Identifying regulatory networks by combinatorial analysis of promoter elements

Yitzhak Pilpel[1]*, Priya Sudarsanam[1]* & George M. Church[1]

*These authors contributed equally to this work.

Published online: 10 September 2001, DOI: 10.1038/ng724

Several computational methods based on microarray data are currently used to study genome-wide transcriptional regulation. Few studies, however, address the combinatorial nature of transcription, a well-established phenomenon in eukaryotes. Here we describe a new approach using microarray data to uncover novel functional motif combinations in the promoters of *Saccharomyces cerevisiae*. In addition to identifying novel motif combinations that affect expression patterns during the cell cycle, sporulation and various stress responses, we observed regulatory cross-talk among several of these processes. We have also generated motif-association maps that provide a global view of transcription networks. The maps are highly connected, suggesting that a small number of transcription factors are responsible for a complex set of expression patterns in diverse conditions. This approach may be useful for modeling transcriptional regulatory networks in more complex eukaryotes.

Introduction

The regulation of gene expression in eukaryotes is highly complex and often occurs through the coordinated action of multiple transcription factors. Examples of this combinatorial transcriptional control have been described for several organisms[1–6]. Combinatorial regulation of transcription has several advantages, including the control of gene expression in response to a variety of signals from the environment and the use of a limited number of transcription factors to create many combinations of regulators whose activities are modulated by diverse sets of conditions.

The customary approach to analyzing microarray data[7–11] does not explicitly address the combinatorial nature of transcriptional regulation. Here, however, we have performed an extensive study to identify synergistic motif combinations that control gene expression patterns in *S. cerevisiae* (Fig. 1*a*). We analyzed microarray expression data to screen for statistically significant motif combinations. This combinatorial analysis was incorporated into a new analytic model that explores the effect on gene expression patterns of adding or subtracting motifs from particular motif combinations. We identified several novel motif combinations that seem to be directly responsible for particular expression patterns during the cell cycle, sporulation and various stress-response conditions. We have also generated motif synergy maps that display the motif associations discovered in this study. These maps provide a global view of the connections between regulators of the transcriptional networks within the cell in different conditions.

Results

Identification and analysis of motif combinations

To identify motif combinations that control gene expression patterns, we first established a database of known and putative regulatory motifs and used ScanACE[12] to identify all the genes in the *S. cerevisiae* genome containing each motif in their promoters (Fig. 1*a*). We then used the expression profiles of genes whose promoters contained the particular motif or motif combination to evaluate the effect of each motif on gene expression. For each motif or combination, we calculated the expression coherence score, a measure of the overall similarity of the expression profiles of all the genes containing that motif, in several different conditions, including different stages of the cell cycle[13], sporulation[14], diauxic shift[15], heat and cold shock[16], and treatment with DTT[16], pheromone[17] and DNA-damaging agents[18] (see Web Table A for a list of expression coherence scores). We used a working statistical definition of motif synergy, distinct from its use in the experimental context, to identify functional motif combinations. A pair of motifs was considered 'synergistic' if the expression coherence score of genes containing both motifs in their promoters was significantly greater than that of genes containing either motif alone (Fig 1*b*). We computed motif synergy scores for all pairs of motif combinations in the current database (Web Table B).

We identified several experimentally established transcriptional motif associations in our analysis. Sites for Mcm1 and SFF, known to control transcription of some G2-specific genes[19,20], are synergistic in the cell-cycle data set at the appropriate phase of the cell cycle (time points 7 and 14; Fig. 1*b*). In addition, the Mcm1-Ste12 (ref. 21), Bas1-Gcn4 (ref. 22) and Mig1-CSRE (ref. 23) motif combinations, known to interact functionally at some promoters, are predicted here to be synergistic. We also observed, by studying the effect of DNA-damaging agents, that the sites for the factors Abf1 and Rpn4 are synergistic. Both factors have previously been independently implicated in regulating transcription during nucleotide excision repair[18,24]; however, there have been no reports of a functional interaction between them.

[1]*Department of Genetics and Lipper Center for Computational Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. Correspondence should be addressed to G.M.C. (e-mail: church@salt2.med.harvard.edu).*

