BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i440–i449
doi:10.1093/bioinformatics/btm183

Nucleotide variation of regulatory motifs may lead to distinct expression patterns

Liat Segal[1,†], Michal Lapidot[2,†], Zach Solan[3], Eytan Ruppin[1,4], Yitzhak Pilpel[2] and David Horn[3,*]

[1]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, [2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, [3]School of Physics and Astronomy and [4]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

ABSTRACT

Motivation: Current methodologies for the 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 position-specific scoring matrices (PSSMs). In order to avoid bias introduced by such assumptions, we apply an unsupervised motif extraction (MEX) algorithm to sequences of promoters. The extracted motifs are assessed for their likely *cis*-regulatory function by calculating the expression coherence (EC) of the corresponding genes, across a set of biological conditions.

Results: Applying MEX to all *Saccharomyces cerevisiae* promoters, followed by EC analysis across 40 biological conditions, we obtained a high percentage of putative *cis*-regulatory motifs. We clustered motifs that obtained highly significant EC scores, based on both their sequence similarity and similarity in the biological conditions these motifs appear to regulate. We describe 20 clusters, some of which regroup known TFBS. The clusters display different mRNA expression 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 additional information, such as 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.

Contact: horn@tau.ac.il

Supplementary information: Supplementary data are available at http://adios.tau.ac.il/regmotifs and at *Bioinformatics* online.

1 INTRODUCTION

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 non-functional with respect to transcription. Furthermore, a single TF 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.

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 (Lawrence *et al*., 1993), which detects over-represented motifs. However since regulatory motifs are very short, while in contrast, the regulatory portion of the genome is very long (e.g. 6 000 000 bp 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 (Brazma *et al*., 1998; Harbison *et al*., 2004; Tavazoie *et al*., 1999).

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 then crucial for obtaining reliable results. Comparing species with a short divergence time may yield false positives while choosing too distant species will fail to recover species-specific sites. For instance, ~40% of human functional TFBS are expected to be non-functional in rodents (Dermitzakis and Clark, 2002). Furthermore, the alignment of orthologous intergenic sequences is nontrivial since well-conserved sequences of different lengths are interspersed with sequences that show little conservation.

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 position. These biases are commonly represented mathematically by position-specific scoring matrices (PSSMs), whose components give the probabilities of finding each

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

nucleotide at each binding site position (Berg and von Hippel, 1987; Stormo, 2000).

Motif representations by PSSM, however, ignore dependencies between nucleotide positions in regulatory motifs, even though such dependencies are known to occur (Benos *et al.*, 2002; Bulyk *et al.*, 2002). Statistical models that account for such dependencies include hidden Markov models and Bayesian networks (Durbin *et al.*, 1998). Yet, even sophisticated models of this kind have relatively low values of sensitivity and specificity when required to represent the known binding sites (Barash *et al.*, 2003).

Here, we employ a different approach that attempts to avoid the limitations and inherent assumptions discussed above. We adapt a recently published unsupervised algorithm (Solan *et al.*, 2005), 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 (Lapidot and Pilpel, 2003; Pilpel *et al.*, 2001) in order to check which of the identified motifs exert significant effects on the expression profiles of their downstream genes. The expression analysis shows an enormous enrichment of highly scoring motifs among MEX's predictions, and 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 methods, which rely on sequence alone. For instance, partially overlapping TFBS that are bound by distinct TFs regulating different biological conditions, may be indistinguishable by sequence, yet appear in separate clusters using our method. Another biological phenomenon we 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; there exist specific structural relations, such as positional dependencies, that correlate with high EC values in particular biological conditions, i.e. they are of functional importance.

2 METHODS

2.1 The motif extraction algorithm (MEX)

MEX is a motif extraction algorithm (Solan *et al.*, 2005) developed as part of another algorithm, ADIOS, which induces grammar from texts. Given a set of DNA sequences, such as the promoters of all genes in *Saccharomyces cerevisiae*, one loads all strings on a graph composed of four vertices corresponding to the 4 nucleotides that form the alphabet of the problem at hand. All these strings form paths over the graph. MEX identifies significant patterns by searching for convergence of multiple paths onto the same subpaths and divergence from them. These substrings are considered to be motifs if they obey two requirements concerning the amount of convergence/divergence observed, specified by a parameter η, and its statistical significance given the number of paths involved, set by a threshold α. Further information can be found in the Supplementary Material and at http://adios.tau.ac.il.

We applied MEX to 4800 promoters of 6300 genes from the genome of *S.cerevisiae*. Some promoters are located in the intergenic region of two oppositely oriented genes and thought to regulate both. Throughout this work, bidirectional promoters were taken twice in different orientations and associated with the corresponding genes. Each promoter sequence, of length up to 1000 bp, is considered as a path on the graph created by MEX. After all information is loaded onto the graph, we use all promoter sequences as trial paths in order to extract motifs.

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

2.2 The expression coherence (EC) score

The EC score of a motif, that appears in the promoters of a set of genes, *S*, is defined as the fraction of gene pairs (*i,j*) in *S*, for which the Euclidean distance between normalized expression profiles falls below a threshold, D_T. D_T is determined 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 converge into one or a few tight clusters in expression space (see Supplementary Material for an intuitive explanation). A sampling-based means exists for the assessment of statistical significance, in terms of *P*-value of EC scores, given the gene set size *N* and the examined expression time series (Lapidot and Pilpel, 2003). When calculating EC scores for multiple motifs, we applied the false discovery rate (FDR) theorem (Benjamini and Hochberg, 1995) in order to account for the testing of multiple hypotheses and to control the amount of false positives. The FDR criterion determines the *P*-value cutoff below which motifs are guaranteed to be statistically significant at a specified false discovery rate.

