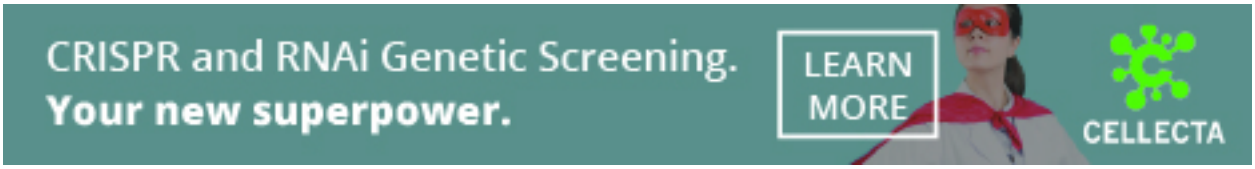
Discovery, validation, and genetic dissection of transcription factor binding sites by comparative and functional genomics

Jason Gertz, Linda Riles, Peter Turnbaugh, et al.

Genome Res. 2005 15: 1145-1152
Access the most recent version at doi:10.1101/gr.3859605



Discovery, validation, and genetic dissection of transcription factor binding sites by comparative and functional genomics

Jason Gertz, Linda Riles, Peter Turnbaugh, Su-Wen Ho, and Barak A. Cohen[1]

Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63108, USA

Completing the annotation of a genome sequence requires identifying the regulatory sequences that control gene expression. To identify these sequences, we developed an algorithm that searches for short, conserved sequence motifs in the genomes of related species. The method is effective in finding motifs de novo and for refining known regulatory motifs in *Saccharomyces cerevisiae*. We tested one novel motif prediction of the algorithm and found it to be the binding site of Stp2; it is significantly different from the previously predicted Stp2 binding site. We show that Stp2 physically interacts with this sequence motif, and that *stp2* mutations affect the expression of genes associated with the motif. We demonstrate that the Stp2 binding site also interacts genetically with Stp1, a regulator of amino acid permease genes and, with Sfp1, a key regulator of cell growth. These results illuminate an important transcriptional circuit that regulates cell growth through external nutrient uptake.

[Supplemental material is available online at www.genome.org.]

With the rate of DNA sequence determination far outpacing our capacity for sequence analysis, identifying the regulatory elements in genomic DNA sequence has become a critical task. In yeast, which has served as a proving ground for genomic techniques, less than half of the transcription factors (TFs) have a *cis*-regulatory site assigned to them (Zhu and Zhang 1999). In a recent study, Harbison et al. (2004) combined TF location data, gene expression data, and comparative genomics to predict TF binding sites. They made many predictions of novel TF–*cis*-regulatory element interactions, and while many of them may be correct, none have been validated by demonstration of a direct DNA–protein interaction. In another study (GuhaThakurta et al. 2004), novel binding sites were predicted computationally and then shown to be important for tissue-specific gene regulation, but the specific binding protein(s) were not identified. Only once, in a prokaryotic system, has a binding site been discovered computationally and its DNA–protein interaction experimentally demonstrated (McCue et al. 2001). Without experimental validation we cannot know how well statistical overrepresentation actually predicts novel sequences that serve as the binding sites for sequence specific DNA binding proteins.

One reason why so few computationally predicted motifs have been experimentally validated is that most motif finding approaches tend to "over predict" potential regulatory sites in genomic sequences. That is, the number of motifs predicted by these algorithms tends to be greater (sometimes vastly greater) than the actual number of regulatory sequences in the genome, making it difficult to choose those motif predictions that merit experimental validation. Even as the speed and sophistication of motif finding algorithms increase, an important question remains unanswered: Given a set of predicted gene regulatory motifs, how do we evaluate them to concentrate our experimental efforts on the most propitious predictions? Here we present an algorithm that predicts regulatory motifs by using only comparative sequence data, allowing the independent evaluation of motif predictions based on gene expression profiling data (Aach et al. 2000), genome-wide chromatin immunoprecipitation (ChIP) data (Lee et al. 2002), and functional data (Mewes et al. 1999). By using this approach, we identified the TF that recognizes one of the sequence motifs we predicted. The binding site for this protein, Stp2, was previously predicted to be different from the one we identify. We show that Stp2 affects the regulation of genes associated with the sequence motif it recognizes, and that the Stp2 binding site also interacts genetically with Stp1 and Sfp1.

[1]Corresponding author.
E-mail cohen@genetics.wustl.edu; fax (314) 362-7855.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3859605.

Results

Specificity and conservation of DNA sequence motifs

New methods of incorporating comparative genome data need to be explored to increase the predictive power of motif finding algorithms (Tompa et al. 2005). We implemented a new motif finding algorithm to address concerns we have with current methods. We wanted to use unaligned, rather than aligned, genomic sequences in order to incorporate information from distantly related species whose intergenic sequences may not align, and to avoid being misled by high rates of binding site turnover in alignments of orthologous promoters (Ludwig et al. 1998, 2000). Although the use of unaligned sequences precludes the use of certain phylogenetic approaches (Hardison et al. 1997), one gains statistical power because the background level of conservation decreases in more distantly related species. The power to detect TF binding sites increases proportionally with both the number of species under consideration and the phylogenetic distance between species (Eddy 2005). Finally, most motif finders accept specific lists of genes as input, such as genes from an expression cluster (Tavazoie et al. 1999), functional cluster (Hughes et al. 2000), or orthologous promoters (McCue et al. 2001; Wang and Stormo 2003). The power of motif finders is therefore limited by the quality of the grouping of the input set of sequences. It would be useful to devise methods of simulta-

neously searching several whole-genome sequences in order to decrease the dependence on the initial groupings of sequences. The approach we present here uses multiple whole-genome sequences in a biologically intuitive way to search for regulatory motifs. In addition, the method does not require the use of aligned promoter sequences.