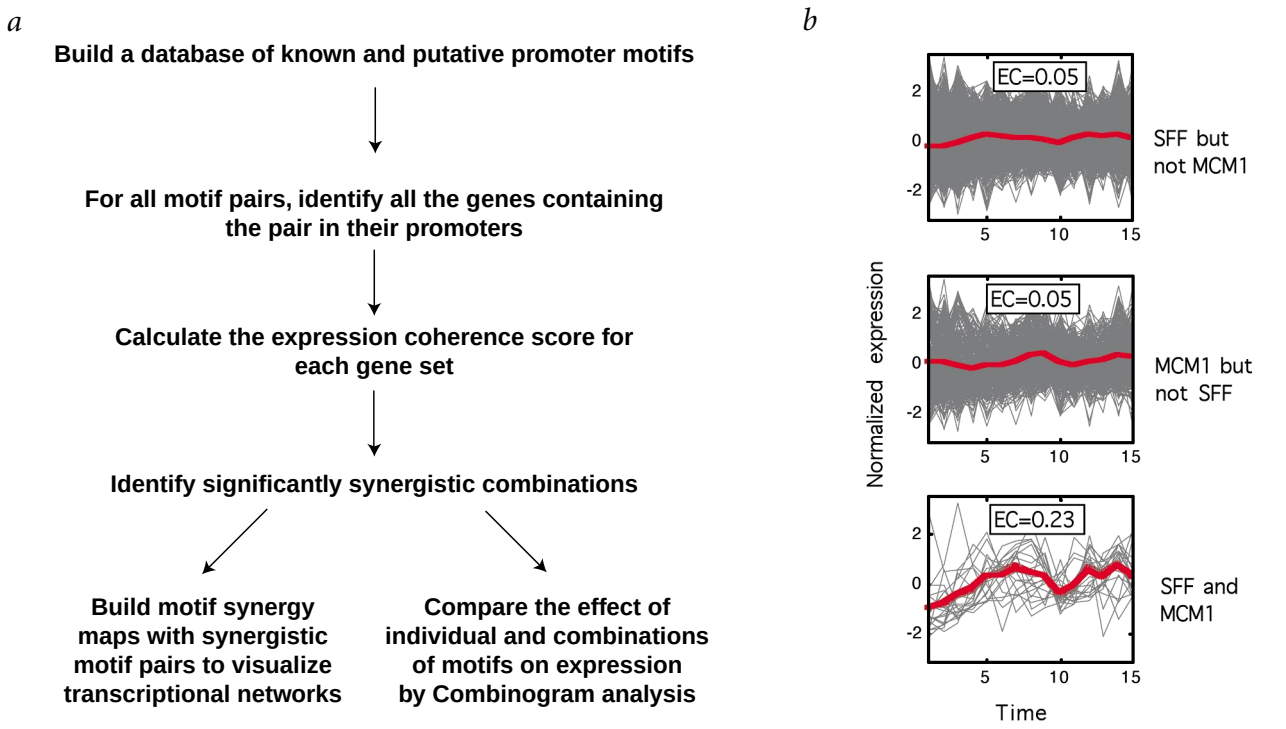

Fig. 1 ***a***, Strategy used to discover and analyze synergistic motif combinations. ***b***, Expression profiles of genes containing the motifs for Mcm1 and/or SFF. In each panel, each grey line represents the normalized expression profiles of an individual gene defined by the indicated motif(s) during the cell cycle. The average expression profile of all the genes displayed in the panel is shown in red. The expression coherence score (EC) for each group of genes is also shown.

Finally, the fact that Rap1 synergizes with different partners in several conditions is consistent with its broad role in controlling transcription in *S. cerevisiae*[25].

Among the new synergistic motif combinations identified in our analysis is a combination composed of the mRRPE (also known as M3a[10]) motif[12], derived from the MIPS rRNA-processing functional category using the motif-finding algorithm AlignACE[12,26], and PAC (also known as M3b[10]), a motif found upstream of many DNA polymerase A and C genes (Table 1)[27]. Both mRRPE and PAC have been identified from the same expression cluster in analyses of cell-cycle[10] and stress response[28,29] microarray data sets, but these studies did not capture the impressive synergy between the two motifs. Our results indicate the power of combinatorial analyses of microarray data compared with the current approach of clustering expression data and then applying motif-finding algorithms[9–11]. As the two motifs also co-occur significantly in the genome, particularly upstream of genes involved in rRNA transcription and processing (Y.P., P.S. and G.M.C., unpublished data), this combination may be biologically significant and worthy of further experimental verification.

To assess the effect of motif combinations on expression coherence, our analysis simply requires that the combinations co-occur in the same promoter; however, it does not address certain other parameters, such as orientation or position of motifs within promoters, that often influence motif function[30]. We further analyzed synergistic motif pairs for preferences in relative locations within promoters. We tested the hypothesis that, for each motif pair, one motif tends to be located closer to the translational Start site than the other. Detailed analysis of the highly synergistic motif pair of PAC and mRRPE (Fig. 2*a*, left) shows that mRRPE is found preferentially closer to the translational Start site to a statistically significant extent (P=0.002). Among the 79 promoters containing a single copy of PAC and mRRPE, mRRPE is closer to the Start site in 51 cases. By contrast, for a random, typical motif pair, designated M1 and M2, we found no such significant bias (P=0.11) in 26 cases studied (Fig. 2*a*, right): M1 is closer to Start in 14 promoters and M2 in 12. We extended this analysis to each of the 115 synergistic pairs identified in the study. Synergistic pairs have a significant tendency to display an orientation bias when compared with a random control set of motif pairs (P=10^{-14}; Fig. 2*b*). For instance, we found a significant (P<0.05) orientation bias in approximately 18% of the synergistic pairs, compared with only approximately 6% of the pairs in the control random sample. These results indicate that motif orientation is important for the function of synergistic motif combinations.

Table 1 • Selected synergistic motif pairs

Motif 1	Motif 2	Conditions
Mcm1	SFF	cc
Mcm1	Ste12	spo
Gcn4	Bas1	hs
Mig1	CSRE	hs
Rap1	mRPE6	cc spo hs dd
Rap1	CCA	spo hs dd
ECB	SFF	cc
MCB	SCB	pr
PAC	mRRPE	cc spo hs dd
PAC	mRRSE3	hs
SCB	SFF	spo
Mcm1	mDNAMetE4	cc
mRRPE	mRRSE3	ds
STRE	mPROT18	hs
Rpn4	Abf1	dd

cc, cell cycle; spo, sporulation; ds, diauxic shift; hs, heat shock; pr, pheromone response; dd, DNA-damaging agents.