2.3 mRNA expression data

Whole-genome mRNA expression data of 40 time series experiments, representing a wide range of natural (e.g. cell cycle) and perturbed conditions in *S.cerevisiae*, were obtained from ExpressDB (http://arep.med.harvard.edu/cgi-bin/ExpressDByeast/EXDStart). For a complete list of conditions see Supplementary Material.

2.4 Clustering motifs by sequence and function

We have formulated an iterative method for clustering motifs, according to both their sequences and the biological conditions in which they operate, as determined by their EC scores across the 40 examined biological conditions. We initiate clusters by gathering motifs that share some building blocks (of length 6 bp), or 'seeds'. Then, a series of iterations improves the clusters, using various procedures detailed below. The clusters refinement steps include addition and removal of motifs from existing clusters and splitting and merging of clusters. We quantified the quality of clusters using several criteria

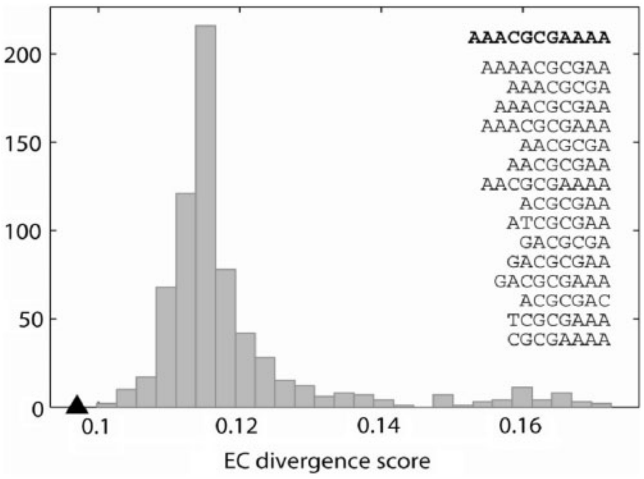


Fig. 1. An example for testing the contribution of a specific motif to the cluster's tightness. The EC divergence score of the cluster including the motif AAACGCGAAAA (black triangle) is compared to the empirical distribution of EC divergence of clusters, in which the motif in question has been replaced 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.

associated with sequence patterns and EC score patterns of the motifs. The refinement steps are listed below.

2.4.1 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.

2.4.2 Pruning clusters to increase EC tightness For each motif, which had a significant EC score (had passed the FDR criterion) in at least one of the examined biological conditions, one defines an EC vector of length 40, whose entries specify the EC *P*-values of the motif across the 40 biological conditions. Such vectors comprise the matrices in Figures 5–7. 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, we compare the EC divergence of M with the empirical distribution of EC divergence scores resulting from replacing m with every one of the motifs that lie outside the cluster M. An example is shown in Figure 1.

2.4.3 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 we compare the EC divergence resulting from its addition ($M+m$) with the empirical EC divergence distribution resulting from additions of each of the motifs lying outside M.

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 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.

2.4.4 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 SDs:

$$F_{\mathrm{A,B}} = \frac{\|\mu_{\mathrm{A}} - \mu_{\mathrm{B}}\|}{\|\sigma_{\mathrm{A}}\| + \|\sigma_{\mathrm{B}}\|}$$

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 40 SDs corresponding to the 40 EC experiments. Clusters will be merged if their F is smaller than some threshold, as long as they also obey the sequential similarity criterion.

2.4.5 Splitting of clusters Clusters are split into K smaller clusters if they exceed a given size. Splitting is done using K-means on the EC space of the cluster (K is a parameter of the algorithm and was set to $K=3$ in this work). After applying this indiscriminative step, however, a fusion step is applied, so that unnecessary splitting will be reversed.

2.4.6 Fine refinement of clusters The former procedures are applied iteratively in a preset order, generating clusters that are rather tight in EC and in sequence and differ from each other in sizes, EC patterns and motif sequences. The clusters are given a cluster score (defined in Supplementary Material), encapsulating the various measures used so far. In a final pruning step, using finer parameters, improvement is tested with respect to the cluster score, and the pruning is accepted or rejected accordingly.

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

2.5 Finding GO annotations of clusters

Co-regulated genes may be involved in similar cellular processes and functions. Information regarding the functional tendencies of genes on the promoters of which the cluster's motifs are found may be helpful in getting a notion about the identity of clusters. We used GO term finder (Boyle *et al*., 2004) in order to test the GO enrichment (Ashburner *et al*., 2000) of such sets of genes.

2.6 Incremental TF binding rates

In order to further validate the identity of clusters with respect to known TFs, we performed a comprehensive estimation of the binding of various *S.cerevisiae* TFs to the promoters on which our motifs were found. For this purpose we employed yeast genome-wide location analysis data (Harbison *et al*., 2004), in which the genomic occupancy of 203 DNA-binding TFs 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 TF. Since some TFs are less specific, and typically bind more genes than others, we defined *incremental binding rates* by subtracting the mean binding rate of each TF from the binding rates of each TF to every cluster. Figure 4 displays the incremental binding rates of each of the 203 tested *S.cerevisiae* TFs to our motif clusters.