While testing different methods for measuring the conservation of regulatory sequences between related species, we noticed an inverse relationship between the information content of a regulatory sequence motif (the specificity of a binding site pattern) and the frequency with which the motif is found in orthologous promoters from related species. If one searches orthologous promoters for exact matches to a consensus sequence, true variants of the sites will be missed. However, if one searches orthologous promoters with a degenerate motif, many false-positive motifs are identified in related genomes.

For example, the Mbp1/Swi4 complex binds to the MCB Box (Mlul Cell Cycle Box) and regulates the expression of genes in the G_1 phase of the cell cycle (Koch et al. 1993). By starting from a matrix of known MCB sites, we iteratively selected for a multiple alignment of MCB sites with high conservation (*C*, measured as the average number of orthologous promoters in which the motif is found across the genome; see Motif Refinement). The increase in conservation resulted in a degenerate motif that embraces many false positives (i.e., are found upstream of genes not regulated by MCB) (Fig. 1A). Conversely, selecting for variants in the MCB alignment with high information content (*I*, measured in bits) produced a very specific motif close to the MCB consensus sequence (Fig. 1B). This increase in information content came at the expense of missing many conserved MCB boxes in orthologous promoters that did not exactly match the consensus sequence. Iteratively optimizing the MCB motif alignment for the sum of the information content and conservation (*I*+*C*) generates an MCB alignment that has both higher information content (is more specific) and shows greater conservation (finds more sites in orthologous promoters) (Fig. 1C). We used this improved scoring scheme in a genetic algorithm (GA) that takes an input binding site alignment and iteratively optimizes its *I*+*C* score, the sum of the information content and conservation (See Methods).

GAs have been used successfully in approaching several problems, including classifying disease based on mass spectroscopy data (Petricoin et al. 2002), predicting protein folds from primary sequence (Dandekar and Argos 1992), associating DNA regulatory motifs with gene expression (Fogel et al. 2004), finding new *cis*-regulatory modules in coexpressed genes (Aerts et al. 2004), finding operons (Jacob et al. 2004) and many other diverse problems (Reggia et al. 1998; Koza et al. 2003; Cho et al. 2004; Pond and Frost 2004; Spalek et al. 2005). While GAs are inherently heuristic, they are a useful class of algorithms to develop and refine new methods because they are simple to understand, are easy to implement, and allow a user to incorporate biological intuition in a straightforward way. A more sophisticated implementation of our algorithm will require a more formal, probabilistic framework, such as an expectation maximization approach, but first we demonstrate that the new features of our approach are indeed useful and have led to the discovery of a novel DNA–protein interaction.

Motif refinement

Compared with the original MCB alignment, the refined MCB alignment finds a larger number of genes that show the characteristic G_1 phase expression in the cell cycle, while finding fewer genes with non–cell cycle-dependent expression (Supplemental Table 1). The expression coherence (EC) (Pilpel et al. 2001; Sudarsanam et al. 2002) in the cell cycle (Cho et al. 1998) of genes found by the new alignment increased from 0.156 to 0.201, with the *P*-value decreasing from 0.0084 to 0.00038. In addition the percentage of genes with promoter regions bound by Mbp1 in ChIP experiments rose 4%. These results suggest that the refined MCB motif matrix is a better description of the true MCB site then the original matrix.

Based on our success with the MCB motif, we used our algorithm to refine a group of 32 motifs derived from performing Gibbs sampling on the promoters of genes with similar annotations (Supplemental Table 1; Hughes et al. 2000). We were able to successfully refine 17 of 32 motifs tested, increasing both their information content and conservation. In no case did we observe a decrease in both information content and conservation. Al-

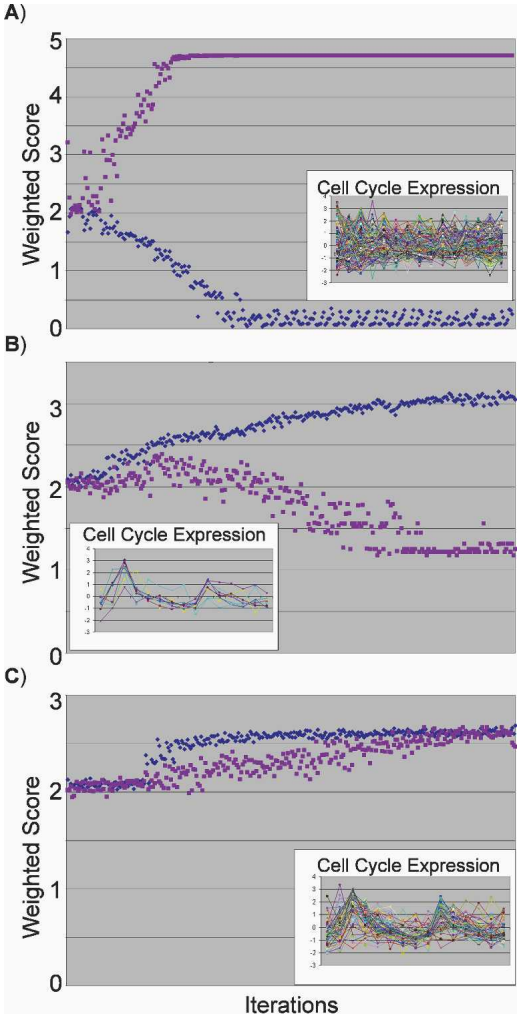


Figure 1. Optimizing the MCB Box motif. The information content scores (blue ♦) and conservation scores (purple ■) are plotted during iterative rounds of optimization while selecting for conservation only (*A*), information content only (*B*), or the sum of information content and conservation (*C*). (*Insets*) The cell-cycle expression (Cho et al. 1998) of promoters containing sequences matching the final optimized motif in each case. The *x*-axis represents experimental timepoints through the cell cycle, and the *y*-axis represents relative gene expression levels.

though neither expression coherence nor %ChIP was explicitly optimized in the refinement process, there was a strong bias toward increases in these scores after refinement ($P = 0.007$ for EC and $P = 0.031$ for %ChIP, nonparametric sign test). These refined motifs are available in Supplemental Table 1.