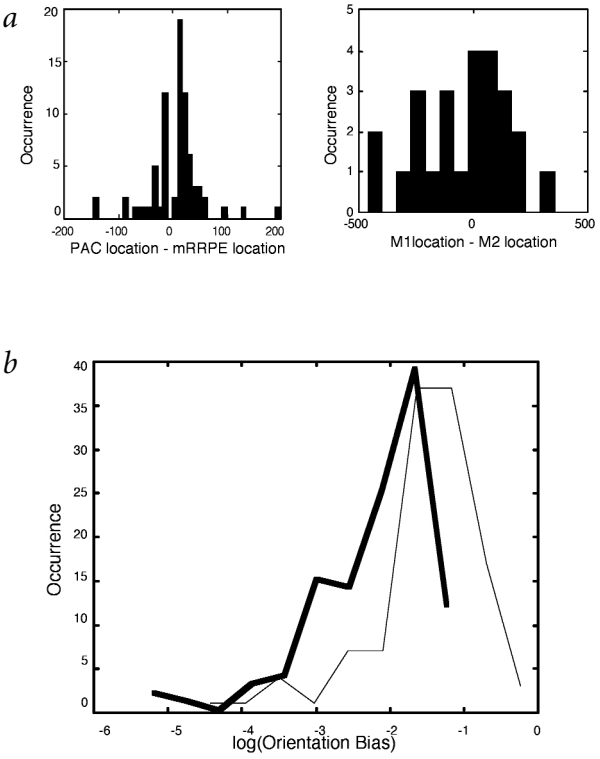

Fig. 2 The effect of relative motif orientation on motif synergy. ***a***, Orientation analysis of the most synergistic motif pair, PAC and mRRPE (left), and of two randomly chosen motifs designated M1 and M2 (corresponding to motifs number 352 and 169 on the motif list at http://genetics.med.harvard.edu/~tpilpel/MotComb.html; right). Shown are the distribution of differences between the locations (relative to the translational Start site) of PAC and mRRPE in promoters containing single copies of each motif (left) and the distribution of differences between the locations of M1 and M2 in promoters containing single copies of each motif (right). In the PAC-mRRPE pair, mRRPE is found preferentially closer to Start, whereas in the random pair, a more balanced distribution is seen. We calculated an orientation bias statistic using a cumulative binomial probability for the probability of obtaining more or the same extent of bias by chance, assuming no *a priori* bias; the probability for the observed orientation bias is 0.002 for mRRPE and PAC and 0.14 for M1 and M2. ***b***, Histograms of the logarithm of the orientation bias scores for all 115 synergistic motif pairs (thick line) and for a random control set (thin line) of 115 motif pairs. The *P* value for the hypothesis that the two histograms are identical is 10^{-14} according to the Wilcoxon rank sum test.

A global map of yeast combinatorial control

To discern higher-order interactions between transcription regulators of different cellular processes, we generated a motif synergy map depicting the functional associations between motifs discovered in this study. The map shows a fairly high degree of connectivity, with all the nodes in one connected cluster (Fig. 3). This is a consequence of the numerous synergistic interactions formed by a few motifs for factors such as Rap1, Abf1, SFF and CCA. This suggests that a small number of transcription factors associating in various combinations may be sufficient to control a wide variety of expression patterns in *S. cerevisiae* under different conditions.

In addition to indicating particular regulatory interactions in a specific condition, the motif synergy maps also display global connections between different experimental conditions. As seen in the map (Fig. 3), some motifs are striking in their ability to synergize with different motifs in many conditions. The Rap1 motif, for example, forms synergistic combinations with different motifs in almost every condition studied here. This finding shows that a single motif can affect transcription in multiple conditions by participating in different combinations in each condition, and is consistent with the broad role of Rap1 in controlling transcription in *S. cerevisiae*[25].

Another interesting property of the motif synergy map is that motifs controlling similar cellular pathways seem to cluster together; that is, they form synergistic combinations with each other (Fig. 3). For example, several cell cycle–specific motifs, including sites for Mcm1, SFF, Swi5 and the ECB box, synergize with one another. Similarly, several motifs that regulate the transcription of amino acid–biosynthetic genes such as Bas1, Gcn4 and Lys14 form functional associations. These results suggest that our approach uncovers motif combinations that are likely to interact functionally with each other by controlling transcription of similar pathways.

Exploring the causal relationship between motifs and expression patterns

In addition to identifying motif associations, we also tried to determine the influence of each motif in a combination on the observed expression pattern. For example, we asked whether motif combinations that have some motifs in common give similar expression patterns, which would suggest that the shared motifs may be important for determining the expression profile. In addition, for a particular synergistic motif pair in a given experimental condition, it was unclear whether one motif is more critical in determining the pattern of expression or if both motifs in the combination contribute equally. One way to investigate the impact of individual motifs is to add or remove motifs from a given combination and assess the effect of each set of motifs on expression coherence. This may predict whether each motif is necessary and/or sufficient for the particular expression pattern and indicate causal links between the motifs and the expression profiles.

To simultaneously assess expression coherence and the similarity between expression patterns of different motif combinations, we developed the Combinogram workbench, an integrated set of computational tools for the analysis and visualization of relationships between regulatory motifs and expression profiles (Figs. 4 and 5). The analysis is initiated with a collection of *n* (usually 5–20) motifs whose effect on gene expression in a particular expression condition needs to be characterized. Each gene in the genome is assigned a binary signature 'a string of 1s and 0s' indicating the presence or absence of each of the *n* motifs in its promoter. All the genes in the genome with the same motif signature are combined into a gene set defined by a motif combination (GMC). A GMC for a particular motif combination is thus defined by all the genes that have the combination but not any of the other motifs in the set. To explore the effects of individual motifs, we generated all possible GMCs in the motif set and calculated the expression coherence score for each GMC. In addition, we determined the average expression profiles of all the GMCs, grouped them in clusters based on the similarity between the profiles, and depicted them in a dendrogram.