2.7 Localization of motifs along promoters

Further analysis of motif clusters refers to their positional distribution along the promoters on which they occur. This is compared with a background model consisting, for each cluster, of 1000 randomly selected groups of motifs (sampled from our final set of motifs) of the same size as that of the cluster. We examined the difference in positional tendency between each cluster and its background model and evaluated its significance. A P-value was calculated using running windows of 50 bp along the promoters. For each such window the average rate of motif occurrence was calculated for the cluster, as well as for each of the groups composing the background model and a P-value was assessed. Note that the background model used here is not a random one, but contains information that is inherent in the final set of motifs. Had we chosen groups of random sequences from the promoters (instead of motifs from our final set), the distribution would be uniform.

3 RESULTS

3.1 Extraction and selection of putative *cis*-regulatory motifs from the promoters of *S.cerevisiae*

Applying MEX to genome-wide promoters of *S.cerevisiae* (see Methods section), using the parameters $\alpha = 0.1$ and $\eta = 1$, we have extracted 9370 putative *cis*-regulatory motifs. As TFs are thought to recognize and bind double-stranded DNA, we unified motifs with their reverse compliments, reducing the total number of extracted motifs to 8498 potential TF-binding sites. To assess the regulatory potential of these putative motifs, we calculated EC scores (see Methods section) and corresponding P-values for these 8498 putative motifs across 40 biological conditions including cell cycle, sporulation and multiple environmental stresses. Each biological condition was represented by a time series experiment, monitoring yeast whole-genome mRNA levels.

Setting the false discovery rate to 0.1, 22% of the MEX-extracted motifs had a significant EC score in at least one of the examined biological conditions. For comparison, in an exhaustive enumeration of all k-mers of length 7–11 residing in *S.cerevisiae* promoters (Shalgi *et al.*, 2005; Lapidot and Pilpel, in preparation), only 0.6% scored significantly, under the same FDR condition. MEX provides striking enrichment in selecting biologically significant motifs. Moreover, when comparing the motif sets that passed FDR in both cases, we observed that 55% of MEX's motifs were not discovered by the exhaustive approach (Supplementary Fig. 10.1). These are mostly weaker motifs that could not be identified within a very noisy background, i.e. MEX increased the signal to noise ratio. In addition, MEX extracted sequence motifs of length up to 19 nt, which is extremely expensive computationally when running over all k-mers. Motifs that were detected by the exhaustive approach, but not by MEX, most likely do not obey the inherent position dependencies, selected for by MEX. It has been reported that some, but not all functional TFBS display such position dependencies (Tomovic and Oakeley, 2007).

We compared both our MEX-extracted motif set and the exhaustive motif set to the well-accepted reference set published by Harbison *et al.* (2004). We applied a scoring method that assesses how likely a given k-mer is to be generated by a given PSSM (see Supplementary Material). Applying a cutoff of 99% identity, 16% of the exhaustive set provides coverage of 91% of Harbison's PSSMs, and 45% of the MEX set, covered 66% of the Harbison set. Namely MEX is not as comprehensive as the exhaustive set, but it is enriched in signal and contains less false positives.

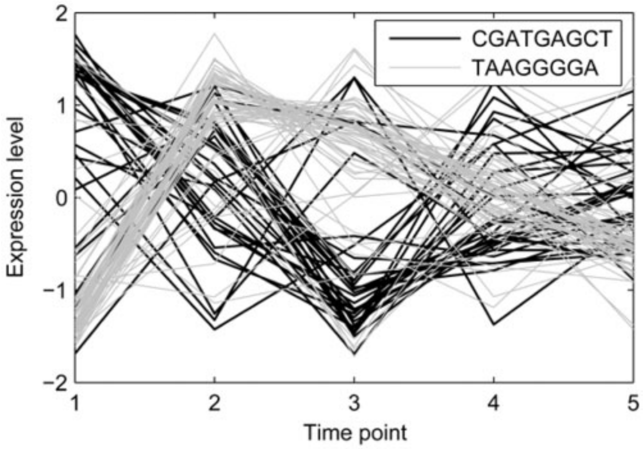


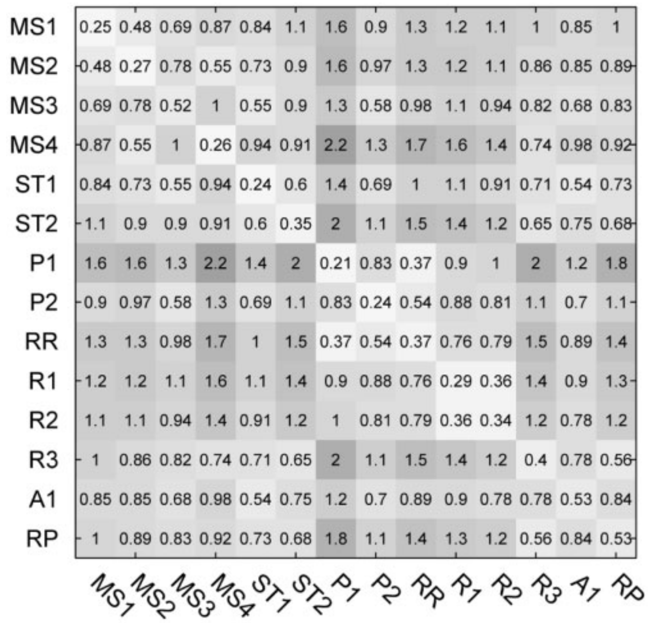
Fig. 2. A semantic characterization of two of the motifs extracted by MEX. MEX has identified two motifs governing opposite responses to hypo-osmotic stress, CGATGAGCT (corresponding to the PAC motif) and TAAGGGGA (corresponding to STRE). Each line represents the expression profile of a single gene that contains the PAC-related motif (black lines) or the STRE-related motif (gray lines) in its promoter, following hypo-osmotic stress (Gasch *et al.*, 2000). Gene sets containing the PAC-related motif in their promoters are coherently expressed (EC = 0.12, P-value < 0.0006). Genes governed by the STRE-related motif, are also coherently expressed (EC = 0.38, P-value 0.00001), yet display an opposing tendency following stress. This illustrates the strength of MEX in identifying sequence motifs corresponding to known *S.cerevisiae* TFBS based on promoter sequence alone and the strength of the EC analysis in assigning a biological function to these motifs.