We tested the stability of the motif refinement process by virtually mutating the sequences within known motif matrices and rerunning the refinement algorithm. The program converged on the same refined motif even with a 50% mutation rate in the starting alignment. The ability of the algorithm to pick up such a distorted signal suggested a way to use the same process for de novo motif discovery.

De novo motif discovery

By using the background nucleotide frequencies in the *Saccharomyces cerevisiae* genome, we generated random, artificial sequences and used them to create random motif "alignments." These random alignments had very low information content and were used as input for our algorithm. We used random sequences instead of sequences from the genome in order to start the algorithm with a diverse set of input matrices. After 25 rounds of optimization based on the *I*+*C* score (see Methods), the algorithm always converged on a high information content matrix that identifies potential regulatory sequences in orthologous promoters. The output binding site alignment is highly dependent on the input random matrix such that the same input alignment always converges on the same output alignment. We performed 220 trials with alignments of random sequences (between 6–12 bp in length) to test the ability of our program to find motifs de novo. From these 220 trials, 159 motifs matched previously identified TF binding sites. In the worst case, assuming all of the unknown motifs are not true motifs, our program has a false-positive rate of 28%. The true false-positive rate is likely to be significantly lower. The 159 known motifs broke down into 33 different TF binding sites. This is only slightly less than the 37 known yeast TF sites for which we have high-quality weight matrices (see Methods), suggesting that the false-negative rate is ~11% (four of 37). Although our method is not directly comparable to those currently in use, our false-positive and false-negative rates are lower than those reported in a recent study that compares various motif finding algorithms (Tompa et al. 2005). There was fairly even representation over the known motifs, with no one motif dominating the output of these initial runs (Fig. 2A). This suggests that we are sampling evenly across the landscape of possible motifs and are not biased toward any particular class of motif. The fact that certain known motifs (e.g., the Cbf1 site) did not show up in these initial runs is likely due to the small number of runs and not to any inherent biases in the algorithm.

From the 220 trials, we found 43 different motifs without known protein binding partners (33 appeared only once; 28 fell into 10 different classes with two or more representatives each) (for logos, see Supplemental Table 2). Of these 43 only two were discovered in other large-scale comparative genomic analyses of the yeast genome (Cliften et al. 2003; Kellis et al. 2003). In addition none of these novel motifs were predicted in high-throughput ChIP studies (Lee et al. 2002; Harbison et al. 2004). None of these novel motif predictions were classified as AT tracts (low complexity motifs comprised of polyA and polyT sequences). Because 43 motifs are too many to assess experimentally, we sought ways of evaluating our predictions to focus our efforts on novel motifs that are most likely to be real.

We used independent sources of high-throughput data to pick promising candidate motifs. The sequence logo (Schneider and Stephens 1990) for one of the unknown motifs we identified is shown in Figure 2B. This motif was subjected to additional rounds of selection in an attempt to refine its sequence (Fig. 2C). The refined motif is conserved in 19.9% of promoters when the *S. cerevisiae* promoter has a site, which is comparable to known binding sites (i.e., Ume6 binding site is conserved 20.0%). The genes whose promoters contain copies of this motif show coherent expression (EC = 0.38, $P < 10^{-6}$) (Pilpel et al. 2001) in cells treated with the DNA damaging agent methyl-methane sulfonate (MMS) (Jelinsky et al. 2000). The promoters that contain this binding site also overlap significantly with those identified in ChIP experiments (Lee et al. 2002) with Sfp1 ($P = 0.00035$), Stp2 ($P = 0.00011$), and Phd1 ($P = 0.00026$).

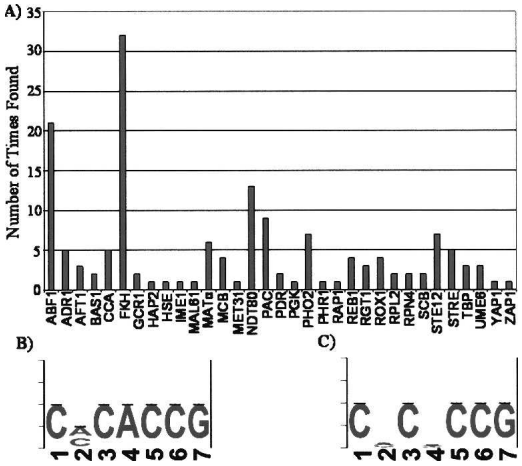


Figure 2. De novo motif identification. (*A*) Histogram of the 159 predictions correlated to known motifs we identified by optimizing randomly generated input alignments. (*B*) Sequence Logo (Schneider and Stephens 1990) of a novel motif generated by using our algorithm. (*C*) The sequence logo for the same motif after undergoing an additional 25 rounds of selection.

Stp2 interacts with the motif

These results lead us to hypothesize that our motif is bound either by Sfp1, Stp2, Phd1, or a combination of these three proteins. We used the "one-hybrid" (Alexander et al. 2001; Li and Herskowitz 1993) assay with 170 different TF fusions (Supplemental Table 3) to identify proteins that bind to this sequence (Fields and Song 1989). We inserted a 31-bp sequence from the *AGP2* promoter containing two conserved instances of the motif in opposite orientations upstream of a *HIS3* reporter gene. We crossed cells containing this reporter gene with an array of strains each carrying a different TF fused to the Gal4 transcriptional activation domain (AD) and scored the resultant diploid cells for histidine prototrophy. Only the strain carrying Stp2-AD yielded His^+ diploids (Fig. 3A). Mutations introduced into the first putative binding site in the reporter gene abolish the His^+ phenotype (Fig. 3B,C). Mutations in the other binding site significantly diminish the His^+ phenotype. These results suggest that Stp2 binds to the motif.

