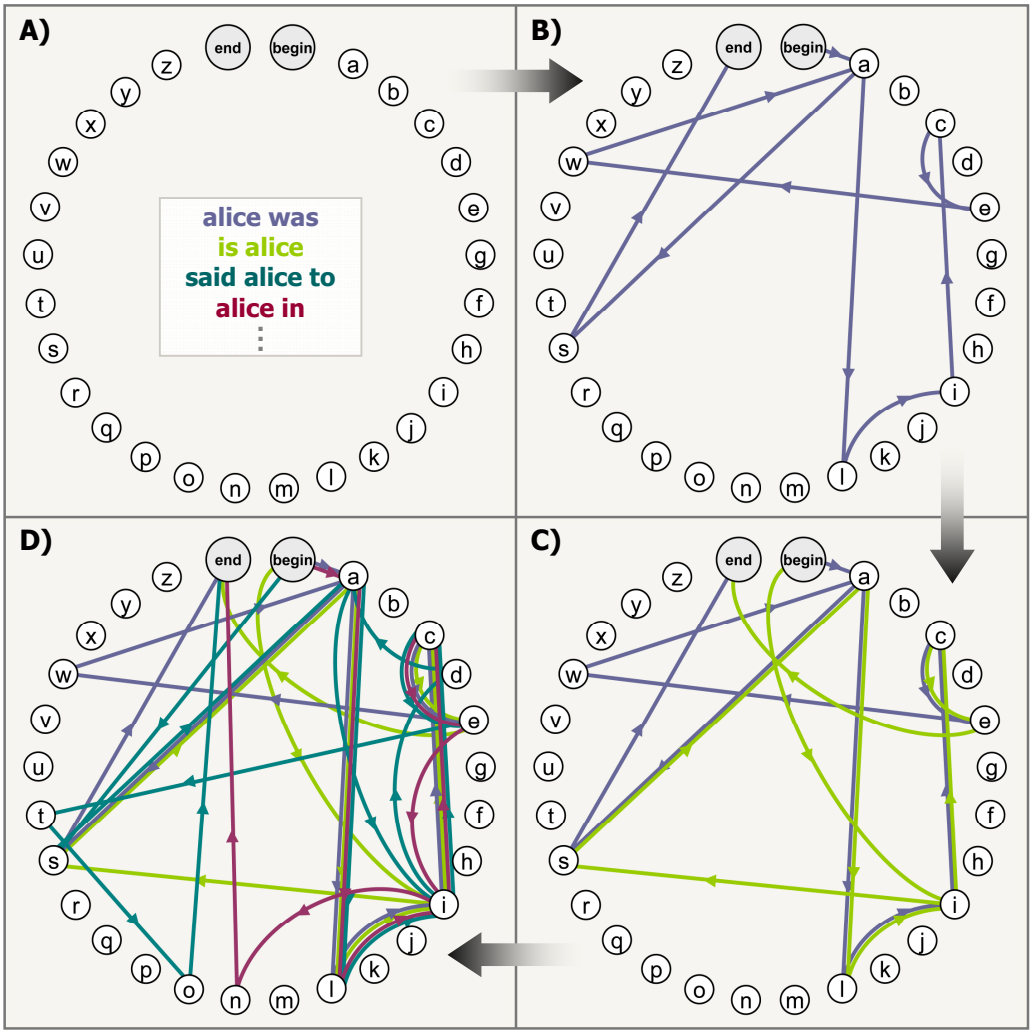


Figure 2.1 MEX loads the corpus onto a directed graph, one sentence after the other. The graph is composed of vertices representing the letters of the given alphabet, in addition to two vertices representing the beginnings and the endings of sentences (A). One sentence at a time, directed edges are added to the graph, representing the order in which letters appear in each sentence (B-D). The ordered edges composing a sentence are considered a path along the graph. In this example, four paths are loaded onto the graph: 'alicewas' (blue path), 'isalice' (light green path), 'saidalice' (turquoise path) and 'alicein' (red path), one after the other.

Once the entire corpus has been represented as search paths on a directed graph, the algorithm starts searching for statistically significant patterns. Intuitively, for each search path MEX looks for sub-paths that may be considered as candidates for being significant patterns. A sub-path that represents a significant pattern is expected to be shared by other paths throughout the graph, such that these paths will converge into the sub-path at its first vertex, form a bundle along the sub-path and scatter after the sub-path's last vertex. This follows the assumption that at different instances of a given word throughout the corpus, after the word ends, it is likely to find many different possible words following it. In such a case many paths will form a bundle along the sub-path representing the word and scatter immediately after it ends. The vertex after which such a divergence occurs may be considered as the last vertex of the pattern. A similar notion underlies the way MEX searches the start points of patterns, by looking for a divergence of a bundle while going leftwards through a search path. Figure 2.2 demonstrates this idea. The four paths in figure 2.2 converge and form a bundle along the sub-path 'a→l→i→c→e', after which they diverge.

This can be rephrased into a probabilistic language; for each search path (sentence) that is to be explored for patterns, two probability functions are defined, based on information inheres in the complete graph. The first one, P_{Right}, is the right moving ratio of the through-going flux of paths to the incoming flux of paths, which varies along the search path. Starting at the vertex e_1 we define P_{Right} at e_2 as:

$$P_{Right}(e_1, e_2) = p(e_2 \mid e_1) = \frac{\text{total no. of paths passing from } e_1 \text{ to } e_2}{\text{total no. of paths entering } e_1}$$

At e_3 P_{Right} becomes:

$$P_{Right}(e_1, e_3) = p(e_3 \mid e_1 e_2) = \frac{\text{total no. of paths passing from } e_1 \text{ through } e_2 \text{ to } e_3}{\text{total no. of paths passing from } e_1 \text{ to } e_2}$$

And generally:

$$P_{Right}(e_i, e_j) = p(e_j \mid e_i e_{i+1} e_{i+2} \cdots e_{j-1}) = \frac{\text{total no. of paths passing from } e_i \text{ up to } e_{j-1} \text{ and continue to } e_j}{\text{total no. of paths passing from } e_i \text{ up to } e_{j-1}}$$

Similarly, a second function, P_{Left}, is defined as we proceed leftward from some vertex e_j down the search path towards the vertex e_i and examine the left-going ratio of the through-going flux of paths to the incoming flux of paths:

$$P_{Left}(e_j, e_i) = p(e_i \mid e_{i+1} e_{i+2} \cdots e_{j-1} e_j) = \frac{\text{total no. of paths passing from } e_i \text{ to } e_j}{\text{total no. of paths passing from } e_{i+1} \text{ to } e_j}$$

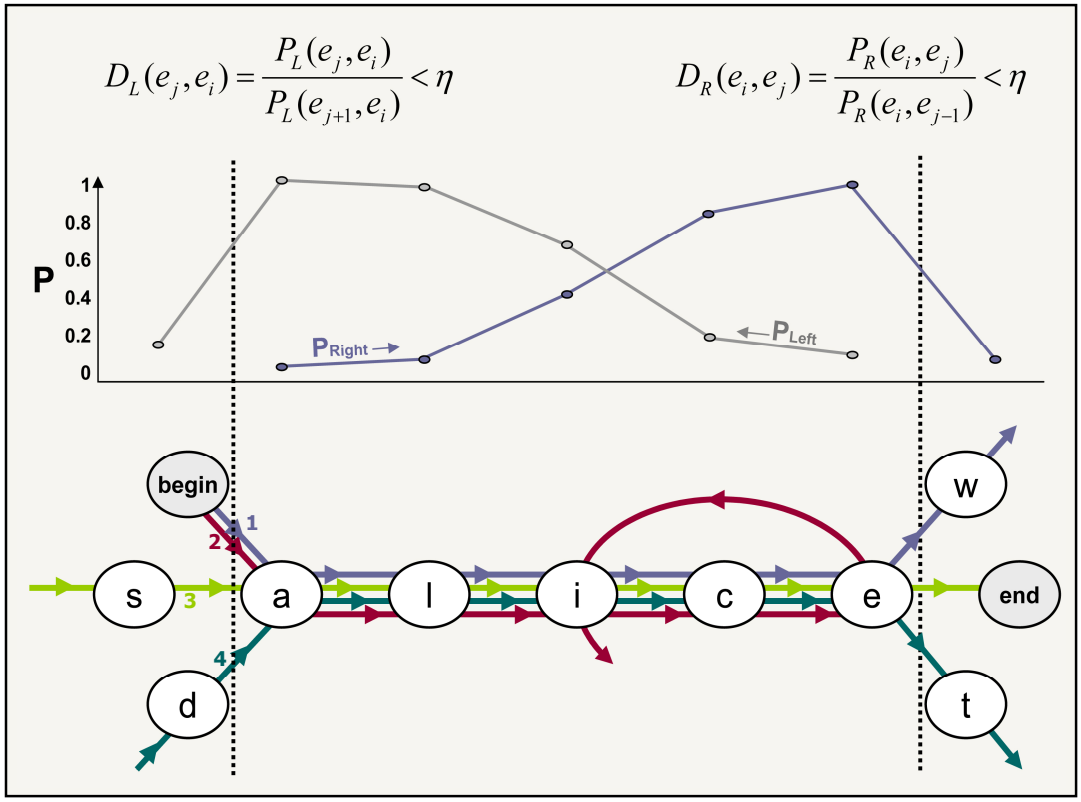


Figure 2.2 A partial view of the graph used by MEX. The search path no. 1, 'alicewas' (blue line), shares the sub-path 'a→l→i→c→e' with three other paths: 'isalice' (2), 'saidalice' (3) and 'alicein' (4). The four paths form a bundle that may constitute a significant pattern. The conditional probabilities P_{Right} and P_{Left}, originating at the vertices 'a' and 'e', respectively, are illustrated for the example shown here. A sharp drop in the right moving probability, P_{Right}, indicates that the paths constructing the bundle have scattered, thus may denote the end of the pattern. Similarly, a sharp drop in P_{Left} may indicate the beginning of the pattern, hence reveal the pattern 'alice'.

Vertex	Conditional Probability Expression	P_{Right}
a	P(a) = 8770 / 109625	0.08
l	P(al | a) = 1046 / 8770	0.12
i	P(ali | al) = 486 / 1046	0.45
c	P(alic | ali) = 397 / 486	0.85
e	P(alice | alic) = 397 / 397	1
w	P(alicew | alice) = 48 / 397	0.12
a	P(alicewa | alicew) = 21 / 48	0.44
s	P(alicewas | alicewa) = 17 / 21	0.81
b	P(alicewasb | alicewas) = 2 / 17	0.12
e	P(alicewasbe | alicewasb) = 2 / 2	1
g	P(alicewasbeg | alicewasbe) = 2 / 2	1
⋮	⋮	⋮

Table 2.1 Calculating right-going conditional probebilities for the search path 'alicewasbegining...'. Probablities are calculated for a given serach path, based on information inheres in the entire graph. The corpus used in this example was the sentences from Alice in wonderland, by Lewis Carroll.

MEX calculates P_{Right} from different starting points to each vertex down the search path. Going rightwards through a sub-path that represents a significant pattern, it is expected that P_{Right} will first increase since other paths join the search path to form a coherent bundle, and then decrease as many paths leave the search path.

In order to demonstrate this, let us examine as a toy problem the corpus of Alice in wonderland, by Lewis Carroll. MEX has received as an input the sentences within Alice in wonderland, after all word delimiters have been removed. Going through the first search path 'alicewasbeginningtogetverytired...' MEX calculates the rightward-going probabilities, P_{Right}, along the path, as demonstrated at table 2.1. MEX starts at the first vertex 'a' and calculates the probability of its appearance in the corpus, $P_{Right}(a)$; as 'a' appears in 8770 cases out of the total of 109625 letters in the corpus, $P_{Right}(a)=\frac{8770}{109625}=0.08$. MEX continues to the next vertex 'l', calculating the probability of its appearance after the previous vertex, i.e. $P_{Right}(al|a)$; in this case, 'l' appears 1046 times after the 8770 instances of 'a', hence $P_{Right}(al|a)=\frac{1046}{8770}=0.12$. MEX continues calculating the rightward-going probabilities $P_{Right}(ali|al)$, $P_{Right}(alic|ali)$ and so on, up to the end of the search path. As can be seen in table 2.1, the rightward-going probabilities initially rise and then drop sharply. Such a dramatic drop may occur owing to the sudden divergence of a coherent bundle, and will be considered as a candidate for terminating a pattern.

We will define the end of a motif as the vertex after which a dramatic drop in the right-moving probabilities is apparent (expressing the divergence of edges from that vertex), and the beginning of a motif as a dramatic drop in the left moving probabilities (expressing the convergence of edges to that vertex).

Formally, let us define a "decrease ratio":

$$D_{Right}\left(e_i, e_j\right) = \frac{P_{Right}\left(e_i, e_j\right)}{P_{Right}\left(e_i, e_{j-1}\right)}$$

$$D_{Left}\left(e_j, e_i\right) = \frac{P_{Left}\left(e_j, e_i\right)}{P_{Left}\left(e_j, e_{i+1}\right)}$$

We will declare e_{j-1} as a candidate end point of the pattern if $D_{Right}(e_i,e_j)$ is smaller than a preset cutoff parameter $\eta<1$. Similarly, e_{i+1} will be declared as candidate start point of a pattern if $D_{Left}(e_j,e_i)<\eta$.

The statistical significance of the decreases in P_{Right} and P_{Left} must be evaluated. P_{Right} and P_{Left} can be regarded as variable-order Markov probability functions. We can

define their significance in terms of a null hypothesis stating that $P_{Right}(e_i,e_j) \geq \eta P_{Right}(e_i,e_{j-1})$ and $P_{Left}(e_j,e_i) \geq \eta P_{Left}(e_j,e_{i+1})$, and require that the p-values of both $D_{Right}(e_i,e_j)<\eta$ and $D_{Left}(e_j,e_i)<\eta$ be, on average smaller than a preset threshold parameter $\alpha<1$.

A bundle of coinciding paths whose end-points obey these significance conditions is declared as a possibly significant pattern. Given a search path, we calculate both P_{Right} and P_{Left} from all of the possible starting points, traversing each path leftward and rightward, correspondingly. This technique defines many search-sections, which may be candidates for significant patterns. The most significant ones of these candidates are returned as the outcome patterns for the search path in question.

2.2 Applying MEX on promoter regions of *S. cerevisiae*

Given a set of DNA sequences, such as the promoters of all genes in *S. cerevisiae*, one may regard each promoter as a sentence with an alphabet of size four, corresponding to the four nucleic acids composing the DNA. This corpus of promoter regions in *S. cerevisiae* was given to MEX as an input, aiming to extract regulatory motifs.

In the yeast problem we apply MEX to 4800 promoters of 6300 genes (some promoters are shared by two genes because they fall within the intergenic region of two genes that are located on opposite strands of the DNA chain). Each promoter sequence, of length up to 1000bp, is considered as a path on the graph. After all information is loaded onto the graph, we use all 4800 sequences as trial-paths in order to extract motifs.

MEX selects motifs according to some edge criteria rather than over-representation in the data set. Nonetheless it can pick up repetitive motifs, in particular those of very high occurrence (in the thousands), that may be completely unrelated to regulatory functions. Hence we limit ourselves to motifs whose occurrence rate is between 5 and 100 per promoter. We also require a lower limit of length 6 for the motifs.