In summary, the strength of MEX is not in its comprehensiveness, but in its scalability, ability to identify inter-position dependencies and to detect weaker motifs.

The EC analysis not only selects the potentially functional motifs, but also assigns them with a semantic description, namely the set of biological condition in which they operate, and the regulatory effect they exert (e.g. increased expression in response to a particular stress, or peak in expression level at a particular stage of the cell cycle). Figure 2 shows such semantic annotation of two high-scoring motifs found by MEX. These motifs govern opposite responses to hypo-osmotic pressure. Analyses of individual motifs can be performed through the Motif Analysis Workbench (Lapidot and Pilpel, 2003) at http://longitude.weizmann.ac.il/services.html.

3.2 Clustering motifs by sequence and function

We applied further screening of our motifs, selecting a 'distilled' set of 694 motifs, which both passed FDR of 0.1 and were assigned an EC score with a P-value of 0.001 or lower in at least one of the examined biological conditions (a complete list of motifs can be found at Supplementary Material). About half of these motifs perfectly match (or are included in) known binding



	MS1	MS2	MS3	MS4	ST1	ST2	P1	P2	RR	R1	R2	R3	A1	RP
MS1	0.25	0.48	0.69	0.87	0.84	1.1	1.6	0.9	1.3	1.2	1.1	1	0.85	1
MS2	0.48	0.27	0.78	0.55	0.73	0.9	1.6	0.97	1.3	1.2	1.1	0.86	0.85	0.89
MS3	0.69	0.78	0.52	1	0.55	0.9	1.3	0.58	0.98	1.1	0.94	0.82	0.68	0.83
MS4	0.87	0.55	1	0.26	0.94	0.91	2.2	1.3	1.7	1.6	1.4	0.74	0.98	0.92
ST1	0.84	0.73	0.55	0.94	0.24	0.6	1.4	0.69	1	1.1	0.91	0.71	0.54	0.73
ST2	1.1	0.9	0.9	0.91	0.6	0.35	2	1.1	1.5	1.4	1.2	0.65	0.75	0.68
P1	1.6	1.6	1.3	2.2	1.4	2	0.21	0.83	0.37	0.9	1	2	1.2	1.8
P2	0.9	0.97	0.58	1.3	0.69	1.1	0.83	0.24	0.54	0.88	0.81	1.1	0.7	1.1
RR	1.3	1.3	0.98	1.7	1	1.5	0.37	0.54	0.37	0.76	0.79	1.5	0.89	1.4
R1	1.2	1.2	1.1	1.6	1.1	1.4	0.9	0.88	0.76	0.29	0.36	1.4	0.9	1.3
R2	1.1	1.1	0.94	1.4	0.91	1.2	1	0.81	0.79	0.36	0.34	1.2	0.78	1.2
R3	1	0.86	0.82	0.74	0.71	0.65	2	1.1	1.5	1.4	1.2	0.4	0.78	0.56
A1	0.85	0.85	0.68	0.98	0.54	0.75	1.2	0.7	0.89	0.9	0.78	0.78	0.53	0.84
RP	1	0.89	0.83	0.92	0.73	0.68	1.8	1.1	1.4	1.3	1.2	0.56	0.84	0.53

Fig. 3. Fisher distances between 14 of 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 100 random divisions for each cluster).

sites of 85 TFs (published by Harbison *et al.*, 2004 and by Pritsker *et al.*, 2004).

We clustered this distilled set according to both sequence similarity and similarity in function, i.e. in the set of conditions each motif appears to regulate, as judged by the EC analysis (see Methods Section). We have detected 20 clusters, covering a total of 182 motifs. 14 of our clusters have large overlaps with known clusters. Figure 3 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.

Below we discuss in detail 8 clusters, belonging to 3 distinct groups. The remaining clusters are presented in the Supplementary Material.

3.3 MCB/SCB clusters

The first four clusters (see Figure 4) have large overlaps with known recognition sites of two related TF complexes, MBF (MluI cell cycle box binding factor) and SBF (Swi4-Swi6 cell cycle box binding factor). Both complexes are heterodimeric and are known to regulate the G1/S transition during cell cycle (Koch *et al.*, 1993). MBF consists of two protein components, Mbp1 and Swi6, and recognizes the binding site MCB (MluI cell cycle box). SBF consists of Swi4 and Swi6 and binds a site called SCB (Swi4-Swi6 cell cycle box). Both MBF and SBF play important roles in the regulation of many processes, including DNA synthesis, DNA repair and budding.

Figure 5 displays our four motif clusters, associated with MCB and SCB, along with their EC patterns. 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 (see Methods section). 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 estimated the incremental rate of binding of TFs to the set of promoters of each cluster (see Methods section and Fig. 4). As before, it appears that the four clusters at hand show a significantly high binding rate to Mbp1, Swi4 and Swi6.