Cell cycle and sporulation combinatorial controls

Our analysis of synergistic motif combinations shows several interesting associations between cell cycle motifs as well as regulatory cross-talk between the two processes of cell cycle and sporulation—for example, the synergy observed for the SCB-SFF motif pair during sporulation (Table 1). We therefore carried out a Combinogram analysis of the known cell-cycle and sporulation motifs to identify their roles in both the cell cycle (Fig. 4*a*) and sporulation (Fig. 4*b*). Expression profiles of GMCs containing the MCB motif, which is known to be important for transcription during G1 (ref. 31), are very similar and cluster together in the dendrogram section of the cell-cycle Combinogram (Fig. 4*a*). The Combinogram predicts that the MCB motif is both necessary and sufficient to invoke the G1-specific expression pattern: MCB is the only motif common to all the GMCs in the G1 expression cluster, and the GMC containing MCB alone is a member of the cluster.

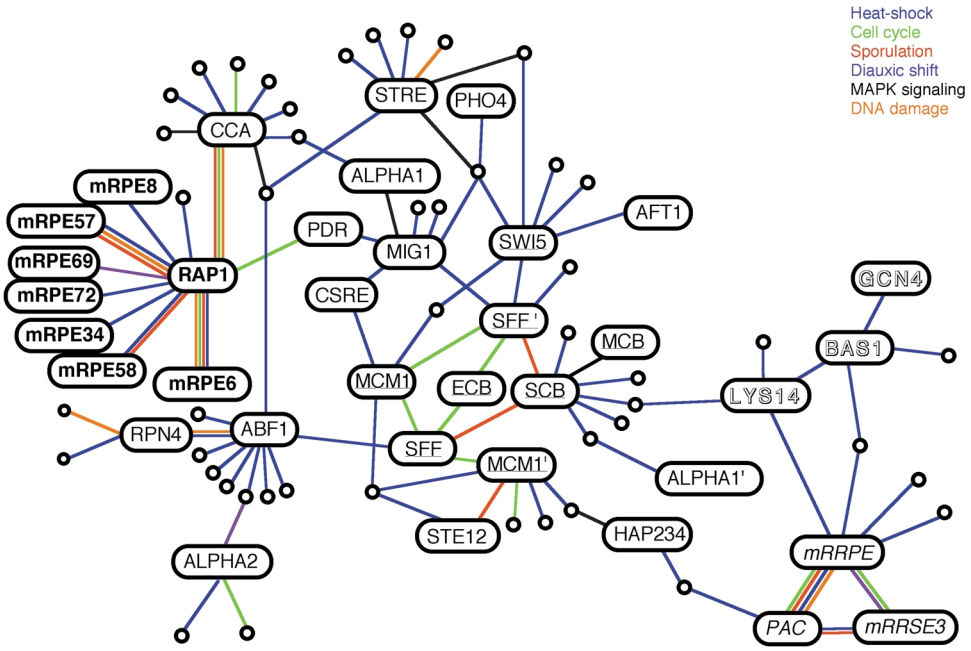

Fig. 3 Global motif synergy map. The nodes in these graphs represent known or putative motifs and are indicated either by a small black circle or by an oval containing the name of the motif. Names of putative motifs begin with the letter "m" and indicate the MIPS functional category from which they were derived: mRPE, ribosomal protein element; mRRPE, rRNA processing element; mRRSE, rRNA synthesis element. The symbol ' following a motif name indicates a variant of the motif found in the literature that was generated by running AlignACE on the promoters of genes known to be regulated by the motif. Motifs bound by a known protein are indicated by the name of the protein in capitals. Lines connect motif pairs that synergized significantly in at least one of the seven expression experiments; line colors indicate the expression condition(s) in which the motif pair had a significantly high synergy score (upper right). Some motif names are marked according to the function of the genes they regulate or the MIPS functional category from which they were derived: bold face, ribosomal proteins; italics, rRNA transcription/processing/synthesis; underlined, cell cycle; shadow, relating to amino-acid biosynthesis.

Combinograms also show the influence of other motifs on the expression pattern characteristic of a particular motif. For example, although most MCB-containing GMCs display a primarily G1-specific profile, they also contain some genes with G2-specific expression (data not shown). However, the MCB-SFF′ (′ indicates a variant of the motif found in the literature; see Fig. 3) GMC is the most coherent combination in the cell-cycle Combinogram (Fig. 4*a*), with almost all the genes peaking only in G1. In genome-wide chromatin immunoprecipitations (ChIp) carried out with the factors Mbp1 and Forkhead1, members of complexes that bind the MCB motif and the SFF motif respectively, a significantly large number of promoters were precipitated by both factors (I. Simon and R. Young, personal communication). This is consistent with the functional interaction between the MCB and SFF motifs predicted by the present study.

GMCs containing the SFF′ motif with other motif partners are grouped away from the G1-specific expression cluster in the Combinogram and have primarily a G2-specific pattern (Fig. 4*a*) that is consistent with previous experimental evidence for the regulatory role of the SFF complex[19,20]. The MCB-SFF' GMC is part of the MCB expression cluster, indicating that the presence of the SFF motif does not change the G1-specific expression pattern defined by the MCB motif. These results suggest that SFF acts as an activator during G1, which is consistent with observations that the SFF complex is constitutively bound to its sites throughout the cell cycle[20]. If the expression profiles observed in the GMCs defined by MCB (G1-specific) or SFF-containing motifs (G2-specific) are characteristic of these motifs, the G1-specific expression profile observed in the MCB-SFF GMC suggests that the MCB motif is more dominant than the SFF motif in determining this expression pattern. Alternatively, it is possible that SFF acts as a repressor of genes containing the MCB-SFF combination during G2 (ref. 20).