2.3 Expression Coherence analysis

In order to check which of the motifs extracted by MEX are likely to function as regulatory elements in yeast, we have used the expression coherence (EC) method [22, 33, 42]. The EC score of a motif that appears in the promoters of N genes is defined as the fraction of gene pairs (i,j) in the set S, such that the Euclidean distance between their mean and variance normalized expression profiles, $D(ij)$, falls below a threshold, D, divided by the total number of gene pairs in the set, $\frac{1}{2}N(N-1)$. The value D is set as a distance at which random gene pairs have a probability p of scoring below. The EC score may range between 0 and 1 and is higher for sets of genes that cluster in one or a few tight clusters.

A sampling-based means exists for the assessment of the statistical significance of EC scores, in terms of p-value, given the gene set size N [22]. In order to account for the testing of multiple hypotheses and to control the amount of false positives, the EC analysis uses the false discovery rate (FDR) theorem [5]. The FDR criterion determines the p-value cutoff below which motifs are guaranteed to be statistically significant at a specified false discovery rate.

Expression analysis of genes that contain regulatory motifs in their promoters allows not only to select potentially functional motifs, but also to decipher their semantics. A comprehensive semantic characterization of a regulatory motif would amount to describing the condition in which it acts, and its regulatory effects, e.g. increase in expression along a particular stress, or peaking of expression profile during a particular phase of the cell cycle. Figure 2.3 shows such semantic annotation of two high scoring sequence-motifs generated by MEX. These motifs govern opposite responses to hypo-osmotic pressure.

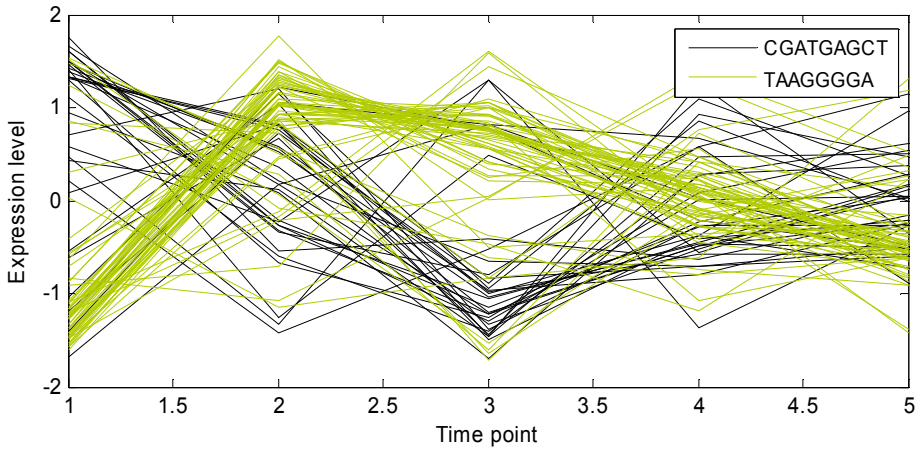


Figure 2.3 A semantic characterization of two of the motifs extracted by MEX. MEX has identified two motifs governing opposite responses to hypo-osmotic stress. As shown by the graph, genes containing the motif CGATGAGCT (corresponding to the PAC motif) in their promoters (black lines) behave similarly in response to hypo-osmotic stress (EC=0.12, p-value < 6*e-4), whereas genes containing the motif TAAGGGGA (corresponding to STRE) in their promoters (green lines), behave similarly to each other (EC=0.38, p-value < 1*e-5), yet differently from the first group of genes. This illustrates the strength of MEX in identifying sequence motifs corresponding to known S. cerevisiae regulatory motifs based on promoter sequence alone. The expression data for the analysis was taken from Gasch et al.[18].

2.4 Finding functional clusters of motifs

We have formulated an iterative method for clustering motifs, according to their sequences and EC scores information. We first initiate clusters by gathering motifs that share some building blocks, or `seeds'. Then, a series of iterations improves the clusters, using various procedures detailed below. The clusters refinement steps include the addition and removal of motifs from existing clusters and splitting and merging of clusters. We have quantified the quality of clusters using several criteria associated with sequential patterns and EC score patterns of the motifs. The metrics as well as the refinement steps are listed below.

The clustering algorithm may be used to cluster any type of sequential data that are linked to numerical data. The input to the algorithm is a set of sequences of a given alphabet (e.g. motifs) and a complementary set of vectors (e.g. EC vectors), holding an additional information that needs to be taken into account in the clustering process.

Our clustering method may be considered 'fuzzy' in the sense that single motifs may belong to several clusters. Additionally, not all motifs must be clustered and may be left as singletons.

Initiating clusters by seeds

Our set of motifs was scanned to find short strings of nucleotides (of length 6) that appear within at least three motifs, to be called `seeds'. Selecting all motifs that contain a given seed defines a preliminary cluster.

Pruning clusters to increase EC tightness

For each motif one defines an EC vector of length 40 whose entries specify the p-values of significantly successful EC experiments (that had passed the FDR criterion). Such vectors comprise the matrices in figures 3.3 to 3.5. Let us define the space of all these vectors as EC space and define an *EC divergence* measure for a cluster of motifs as the average distance of all pairs of its EC vectors. In order to decide whether to eliminate a motif from a given cluster, we ask whether its presence increases the divergence of the cluster. To decide whether a motif m should be eliminated from a cluster M_C, we compare the EC divergence of M_C with the empirical distribution of EC divergence scores resulting from replacing m with every one of the motifs that lie outside the cluster M_C (that is, with a background sample M_B). The motif will be pruned from the cluster if it does not significantly reduce the cluster's EC divergence, in comparison to motifs from the random background. The deletion of motifs from a

cluster occurs after all motifs have been tested, thus the order of tested motifs does not affect their chances of remaining in the cluster. A pseudo code describing this procedure is available in box 2.1. An example is shown in Figure 2.4.

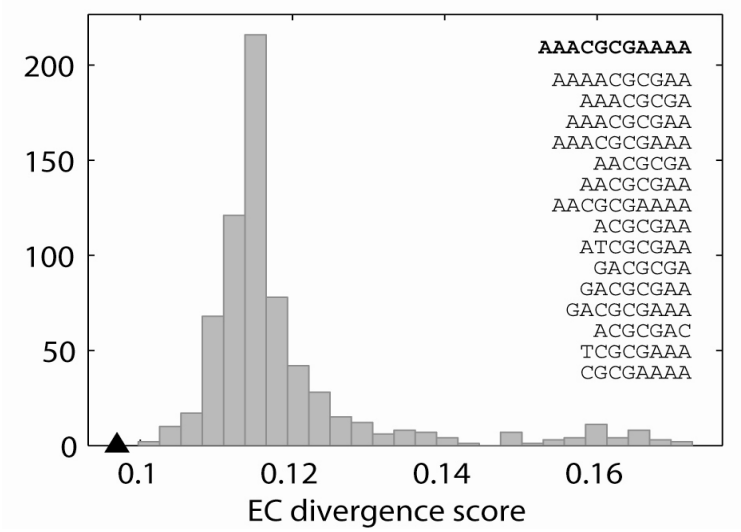


Figure 2.4 An example for testing the contribution of a specific motif to the cluster's tightness. The EC-divergence score of the cluster including the motif AAACGCGAAAA (black triangle) is compared to the empirical distribution of EC-divergence of clusters, in which the motif in question has been replaced with random motifs (histogram). Our null hypothesis claims that the motif does not reduce EC-divergence of the group (which is equivalent to saying that the motif harms the tightness of the cluster). In this example, however, the divergence-score of the cluster with the motif included in it is very small. Hence, we can reject the null hypothesis with a probability value of 0.001 and include the motif in the cluster.

Expanding clusters

We search for new motifs to be added to the cluster without increasing its EC divergence. To decide whether a motif *m* should be added to a cluster M_C we compare the EC divergence resulting from its addition (M_C+*m*) with the empirical EC divergence distribution resulting from additions of each of the motifs lying outside M_C (that is, in a background sample M_B), one at a time.

At the same time we also require sequential similarity of the new motif to the ones that belong to the cluster. The sequential distance between motifs is defined as the edit distance of their best alignment, not allowing gaps. The sequential distance score, *D*, is normalized between 0 and 1, such that *D*=0 if the short motif is fully contained in the long one and *D*=1 if the motifs have no match at all.

A cluster will be expanded by motifs that keep its tightness, as well as being strongly similar to the cluster by sequence. The addition of motifs to a cluster occurs after all motifs in M_B have been tested, thus the order of tested motifs does not affect their chances of being added to the cluster. A pseudo code is available in box 2.2.

Box 2.1: Pruning clusters to increase EC tightness

Given a set M of motifs, their EC scores vectors, $EC_{i\in M} \in \Re^{40}$, and two disjoint subsets, $M_C, M_B \subset M$ (the cluster in question and a background subset of motifs, respectively), we wish to eliminate from cluster M_C motifs that increase its EC divergence. We will test the contribution of M_C's motifs to its EC divergence, by comparing them to M_B's motifs contribution to M_C's EC divergence.

Pseudo code:

1. Calculate the EC distance between every pair of *EC* vectors in $M_C \bigcup M_B$:
$$ECdist_{ij} = avg(|\, EC_i - EC_j \,|)\ ;\ i, j \in M_C \bigcup M_B$$
2. Calculate M_C's divergence score: $DivScore_{M_C} = avg(ECdist_{ij})\ ;\ i < j \in M_C$.
3. Create a new subgroup M_{ij} by replacing the *i*'th motif of M_C with the *j*'th motif of M_B.
4. Calculate M_{ij}'s divergence score.
5. Repeat steps 3, 4 for all $i \in M_C$, $j \in M_B$ in order to examine the effect of each motif *i* on M_C's tightness.
6. For each motif $i \in M_C$, generate the empirical distribution of divergence scores, as found in the replacement of motif *i* with every motif $j \in M_B$.
7. For each motif $i \in M_C$, calculate the p-value of getting the divergence score of M_C by chance.
8. Compare each p-value to a preset significance value α.
9. Eliminate motifs that are not significantly reducing the divergence of the group in comparison to the randomly sampled motifs.

Box 2.2: Expanding clusters

Given a set *M* of motifs, their EC scores vectors, $EC_{i\in M} \in \Re^{40}$ and two disjoint subsets, $M_C, M_B \subset M$, we wish to expand M_C by similar motifs from a background set M_B that do not increase its EC divergence.

Pseudo code:

1. Find candidates motifs for addition, $\mathrm{M_{cand}} \subset \mathrm{M_B}$, that show strong similarity by sequence to at least one motif in M_C.
2. Calculate the EC distance between every pair of *EC* vectors in $M_C \bigcup M_B$:
$$ECdist_{ij} = avg(|\, EC_i - EC_j \,|)\ ;\ i, j \in M_C \bigcup M_B.$$
3. Create a new candidate subgroup $M_{C,cand}$ by adding M_C a single motif from M_{cand}.
4. Calculate $M_{C,cand}$'s divergence score: $DivScore_{M_{C,cand}} = avg(ECdist_{ij})\ ;\ i < j \in M_{C,cand}$.
5. Create a new test group $M_{C,j}$ by adding M_C a single motif from $M_B \neq M_{cand}$
6. Calculate $M_{C,j}$'s divergence score.
7. Repeat steps 5, 6 for all $j \in M_B \neq M_{cand}$ to generate the empirical distribution of divergence scores of the test groups.
8. Calculate the probability value for getting the divergence score of $M_{C,cand}$ by chance.
9. Expand the group by the current motif candidate if it produces a significantly low divergence score (lower than some preset significance value, α).
10. Repeat steps 3-9 for every motif candidate in respect to the original cluster M_C.

Fusion of clusters

Clusters will be merged if they share a minimum percentage of motifs and are also found to be similar in EC. EC distance between two clusters A and B is defined by a *Fisher* criterion, as the distance between the centers of the clusters, divided by the sum of their standard deviations:

$$F_{A,B} = \frac{\|\mu_A - \mu_B\|}{\|\sigma_A\| + \|\sigma_B\|}$$

μ_A and μ_B are the mean EC vectors of the two EC matrices (the center of each cluster). For each cluster we define σ as the vector of the 40 standard deviations corresponding to the 40 EC experiments. Clusters will be merged if their *Fisher distance*, F, is smaller than some threshold, as long as they also obey the sequential similarity criterion.

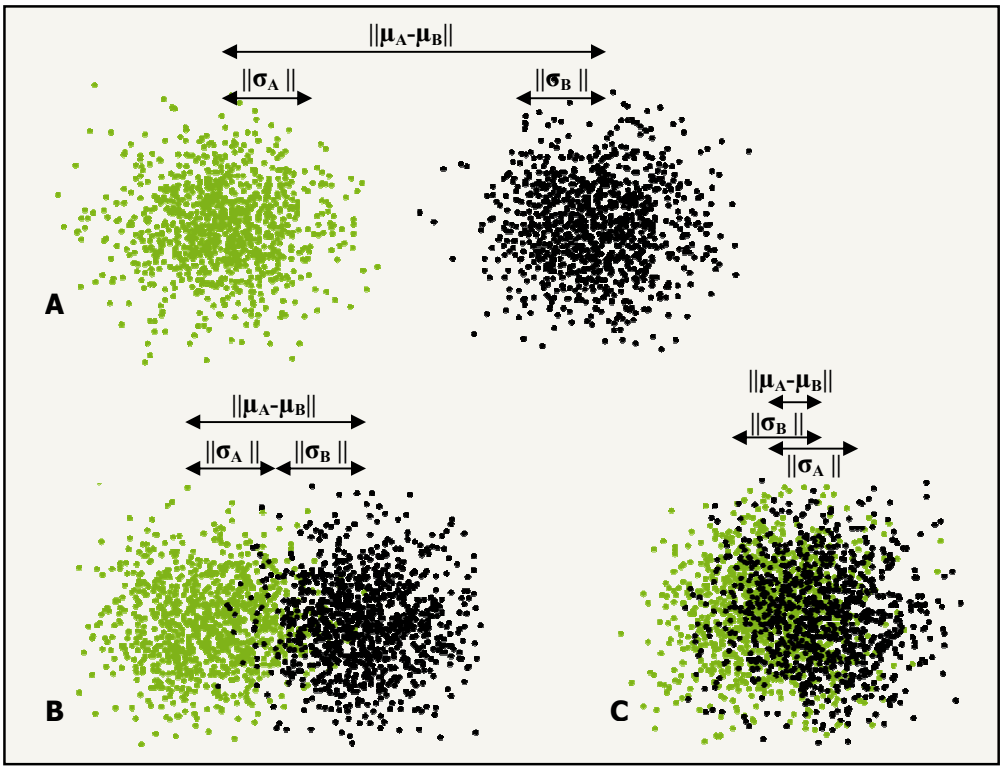


Figure 2.5 A demonstration of the Fisher criterion. The fisher distance for distant clusters exceeds the value 1 (A). The smaller the fisher distance is, the more difficult it gets to distinguish between the clusters (B, C).

Splitting of clusters

Clusters will be split into K smaller clusters if they exceed a given size. Splitting is done using the K-means algorithm on the EC space of the cluster. After applying this indiscriminative step, however, a fusion step is applied, so that unnecessary splitting will be reversed.