Electrophoretic mobility shift assays using whole-cell extracts from *stp2*Δ, wild-type, and *STP2* overexpressing cells confirmed that Stp2 binds to the sequence motif (Fig. 4A). The presumed DNA–protein complex was super-shifted upon incubation with an antibody specific to overexpressed *STP2*. Only a twofold

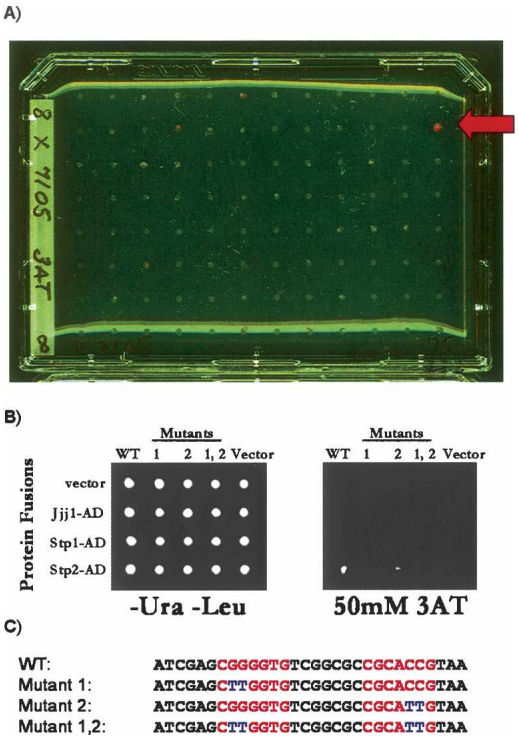


Figure 3. One-hybrid assay. (*A*) Ninety-six AD fusions mated to *HIS3* reporter plasmid grown on 75 mM 3-amino triazole (3AT), a competitive inhibitor of the *HIS3* gene product. The arrow points to the Stp2-AD fusion. (*B*) AD fusions mated to mutant versions of a *HIS3* reporter plasmid. (C) Sequences used as promoters in one-hybrid assay. Red indicates the motif, and blue indicates mutations.

excess of the unlabeled sequence motif competes for binding to the labeled DNA, while binding is still detected in the presence of a fourfold excess of unlabeled double mutant probe (Fig. 4B). These results suggest that Stp2 binds specifically to DNA containing the motif.

Expression of genes with the motif in their promoters is altered in *stp2*Δ and *sfp1*Δ mutants

To test if Stp2 regulates the expression of genes that contain the motif in their promoters, we compared expression profiles from wild-type cells, *stp2*Δ cells, and cells engineered to overexpress *STP2*. Genes that were down-regulated in an *stp2*Δ strain and up-regulated in the *STP2* overexpression strain are significantly enriched for the presence of our motif in their promoters ($P = 1.57 \times 10^{-6}$). This suggests that Stp2 is a transcriptional activator that acts through the motif. Genes that contain this motif in their promoter and whose expression is affected by changes in Stp2 levels are enriched for cellular transport and transport mechanisms ($P = 0.00539$), including genes involved in mitochondrial, vacuolar, golgi, endoplasmic reticulum, and membrane transport. Previous work has shown that Stp2 regulates amino acid permease genes (de Boer et al. 2000; Nielsen et al. 2001). Our results suggest that Stp2 has a broader role in the regulation of transporters in general.

Our analysis of ChIP data suggested that Sfp1 also binds promoters containing this motif. We decided to investigate this connection further because *SFP1* expression increases in the presence of MMS and our analysis indicated that promoters that contain the motif show coordinate expression in response to MMS. To determine whether Sfp1 plays a role in the regulation of the expression of genes containing the motif, we compared expression profiles of wild-type and *sfp1*Δ cells. We found significant overlap between promoters that show expression differences in *sfp1*Δ cells and promoters that contain the motif ($P = 10^{-4}$), suggesting that Sfp1 may have an overlapping function with Stp2 in the activation of these promoters. Alternatively Sfp1 may play a role in activating Stp2.

To test these models, we performed gene expression profiling on wild-type, *sfp1*Δ, and *stp2*Δ strains grown in rich medium or in rich medium treated with MMS. Both *SFP1* and *STP2* are expressed at slightly higher levels after treatment with MMS (Fig. 5A). The increased expression of *STP2* in MMS is not observed in *sfp1*Δ cells. However, the expression of *SFP1* in MMS is unaffected by the absence of *STP2* (Fig. 5A). These results suggest that *SFP1* is an upstream regulator of *STP2*.

Our results suggest that Sfp1 acts upstream of Stp2. Analysis of our microarray data revealed that a large fraction of genes that contain the motif were misexpressed in *sfp1*Δ cells. Many of these genes were also misexpressed in *stp2*Δ cells. To determine whether Sfp1 and Stp2 act through the same promoters, we compared the sets of genes that were misexpressed in each mutant (Fig. 5B). We found a significant overlap ($P = 1.6 \times 10^{-5}$) of 11 genes that contain the motif and are misexpressed in both *sfp1*Δ and *stp2*Δ cells, suggesting that both Sfp1 and Stp2 are necessary for the proper expression of certain genes (Supplemental Table 4).

Stp1 acts upstream of Stp2

Stp2 is 42% identical to its paralog Stp1 over 365 amino acids. Both proteins are involved in the induction of *BAP2* and *BAP3*,