Combining the information of known motifs, GO annotation enrichment and the binding of TFs to the genome, we concluded the following; The first cluster, MS1, contains 'classic' MCB and SCB binding sites 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 binding sites, 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. 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. We conclude that 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 (Fig. 5) and for leading MS2 in its effect on the menadione and hydrogen peroxide experiments. Clusters MS1 and MS2 provide support for a known difference in binding preferences between MBF and SBF, and proof of concept for our method's ability to distinguish between two highly similar motif clusters.

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

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.

3.4 STRE clusters

Another demonstration of the importance of specific sequences of TFBS is observed in two clusters, ST1 and ST2, which have

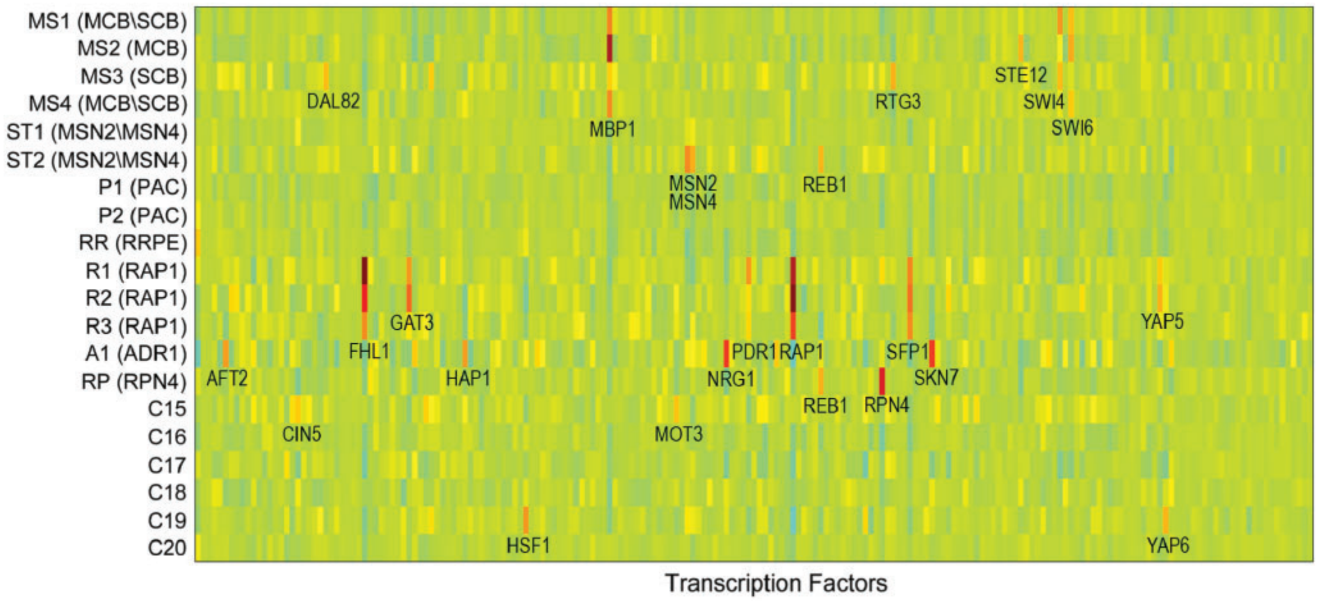


Fig. 4. Incremental binding rates of each of the 203 transcription factors (columns) to every cluster (rows). Hot colors (dark red) represent high incremental binding rates. The first four clusters (MS1–MS4) have large overlaps with well-known TFBS, such as those bound by MBF and SBF. The first is a 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 and RRPE have not yet been discovered, as is also reflected by the lack of signal for the clusters P1, P2 and RR.

been identified as STREs (Stress Response Elements). STREs are known to be bound by two related TFs, MSN2p and MSN4p. These two Cys2His2 zinc finger proteins are known to take part in regulating the expression of many stress-related genes.

Cluster ST1 has high overlap with well-known binding sites of MSN2p and MSN4p. The sequences composing cluster ST2 show sequential similarity to the known binding sites of MSN2p and MSN4p yet have not been previously identified as STRE.

Furthermore, the genes belonging to the promoters on which the two clusters are found are highly enriched with GO annotations such as stress response, energy reserve metabolism, sporulation and more. This is in agreement with the known roles of MSN2p and MSN4p in the regulation of stress-related genes. It appears (Fig. 4) that while ST2 shows high binding rates to both MSN2p and 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.

As expected, the EC pattern of ST1 is especially rich for stress-related conditions (Fig. 6). Although similar in tendency to ST1, the EC pattern of ST2 is not as strong as that of ST1.

3.5 PAC clusters

The third group of clusters contains P1 and P2. P1 has large overlap with Polymerase A and C (PAC) motifs, which are known to be involved in the regulation of ribosomal genes. P2 contains motifs bearing some similarity to known PAC motifs, though some of them have not been previously identified as such. The identity of both P1 and P2 was further validated through the GO annotations analysis of the relevant genes, pointing mainly to the biogenesis of the ribosome. The genes associated with both clusters have not been found to be significantly bound by any of the 203 transcription factors tested by Harbison *et al* (Fig. 4). This is not surprising, since the TF that binds PAC motifs is unknown.

The EC patterns of both P1 and P2 (Fig. 7) are extremely rich in significance at most conditions. This agrees with the fact that PAC regulates many ribosomal genes, thus affecting numerous cellular processes. Yet, while the EC patterns of P1 and P2 are similar in their tendencies, they are different in potency (Fig. 7).