Fine refinement of clusters

The former procedures are applied iteratively in a preset order, to generate clusters that are rather tight in EC and in sequence and differ from each other in sizes, EC patterns and motif sequences. In a final pruning step finer parameters are used. Then the improvement of each cluster is tested with respect to a *cluster score*, assessing the quality of the cluster, and the pruning is accepted or rejected accordingly. The clusters are given a *cluster score*, a heuristic function encapsulating the various measures used in the analysis:

$$ClusterScore = \frac{\sqrt{\frac{SeqDivScore_{cluster}}{SeqDivScore_{others}} \cdot \frac{DivScore_{cluster}}{DivScore_{others}}}}{\left(MC \cdot F_{cluster,others}\right)^2}$$

$SeqDivScore_{cluster}$, $DivScore_{cluster}$ are the cluster's sequential and EC divergence scores, respectively (the former is defined similarly to the latter, as the average sequential distance of all pairs of motifs within the cluster). $SeqDivScore_{others}$, $DivScore_{others}$ are the sequential and EC divergence scores of all the motifs outside the cluster, respectively. $F_{cluster,others}$ is the EC fisher distance between the cluster and the rest of the motifs, and MC is the number of motifs within the cluster. The smaller the cluster score is, the better the quality of the cluster is considered.

The cluster score quantifies the quality of a cluster in terms of its internal tightness relatively to the background. As affected by many different factors, the cluster score is sensitive to noise. Hence it is only used at a late stage along the algorithm, when clusters are already coherent to a great extent.

Flow of the algorithm

After initiation, cycles of the various iterations occur, gradually improving the clusters with respect to their sequences and EC patterns. The algorithm stops when the rate of change of the clusters falls below a certain cutoff (a stopping criterion) or if no clusters are found. Clusters that are too small (below a preset threshold) are disregarded.

2.5 Finding GO annotations of clusters

Co-regulated genes might be involved in similar cellular processes and functions. Information regarding the functional tendencies of the genes on the promoters of which the cluster's motifs are found may be helpful in getting a notion about the identity of clusters. Hence, we have used GO TermFinder [8] in order to test the GO enrichment [2] of the sets of genes that are relevant to the our clusters.

GO (gene ontology) is a project aimed to provide a common language for describing aspects of a gene product's biology. GO provides annotations for genes in three categories: the *molecular functions* of the gene product (e.g. transporter activity, kinase activity, transcription factor, etc.), the *biological process* as part of which the gene product acts (e.g. mitosis or protein metabolism) and the *cellular component* in which it acts (e.g. sub-cellular structures, locations, and macromolecular complexes).

GO TermFinder looks for significant enrichments of GO terms that are used to describe a given set of genes. This tool gives an insight on what is common among the genes within a set, in terms of GO annotations. GO TermFinder calculates p-values, using a hyper-geometric distribution, as the probability of x or more out of n genes having a given annotation, given that K of N have that annotation in the genome in general. A corrected p-value cutoff is calculated to account for multiple hypotheses.

2.6 TF binding rates

In order to further validate the identity of clusters with respect to known TFs, we have performed a comprehensive estimation of the binding of various *S. cerevisiae* TFs to the promoters on which our motifs are found. For that purpose we have employed yeast genome-wide location analysis data [19], in which the genomic occupancy of 203 DNA-binding transcriptional regulators had been measured *in vivo* via ChIP-on-chip experiments at various environmental conditions. We have calculated the binding rates, i.e. the percentage of promoters within each cluster that are bound by each transcription factor. Since some transcription factors are less specific, and typically bind more genes than other factors, we define *incremental binding rates* by subtracting the mean binding rate of each TF from the binding rates of each TF to every cluster. For this analysis we have used a p-value cutoff smaller or equal to 0.005 at the original TF binding data, to decide whether a TF binds a given promoter.

The significance of the observed binding rates was tested using a hyper-geometric distribution and probability of getting at least the observed number of bound promoters (for each cluster and a given TF) by chance was estimated. P-values were calculated as the probability that x or more, out of n promoters, are observed to be bound by a particular TF, given that K of N promoters are bound by that TF in general.

Figure 3.2 displays the incremental binding rates of each of the 203 tested TFs to our clusters of motifs and the results of the hyper-geometric tests.

RESULTS

2.7 Extraction of motifs from promoters of *S. cerevisiae*

We have applied MEX to 4800 promoters of 6300 genes (as some promoters fall within the intergenic regions of couples of genes). We have chosen rather permissive parameters for MEX in order to attain many putative motifs that will be further screened based on their regulatory activities. Using the parameters α=0.1 and η=0.99 MEX has extracted 9370 motifs. Considering the occurrences of motifs on both strands of the DNA as identical objects, we identify motifs with their reverse complements. Hence, the set has been reduced to 8498 unique putative motifs.

2.8 Testing Expression Coherence

We have calculated EC scores and their p-values for each of our 8498 putative motifs in 40 experiments where whole-genome mRNA expression of *S. cerevisiae* had been monitored using DNA chips[*]. Setting the false discovery rate to 0.1, we have discovered that 25% of the sequence motifs have a significant EC score in at least one of the experiments. This should be contrasted with a 0.6% success rate, under the same FDR condition, for random sequences of lengths between 7 and 11 nucleotides. In other words, MEX does a good job of selecting motifs that are relevant to the problem at hand.

In order to lower the chances for false positives, we have applied further screening to our motifs, requiring each one to exhibit at least one EC success with a p-value of 0.001 or lower. This distilled set contains 694 significant regulatory motifs. Almost half of these motifs match perfectly (or are included in) known binding sites of 85 transcription factors (motifs published by Harbison et al. and Pritsker et al. [19, 35]).

2.9 Clustering motifs

Our algorithm finds 20 clusters, covering a total of 182 motifs. 14 of our clusters have large overlaps with known motifs. Figure 3.1 displays the Fisher distance matrix of these 14 clusters. On the diagonal (where F=0) we have added F-values that are obtained by randomly dividing each of the given clusters into two arbitrary ones, in order to provide some examples when F values are too low to serve as a criterion for separation among clusters. We clearly obtain groups of related clusters, and we will study and name them accordingly. In the following we will discuss in detail 8 of our clusters.

[*] EC analysis has been performed by Prof. Yitzchak Pilpel and his student Michal Lapidot.

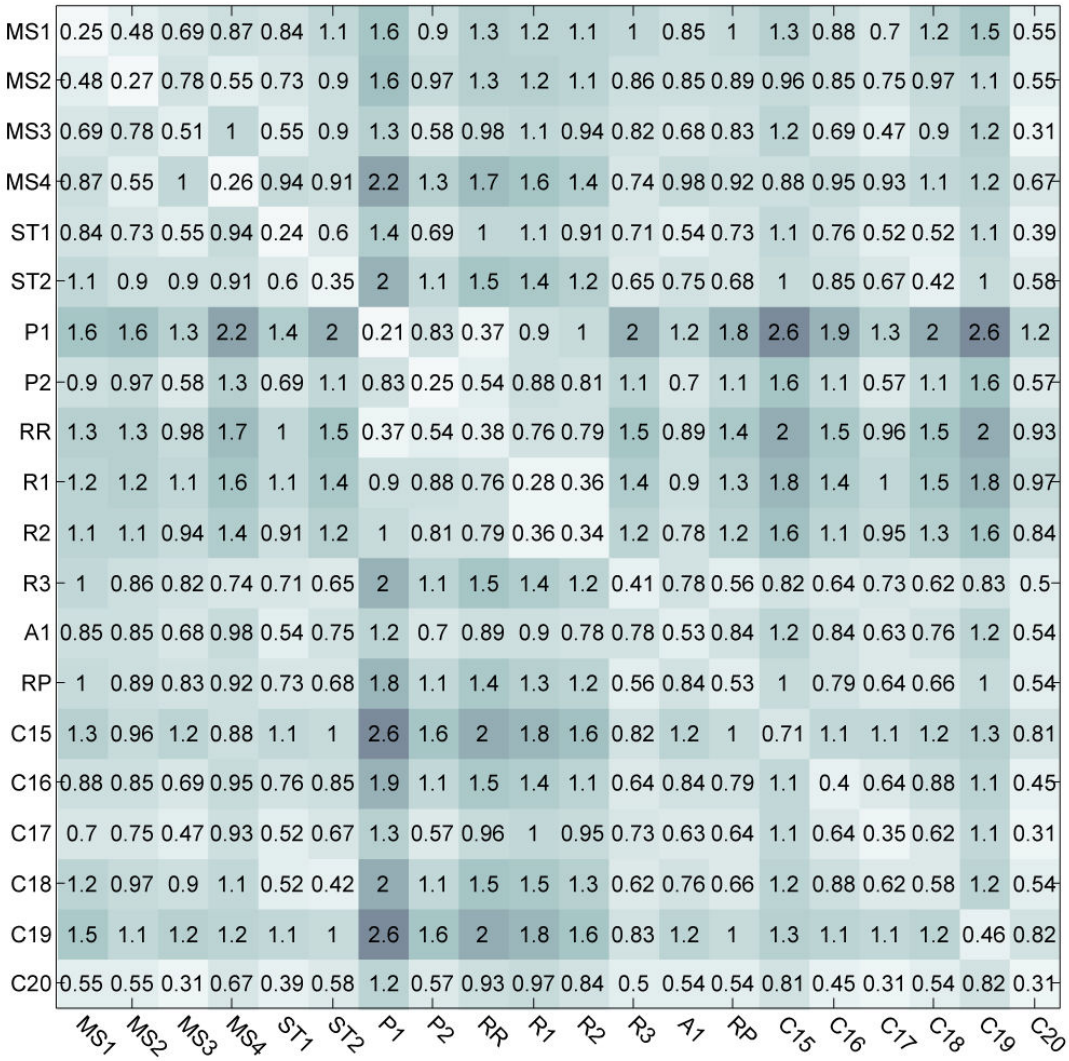


Figure 0.1 Fisher distances between our final clusters. On the diagonal (where F=0) we have added the mean F-values obtained by randomly dividing each of the clusters into two arbitrary ones (mean over 1000 random divisions for each cluster). The values along the diagonal are anti-correlated to the sizes of the clusters (with correlation coefficient of -0.85). These values give a notion for cases when the F-values are too small to serve as a criterion for separation between clusters. It appears that most clusters' EC patterns are rather distant from each other (F≈1). Yet, we clearly obtain groups of related clusters (e.g MS1-MS4).

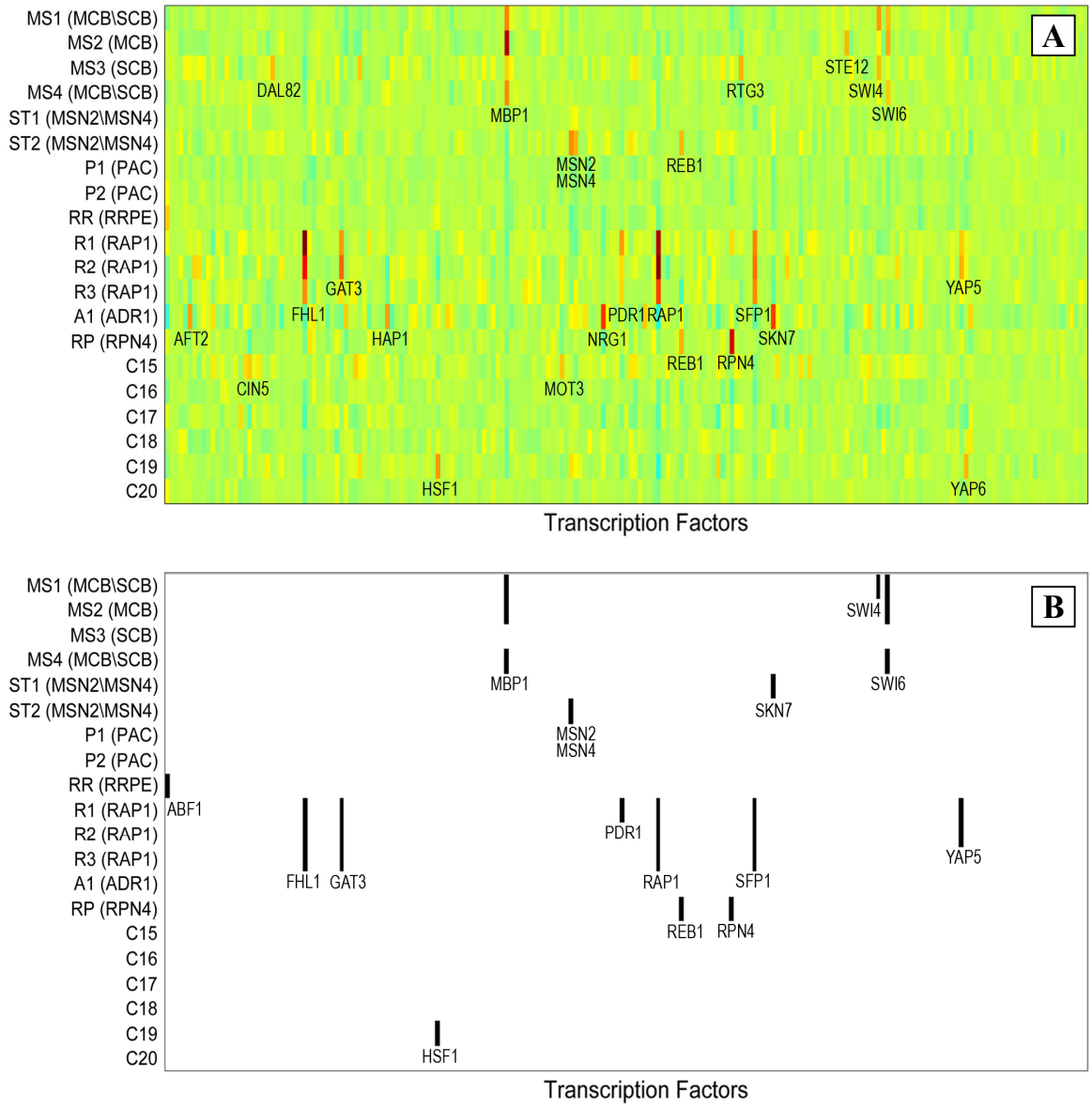


Figure 0.2 Binding of transcription factors to promoters that carry our clusters' motifs. A. Incremental binding rates for each of the 203 transcription factors (columns) to every cluster (rows). Hot colors (dark red) represent high Incremental binding rates. B. The results of hyper-geometric tests, calculating the probability of getting at least the observed number of bound promoters (for each cluster and every given TF) by chance, with respect to a given TF. Black indicates that the test's p-value equals 0.001 at most.

The first four clusters (MS1-MS4) have large overlaps with well-known TFBS, such as those bound by MCB (MluI cell cycle box) and SCB (SWI4-SWI6 cell cycle box). The first is a well known complex, formed by the proteins MBP1 and SWI6, while the latter consists of SWI4 and SWI6. This reassures the identity of clusters MS1-MS4, as the highest incremental binding rates attained for these clusters are of MBP1, SWI4 and SWI6. A similar validation arises for other clusters as well. Note that the TFs which bind the sites known as PAC have not yet been discovered, as is also reflected by the lack of signal for the clusters P1 and P2.