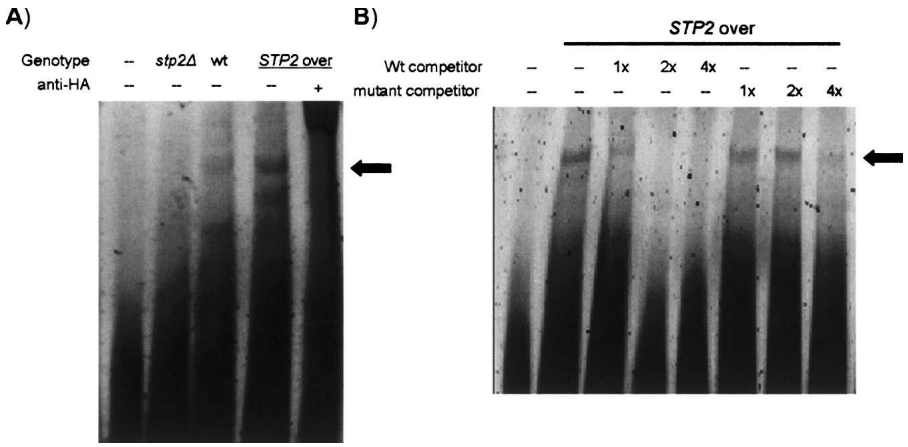


Figure 4. Gel shift assays. (*A*) Gel shift showing the Cy3 labeled probe incubated with whole-cell extracts from *stp2*Δ cells, wild-type cells, or cells engineered to overexpress HA-tagged *STP2*. The black arrow represents the reproducible shift involving Stp2. Faster running bands represent possible cleavage or degradation products of Stp2. The last lane shows a super-shift with anti-HA antibody. (*B*) Gel shift showing *STP2* overexpressing whole-cell extract with varying amounts of unlabeled wild type and mutant competitor.

branch chain amino acid permease genes (de Boer et al. 2000; Nielsen et al. 2001). Although Stp1 does not interact with the motif in the one-hybrid assay, the sequence similarity and phenotypic similarities suggested that Stp1 may have a role in activation through the motif. We therefore tested whether Stp1 is required for the activation of gene expression through the motif. A reporter gene with two copies of the motif upstream of *HIS3* is activated in wild-type cells in the absence of any AD fusion protein (presumably by endogenous Stp2). In an *stp2*Δ strain, activation through the motif is dramatically reduced, demonstrating that Stp2 indeed plays a role in activation through this motif (Fig. 6). Strikingly, in an *stp1*Δ strain activation is abolished. Since our one-hybrid results showed that Stp1 does not directly interact with the motif, these results suggest that Stp1 is an upstream regulator of Stp2.

Discussion

Given the number of studies that have used computational methods to identify regulatory motifs in the yeast genome, it is remarkable that our algorithm identified a biologically important sequence motif missed in previous studies. These results suggest that our method is a viable alternative to those currently available and, in some cases, detects true motifs missed by other algorithms. We believe the key component of our algorithm is the incorporation of evolutionary conservation of sequence in an intuitive fashion. To avoid missing regulatory sites that rapidly evolve, we did not require motifs to be positionally conserved in alignments of orthologous promoters. In addition, by scoring motifs across the whole genome, instead of in a specific input set of genes (e.g., a cluster of similarly expressed genes), we reduced our dependence on expression data and other functional data in identifying potential motifs. This allowed us to evaluate our predicted motifs by using these high-throughput data and focus our attention on the best candidate motif from a list of 43 potential motifs, a strategy that resulted in the identification of the Stp2 binding site.

Stp2's binding site was previously predicted to be CGGCTC (de Boer et al. 2000), which is different from the motif that we identified (Fig. 2C). Genes that contain matches to the previous description of the Stp2 binding site do tend to change in expression profiles of cells that misexpress *STP2* ($P = 0.002$). However the overlap of promoters that contain the site we identified is far more significant ($P = 1.57 \times 10^{-6}$). We therefore believe that the motif we describe here is a more accurate description of the Stp2 binding site.

In addition to discovering the binding motif for Stp2, we also discovered that both Sfp1 and Stp1 might play important roles in regulating Stp2 action through the motif. We have not elucidated the mechanisms for these effects, but our data suggest that both Sfp1 and Stp1 act upstream of Stp2's regulatory role. Sfp1 has been predicted to bind to the RRPE site (Fingerman et al. 2003) and Stp1 has been predicted to bind to the UAS_{aa} (GCCGPy-N_4-PuCGGC) (De Boer et al. 1998). In the Stp2 promoter, there is an RRPE site and a similar sequence to the UAS_{aa} (CGGC-N_{14}-PuCGGC), suggesting that *STP2* expression may be under the dual transcriptional regulation of Sfp1 and Stp1. Sfp1 is thought to play a role in the transcription of genes involved in growth (Blumberg and Silver 1991; Xu and Norris 1998; Jorgensen et al. 2002),while Stp1 is thought to play a role in amino acid transport, suggesting that this is an interesting example of cell growth and external nutrients combining to make transcriptional "decisions."

The continued acquisition of genome sequences from diverse species will enable us to search for regulatory motifs that control transcription in other systems. In addition, we are interested in developing methods to specifically incorporate phylogenetic distances into our strategy (Moses et al. 2004) without explicitly incorporating alignments of genomic sequences.

Methods

Definition of information (*I*)

Information content is a measure of the specificity of a motif matrix (a multiple alignment of potential binding sites). Formally, the information content for a position in the motif matrix is defined as

$$I_{seq}(i) = \sum_{b=A}^{T} f_{b,i} * \log_2 \frac{f_{b,i}}{p_b}$$

where i is the position of the motif, b represents a base (either A, C, G, or T), $f_{b,i}$ is the frequency of a particular base at position i, and p_b is the genomic frequency of a base (Schneider et al. 1986). The information content per position is averaged over each column to give the information component I.

Definition of conservation (*C*)

For a given binding site matrix we find all of the intergenic sites in *S. cerevisiae* that score better than a cut-off established by the program Patser (version 3b) (Hertz and Stormo 1999). For each site found by Patser, we find the gene corresponding to the promoter region in which the site was found. For all genes that have sites, we look at the promoter region of its orthologs in other

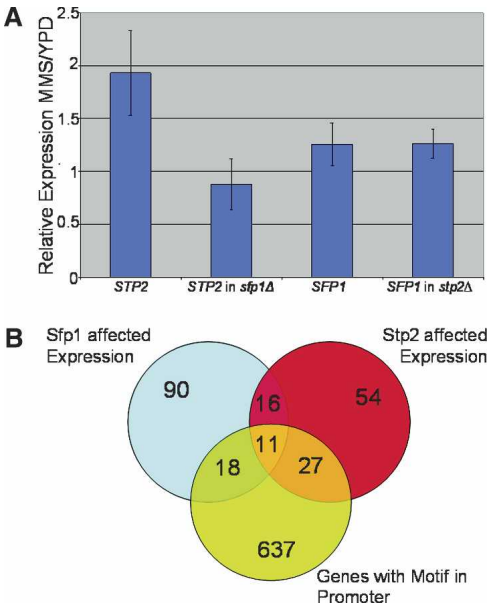


Figure 5. Microarray experiments. (*A*) Comparison of *STP2* and *SFP1* gene expression. Bar height shows the expression in MMS relative to a reference sample divided by expression in YPD compared with the same reference sample; error bars represent 1 SD. (*B*) Venn diagram showing the overlap of genes whose expression is significantly altered in an *stp2*Δ mutant, genes whose expression is significantly altered in an *sfp1*Δ mutant, and genes that contain a copy of the motif in their promoter.