3.6 Mechanisms determining strength of regulation

Genes that are regulated by the same TF often display various expression levels. This is motivated biologically 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 and (4) interactions between different TFs (Sudarsanam *et al*., 2002). A combination of these causes may control the high variability in gene expression as well as act as a fine tuner.

We expect the last cause to be of secondary importance in our clusters analysis, since there exist only small overlaps between genes that carry motifs of two different clusters (Supplementary Fig. 9.1). We further analyzed the first three possible causes for the clusters at hand, to determine which is

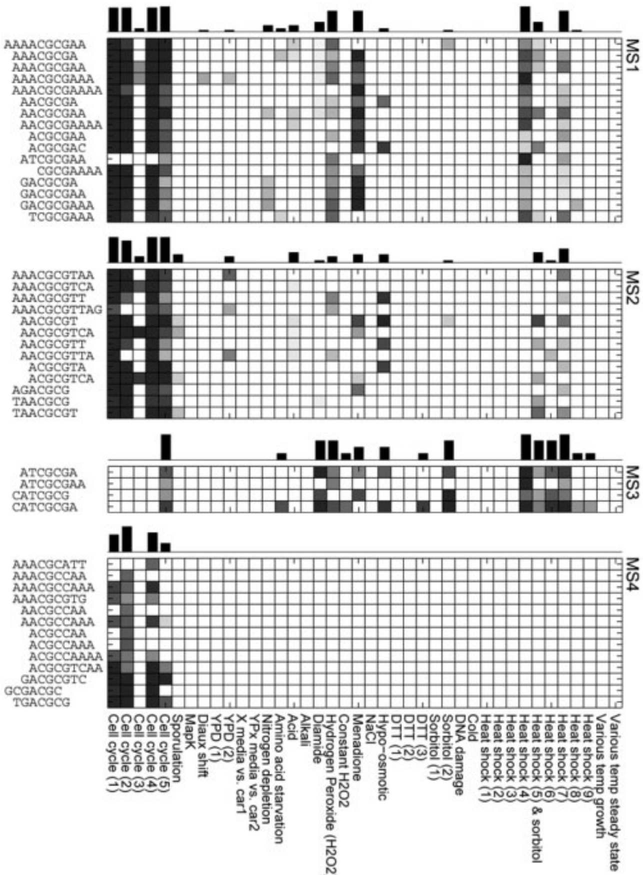


Fig. 5. Four of our clusters contain motifs that are known MBF and SBF recognition elements (Top to bottom: MS1, MS2, MS3 and MS4). Each matrix represents the EC patterns of the motifs within one cluster. The EC pattern of a motif is a vector of 40 *P*-values of EC scores for 40 biological conditions (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 EC scores in each biological condition.

relevant to the different regulation effects observed in Figures 5–7.

For each motif within each cluster, we tested the distribution of the number of its appearances on the promoters in comparison to the distribution of randomly sampled motifs. The background model was based on our 694 motifs. It appears that the distributions of number of appearances of the clusters' motifs on the promoters are not significantly different from that of the background model. Furthermore, no significant differences in motif appearance numbers were detected between clusters.

A second analysis tested the localization of motifs of each cluster along the promoters. Figure 8 displays histograms of motif distances from the translation start site. These distributions are compared with a background model (see Methods section). In most cases, both the clusters and the background model display peaks at about 140 bp upstream to the genes' translation start site. This position preference is characteristic to our set of motifs and is not apparent in sets of random sequences sampled from the promoters.

The motifs of clusters MS1 and MS2, for example, have a similar number of appearances per promoter. Furthermore,

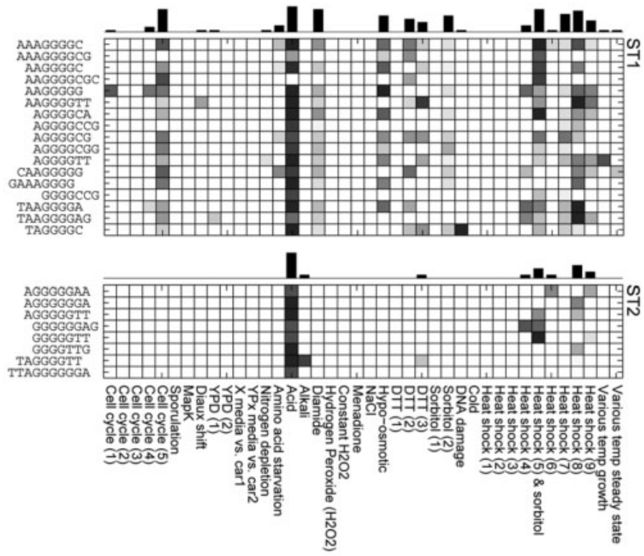


Fig. 6. Matrices of EC patterns for the two clusters ST1 and ST2. These clusters contain motifs that are identified as STRE, to which MSN2p and MSN4p bind, regulating the expression of stress-related genes.