2.10 MCB/SCB clusters

The first four clusters shown in Figure 3.2 have large sequential overlaps with well known TFBS, such as the MCB (MluI cell cycle box) and SCB (SWI4-SWI6 cell cycle box) clusters. MBF (MluI cell cycle box binding factor) and SBF (SWI4-SWI6 cell cycle box binding factor) are two related protein complexes involved in transcriptional regulation of the transition from the G1 to S phase of the cell cycle. The two DNA binding complexes are heterodimeric and contain the regulatory protein SWI6 as a subcomponent. MBF contains the DNA binding protein MBP1, to which SWI6 is bound, while the DNA binding subunit in SBF is SWI4. MBF and SBF play important roles in the regulation of many processes, such as DNA synthesis, DNA repair and budding [1, 31, 34].

We have found four clusters associated with MCB and SCB known motifs. These clusters and their EC patterns are provided in Figure 3.3. The identity of our four clusters was further validated in two manners. First, we have tested the GO enrichment of the set of genes on the promoters of which the cluster's motifs are found. Indeed, the four clusters are found to be significantly enriched with GO terms such as DNA metabolism, DNA repair and response to various types of stress. This analysis provides some information regarding the functional tendencies of the four clusters. It does not, however, provide a high enough resolution for discriminating between them, in terms of specific cellular processes and functions of the genes associated with those clusters.

A second analysis has estimated the incremental rate of binding of transcription factors to the set of promoters of each cluster (See Methods and figure 3.2). With agreement to the results of the previous analysis, it appears that the four clusters at hand show a significantly high incremental binding rate to MBP1, SWI4 and SWI6.

Combining the information of known motifs, GO annotation enrichment and the binding of transcription factors to the genome, we have concluded the following: The first cluster, MS1, contains "classic" MCB and SCB elements bound by MBP1, SWI4 and SWI6. The cluster is very significant in experiments testing the cell cycle and various environmental stresses. The motifs of cluster MS2 are identified as MCB elements, while those of MS3 are identified as SCB motifs. It appears that MS2 is particularly important in cell cycle experiments, whereas MS3 is significant in stress related experiments and not as much in cell cycle ones. Cluster MS4, whose motifs

are functional at cell cycle experiments, is identified mostly as MCB, though some of its motifs fit SCB as well.

The EC patterns of the four clusters show clear differences (Figure 3.3). The latter can be correlated with the detailed nucleic acid decomposition of their motifs. Motifs of MS1 and MS2 have different common cores, ACGCGA and ACGCGT respectively. Hence, the single adenine to thymine substitution in the core of these motifs may be responsible for the relevance of MS1 to a particular heat shock experiment (Figure 3.3) and for leading MS2 in its effect on the menadione and hydrogen peroxide experiments.

The MS3 cluster displays a core of TCGCGA, differing from MS1 at another position within the motif cores. Here again it appears that the particular sequence to which a transcription factor is bound plays an important role in the regulation of gene expression. In particular, note the absence of significance of the MS3's motifs in most cell cycle experiments and their relative importance in the heat-shock ones.

MS4 displays a complementary behavior to MS3, relevant only to cell cycle experiments. Most of its motifs have a core of ACGCCA. Thus we show that the avidity of clusters, and the TFBS that they contain, is strongly dependent on particular details of their motifs.

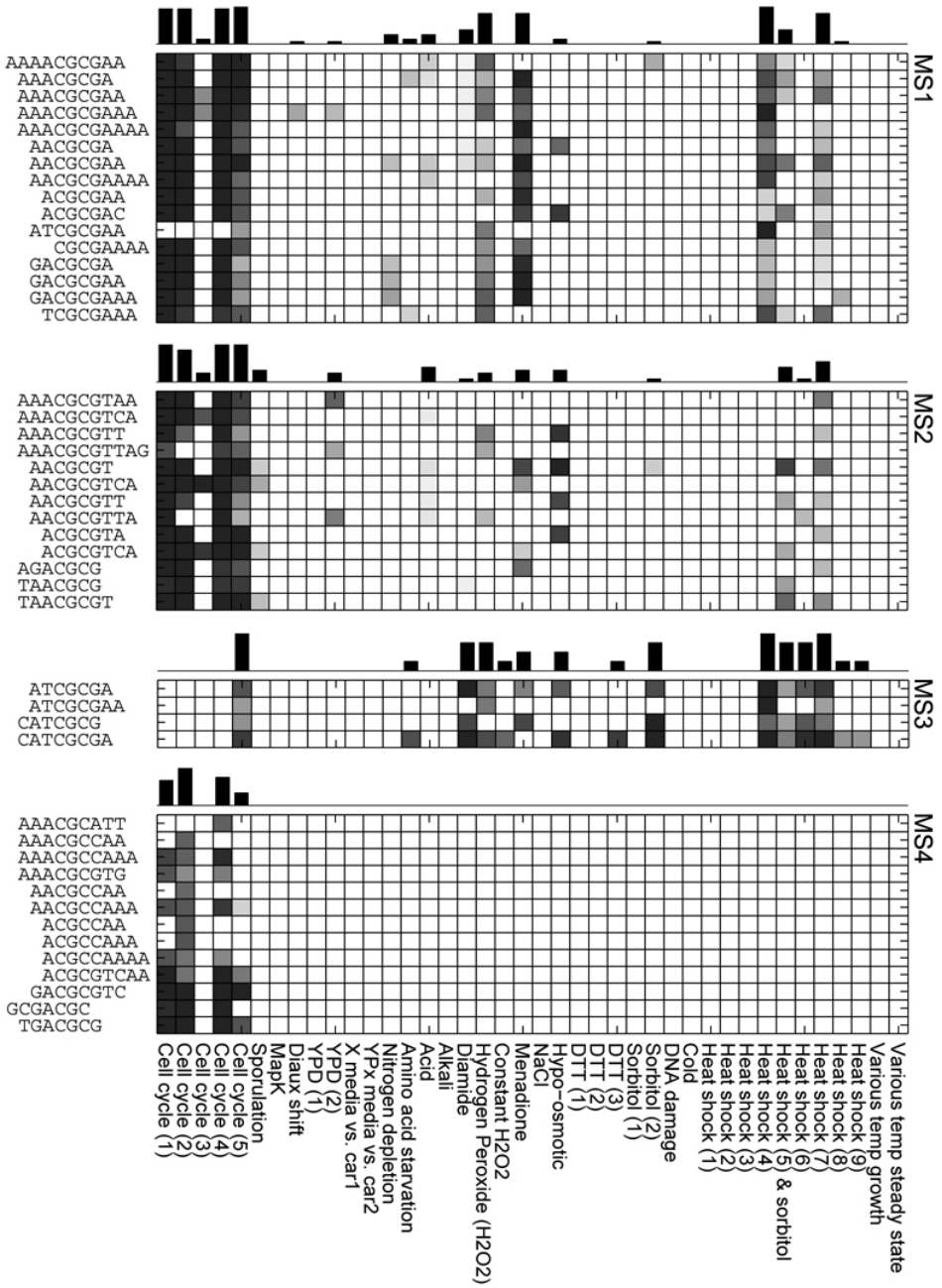


Figure 0.3 Four of our clusters contain motifs that are known MCB and SCB elements (Top to bottom: MS1, MS2, MS3 and MS4). Each matrix represents the EC patterns of the motifs within one cluster. The EC pattern of a motif is a vector of 40 p-values of EC tests for 40 environmental experiments (low p-values are represented by dark colors, with a grayscale proportional to -log(p-values), white implies FDR>0.1). The bars indicate the percentage of motifs that had significant success in each experiment.

2.11 STRE clusters

Another demonstration of the importance of specific sequences of TFBS can be seen in two clusters that have been identified as STRE (Stress Response Elements). STRE are known to be bound by two related transcription factors, MSN2p and MSN4p. These two Cys2His2 zinc finger proteins are known to take part in regulating the expression of many stress related genes [29].

The first cluster associated with STRE (ST1) has high overlap with well known binding sites of MSN2p and MSN4p. The sequences composing the second cluster (ST2) show sequential similarity to the known binding sites of MSN2p and MSN4p though have not been identified as STRE in previous studies.

The genes belonging to the promoters on which the two clusters are found are highly enriched with GO annotations such as response to stress, energy reserve metabolism, sporulation and more. This agrees with the fact that MSN2p and MSN4p regulate the expression of stress related genes.

It appears (Figure 3.2) that while ST2 shows high incremental binding rates to MSN2p and to MSN4p, the well known STRE sequences of ST1 show lower binding rates to these TFs. Note, though, that the incremental binding rates of ST1 to all the other tested TFs are even lower.

Furthermore, as can be seen in Figure 3.2, the incremental binding rate of ST1 to MSN4p is higher than its incremental binding rate to MSN2p, whereas in cluster ST2 the opposite is the case.

As expected, the EC pattern of ST1 is especially rich for stress related conditions (Figure 3.4). Although similar in tendency to ST1, the EC pattern of ST2 is not as strong as that of ST1.

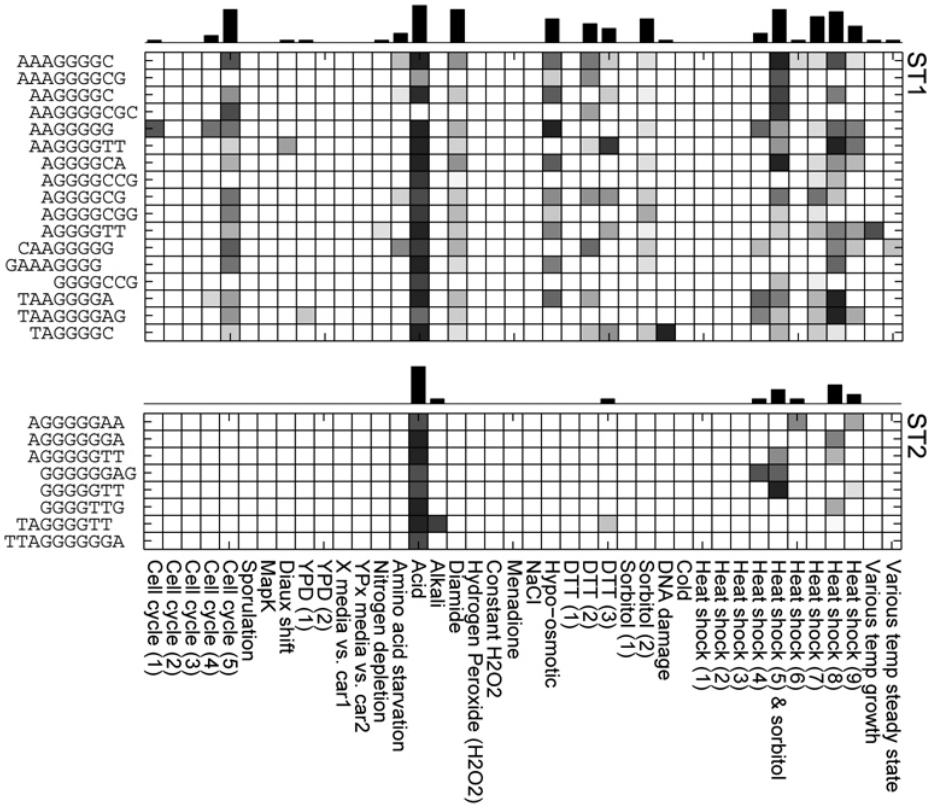


Figure 0.4 Matrices of EC patterns for the two clusters ST1 and ST2. These clusters contain motifs that are identified as STRE, to which MSN2p and MSN4p bind, regulating the expression of stress related genes.

2.12 PAC clusters

A third group of clusters contains P1 and P2. P1 has a large overlap with Polymerase A and C (PAC) motifs. The EC pattern of P1 (Figure 3.5) is extremely rich in significance for a vast majority of the experiments. This agrees with the fact that PAC regulates many ribosomal genes, thus affecting numerous cellular processes [12]. Regulation of ribosomal biogenesis is of major importance to the cell; e.g. more than 50% of the growing cell's total transcription is devoted to the biogenesis of the ribosome [30, 45].

The identity of cluster P1 was further validated through the GO annotations analysis of the relevant genes, pointing mainly to the biogenesis of the ribosome. The genes associated with P1 have not been found to be significantly bound by any of the 203 transcription factors tested by Harbison et al (Figure 3.2). This is not surprising, however, since the transcription factor binding PAC motifs is unknown.

The motifs of the second cluster, P2, show some similarity to known PAC motifs, though some of them have not been previously identified as such. Here as well, we find the relevant genes to be significantly enriched with GO annotations associated with the biosynthesis of the ribosome. Similarly to P1, no transcription factor was found to bind the motifs of P2, and many of the EC patterns are highly significant for many experiments, as in P1. Although the EC patterns of P1 and P2 are similar in their tendencies, they are different in potency (Figure 3.5).

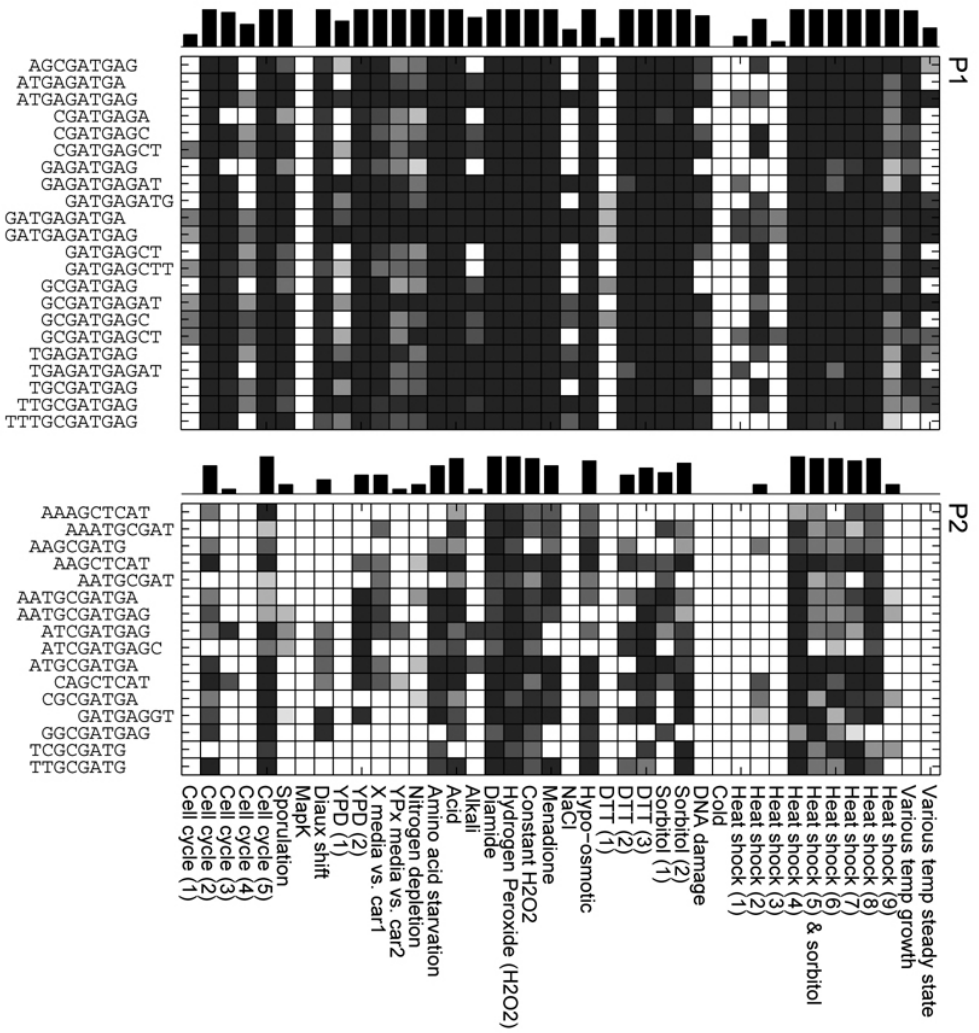


Figure 0.5 Matrices of EC patterns for clusters P1 and P2. The upper cluster (P1) contains known PAC motifs, while most of the motifs of the lower cluster (P2) have not yet been described. The EC patterns of the two clusters are significantly rich. This agrees with the fact that PAC regulates many ribosomal genes, hence affect numerous cellular processes.