In the sporulation data set, the MSE, a motif bound by the sporulation factor Ndt80, seems to be a major determinant of expression patterns. Combinogram analyses also reveal the unexpected influence of cell cycle motifs in controlling transcription during sporulation (Fig. 4*b*). The expression profiles of three out of the four MSE-containing GMCs are tightly clustered, with profiles characteristic of mid-sporulation. This cluster includes the GMC containing MSE alone, indicating that the MSE site is sufficient for establishing the particular expression pattern. However, the same cluster includes a GMC defined by the SCB and SFF′ sites (Fig. 4*b*), suggesting that the factors that bind the SCB and SFF sites can serve as alternative regulators of the mid-sporulation response.

The sporulation Combinogram also displays the effect of two other cell cycle motifs, MCB and SCB, on expression during sporulation. The MCB alone and the MCB-SCB GMCs cluster together with highly similar expression patterns that peak at 2 h into sporulation (Fig. 4*b*). We used the motif-finding algorithm AlignACE[12] to analyze the promoters of all the genes with expression profiles similar to genes in the MCB-SCB GMC (data not shown). We found that MCB is the only significant motif, which suggests that no other motif contributes as much as MCB to this expression pattern. (Our criteria for significance are that both the MAP and –log (group specificity score) exceed 10; ref. 12.) Although we predict that the MCB motif is both necessary and sufficient for this pattern, the presence of the SCB motif seems to substantially improve the expression coherence of the gene set. Our results are consistent with two recent studies that also suggest that these motifs control gene expression during sporulation[32,33] and with previous evidence of a role for Swi6, a member of transcription complexes that bind MCB and SCB, during meiotic recombination[34].

Stress response regulators

Combinograms can also be used to explore the regulation of gene expression in experimental conditions where there is limited knowledge about relevant regulatory motifs. We used Combinograms to analyze motifs involved in synergistic

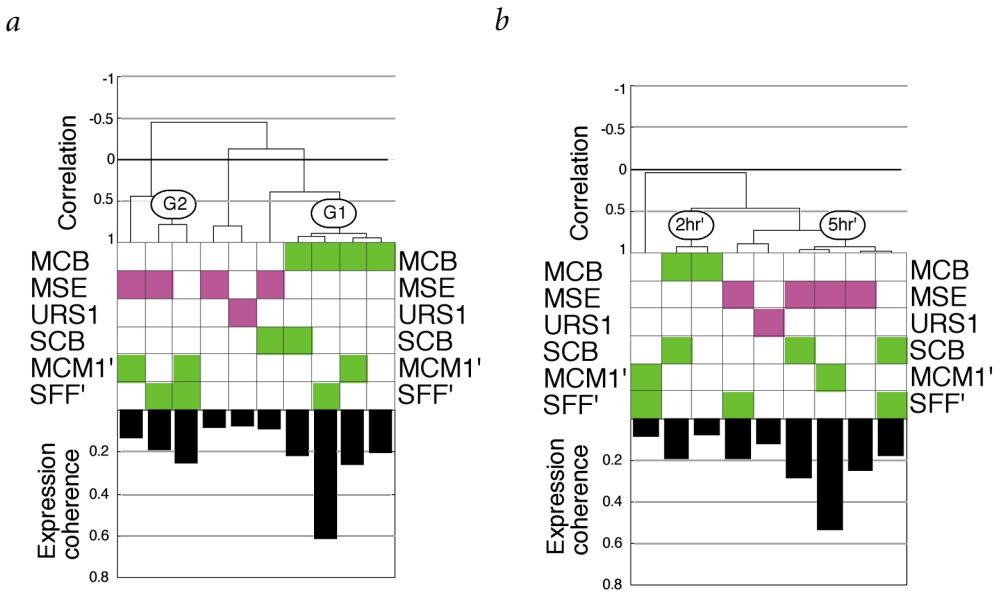

Fig. 4 Combinograms of cell-cycle- and sporulation-related motifs. ***a***, Cell-cycle[13] data set. ***b***, Sporulation[14] data set. The middle section of the Combinogram shows the motif composition of each GMC. Each vertical column represents a single GMC. A colored square indicates that the particular motif is present in the promoters of all the genes in that GMC. A white square indicates that none of the genes in the GMC contain the particular motif. Motifs known to control transcription during the cell cycle or sporulation are green or magenta, respectively. Only GMCs that passed the thresholds imposed for expression coherence score (EC=0.075) and the number of genes in the GMC (at least 10 genes) are shown, to balance the sensitivity and specificity of the Combinogram displays. The top section of the graph shows the dendrogram analysis that assesses the similarity in expression profiles of each GMC using Pearson correlation coefficients between the average expression profile of the genes in the GMC as a measure of distance. G1 and G2, and 2 h and 5 h, indicate GMC clusters that predominantly peak in the G1 and G2 phases of the cell cycle and at 2 h and 5 h into sporulation, respectively. The bottom section of each graph shows the expression coherence scores for each GMC. GMCs containing the cell cycle motifs Swi5 and ECB were included in the analysis but did not pass the thresholds.

combinations in two different stress response conditions, heat shock[16] and treatment with DNA-damaging agents[18] (Fig. 5). In both cases, GMCs with similar motif composition are clustered in the dendrogram. For example, both Combinograms show three distinct clusters (Fig. 5) consisting of GMCs defined (i) by ribosomal protein motifs (Rap1 and mRPE6), (ii) by rRNA transcription and processing motifs (PAC, mRRPE, mRRSE3 and mRRSE10), and (iii) by environment-specific elements such as STRE and HSE during heat-shock (Fig. 5*a*) or by the motifs for the proteasome regulator Rpn4 and the activator Abf1 during DNA damage (Fig. 5*b*). The correspondence between expression clusters and the motif compositions of the GMCs indicates that under these conditions, the expression patterns observed result from the presence of these motif combinations.