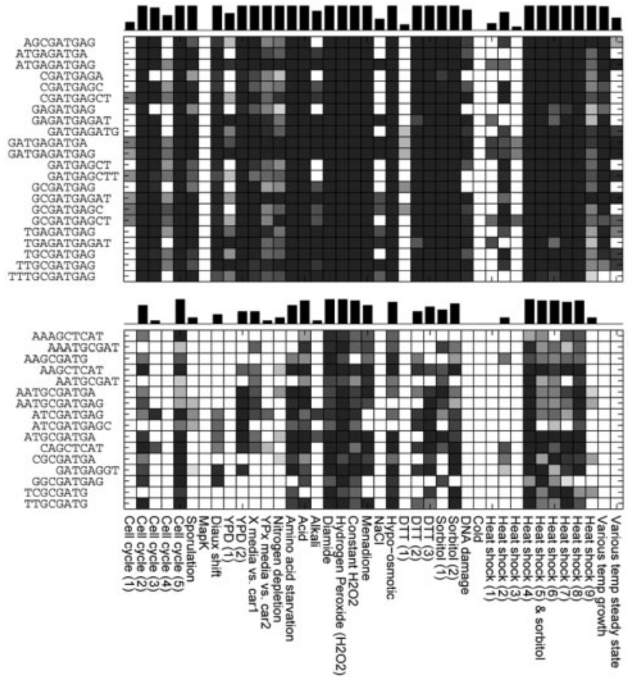


Fig. 7. Matrices of EC patterns for clusters P1 and P2. The upper cluster (P1) contains known PAC motifs, while most of the motifs of the lower cluster (P2) have not yet been described. The EC patterns of the two clusters are significantly rich. This agrees with the fact that PAC regulates many ribosomal genes, hence affect numerous cellular processes.

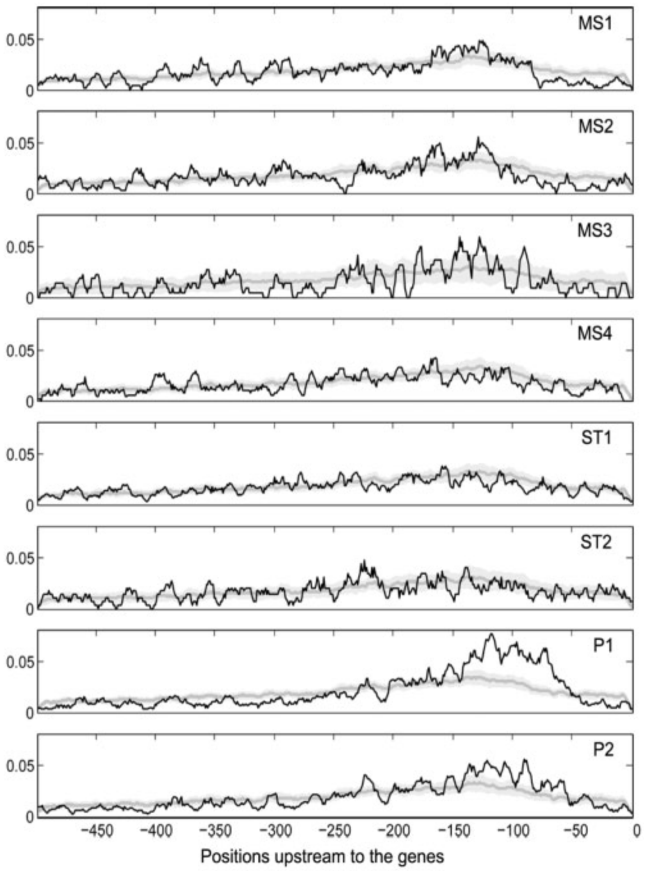


Fig. 8. Localization of motifs on the promoters of several clusters. The black lines indicate, for each position upstream to the genes (up to −500 bp), the percentage of promoters on which the cluster's sequences have been found. This can be compared to the localization of groups of motifs (of the same sizes as those of the clusters in question), randomly sampled from the set of 694 motifs. For each cluster, the dark gray line shows the mean motif occurrence per position over 1000 such sampled groups, while the light gray area represents the samples' SD of occurrences per position.

these motif occurrences are distributed similarly to the background model. Additionally, the analysis of motifs' localization, depicted in Figure 8, does not provide any distinction between these two clusters. Hence, we infer that the difference in their functional behavior in some of the tested conditions, displayed in Figure 5, is caused by stronger binding mechanisms of motifs in MS1, i.e. of motifs with the core ACGCGA. Such binding mechanism may be the result of either a specific binding affinity of the TF to the TFBS, or due to conformational changes of the TF while bound to a specific TFBS (Leung *et al.*, 2004). 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. We conclude that for these four clusters, changing a single nucleotide in a TFBS may have a strong impact on the binding mechanism of the TF to the promoter.

A similar story seems to hold for the STRE clusters. Once again, their differences are neither due to different motif copy numbers nor due to positional preferences across these promoters (Fig. 8). Hence, once again, we attribute slight functional differences to small changes in motif nucleotide compositions which may lead to differences in binding mechanisms of the TFBS.

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

4 DISCUSSION

The conventional representation of motifs by PSSMs encapsulates the sequential information of a cluster of aligned motifs. The simplicity of such representation leads, however, to possibly wrong conclusions. Mononucleotide frequency weight matrices cannot accurately depict the binding site specificities of their included motifs (Bulyk *et al.*, 2002).

Here we started out with single motifs, as extracted by MEX from sequential data, and assessed for potential regulatory function by the EC analysis. MEX does not use either PSSM nor 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.

The main strengths of the method presented here are: (i) it is an *ab initio* motif finder, which infers specificity from sequence alone without prior TFBS knowledge, (ii) it is applicable to whole genomes and does not require prior grouping of promoters, or their alignment, (iii) it captures inter-position dependencies, (iv) it is scalable to larger genomes and (v) it combines sequence with expression data, assigning likely biological functions to the extracted motifs. As such, the current methodology may be complementary to other *ab initio* algorithms like MEME (Bailey and Elkan, 1995), AlignAce (Roth *et al.*, 1998), Weeder (Pavesi *et al.*, 2004), a leading *k*-mer based approach and YMF (Sinha and Tompa, 2003), which search for over-represented motifs in pre-defined clusters of co-regulated genes.