2.13 Other clusters

Apart from the three groups of clusters discussed above, twelve more clusters of motifs have been found via our analysis. These clusters' identities, EC patterns and other analyses are described in appendix B. Among these clusters, six have large sequential overlaps with well known TFBS. Some of the motifs composing cluster RR, for example, are known as RRPE (Ribosomal RNA processing elements), while the motifs of clusters R1, R2 and R3 show high sequential similarity to known RAP1 (Repressor activator protein 1) binding sites, as well as to FHL1 known motifs (R1), to SFP1 motifs (R2) and to AFT1 (R3) [19, 35]. The EC patterns of clusters RR, R1 and R2 are extremely rich in significant experiments and show resemblance to that of P1. This makes sense, given that PAC, RRPE, RAP1, SFP1 and FHL1 are known to be involved in the regulation of ribosomal genes, hence affect many biological processes [12, 21, 27, 28, 44, 17].

The RR cluster shows significant GO enrichments at processes such as the biogenesis, assembly and maintenance of the ribosome, transport from the nucleus, tRNA metabolism, etc. The R1-R3 clusters are found to be enriched with GO processes such as ribosome biogenesis and assembly, chromosome organization and biogenesis, telomere organization and biogenesis, histone modification and others, in agreement to the sequential and EC information.

The only significant binding rate of cluster RR is to ABF1, a chromatin reorganizer transcription factor. This is interesting as the TF that binds RRPE has not yet been described. Furthermore, the RRPE motifs are not included in the set of motifs published by Harbison.

The binding rates of clusters R1-R3 are particularly high for RAP1 as well as for FHL1 and SFP1. In addition, the incremental binding rates of clusters R1 and R2 are high for GAT3 and YAP5, which are assumed to be involved in cell cycle progression and stress related regulation. This is not unexpected, as ribosome biogenesis is known to be tightly coupled to cell cycle progression as well as to environmental changes that affect growth rate [21].

Interestingly, in the case of the ribosomal clusters, P1, P2, RR, R1, R2 and R3, about 40% of the genes that contain RRPE on their promoters (RR) and 20% of the promoters containing RAP1\SFP1\FHL1 (R1-R3) also have PAC motifs (P1, P2). This implies that the regulation of ribosomal genes is complex and involves the

cooperation of several TFs. In fact, it has been recently shown *in vivo* that regulation by RRPE might be SFP1 dependent [17]. It has been further suggested that PAC motifs serve as a repressing element of ribosomal genes, while RRPE motifs activate transcription of those genes in a SFP1 dependent manner.

Although genes coding for ribosomal proteins represent only 2% of yeast genes, they contain about one third of all *S. cerevisiae*'s annotated introns [11, 25, 40]. In particular, a major class of intron containing genes is that encoding ribosomal proteins, tRNA, translation factors and factors involved in ribosome biogenesis. We have tested the enrichment of intron presence on genes on the promoter of which the clusters' motifs are found, in comparison to a background random model (of random groups of motifs of the same sizes as those of the clusters in question). It appears that the three RAP1 clusters, R1-R3, are significantly enriched with genes that contain introns, with a p-value smaller than 0.001 in the case of R1 and R2, and with a p-value smaller than 0.01 at cluster R3. In contrast, clusters P1, P2 and RR have not shown a significant over representation of intronic genes, nor have any of our other clusters. Moreover, clusters P1, P2, C16, C17 and C19 have shown a significant under representation of intronic genes.

The motifs of two more clusters have been identified as known TFBS. Cluster A1's motifs are similar to ADR1 and STRE motifs, while those of cluster RP have been identified as RPN4 motifs. Six other clusters, C15-C20, have not been identified as known clusters of motifs (Appendix B) [19, 35]. In the following analyses we will mainly discuss the first three groups of clusters introduced in the previous sections.

2.14 Mechanisms determining strength of regulation

Genes that are regulated by the same transcription factor are often found to display various levels of expression. This is biologically motivated by the need to provide a wide range of behavior, allowing sub-groups of genes to be regulated in different manners.

Variability of regulation may arise through four major causes: (1) specific TFBS binding mechanism, (2) different numbers of TFBS occurrences on the promoters, (3) specific localizations of the TFBS along the promoter [36] and (4) interactions between different transcription factors [42, 4]. A combination of these causes may control the high variability in gene expression as well as act as a fine tuner of gene regulation.

We expect the last cause to be of secondary importance in our analysis of clusters, since there exist only small overlaps between genes that carry motifs of two different clusters (Appendix C), or between every cluster and each of our single motifs. Furthermore, as the EC analysis was conducted one motif at a time, we do not expect such effects to be visible through our EC results.

We have further analyzed the first three possible causes for the clusters at hand, to decide which is relevant to the different regulation effects that we have seen in figures 3.3, 3.4 and 3.5.

For each motif within a cluster, we have tested the distribution of its appearances on the promoters. This was compared with the distribution of randomly sampled motifs. The random background model was based on all 694 motifs. It appears that the distributions of number of appearances of the clusters' motifs on the promoters have not been found significantly different from the background model for all clusters. Furthermore, no significant differences have been detected in the number of appearances on motifs between the different clusters.

A second analysis tested the localization of motifs of each cluster along the promoters. In Figure 3.6 we provide histograms of motif distances from the translation start site. This was compared with the localizations of randomly sampled groups of motifs (of the same sizes as those of the clusters in question).

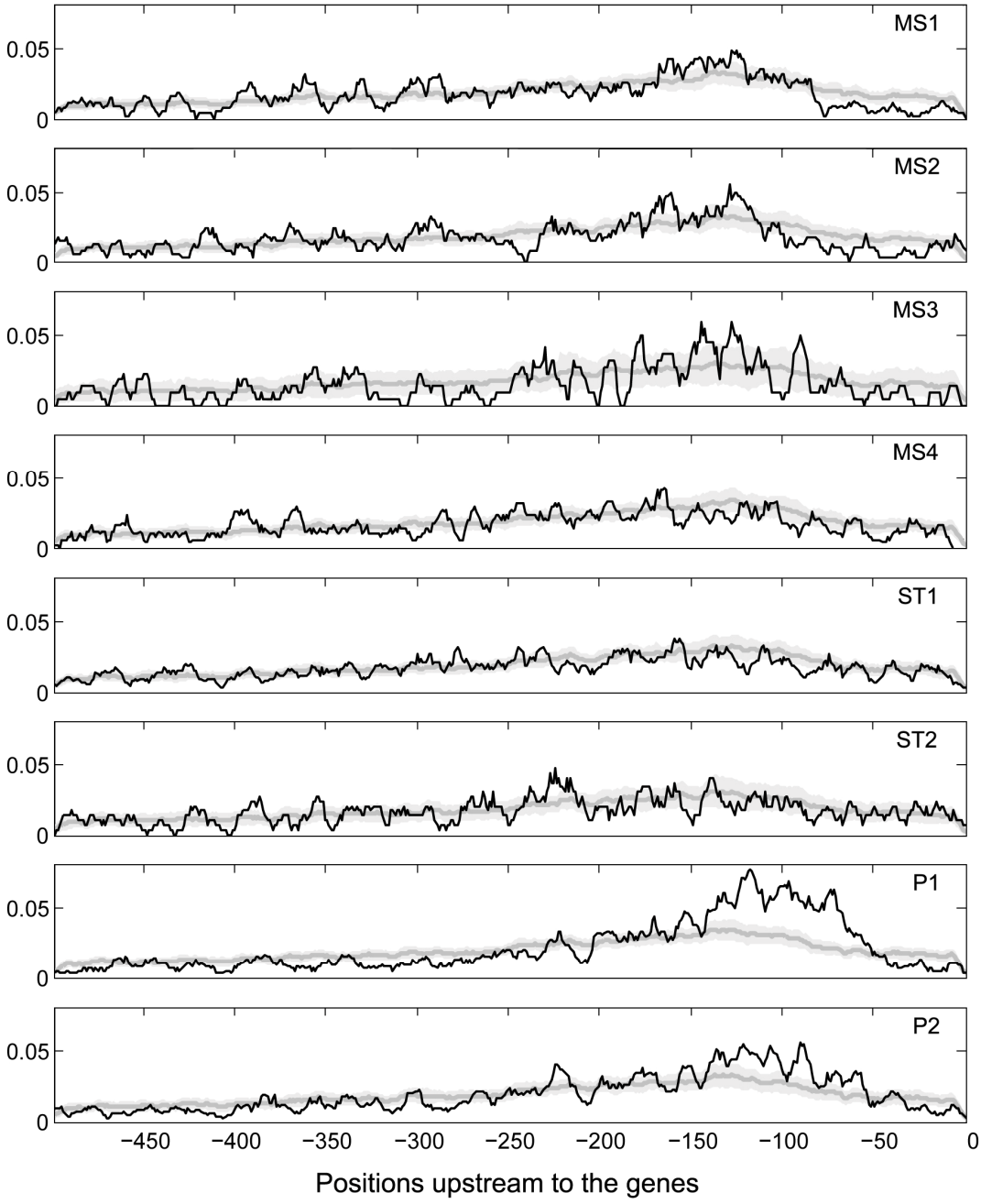


Figure 0.6 Localization of motifs on the promoters of several clusters. The black lines indicate, for each position upstream to the genes (up to -500bp), the percentage of promoters on which the cluster's sequences have been found. This can be compared to the localization of randomly sampled groups of motifs (of the same sizes as those of the clusters in question). For each cluster, the dark gray line shows the mean motif occurrence per position over 1000 such randomly sampled groups, while the light gray area represents the samples' standard deviation of occurrences per position.

The motifs of clusters MS1 and MS2, for example, have a similar number of appearances per promoter. Furthermore, the number of appearances of the motifs of these clusters on the promoters is distributed in a similar manner to that of a background model. In addition, the analysis of motifs' localization, described at Figure 3.6, does not provide any distinction between these two clusters. Hence, we infer that the difference in their functional behavior in some of the tested experiments (Figure 3.3) is caused by stronger binding mechanisms of the motifs in MS1.

Changes in magnitude of the binding mechanism may result from a specific binding affinity of the TF to the TFBS, causing varying preferences of the TF to various TFBS or affecting possible competitions between more than one TF over similar TFBS. Alternatively, such effects can result from conformational changes of the TF while bound to a specific TFBS [24, 26]. Conformational changes may also affect the recruitment of cofactors, thus alter regulation.

The same holds also for comparisons of MS1 with MS3 and MS4. In all these cases the changes in regulation strength seem to be caused by variations in the binding mechanisms of TFs to the relevant TFBS. Thus we conclude that in the case of these four clusters, changing a single nucleotide in a TFBS have a strong impact on the binding mechanism of the TF to the promoter.

A similar trend is observed at the STRE clusters. Once again, their differences are caused neither due to different numbers of copies of motifs on promoters nor due to specific localizations along these promoters (Figure 3.6). Hence we conclude once again that the small changes in nucleotide compositions of the relevant motifs lead to differences in binding mechanisms of the TFBS.

Clusters P1 and P2 tell a different story. As in the previous examples, the motifs of the two clusters appear with similar rates on the promoters. However, in the case of P1, motifs strongly tend to occupy the region between -60bp to -150bp upstream to the genes. This tendency is significantly different from the background model, with a p-value smaller than 0.001. Thus, at the PAC clusters the whereabouts of the motifs along the promoters have strong effects on regulation.

An apparently similar phenomenon is observed at cluster RR, whose motifs show a significant tendency to occupy the region between -80bp and -190bp upstream to the genes (Figure 6.2). As in cluster P1, the EC pattern of RR is extremely rich with significant experiments. Cluster RR, however, is the only cluster whose motifs have been identified as RRPE. Hence, RR’s potency cannot be compared to other clusters, in terms of the mechanisms affecting the strength of its regulation.

3 DISCUSSION

3.1 Motif clustering

We have clustered motifs according to both their sequences and their regulatory semantics, as reflected in the motifs' EC patterns. Such grouping reveals biological insights that are easily missed by conservative clustering methods, which rely either on sequence or on numerical data alone.

The resulting clusters of regulatory motifs and their relationships to known TFs have been analyzed in various manners. In several cases we have obtained few clusters of motifs that contain elements of several known TFBS groups. Examples are clusters MS1-MS4 that contain motifs traditionally labeled as MCB and SCB (bound by the protein complexes MBF and SBF correspondingly). Our clustering does not necessarily follow conventional labeling; e.g. all MCB motifs belong to one PSSM in Harbison et al [19], whereas they are scattered among all of our clusters MS1-MS4.

3.2 Mechanisms determining strength of regulation

Differences in EC patterns imply different regulation strengths associated with the relevant motifs in various sets of experiments. Regulation strength may depend on various mechanisms. We have looked at repetition rates and loci of motifs on promoters to decide whether any of them should carry the burden for higher or lower regulation strength, or whether it is the binding mechanism of the TF to the motif that does it.

In both the MCB/SCB and STRE clusters we have concluded that the latter is the case. Different binding mechanisms may occur due to specific TF-TFBS binding affinity or conformational changes of the TF while bound to a specific TFBS, but may also come about because of the existence of different TFs competing for similar TFBS. Comparing Figure 3.2 to Figure 3.3 one can reach very interesting tentative conclusions: MS4 has very weak or no binding to SWI4, and this may be the reason why no effect is observed in all stress experiments. MS3 has weak binding to MBP1 and this may be the reason for the absence of effects on four of the cell cycle experiments.

In the case of our STRE clusters, their differences in regulation may result from the different tendencies of the clusters' motifs to be bound by MSN2p and MSN4p. As can be seen in Figure 3.2, the preferred binding factor of ST1 is MSN4p, while in the

case of ST2 MSN2p binds the cluster's motifs with a higher incremental binding rate than MSN4p.

3.3 Variations in regulatory motifs lead to high magnitude effects on regulation

Both the MCB/SCB and the STRE clusters demonstrate that small variations in the regulatory motifs lead to high magnitude effects on regulation. It has been shown that even a single nucleotide substitution at the motifs of those clusters is sufficient for such effects. This has been demonstrated between clusters, and can also be seen within clusters. At the latter case, variation of motifs may act as a fine tuner of regulation.

Our PAC clusters P1 and P2 show a different behavior. P1 shows higher EC significance and also has an enhanced spatial distribution within a specific range along the promoters. The latter may perhaps be correlated with the loci of nucleosomes on the DNA, affecting the strength of the regulation [36]. We presume that in this case this is one of the reasons for the much higher regulation strength of P1 motifs.