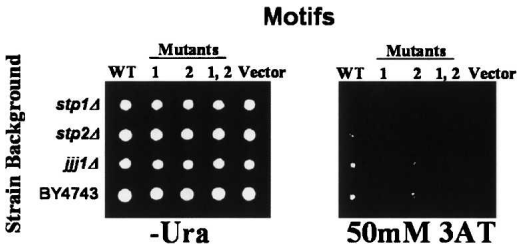


Figure 6. Reporter gene assays of endogenous Stp2 activity. Deletion strains (Brachmann et al. 1998) were mated to strains carrying variants of a *HIS3* reporter plasmid and then grown on 50 mM 3AT. Mutant *HIS3* reporter genes are the same as in Figure 3.

yeast species to identify orthologous promoters with sites that surpass the cut-off score. The genome sequences we use in addition to that of *S. cerevisiae* are three closely related species in the sensu stricto group (*S. mikatae*, *S. kudriavzevii*, and *S. bayanus*), and two more distantly related species in the sensu lato group (*S. castellii* and *S. kluyveri*) (Cliften et al. 2003). Each gene receives a score that is equal to the number of species in which the ortholog contained a site in its promoter region. The scores for all the genes that contained at least one site in *S. cerevisiae* are averaged together to give the conservation component *C*.

Definition of fitness (*F*)

To determine an effective fitness function that incorporates both information content and conservation, we tested different weightings of the function $F = \beta * I + C$, where β is the weight of I (Supplemental Fig. 1). To use comparable quantities for I and C, we first multiply I by 5/2. This has the effect of putting both I and C on a scale of 0–5. From this analysis we concluded that $\beta = 1$, equal weighting, was the most effective overall weighting (Supplemental Fig. 1). Therefore we used the fitness function:

$$F = I + C$$

Genetic algorithm

We constructed a GA that scores regulatory motifs based on their information content and conservation. A flow chart of the algorithm is available in Supplemental Figure 2. The algorithm accepts a multiple alignment of potential regulatory sites as input. The program converts this alignment into a weight matrix and scans the *S. cerevisiae* genome for the 400 best scoring intergenic sites. These sites are randomly divided into four equal groups that form the "parents" of the first generation. Every combination of parents is combined and their sites are mutated (rate = 1 in 1000 bases), duplicated (rate = 1 in 20 sequences), and recombined (rate = 1 event in 1000 bases). The sites are then randomly segregated into two equal groups, or "progeny." Thus 12 progeny, where each progeny is a multiple alignment of potential regulatory sites, are created every generation. The fitness of each progeny matrix is scored, and the top four progeny are selected to become the parents of the next generation. This cycle continues for a specified number of generations.

To introduce new sequences from the genome into the population, we perform a migration step on one randomly selected progeny matrix in each generation. If a progeny matrix is selected for migration, we scan the *S. cerevisiae* genome with that progeny's weight matrix and replace it with the top 100 scoring sites. That progeny's fitness is still scored on the basis of its original sites, but if that progeny is selected, the updated matrix is used as a parent in the next generation. This step allows the algorithm to accept parents that are not always the most fit and introduces variability in the matrices that may help avoid local optimums.

Motif refinement

All of the input sequences for motif refinement were obtained from http://atlas.med.harvard.edu/motifs/. These motifs were generated by running the program AlignACE on groups of genes with similar functional annotations (Hughes et al. 2000). All of these motifs were used as input matrices to our algorithm.

Motif discovery

For motif discovery we generate fixed length random sequences by using the background nucleotide frequencies in the *S. cerevisiae* genome. These sequences are combined to form random matrices that seed the algorithm. Because a motif is a model of the binding site specificity of a TF and not necessarily a collection of actual occurrences of sites in the genome, the use of randomly generated input sequences allows us to search a more diverse set of input matrices that still find matches to actual occurrences of TF sites in the genome. The same algorithm that we use for motif refinement is then used to optimize these random alignments into motif models with high information content and conservation scores across species. To determine which motifs were correlated to known motifs, we used CompareACE (Hughes et al. 2000) with a cutoff of 0.7 correlation coefficient. To cluster unknown motifs, we used the method implemented in Hughes et al. (2000) with a CompareACE cutoff of ≥0.7.