The previously reported approach that perhaps most resembles ours is MobyDick (Bussemaker *et al.*, 2000), which identifies significant 'words' in genome-wide promoters, based on a maximum-likelihood search for over-represented *k*-mers. Whereas we have concentrated on deterministic motifs with a minimal length, they have set an upper-bound on their *k*-mers, which were later expanded by using gapped motifs. Such higher motif structures can also be obtained using the ADIOS algorithm (Solan *et al.*, 2005) that generalizes MEX. Limiting ourselves to deterministic ungapped motifs allowed us to reach our main conclusion regarding the importance of single nucleotide variations.

Our motif clusters were analyzed in various manners and their relationships to known TFs were tested. We have left these clusters in the form of groups of motifs, rather than combining them into PSSM or consensus representations, because we learned from our analysis that single changes of a nucleotide in a motif can go a long way in affecting the biological behavior.

In several cases, we obtain few motif clusters that contain elements of several known TFBS groups. Examples are clusters MS1 to MS4 that contain motifs traditionally labeled as MCB and SCB (belonging to the TFs 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*., whereas they are scattered among four of our clusters, MS1–MS4.

Differences in EC patterns imply different regulation strengths associated with the relevant motifs in various sets of biological conditions. Regulation strength depends on various mechanisms. We 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 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 (or of its co-factors) while bound to a specific TFBS, but may also come about because of the existence of different TFs (or co-factors) competing over similar TFBS. For instance, it has been reported that binding site sequence variations may cause the bound TF to adopt different conformations, directing interactions with specific co-factors and resulting in different expression responses (Lefstin and Yamamoto, 1998).

Since both MCB and SCB are bound by protein complexes, one may hypothesize that the differences in the biological conditions regulated by clusters MS3 and MS4 result from different compositions of the complexes. Comparing Figure 5 to Figure 4, one may suggest that MS4 has very weak or no binding to Swi4 and this may be the reason why no effect is observed in all stress conditions. MS3 has weak binding to Mbp1 and this may be the reason for the absence of effects on four of the cell cycle experiments.

Our PAC clusters, P1 and P2, show different EC patterns. P1 both shows higher EC significance and displays positional bias along the promoter. The latter may perhaps be correlated with the loci of nucleosomes on the DNA (Segal *et al*., 2006), affecting the strength of the regulation. We presume that in this case this is one of the reasons for the much higher regulation strength of P1 motifs.

Most of the P2 motifs were not mentioned in the literature, presumably because the effects of P1 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-representation within a given class of genes. Hence our method was able to uncover clusters of motifs that have been overlooked by others.

ACKNOWLEDGEMENTS

This research was partially supported by the US Israel Binational Science Foundation and by the Israel Science Foundation. Y.P. is an incumbent of the Rothstein Career Development Chair in Genetic Diseases. M.L. is a Fellow of the Horowitz Foundation for Complexity Sciences. Y.P. acknowledges support from the Ben May Charitable Trust. L.S. and D.H. thank M. Kupiec for helpful discussions.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology, **25**, 25–29.

Bailey,TL. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **3**, 21–29.

Barash,Y. *et al*. (2003) *Modeling Dependencies in Protein-DNA Binding Sites*. RECOMB, Berlin, Germany, pp. 28–37.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., Series B (Methodological)*, **57**, 289–300.

Benos,P.V. *et al*. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*., **30**, 4442–4451.

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins: statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol*., **193**, 723–743.

Boyle,E.I. *et al*. (2004) 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**, 3710–3715.

Brazma,A. *et al*. (1998) Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res*., **8**, 1202–1215.

Bulyk,M.L. *et al*. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*., **30**, 1255–1261.

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

Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol*., **19**, 1114–1121.

Durbin,R. *et al*. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Gasch,A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Koch,C. *et al*. (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science*, **261**, 1551–1557.

Lawrence,C. *et al*. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lapidot,M. and Pilpel,Y. (2003) Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription. *Nucleic Acids Res*., **31**, 3824–3828.

Lefstin,J.A. and Yamamoto,K.R. (1998) Allosteric effects of DNA on transcriptional regulators. *Nature*, **392**, 885–888.

Leung,T.H. *et al*. (2004) One nucleotide in a [kappa]B site can determine cofactor specificity for NF-[kappa]B dimers. *Cell*, **118**, 453–464.

Pavesi,G. *et al*. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res*., **32**, W199–W203.

Pilpel,Y. *et al*. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet*., **29**, 153–159.

Pritsker,M. *et al*. (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation 10.1101/gr.1739204. *Genome Res*., **14**, 99–108.

Roth,F.P. *et al*. (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol*., **16**, 939–945.

Segal,E. *et al*. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.

Shalgi,R. *et al*. (2005) A catalog of stability-associated sequence elements in 3′ UTRs of yeast mRNAs. *Genome Biol*., **6**, R86.

Sinha,S. and Tompa,M. (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*., **31**, 3586–3588.

Solan,Z. *et al*. (2005) Unsupervised learning of natural languages. *Proc. Natl Acad. Sci. USA*, **102**, 11629–11634.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Sudarsanam,P. *et al*. (2002) Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Res*., **12**, 1723–1731.

Tavazoie,S. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet*., **22**, 281–285.

Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–41.