3.4 Motif representations

The conventional representations of motifs by Position Specific Scoring Matrices (PSSM) or via consensus sequences encapsulate the sequential information of a group of aligned motifs. The simplicity of such representations involves the loss of information and leads to possibly wrong conclusions. Mononucleotide frequency weight matrices cannot depict accurately the binding site specificities of their included motifs [10]. Even though some positions show distinct preferences to certain nucleotides, such preferences may depend on the nucleotides occupying other positions. Inter-dependencies between positions within the binding sites may affect the binding of the TF to the DNA, hence the regulation.

Consensus and PSSM representations are inherent in many motif extraction algorithms [23, 20, 9, 19, 43, 32]. Even though these methods seem to capture a large share of the transcription factor binding sites, the predictions of such methods depend on the way they represent motifs and inherit the assumptions and faults of those representations.

Here we have started out with single motifs, as extracted by MEX from sequence data, and filtered by the EC analysis. MEX does not use motif representations such as PSSM or consensus sequences in its search for motifs. This allows us to analyze each

sequence independently, and only then generate clusters of motifs, gaining a better understanding of the regulation without reducing the sequence information. As a result, inter-dependencies within the sequences are not lost. Furthermore, we have left our clusters in the form of groups of motifs, rather than combining them into PSSM representations, as we have learned from our analysis that single changes of a nucleotide in a motif can go a long way in affecting the biological behavior.

Another problem is that a PSSM is built using a finite set of known sequences, which may be incomplete and biased, hence resulting in biased predictions. For instance, most of the P2 motifs have not been mentioned in the literature, presumably because the effects of P1's TFBS overshadow them. This demonstrates that one needs a discriminating analysis to distinguish the P2 motifs from their stronger P1 relatives. MEX tests the significance of each motif in an independent manner, and is not limited by statistical considerations such as over-expression or over-representation within a given class of genes or a given class of motifs. Hence MEX may uncover TFBS, such as those of P2 that have been overlooked by other methods.

4 APPENDIX A – EC EXPERIMENTS

Experiment short name	Experiment name[1]
Cell cycle (1)	ExpressDB Cho - cell cycle
Cell cycle (2)	ExpressDB Spellman - cell-cycle alpha
Cell cycle (3)	ExpressDB Spellman - cell-cycle cdc15
Cell cycle (4)	ExpressDB Spellman - cell-cycle cdc28
Cell cycle (5)	ExpressDB Spellman - cell-cycle eluteration
Sporulation	ExpressDB Chu - sporulation
MapK	ExpressDB - MapK
Diaux shift	ExpressDB Gasch environmental response - diaux shift
YPD (1)	ExpressDB Gasch environmental response - YPD1
YPD (2)	ExpressDB Gasch environmental response - YPD2
X media vs. car1	ExpressDB Gasch environmental response - x media vrs car1
YPx media vs. car2	ExpressDB Gasch environmental response - YPx media vrs car2
Nitrogen depletion	ExpressDB Gasch environmental response - Nitrogen Deplation
Amino acid starvation	ExpressDB Gasch environmental response - Amino Acid starv
Acid	ExpressDB Environmental response - Acid
Alkali	ExpressDB Environmental response - Alkali
Diamide	ExpressDB Gasch environmental response - diamide
Hydrogen Peroxide (H2O2)	ExpressDB Environmental response - Peroxide
Constant H2O2	ExpressDB Gasch environmental response - constatnt h2o2
Menadione	ExpressDB Gasch environmental response - Menadione
NaCl	ExpressDB Environmental response - NaCl
Hypo-osmotic	ExpressDB Gasch environmental response - Hypo-osmotic
DTT (1)	Eisen - dtt
DTT (2)	ExpressDB Gasch environmental response - DTT1
DTT (3)	ExpressDB Gasch environmental response - DTT2
Sorbitol (1)	ExpressDB Environmental response - Sorbitol
Sorbitol (2)	ExpressDB Gasch environmental response - sorbitol
DNA damage	Jelinsky - DNA Damage
Cold	Eisen - cold
Heat shock (1)	ExpressDB Environmental response - Heat
Heat shock (2)	Eisen - heat
Heat shock (3)	ExpressDB Gasch environmental response - 37-25 shock
Heat shock (4)	ExpressDB Gasch environmental response - Heat Shock 1
Heat shock (5) & sorbitol	ExpressDB Gasch environmental response - hs 29-33 1m sorbitol
Heat shock (6)	ExpressDB Gasch environmental response - hs 29-33
Heat shock (7)	ExpressDB Gasch environmental response - hs 29-33 No sorbitol
Heat shock (8)	ExpressDB Gasch environmental response - Heat Shock2 (3 time zero)
Heat shock (9)	ExpressDB Gasch environmental response - hs various temp to 37c
Various temp growth	ExpressDB Gasch environmental response - various temp growth
Various temp steady state	ExpressDB Gasch environmental response - var temp steady state

Table 4.1 List of the 40 EC experiments

[*] Expression data was collected at: Pilpel et al. 2001, Sudarsanam et al. 2002 and Lapidot and Pilpel 2003 [22, 33, 42]
Data are located at: **http://arep.med.harvard.edu/ExpressDB/yeastindex.html**
http://www-genome.stanford.edu/yeast_stress/

5 APPENDIX B – ALL CLUSTERS

5.1 List of clusters

Our algorithm finds 20 clusters, covering a total of 182 motifs. 14 of our clusters have large overlaps with known clusters:

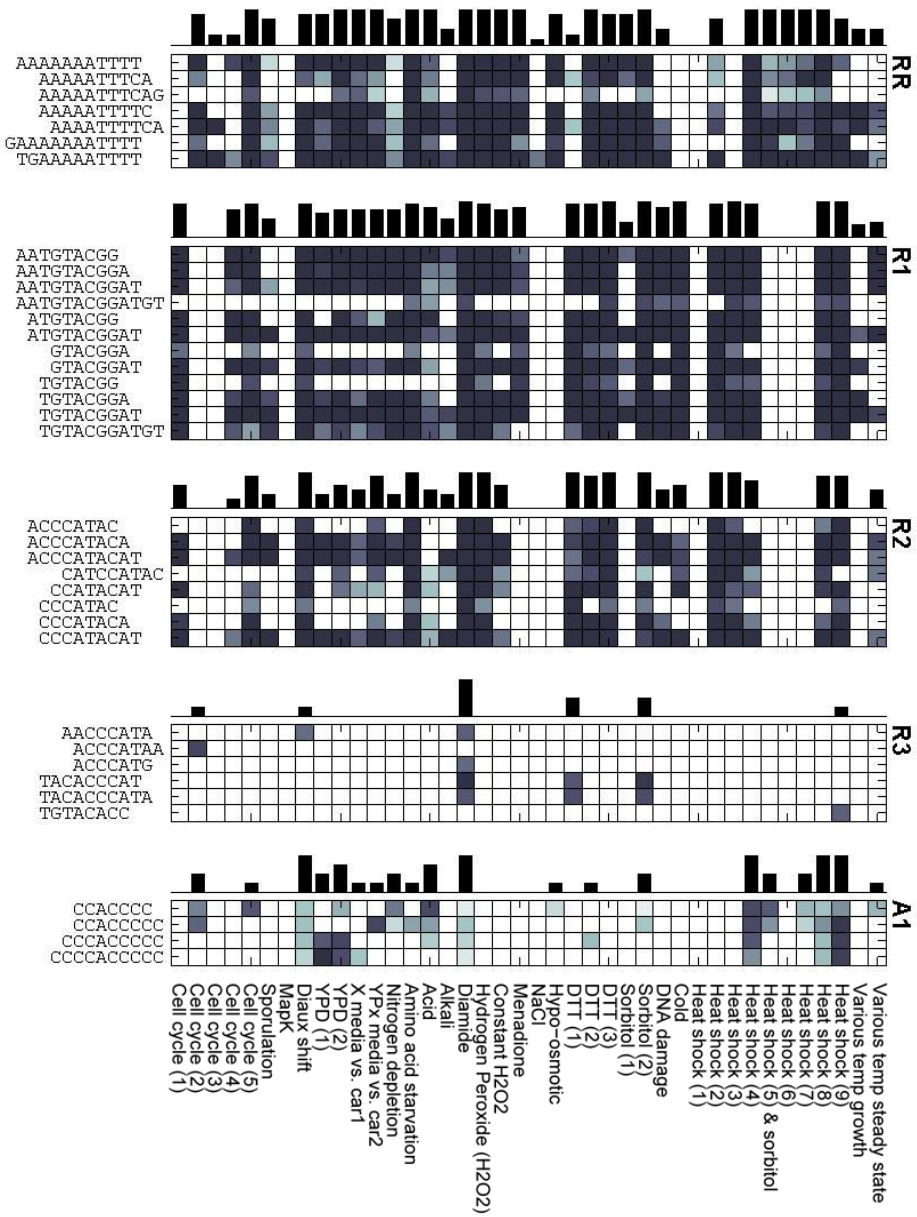
Cluster name	Identified as
MS1	MBF / SBF
MS2	MBF / SBF
MS3	MBF / SBF
MS4	MBF / SBF
ST1	STRE
ST2	STRE
P1	PAC
P2	PAC
RR	RRPE
R1	RAP1
R2	RAP1
R3	RAP1
A1	ADR1 / STRE
RP	RPN4
C15	Unknown
C16	Unknown
C17	Unknown
C18	Unknown
C19	Unknown
C20	Unknown

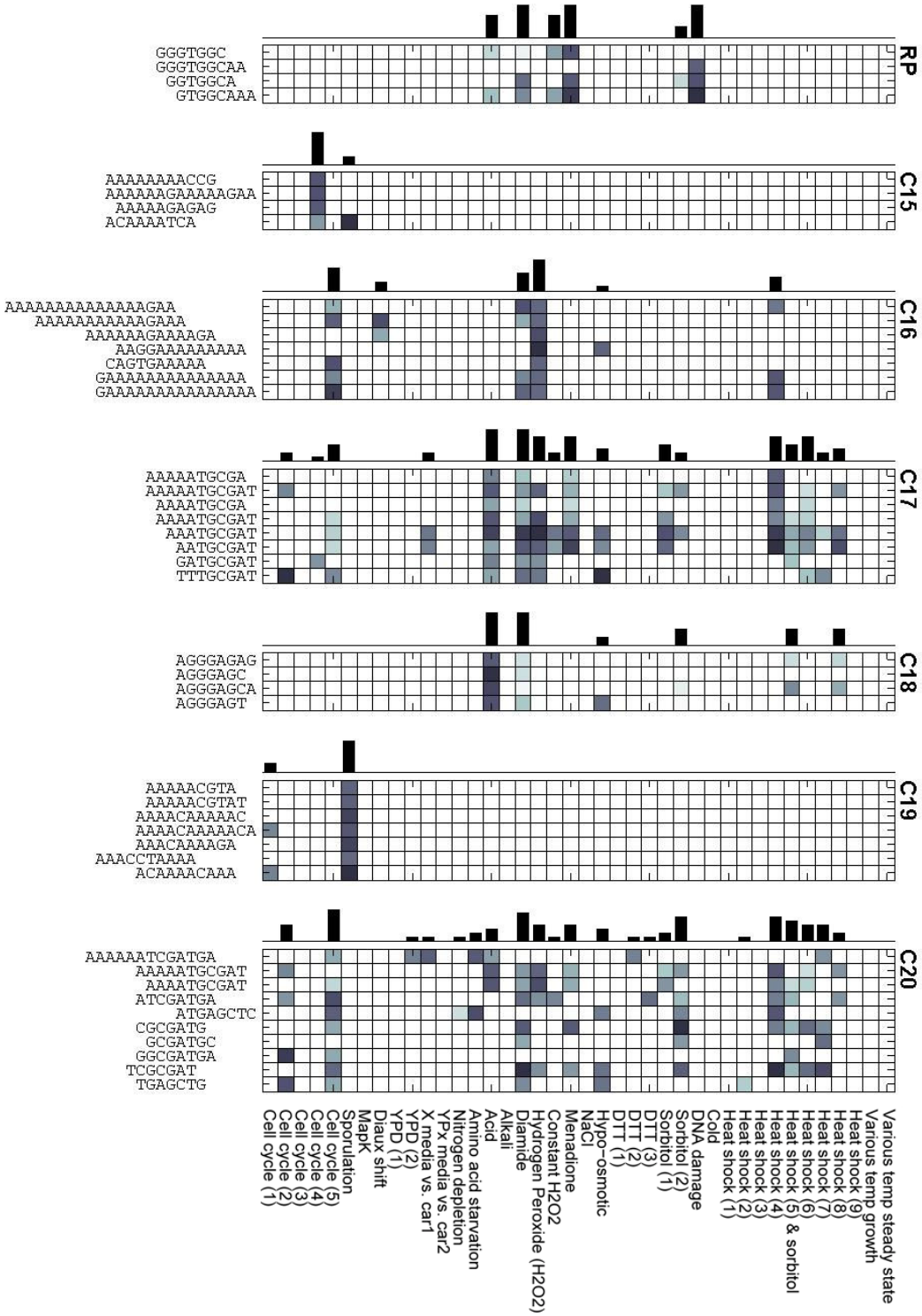
Table 5.1 List of clusters

5.2 Clusters' EC patterns

Within this dissertation we have put our focus on three groups of clusters: the four clusters that correspond to the MCB/SCB binding sites, those that match STRE and the two clusters that correspond to the PAC binding sites. In addition to the clusters that have been discussed previously, 12 more clusters of motifs have been found via our analysis. Following are the EC patterns of those 12 clusters:

Figure 5.1 EC patterns of clusters RR, R1, R2, R3, A1, RP, C15, C16, C17, C18, C19 and C20