The expression patterns of the ribosomal protein and the rRNA regulatory motif clusters are similar, and both are negatively correlated to that of the environment-specific cluster (Fig. 5*a*). It is possible that the corresponding protein profiles, those of ribosomal and heat-shock or proteasome proteins, respectively, are also negatively correlated. This pattern may be expected, given the opposing cellular roles of these complexes in protein synthesis and proteolytic degradation, respectively. The dichotomy in the expression response between ribosomal motifs and motifs specific to environmental condition seems to be a broad phenomenon and has been observed in several microarray experiments (http://genetics.med.harvard.edu/~tpilpel/MotComb.html). Similar observations have been made by analyzing gene expression clusters in other stress-inducing microarray studies[28].

The Combinograms also demonstrate the importance of a new motif, mRPE6. This motif is derived from the MIPS ribosomal protein category[12] and shows a high degree of expression coherence in combination with the Rap1 site in both the heat-shock (Fig. 5*a*) and DNA damage (Fig. 5*b*) data sets. In addition, it synergizes with Rap1 in multiple conditions, suggesting a potential new motif partner for modulating Rap1 function.

Discussion

The recent accumulation of microarray data has led to the development of several computational approaches for studying genome-wide transcriptional regulation. However, very few studies have addressed the combinatorial nature of eukaryotic transcription[35]. A recent study used *S. cerevisiae* microarray data to fit a linear model that describes the additive effect of oligomers on the expression levels of individual genes at particular time points[33]. The study did not, however, implement a necessary criterion for establishing synergy between motifs: comparing the expression of genes containing each motif combination with gene sets containing each of the individual motifs alone. While this criterion was not implemented in the previous study, it was instrumental in our discovery of statistically significant motif combinations. Therefore, other methods for detecting motif combinations[33,35] may uncover different types of associations than those described here.

Because our analysis provided specific examples of synergistic motif combinations, it enabled us to generate a motif synergy map that provides a global view of the functional interactions between regulators of transcription in gene networks in *S. cerevisiae*. The map contains a few 'hubs,' or nodes with many interactions, indicating that certain factors may act as global 'facilitator proteins' that assist their gene-specific partners in their function, possibly by modifying chromatin structure or targeting their partners to the promoters. Such factors may activate or repress transcription depending on the partner motif or factor and the condition, enabling a transcriptional response that integrates multiple environmental signals and pathways.

The process of deriving all the predictions in this study, including methodological and threshold choices, was unbiased by previous experimentally or computationally derived knowledge. The predictions that are confirmed by the literature may therefore be considered true positive controls. It is clear, however, that these hypotheses merit further confirmation by additional experiments; we hope that our predictions will aid future experiments. In addition, our

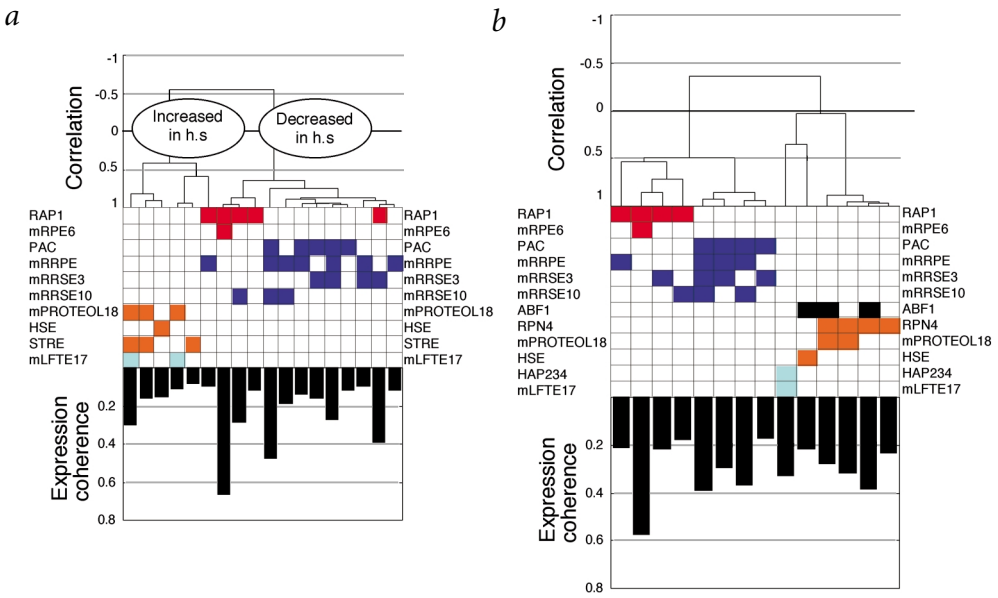

Fig. 5 Combinograms of the heat-shock[16] and the nucleotide excision repair[18] experiments. ***a***, Heat-shock data set. ***b***, Nucleotide excision repair data set. The names of putative motifs start with the letter "m" and indicate the MIPS functional category from which they were derived: mRPE, ribosomal protein element; mRRPE, rRNA processing element; mRRSE, rRNA synthesis element; mLFTE, lipid and fatty acid transport element; mPROTEOL, proteolysis. Motifs in the middle section of the diagram are colored according to the function of the genes they regulate or the MIPS functional category from which they were derived: red, ribosomal proteins; blue, rRNA transcription motifs; orange, stress related motifs; turquoise, energy production-related; black, miscellaneous functions.

use of motifs derived independently of the MIPS categories, such as motifs assembled from the literature, in many of the synergistic pairs controls for a potential circularity resulting from the fact that genes within the same MIPS category are often co-expressed[10].