Strains and plasmids

The deletion, wild-type, and *STP2* overexpression strains were derived from BY4743 (*MATa/α his3Δ1/his3Δ1 leu2Δ0/leu2Δ0 lys2Δ0/LYS2 MET15/met15Δ0 ura3Δ0/ura3Δ0*) as described in Brachmann et al. (1998). To overexpress *STP2*, we created the plasmid pPT100 that contains a 2μn origin and a *GAL1* promoter in front of an N-terminal HA-tagged *STP2* ORF. pPT100 was created from YEplac181 (Gietz and Sugino 1988), by digesting it with EcoR1, filling in the ends with Klenow, and ligating the ends back together to give pBC100. Then the Kpn1/BamH1 fragment from pRF4-6o (gift from Russ Finley, http://www.proteome.wayne.edu/vectorsandstrains.html) was inserted into pBC100 by using Kpn1 and BamH1 giving pBC103. The *STP2* gene was amplified from genomic DNA by PCR using the following primers: forward, 5′-CCCTTATGATGTGCCAGATTATGCCTCTCCC GAATTCATGCCTATCTTATCACTATCTTCAACACGG-3′; reverse, 5′-CCAAACCTCTGGCGAAGAAGTCCAAAGCTTCTCG AGCTATTAAAATTCTATCCCATAAGCTTTTTTGTAAAGGGCC. The *STP2* gene was combined with pBC103 by homologous recombination, yielding pPT100. For the one-hybrid assays, we used PJ69-4a (*MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4 gal80 LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ*) to carry the AD fusions. These strains were mated to *HIS3* reporter strains derived from PJ69-4α (*MATα trp1-901 leu2-3,112 ura3-52 his3-200 gal4 gal80 LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ*) (James et al. 1996). *HIS3* reporter plasmid pBM4429 was based on pRS316 (backbone CEN plasmid with *URA3*) (Sikorski and Hieter 1989). Three overlapping PCR products were inserted into the backbone by gap-repair to produce a *MEL1* minimal promoter interrupted by *TRP1*, flanked by Spe1 and Xho1 sites. The resulting plasmid was cut with Spe1 and Xho1 and gel-purified for gap repair with the double-stranded motif. Plasmids containing TF AD fusions were given as a gift from Stan Fields, University of Washington.

One-hybrid assay

A modified one-hybrid assay (Li and Herskowitz 1993; Alexander et al. 2001) was used to screen 170 TF AD fusions. The TF strains were pinned to lawns of the *HIS3* reporter strain. The resulting diploids were assayed for growth on various concentrations of 3AT.

Gel shift assays

Electromobility shift assays were performed by using whole-cell yeast extract from *stp2*Δ, wild-type, and *STP2* overexpressing cells. In all cases, the cells were grown to log phase and then transferred to 2% galactose for 2 h. Cells were pelleted and then resuspended in 250 µL of binding buffer (20 mM Tris-HCl at pH 8, 200 mM KCl, 5 mM $MgCl_2$, 0.5 mM $CaCl_2$, 0.1 mM EDTA, 0.5 mM DTT, 11% glycerol), supplemented with 2.5 µL of yeast protease inhibitor cocktail (Sigma-Aldrich). PMSF was added to a final concentration of 1 mM. Cells were lysed by vortexing the suspension with glass beads for 5 min. The lysed suspension was centrifuged for 15 min at full speed. The supernatant from the centrifugation was used in the binding reactions. For the binding reactions, 10 µL of the extract was mixed with 1 µL of 1 µg/µL poly dI/dC and 1 µL of 50 µM Cy3 labeled double-stranded probe. The binding reaction was incubated for 30 min at room temperature and then for 10 min on ice. The mixture was then electrophoresed for 1 h at 4°C at 100 V on a 5% polyacrylamide gel. The gels were scanned by using a Molecular Dynamics Typhoon 8600 (Amersham Biosciences).

Microarray experiments

Microarray experiments were done with three biological replicates. For the MMS microarray experiments, cells were grown to log phase in YPD and split into two aliquots, and 0.01% MMS was added for 1 h to one culture. For *STP2* overexpression, the cells were grown to log phase in 2% glucose and split into two aliquots, one of which was then transferred to 2% galactose. Both cultures were then grown for an additional 2 h. RNA extraction, labeling, and hybridization was done as described in Dudley et al. (2002).

We printed microarrays by using the *S. cerevisiae* oligonucleotide set manufactured by Qiagen-Operon (http://oligos.qiagen.com/arrays/oligosets_yeast.php). This oligo set contains 6388 70mer probes representing 6388 predicted *S. cerevisiae* ORFs. We printed each oligo in duplicate onto epoxy-coated slides (MWG Biotech).

Microarray analysis

The microarrays were scanned by using GenePix 4000B, and the spots were analyzed by using GenePix 4.0 image analysis software (Axon Instruments). To find the significantly changing genes in each set of experiments, Fisher's linear discriminant (Fisher 1936) was computed for Stp2 by using *STP2* overexpression compared with wild-type and *stp2*Δ compared with wild type, and for Sfp1 by using *sfp1*Δ compared with a reference and wild type compared with a reference. We set a confidence cutoff of 99.9% when using a paired *t*-test. To calculate the *P*-value of an overlap, the hypergeometric distribution was used with 6000 (the number of genes) used as the size of the universal set.

Acknowledgments

We thank Ting Wang and Gary Stormo for providing the Patser program and for helpful discussions, Jon Armstrong and Elaine Mardis for printing our microarrays, Stan Fields for providing the AD fusions, and Ashwin Desikan for technical help. We also thank Mark Johnston, Robi Mitra, Katherine Varley, and members of the Cohen Lab for helpful insights and critical readings of the manuscript. J.G. is supported by the NSF's Graduate Research Fellowship DGE-0202737.

References

Aach, J., Rindone, W., and Church, G.M. 2000. Systematic management and analysis of yeast gene expression data. *Genome Res.* **10:** 431–445.

Aerts, S., Van Loo, P., Moreau, Y., and De Moor, B. 2004. A genetic algorithm for the detection of new *cis*-regulatory modules in sets of coregulated genes. *Bioinformatics* **20:** 1974–1976.

Alexander, M.K., Bourns, B.D., and Zakian, V.A. 2001. One-hybrid systems for detecting protein–DNA interactions. *Methods Mol. Biol.* **177:** 241–259.

Blumberg, H. and Silver, P. 1991. A split zinc-finger protein is required for normal yeast growth. *Gene* **107:** 101–110.

Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J., Hieter, P., and Boeke, J.D. 1998. Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14:** 115–132.

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

Cho, J.H., Seok Sung, K., and Ryong Ha, S. 2004. A river water quality management model for optimising regional wastewater treatment using a genetic algorithm. *J. Environ. Manage.* **73:** 229–242.

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

Dandekar, T. and Argos, P. 1992. Potential of genetic algorithms in protein folding and protein engineering simulations. *Protein Eng.* **5:** 637–645.