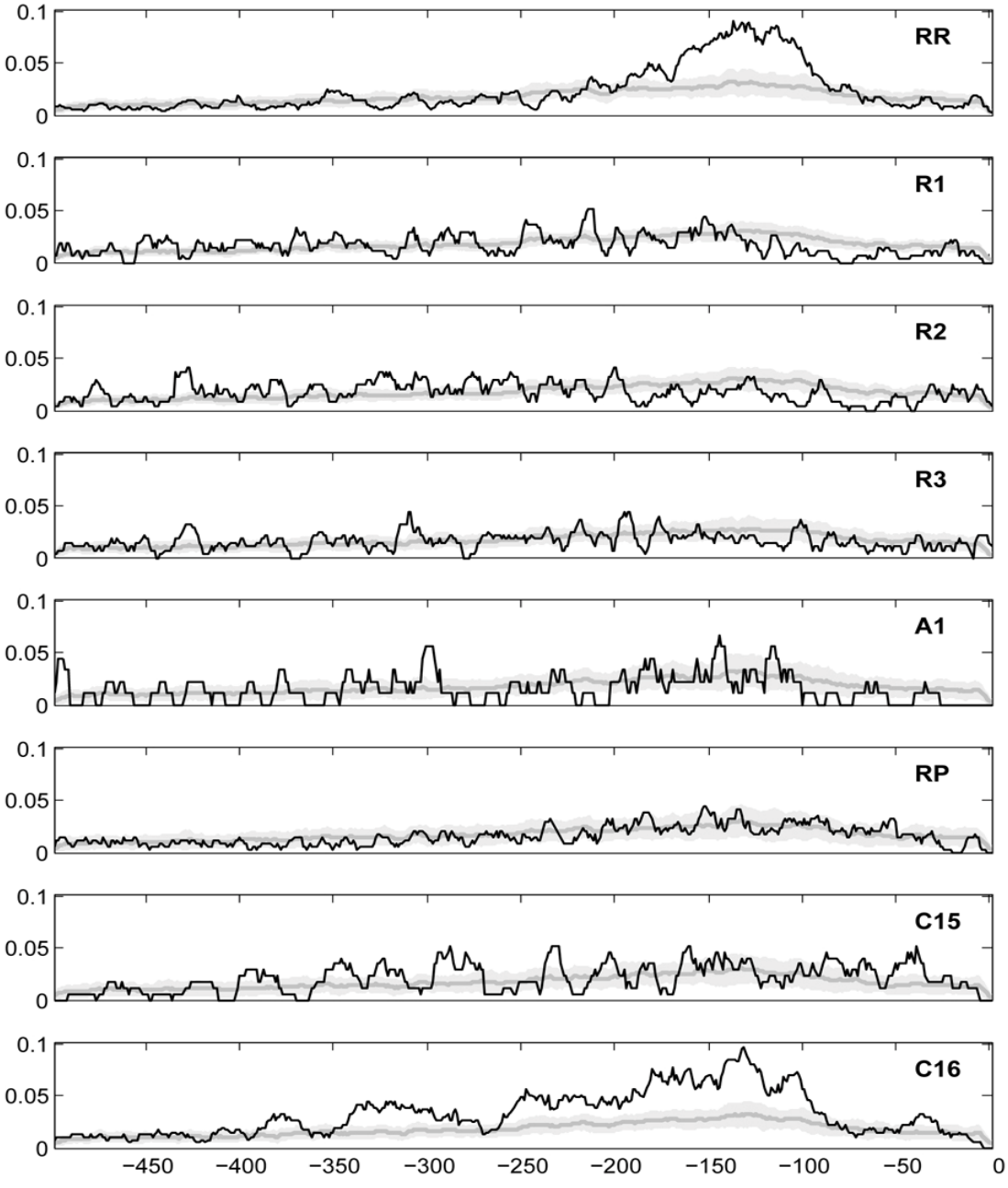


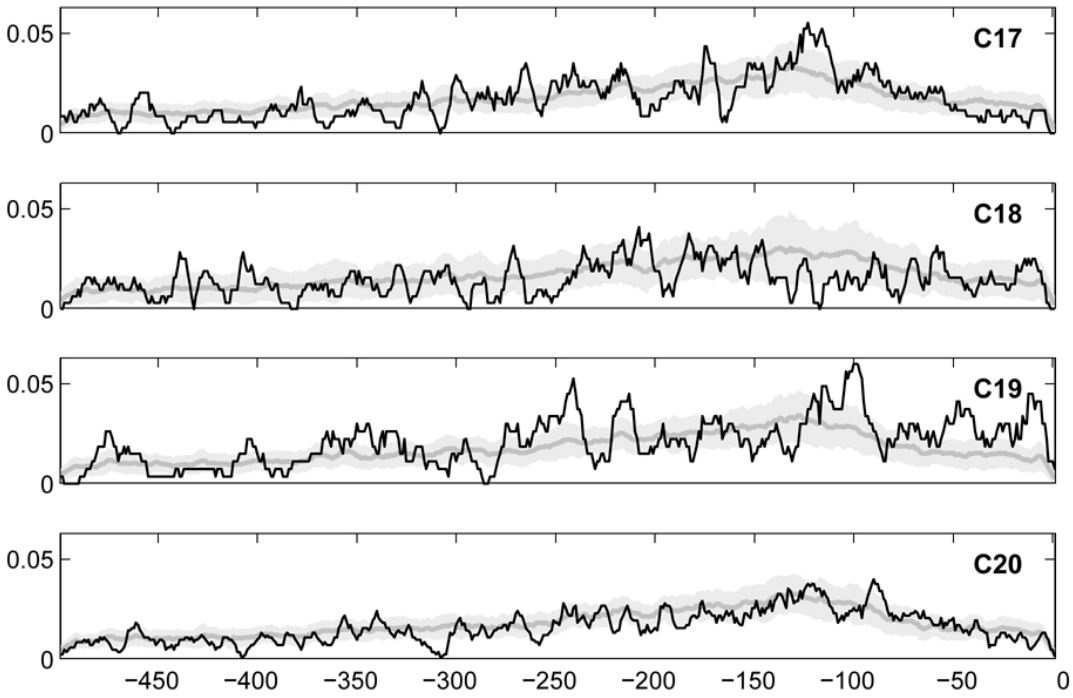

RP
GGGTGGC
GGGTGGCAA
GGTGGCA
GTGGCAAA
C15
AAAAAAAACCG
AAAAAAGAAAAAGAA
AAAAAGAGAG
ACAAAATCA
C16
AAAAAAAAAAAAAAGAA
AAAAAAAAAAAGAAA
AAAAAAGAAAAGA
AAGGAAAAAAAAA
CAGTGAAAAA
GAAAAAAAAAAAAAA
GAAAAAAAAAAAAAAA
C17
AAAAATGCGA
AAAAATGCGAT
AAAATGCGA
AAAATGCGAT
AAATGCGAT
AATGCGAT
GATGCGAT
TTTGCGAT
C18
AGGGAGAG
AGGGAGC
AGGGAGCA
AGGGAGT
C19
AAAAACGTA
AAAAACGTAT
AAAACAAAAAC
AAAACAAAAACA
AAACAAAAGA
AAACCTAAAA
ACAAAACAAA
C20
AAAAAATCGATGA
AAAAATGCGAT
AAAATGCGAT
ATCGATGA
ATGAGCTC
CGCGATG
GCGATGC
GGCGATGA
TCGCGAT
TGAGCTG
Cell cycle (1)
Cell cycle (2)
Cell cycle (3)
Cell cycle (4)
Cell cycle (5)
Sporulation
MapK
Diaux shift
YPD (1)
YPD (2)
X media vs. car1
YPx media vs. car2
Nitrogen depletion
Amino acid starvation
Acid
Alkali
Diamide
Hydrogen Peroxide (H2O2)
Constant H2O2
Menadione
NaCl
Hypo−osmotic
DTT (1)
DTT (2)
DTT (3)
Sorbitol (1)
Sorbitol (2)
DNA damage
Cold
Heat shock (1)
Heat shock (2)
Heat shock (3)
Heat shock (4)
Heat shock (5) & sorbitol
Heat shock (6)
Heat shock (7)
Heat shock (8)
Heat shock (9)
Various temp growth
Various temp steady state

5.3 Localization of motifs on the promoters of clusters

Following are localization analyses done for our 12 clusters that have not been previously shown within the body of this dissertation:

Figure 5.2 Localization of motifs along the promoters: clusters RR, R1, R2, R3, A1, RP, C15, C16, C17, C18, C19, and C20




0.05
0
C17
0.05
0
C18
0.05
0
C19
0.05
0
C20
−450
−400
−350
−300
−250
−200
−150
−100
−50
0

6 APPENDIX C – INTERSECTIONS BETWEEN CLUSTERS

We have examined the number of genes that are shared by every couple of clusters:

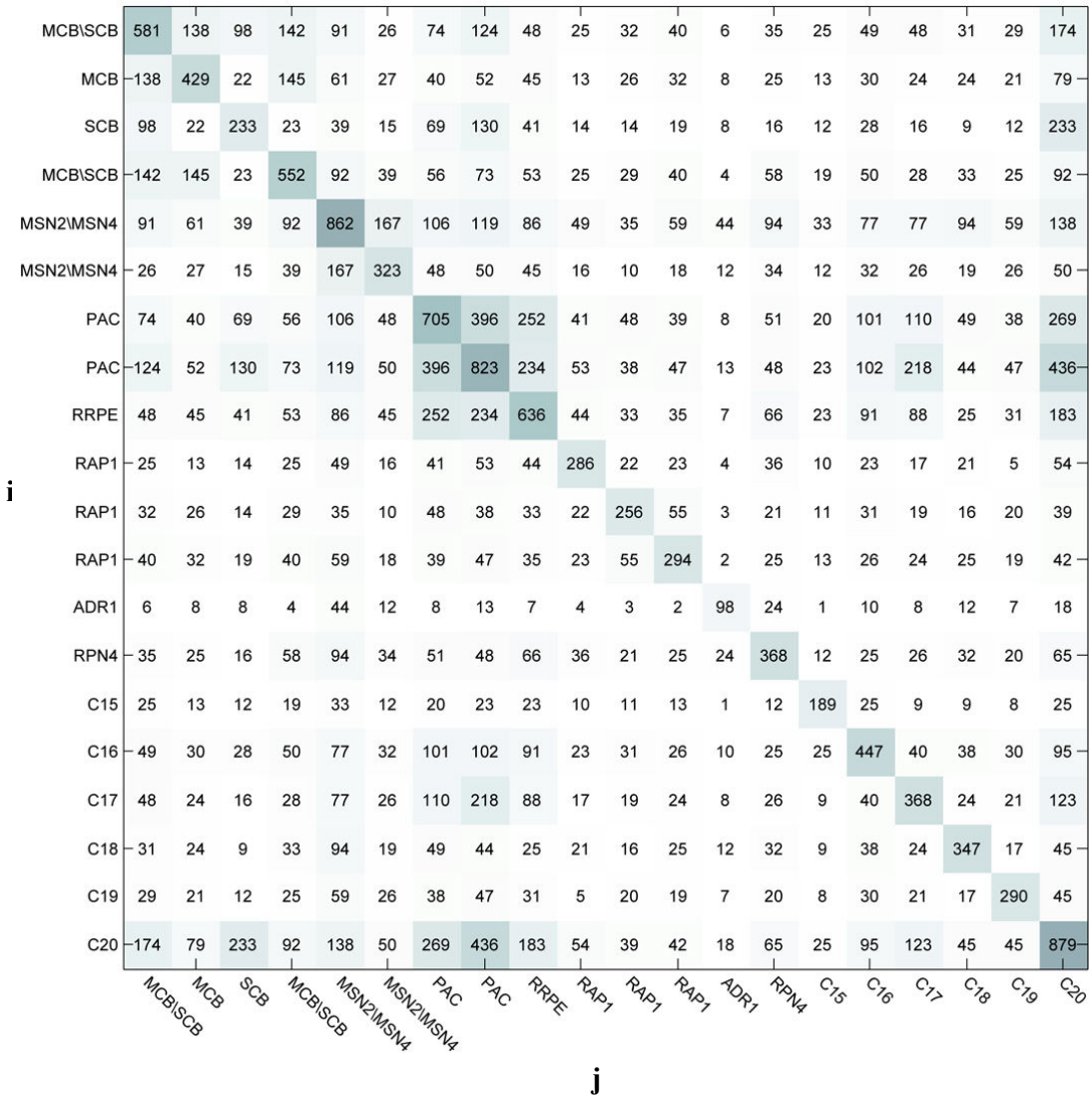


Figure 6.1 Gene intersections between clusters – absolute numbers. For i≠j, each value represents the number of genes that are shared by cluster i and cluster j. On the diagonal appear the numbers of genes on the promoter of which the clusters motifs are found.

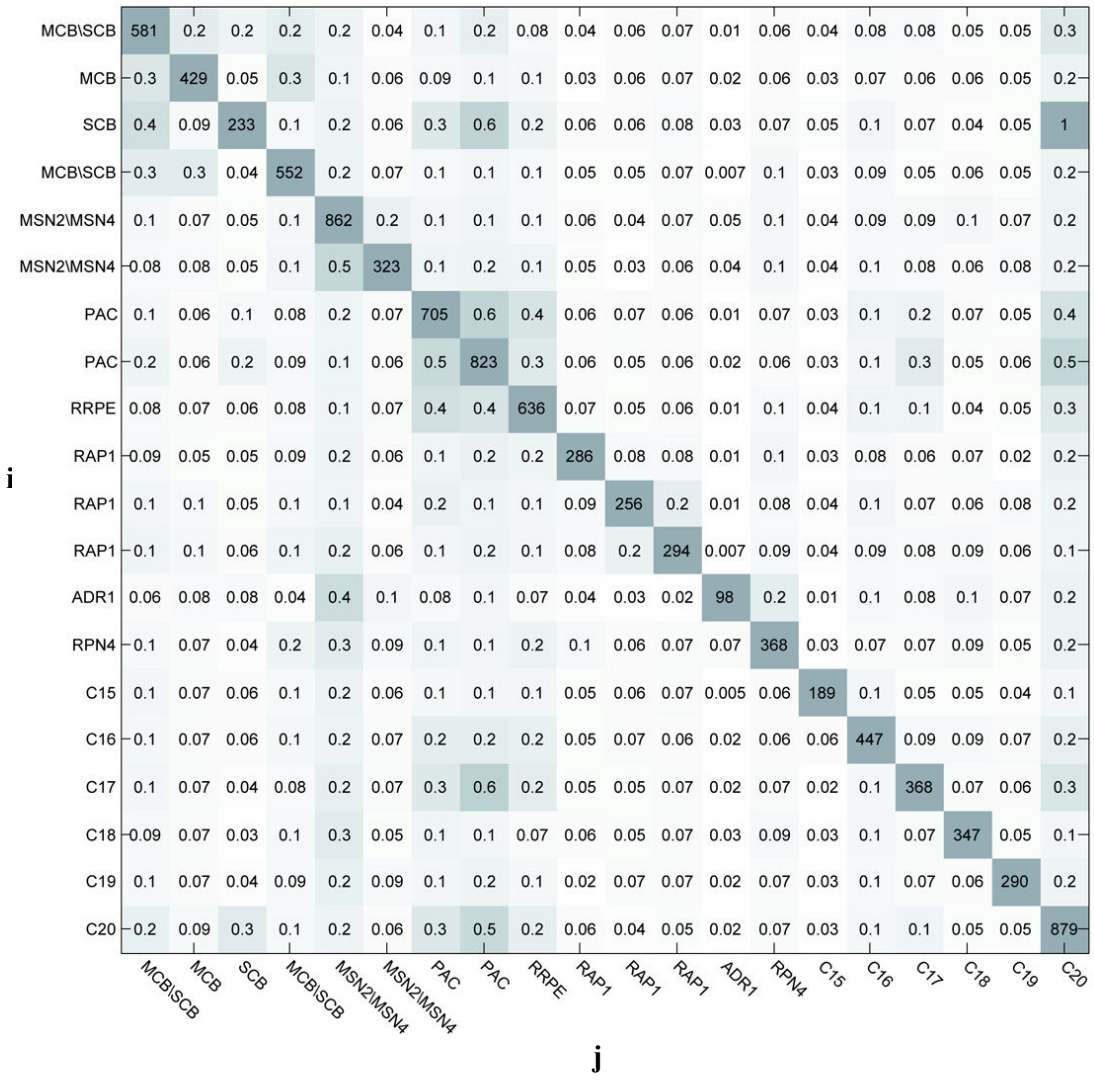


Figure 6.2 Gene intersections between clusters – rates. The matrix above specifies the fractions of genes that are shared by every cluster i and cluster j, out of cluster i (i in rows, j in columns). In the cases of i=j, each value indicates the amount of genes on the promoter of which the clusters motifs are found (the color on the diagonal is set as 100%, in respect to the grayscale of the rest of the matrix).