The approach used in this study has several useful outcomes. First, the criterion for finding synergistic motif combinations should ensure a lower rate of false positives in defining the genes controlled by each motif. Second, the motif synergy map described here may be important for annotating the regulatory role of new motifs that co-cluster with known motifs, as motifs that affect the same cellular processes often synergize together. Third, the use of Combinograms to determine the role of each motif in a combination strengthens the link between the motif composition of promoters and the particular expression pattern. This kind of approach may be applied to predict the expression profiles of genes for which microarray data is unavailable, as is true for significant portions of the human or mouse genomes, based on similarities in promoter-motif composition. Finally, we anticipate that such combinatorial approaches will be critical for dissecting the complex architecture of transcriptional networks in more complex eukaryotes, in anticipation of an avalanche of microarray data from the human and mouse genomes.

Methods

A data set of known and putative yeast regulatory motifs. We used 356 DNA motifs, including 37 known motifs. We derived 329 motif matrices by applying AlignACE[12] to the upstream regions of genes in the MIPS[36] functional categories. The 329 motifs represent a nonredundant set selected from an initial set of 819 motifs[12] using hierarchical clustering and the requirement that the CompareACE score[12] for similarity between pairs of motifs not exceed 0.5. We chose the motif with the highest group specificity score[12] in each cluster. This set includes 25 of the known motifs. We collected the remaining known motifs from the literature and the SCPD database[37].

For each motif, we calculated the mean (M) and standard deviation (SD) of the ScanACE scores[12] of the genes used to derive the motif. We assigned motifs to the 4,483 upstream regions (URs) in the *S. cerevisiae* genome by including only those URs that score higher than M–2×SD. If more than 300 URs contained the motif, we chose only the 300 top-scoring URs. Although the choice of these particular settings is somewhat arbitrary, a detailed parameter landscape analysis indicates that choice of other threshold values from a wide range of potential settings would have had relatively little effect on the final results (http://genetics.med.harvard.edu/~tpilpel/MotComb.html). Experimental results from genome-wide DNA-protein interaction studies[32,38] may help to refine these settings in the future.

Expression coherence score. Expression data was downloaded from the expression database ExpressDB[39]. Using a given set of K genes containing a particular motif or motif combination in their promoters and an expression data set, we calculated the Euclidean distances between the mean and variance-normalized expression profiles of each of the P=0.5*K*(K–1) pairs of genes. In the case of divergently transcribed genes, both transcripts were considered. The expression coherence score, EC, associated with a motif/motif combination, is defined as p/P, where p is the number of gene pairs whose Euclidean distance is smaller than a threshold distance (D). We determined the value of D as follows: we randomly sampled 100 genes from the entire genome and calculated the Euclidean distances between their normalized expression profiles for all possible 100×99×0.5 gene pairs for a given expression data set, and then defined D as the lowest value in the fifth percentile of the distribution of these distances. Alternative thresholds give rise to qualitatively similar results (http://genetics.med.harvard.edu/~tpilpel/MotComb.html).

Synergy of motif combinations. We calculated the expression coherence (EC_L) score for genes containing L motifs in their promoters, including only combinations that occur in at least 10 genes. We calculated similar EC scores for the GMCs containing all possible subsets of L–1 motifs (excluding one motif in each iteration) and determined the maximum score ($MaxEC_{L-1}$). We used a statistical definition of motif synergy to characterize the combinations: a motif combination was 'synergistic' if EC_L was significantly higher than $MaxEC_{L-1}$. For example, motifs A and B (L=2) are 'synergistic' if genes containing motifs A and B have a significantly higher EC score than the GMC containing motif A but not motif B and the GMC containing motif B but not motif A (Fig. 1*b*).

We tested the null hypothesis that EC_L is less than or equal to $MaxEC_{L-1}$. We used a Monte Carlo procedure for two motifs A and B, where it is assumed that the gene set containing motif A has a higher EC score than that containing motif B. S(AB) and S(A/B) are the sizes of the gene sets containing motifs A and B, and A but not B, respectively, and EC(AB) and EC(A/B) are their respective EC scores. To test the corresponding null hypothesis that EC(AB) is less than or equal to EC(A/B), we randomly partitioned the gene set containing motif A (with or without motif B) into two sets (s1 and s2) of sizes S(AB) and S(A/B), respectively; we then calculated the EC score of each partition (EC(s1) and EC(s2)). We repeated the random partitioning procedure T times and obtained a distribution for EC score differences (EC(s1)–EC(s2)). If the observed difference, EC(AB)–EC(A/B), was at the top of the random distribution, we estimated an upper bound of $1/T$ for the P value of the null hypothesis. In the examination of multiple motif pairs, the evaluation of the significance of the best pair may be overestimated. To avoid this, we set the value of T to the number of motif pairs examined (the number of hypotheses generated). This procedure can be extended to L>2 motifs.

Combinogram analyses. We started the analysis with a set of N motifs from synergistic motif combinations in a given expression experiment. We assigned each gene in the genome a binary signature of length N, placing a 1 at the i^{th} position if the gene contained motif i in its promoter and a 0 otherwise. We thus generated 2^N gene sets, termed 'genes defined by motif combinations' (GMCs), where all the genes in a given GMC shared the same motif signature. We determined the expression coherence score and the averaged expression profile of all the genes in each GMC. We calculated the Pearson correlation coefficients between averaged expression profiles for all pairs of GMCs; this was input in the dendrogram analyses generated with the Cluster Analysis module in Matlab 5 (Mathworks) using the average-linkage option.