De Boer, M., Bebelman, J.P., Goncalves, P.M., Maat, J., Van Heerikhuizen, H., and Planta, R.J. 1998. Regulation of expression of the amino acid transporter gene BAP3 in *Saccharomyces cerevisiae*. *Mol. Microbiol.* **30:** 603–613.

de Boer, M., Nielsen, P.S., Bebelman, J.P., Heerikhuizen, H., Andersen, H.A., and Planta, R.J. 2000. Stp1p, Stp2p and Abf1p are involved in regulation of expression of the amino acid transporter gene BAP3 of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **28:** 974–981.

Dudley, A.M., Aach, J., Steffen, M.A., and Church, G.M. 2002. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl. Acad. Sci.* **99:** 7554–7559.

Eddy, S.R. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3:** e10.

Fields, S. and Song, O. 1989. A novel genetic system to detect protein–protein interactions. *Nature* **340:** 245–246.

Fingerman, I., Nagaraj, V., Norris, D., and Vershon, A.K. 2003. Sfp1 plays a key role in yeast ribosome biogenesis. *Eukaryot. Cell* **2:** 1061–1068.

Fisher, R.A. 1936. The use of multiple measures in taxonomic problems. *Ann. Eugenics* **7:** 179–188.

Fogel, G.B., Weekes, D.G., Varga, G., Dow, E.R., Harlow, H.B., Onyia, J.E., and Su, C. 2004. Discovery of sequence motifs related to coexpression of genes using evolutionary computation. *Nucleic Acids Res.* **32:** 3826–3835.

Gietz, R.D. and Sugino, A. 1988. New yeast-*Escherichia coli* shuttle vectors constructed with in vitro mutagenized yeast genes lacking six-base pair restriction sites. *Gene* **74:** 527–534.

GuhaThakurta, D., Schriefer, L.A., Waterston, R.H., and Stormo, G.D. 2004. Novel transcription regulatory elements in *Caenorhabditis elegans* muscle genes. *Genome Res.* **14:** 2457–2468.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104.

Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.

Hertz, G.Z. and Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15:** 563–577.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. 2000. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296:** 1205–1214.
Jacob, E., Sasikumar, R., and Nair, K.N. 2004. A fuzzy guided genetic algorithm for operon prediction. *Bioinformatics* **21:** 1403–1407.
James, P., Halladay, J., and Craig, E.A. 1996. Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics* **144:** 1425–1436.
Jelinsky, S.A., Estep, P., Church, G.M., and Samson, L.D. 2000. Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Mol. Cell Biol.* **20:** 8157–8167.
Jorgensen, P., Nishikawa, J.L., Breitkreutz, B.J., and Tyers, M. 2002. Systematic identification of pathways that couple cell growth and division in yeast. *Science* **297:** 395–400.
Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Koch, C., Moll, T., Neuberg, M., Ahorn, H., and Nasmyth, K. 1993. A role for the transcription factors Mbp1 and Swi4 in progression from G_1 to S phase. *Science* **261:** 1551–1557.
Koza, J.R., Keane, M.A., and Streeter, M.J. 2003. Evolving inventions. *Sci. Am.* **288:** 52–59.
Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804.
Li, J.J. and Herskowitz, I. 1993. Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system. *Science* **262:** 1870–1874.
Ludwig, M.Z., Patel, N.H., and Kreitman, M. 1998. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: Rules governing conservation and change. *Development* **125:** 949–958.
Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403:** 564–567.
McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V., and Lawrence, C.E. 2001. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* **29:** 774–782.
Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. 1999. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **27:** 44–48.
Moses, A.M., Chiang, D.Y., Pollard, D.A., Iyer, V.N., and Eisen, M.B. 2004. MONKEY: Identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.* **5:** R98.
Nielsen, P.S., van den Hazel, B., Didion, T., de Boer, M., Jorgensen, M., Planta, R.J., Kielland-Brandt, M.C., and Andersen, H.A. 2001. Transcriptional regulation of the *Saccharomyces cerevisiae* amino acid permease gene BAP2. *Mol. Gen. Genet.* **264:** 613–622.
Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., et al. 2002. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359:** 572–577.
Pilpel, Y., Sudarsanam, P., and Church, G.M. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* **29:** 153–159.
Pond, S.L. and Frost, S.D. 2004. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol. Biol. Evol.* **22:** 478–485.
Reggia, J.A., Lohn, J.D., and Chou, H.H. 1998. Self-replicating structures: Evolution, emergence and computation. *Artif. Life* **4:** 283–302.
Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18:** 6097–6100.
Schneider, T.D., Stormo, G.D., Gold, L., and Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188:** 415–431.
Sikorski, R.S. and Hieter, P. 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* **122:** 19–27.
Spalek, T., Pietrzyk, P., and Sojka, Z. 2005. Application of the genetic algorithm joint with the Powell method to nonlinear least-squares fitting of powder EPR spectra. *J. Chem. Inf. Comput. Sci.* **45:** 18–29.
Sudarsanam, P., Pilpel, Y., and Church, G.M. 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., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* **22:** 281–285.
Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23:** 137–144.
Wang, T. and Stormo, G.D. 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19:** 2369–2380.
Xu, Z. and Norris, D. 1998. The SFP1 gene product of *Saccharomyces cerevisiae* regulates G_2/M transitions during the mitotic cell cycle and DNA-damage response. *Genetics* **150:** 1419–1428.
Zhu, J. and Zhang, M.Q. 1999. SCPD: A promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* **15:** 607–611.

Web site references

http://atlas.med.harvard.edu/motifs/; input sequences for motif refinement.
http://www.proteome.wayne.edu/vectorsandstrains.html; Kpn1/BamH1 fragment from pRF4-6o.
http://oligos.qiagen.com/arrays/oligosets_yeast.php; *S. cerevisiae* oligonucleotide set manufactured by Qiagen-Operon.

Received February 21, 2005; accepted in revised form May 3, 2005.