7 LIST OF ABBREVIATIONS

AP-1	Activator protein-1
EC	Expression Coherence
FDR	False discovery rate
FHL1	Fork-head like 1 (transcription factor)
GO	Gene Ontology
MBF	MCB-binding factor (transcription factor)
MCB	MluI cell cycle box (transcription factor binding site)
MEX	Motif Extraction algorithm
PAC	Polymerase A and C (transcription factor binding site)
PSSM	Position specific scoring matrix
RAP1	Repressor activator protein 1 (transcription factor)
RRPE	Ribosomal RNA processing elements (transcription factor)
SBF	SCB-binding factor (transcription factor)
SCB	SWI4/6 dependent cell cycle box (transcription factor binding site)
SFP1	Split finger protein (transcription factor)
STRE	Stress response elements (transcription factor binding site)
TF	Transcription factor
TFBS	Transcription factor's binding site
YAP5	Yeast AP-1 (transcription factor)

8 BIBLIOGRAPHY

[1] B. J. Andrews and I. Herskowitz, *The yeast SWI4 protein contains a motif present in developmental regulators and is part of a complex involved in cell-cycle-dependent transcription*, Nature, 342 (1989), pp. 830-833.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Gene Ontology: tool for the unification of biology*, 25 (2000), pp. 25-29.

[3] Y. Barash, G. Elidan, N. Friedman and T. Kaplan, *Modeling Dependencies in Protein-DNA Binding Sites*, *RECOMB*, Berlin, Germany, 2003, pp. 28-37.

[4] M. A. Beer and S. Tavazoie, *Predicting Gene Expression from Sequence*, Cell, 117 (2004), pp. 185-198.

[5] Y. Benjamini and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society. Series B (Methodological), 57 (1995), pp. 289-300.

[6] P. V. Benos, M. L. Bulyk and G. D. Stormo, *Additivity in protein-DNA interactions: how good an approximation is it?*, Nucl. Acids Res., 30 (2002), pp. 4442-4451.

[7] O. G. Berg and P. H. von Hippel, *Selection of DNA binding sites by regulatory proteins : Statistical-mechanical theory and application to operators and promoters*, Journal of Molecular Biology, 193 (1987), pp. 723-743.

[8] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry and G. Sherlock, *GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes*, Bioinformatics, 20 (2004), pp. 3710-3715.

[9] A. Brazma, I. Jonassen, J. Vilo and E. Ukkonen, *Predicting Gene Regulatory Elements in Silico on a Genomic Scale*, Genome Res., 8 (1998), pp. 1202-1215.

[10] M. L. Bulyk, P. L. F. Johnson and G. M. Church, *Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors*, Nucl. Acids Res., 30 (2002), pp. 1255-1261.

[11] T. A. Clark, C. W. Sugnet and M. Ares, Jr., *Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays*, Science, 296 (2002), pp. 907-910.

[12] M. Dequard-Chablat, M. Riva, C. Carles and A. Sentenac, *RPC19, the gene for a subunit common to yeast RNA polymerases A (I) and C (III)*, J. Biol. Chem., 266 (1991), pp. 15300-15307.

[13] E. T. Dermitzakis and A. G. Clark, *Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover*, Mol Biol Evol, 19 (2002), pp. 1114-1121.

[14] E. T. Dermitzakis, A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier and S. E. Antonarakis, *Evolutionary Discrimination of Mammalian Conserved Non-Genic Sequences (CNGs)*, Science, 302 (2003), pp. 1033-1035.

[15] R. Durbin, S. R. Eddy, A. Krogh and G. J. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[16] E. Emberly, N. Rajewsky and E. Siggia, *Conservation of regulatory elements between two species of Drosophila*, BMC Bioinformatics, 4 (2003), pp. 57.

[17] I. Fingerman, V. Nagaraj, D. Norris and A. K. Vershon, *Sfp1 Plays a Key Role in Yeast Ribosome Biogenesis*, Eukaryotic Cell, 2 (2003), pp. 1061-1068.

[18] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein and P. O. Brown, *Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes*, Mol. Biol. Cell, 11 (2000), pp. 4241-4257.

[19] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel and R. A. Young, *Transcriptional regulatory code of a eukaryotic genome*, 431 (2004), pp. 99-104.

[20] 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 (2000), pp. 1205-1214.

[21] P. Jorgensen, I. Rupes, J. R. Sharom, L. Schneper, J. R. Broach and M. Tyers, *A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size*, Genes Dev., 18 (2004), pp. 2491-2505.

[22] M. Lapidot and Y. Pilpel, *Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription*, Nucl. Acids Res., 31 (2003), pp. 3824-3828.

[23] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald and J. C. Wootton, *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment*, Science, 262 (1993), pp. 208-214.

[24] T. H. Leung, A. Hoffmann and D. Baltimore, *One Nucleotide in a [kappa]B Site Can Determine Cofactor Specificity for NF-[kappa]B Dimers*, Cell, 118 (2004), pp. 453-464.

[25] P. J. Lopez and B. Seraphin, *Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition*, RNA, 5 (1999), pp. 1135-1137.

[26] S. J. Maerkl and S. R. Quake, *A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors*, Science, 315 (2007), pp. 233-237.

[27] R. M. Marion, A. Regev, E. Segal, Y. Barash, D. Koller, N. Friedman and E. K. O'Shea, *Inaugural Article: Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression*, PNAS, 101 (2004), pp. 14315-14322.

[28] D. E. Martin, A. Soulard and M. N. Hall, *TOR Regulates Ribosomal Protein Gene Expression via PKA and the Forkhead Transcription Factor FHL1*, Cell, 119 (2004), pp. 969-979.

[29] M. T. Martínez-Pastor, G. Marchler, C. Schüller, A. Marchler-Bauer, H. Ruis and F. Estruch, *The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE).* EMBO, 15 (1996), pp. 2227–2235.

[30] T. Moss and V. Y. Stefanovsky, *At the Center of Eukaryotic Life*, Cell, 109 (2002), pp. 545-548.

[31] K. Nasmyth and L. Dirick, *The role of SWI4 and SWI6 in the activity of G1 cyclins in yeast*, Cell, 66 (1991), pp. 995-1013.

[32] G. Pavesi, P. Mereghetti, G. Mauri and G. Pesole, *Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes*, Nucl. Acids Res., 32 (2004), pp. W199-203.

[33] Y. Pilpel, P. Sudarsanam and G. M. Church, *Identifying regulatory networks by combinatorial analysis of promoter elements*, 29 (2001), pp. 153-159.

[34] M. Primig, S. Sockanathan, H. Auer and K. Nasmyth, *Anatomy of a transcription factor important for the Start of the cell cycle in Saccharomyces cerevisiae*, Nature, 358 (1992), pp. 593-597.

[35] M. Pritsker, Y.-C. Liu, M. A. Beer and S. Tavazoie, *Whole-Genome Discovery of Transcription Factor Binding Sites by Network-Level Conservation*, Genome Res., 14 (2004), pp. 99-108.

[36] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J.-P. Z. Wang and J. Widom, *A genomic code for nucleosome positioning*, Nature, 442 (2006), pp. 772-778.

[37] L. Segal, M. Lapidot, Z. Solan, E. Ruppin, Y. Pilpel and D. Horn, *Nucleotide variation of regulatory motifs may lead to distinct expression patterns* submited to ISMB (2007).

[38] E. D. Siggia, *Computational methods for transcriptional regulation*, Current Opinion in Genetics & Development, 15 (2005), pp. 214-221.

[39] Z. Solan, D. Horn, E. Ruppin and S. Edelman, *Unsupervised learning of natural languages*, PNAS, 102 (2005), pp. 11629-11634.

[40] M. Spingola, L. Grate, D. Haussler and M. Ares, Jr., *Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae*, RNA, 5 (1999), pp. 221-234.

[41] G. D. Stormo, *DNA binding sites: representation and discovery*, Bioinformatics, 16 (2000), pp. 16-23.

[42] P. Sudarsanam, Y. Pilpel and G. M. Church, *Genome-wide Co-occurrence of Promoter Elements Reveals a cis-Regulatory Cassette of rRNA Transcription Motifs in Saccharomyces cerevisiae*, Genome Res., 12 (2002), pp. 1723-1731.

[43] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church, *Systematic determination of genetic network architecture*, Nature Genetics, 22 (1999), pp. 281-285.

[44] M. L. Vignais, L. P. Woudt, G. M. Wassenaar, W. H. Mager, A. Sentenac and R. J. Planta, *Specific binding of TUF factor to upstream activation sites of yeast ribosomal protein genes.*, EMBO J., 6 (1987), pp. 1451–1457.

[45] J. R. Warner, *The economics of ribosome biosynthesis in yeast*, Trends in Biochemical Sciences, 24 (1999), pp. 437-440.

אוניברסיטת תל אביב TEL AVIV UNIVERSITY

הפקולטה לרפואה ע"ש סאקלר
המדרשה לתארים מתקדמים
ע"ש ד"ר מרים ושלדון ג' אדלסון

החוג לפיזיולוגיה ופרמקולוגיה

SACKLER FACULTY OF MEDICINE
DR. MIRIAM AND SHELDON G. ADELSON
GRADUATE SCHOOL OF MEDICINE

Department of Physiology and Pharmacology

הבדלים קטנים במוטיבים בקרתיים עשויים להביא להבדלים משמעותיים בביטוי הגנים

חיבור זה נכתב כמילוי חלקי של הדרישות לקבלת

תואר מוסמך

בפקולטה לרפואה ע"ש סאקלר

אוניברסיטת תל אביב

על ידי

ליאת סגל

ת.ז. 036173334

העבודה נעשתה בהנחייתם של

פרופסור דוד הורן
בית הספר לפיזיקה ולאסטרונומיה

פרופסור איתן רופין
החוג לפיזיולוגיה ופרמקולוגיה, בית הספר לרפואה
בית הספר למדעי המחשב

אדר תשס"ז

תקציר

מטרות המחקר

שיטות מקובלות למציאת אתרי קישור אפשריים של גורמי שעתוק (transcription factor binding sites) פועלות תחת הנחות ודרישות, כגון, קיום ביטוי יתר של המוטיבים על פני הפרומוטרים וייצוג המוטיבים באמצעות מטריצת הסתברויות (PSSM) או בעזרת ביטויי קונצנזוס של המוטיבים. מכיוון שמוטיבים בקרתיים (regulatory motifs) אינם בהכרח מצויים בביטוי יתר על פני הגנום כולו, שיטות אלו דורשות עיבוד מקדים של הנתונים, על מנת לקבץ גנים לקבוצות המבוקרות באופן דומה, תוך שימוש במידע נוסף מלבד המידע הרצפי. במטרה להימנע מהטיות הנובעות מהנחות מעיין אלה, הפעלנו על רצפי הפרומוטרים את האלגוריתם הלא מוכוון (unsupervised) למיצוי מוטיבים מרצפים, MEX (Motif Extraction), אשר יועד, במקור, למציאת מילים מאוסף משפטים בשפות טבעיות. העבודה המוצגת בחיבור זה מבוססת על מאמר, שהוגש לאחרונה לפרסום.

שיטות

הפעלנו את MEX על רצפי הפרומוטרים של השמר *S. cerevisiae*, על מנת לזהות מוטיבים בקרתיים אפשריים. MEX אינו מסתמך על ביטוי יתר של המוטיבים על פני הפרומוטרים ואינו דורש עיבוד מקדים של רצפי הפרומוטרים, לרבות קיבוצם לקבוצות המבוקרות יחדיו.

תוצרי האלגוריתם MEX עברו סינון נוסף, המתייחס למשמעות הבקרתית שלהם, באמצעות בדיקת תאימות הביטוי (Expression Coherence, EC) של הגנים, אשר מוטיב כלשהו מצוי בפרומוטרים שלהם. ניתוח ה- EC של המוטיבים התבצע על פני ארבעים ניסיים ביולוגים שונים.

את המוטיבים הבקרתיים קיבצנו (clustered) על פי הרצפים שלהם ובו זמנית על סמך הניסויים בהם הם הביאו לביטוי מתואם של הגנים. מציאת המוטיבים וקיבוצם בשיטות אלו חושפים תובנות ביולוגיות חדשות, אשר קשה היה להבחין בהן באמצעות השיטות המקובלות.

תוצאות

הפעלת MEX על רצפי הפרומוטרים של השמר *S. cerevisiae* נמצאה כמועילה מאד במציאת מוטיבים בקרתיים על סמך מבחן ה- EC, בארבעים ניסיונות ביולוגים שונים. בקיבוץ המוטיבים בעלי ציוני ה- EC המשמעותיים ביותר, מתקבלים עשרים צבירים (clusters), אשר חלקם מתאימים לאתרי קישור מוכרים של גורמי שעתוק. לצבירים שנמצאו יש תבניות EC ייחודיות, המעידות על האופי הבקרתי של המוטיבים, במקביל להבדלים ספציפיים ברצפם. במקרים מסוימים, שונות של חומצת בסיס יחידה על פני המוטיב מביאה להבדל מובהק בתבנית ה- EC של המוטיבים בקבוצה. תוצאות אלה נבדקו ביחס למידע זמין נוסף, כגון, מדידת הקישור באורגניזמים חיים (*in vivo*) של גורמי שעתוק אל הפרומוטרים המשתייכים לצבירים השונים. ביצענו ניתוח מקיף של הצברים המתאימים לאתרי הקישור של MCB/SCB, STRE ושל PAC.

בשני המקרים הראשונים אנו מראים כי לצבירים השונים יש אופני קישור שונים בין פקטור השעתוק ל- DNA. במקרה של PAC גילינו צביר חדש של מוטיבים, אשר ככל הנראה מוסכו עד כה בשל האפקט החזק של המוטיבים הכלולים בצביר המוכר של PAC.

מסקנות

בעוד שייצוג מוטיבים באמצעות קונצנזוס או מטריצת הסתברויות (PSSM) הוא נפוץ ופשוט, שיטות ייצוג שכאלה מביאות לאיבוד מידע ועשויות להביא לטעויות. מכיוון ש- MEX אינו משתמש בשיטות ייצוג של מוטיבים אנו יכולים לבחון כל מוטיב באופן בלתי תלוי, ורק אז לקבץ את המוטיבים הבקרתיים לצבירים, לקבלת הבנה ביולוגית טובה יותר, מבלי לאבד מידע או להטות את התוצאות. מתוצאות האנאליזה שביצענו, ניתן לראות כי שינוי בחומצת בסיס יחידה על פני המוטיב, עשוי להשפיע באופן דרמטי על בקרת הגנים. חוזק הבקרה עשוי לנבוע ממספר גורמים. בחנו את שיעור החזרות של המוטיבים הבקרתיים על פני הפרומוטרים ואת מיקומם במעלה הפרומוטר, על מנת לבדוק האם הם הגורמים להבדל בחוזק הבקרה או שמא מדובר בשינוי באופי הקישור בין גורם השעתוק לרצף המוטיב המסוים על פני DNA. במקרים של הצבירים אשר זוהו כ- MCB/SCB וכ- STRE הסקנו כי האחרון הוא המקרה. במקרים אלו הראנו כי הבדל יחיד ברצף המוטיב מביא להבדלים משמעותיים בבקרת הגנים.