Motif synergy maps. We generated motif interaction graphs using the Brown University GeomNet server (http://loki.cs.brown.edu:8081/graph-server/gds/gds-home.shtml). We used the GEM algorithm option, because this seems to be superior to others in terms of graph clarity. The input for the server is a set of synergistic motif pairs; only motif pairs in which at least one of the two members is a known motif are analyzed. The output is a set of node locations (motifs) in a plane. A pair of nodes is connected by an edge if the synergy score of the two motifs is lower than a P value threshold, P_t, which was set at 1/Pairs, where Pairs (the number of motif pairs tested) equals (total number of motifs)×(number of known regulatory motifs)/2. We used a Matlab script to render the graph, followed by manual manipulation in Canvas 3.5 to minimize the number of lines crossing each other.

Note: Supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Acknowledgments

We thank J. Hughes for providing many of the motifs used in these analyses and U. Keich for assistance with the statistical analyses of motif synergies. We are grateful to J. Aach, B. Cohen, A. Derti, P. D'Haeseleer, A. Dudley, M. Kupiec, R. Mitra, F. Roth, D. Segré and M. Wright for advice and suggestions. Y.P. was a scholar of the Fulbright Foundation. We are grateful to the US Department of Energy and National Science Foundation and to the Lipper Foundation for grant support.

Received 17 April; accepted 31 July 2001.

1. Kel, O.V., Romaschenko, A.G., Kel, A.E., Wingender, E. & Kolchanov, N.A. A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res.* **23**, 4097–4103 (1995).
2. Quandt, K., Grote, K. & Werner, T. GenomeInspector: basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. *Genomics* **33**, 301–304 (1996).
3. Yuh, C.H., Bolouri, H. & Davidson, E.H. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**, 1896–1902 (1998).
4. Wang, J., Ellwood, K., Lehman, A., Carey, M.F. & She, Z.S. A mathematical model for synergistic eukaryotic gene activation. *J. Mol. Biol.* **286**, 315–325 (1999).
5. Halfon, M.S. *et al.* Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. *Cell* **103**, 63–74 (2000).
6. Fickett, J.W. & Wasserman, W.W. Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.* **11**, 19–24 (2000).
7. Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.* **8**, 1202–1215 (1998).
8. van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827–842 (1998).
9. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.* **9**, 3273–3297 (1998).
10. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
11. Wolfsberg, T.G. *et al.* Candidate regulatory sequence elements for cell cycle–dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.* **9**, 775–792 (1999).
12. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214 (2000).
13. Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
14. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
15. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
16. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
17. Roberts, C.J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
18. Jelinsky, S.A., Estep, P., Church, G.M. & Samson, L.D. Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Mol. Cell. Biol.* **20**, 8157–8167 (2000).
19. Zhu, G. *et al.* Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature* **406**, 90–94 (2000).
20. Koranda, M., Schleiffer, A., Endler, L. & Ammerer, G. Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. *Nature* **406**, 94–98 (2000).
21. Oehlen, L.J., McKinney, J.D. & Cross, F.R. Ste12 and Mcm1 regulate cell cycle–dependent transcription of FAR1. *Mol. Cell. Biol.* **16**, 2830–2837 (1996).
22. Arndt, K.T., Styles, C. & Fink, G.R. Multiple global regulators control HIS4 transcription in yeast. *Science* **237**, 874–880 (1987).
23. Umemura, K. *et al.* Derepression of gene expression mediated by the 5′ upstream region of the isocitrate lyase gene of *Candida tropicalis* is controlled by two distinct regulatory pathways in *Saccharomyces cerevisiae*. *Eur. J. Biochem.* **243**, 748–752 (1997).
24. Reed, S.H., Akiyama, M., Stillman, B. & Friedberg, E.C. Yeast autonomously replicating sequence binding factor is involved in nucleotide excision repair. *Genes Dev.* **13**, 3052–3058 (1999).
25. Morse, R.H. RAP, RAP, open up! New wrinkles for RAP1 in yeast. *Trends Genet.* **16**, 51–53 (2000).
26. Roth, F.P., Hughes, J.D., Estep, P.W. & Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16**, 939–945 (1998).
27. Dequard-Chablat, M., Riva, M., Carles, C. & Sentenac, A. RPC19, the gene for a subunit common to yeast RNA polymerases A (I) and C (III). *J. Biol. Chem.* **266**, 15300–15307 (1991).
28. Gasch, A.P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
29. Causton, H.C. *et al.* Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* **12**, 323–337 (2001).
30. Werner, T. Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome* **10**, 168–175 (1999).
31. Koch, C., Moll, T., Neuberg, M., Ahorn, H. & Nasmyth, K. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science* **261**, 1551–1557 (1993).
32. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
33. Bussemaker, H.J., Li, H. & Siggia, E.D. Regulatory element detection using correlation with expression. *Nature Genet.* **27**, 167–171 (2001).
34. Leem, S.H., Chung, C.N., Sunwoo, Y. & Araki, H. Meiotic role of SWI6 in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **26**, 3154–3158 (1998).
35. Wagner, A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **15**, 776–784 (1999).
36. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
37. Zhu, J. & Zhang, M.Q. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* **15**, 607–611 (1999).
38. Bulyk, M.L., Huang, X., Choo, Y. & Church, G.M. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA* **2001**, 12 (2001).
39. Aach, J., Rindone, W. & Church, G.M. Systematic management and analysis of yeast gene expression data. *Genome Res.* **10**, 431–445 (2000).