Phylogenetically and Spatially Conserved Word Pairs Associated with Gene Expression Changes in Yeasts

Derek Y. Chiang
Dept of Molecular and Cell Biology,
UC Berkeley
dchiang@ocf.berkeley.edu

Alan M. Moses
Graduate Group in Biophysics,
UC Berkeley
amoses@ocf.berkeley.edu

Manolis Kamvysselis
MIT Laboratory of Computer Science
manolis@mit.edu

Eric S. Lander
MIT/Whitehead Institute Center for Genome Research
lander@genome.wi.mit.edu

Michael B. Eisen
Dept of Molecular and Cell Biology, UC Berkeley
Division of Genome Sciences, Lawrence Berkeley National Lab
mbeisen@lbl.gov

ABSTRACT

Background. Transcriptional regulation in eukaryotes is often multifactorial, involving multiple transcription factors binding to the same transcription control region (*e.g.*, upstream activating sequences and enhancers), and to understand the regulatory content of eukaryotic genomes it is necessary to consider the co-occurrence and spatial relationships of individual binding sites. The identification of sequences conserved among related species (often known as phylogenetic footprinting) has been successfully used to identify individual transcription factor binding sites. Here, we extend this concept of functional conservation to higher-order features of transcription control regions involved in the multifactorial control of gene expression.

Results. We used the genome sequences of four yeast species of the genus *Saccharomyces* to identify sequences potentially involved in multifactorial control of gene expression. We found 1,117 potential regulatory "templates": pairs of hexameric sequences that are jointly conserved in transcription regulatory regions and also exhibit non-random relative spacing. Many of the individual sequences in these templates correspond to known transcription factor binding sites, and the sets of genes containing a particular template in their transcription control regions tend to be differentially expressed in conditions where the corresponding transcription factors are known to be active.

Conclusions. The incorporation of both joint conservation and spacing constraints of sequence pairs predicts groups of target genes that were specific for common patterns of gene expression. Our work suggests that positional information, especially the relative spacing between transcription factor binding sites, may represent a common organizing principle of transcription control regions.

RECOMB'03, April 10-13, 2003, Berlin, Germany.

Categories and Subject Descriptors
J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms
Algorithms

Keywords
Phylogenetic footprinting, comparative genomics, multifactorial regulation, transcription regulation, promoter structure

1. INTRODUCTION

All organisms have evolved intricate signaling networks that sense and respond to their environment. At a cellular level, the activation of one or more signaling networks often leads to coordinated changes in gene expression, via the regulated activity and binding of transcription factors to transcription control regions (TCR's) of genes (e.g. enhancers and upstream activating sequences). In yeast and most other eukaryotes, the transcriptional regulation of individual genes is often multifactorial, as multiple transcription factors may bind to a single TCR [1] [2] [3]. In some cases, multiple transcription factors bind to a TCR and act independently of one another to alter gene expression in response to distinct cellular cues [4]; in other examples, multiple factors bind and/or act cooperatively to modulate gene expression via direct or indirect physical interactions with each other [5] [6] [7].

The challenges in understanding how regulatory information is encoded in genomes include both the identification of regulatory sequences in TCR's, and the elucidation of the constraints on productive multifactorial regulation. Many experiments have shown that specific pairs of factors must be bound near each other in order to act cooperatively [8] [9] [10], and it is on these spatial constraints that we focus here.

Previous computational work has been devoted to identifying putative transcription factor binding sites. A plethora of computational methods has been developed to find over-represented sequences in a subset of genes believed to contain a common transcription factor binding site (reviewed in [11]). The rapid pace of genome sequencing has enabled a complementary approach – phylogenetic footprinting (reviewed in [12] [13]) – that recognizes that the conservation of sequences across related organisms often reflects evolutionary selection for their presence in TCR's. Several algorithms have been developed to perform phylogenetic footprinting analyses systematically [14] [15] [16].

After compiling a collection of putative binding sites, associations can be made between various binding site assortments and gene expression. Some recent approaches include Boolean logic [17], regression methods [18] [19] [20], spatial clustering [21], and multiple binding site matrix classifiers [22] [23] [24]. Spatial information on the relative locations of binding sites is ignored in all but the last two classes of approaches. Yet even these methods, which often search for fixed arrangements among the individual binding sites, may miss permutations of binding sites within TCR's that may still be bound and regulated by their corresponding transcription factors.

The primary aim of this work was to incorporate positional information and phylogenetic footprinting to identify sequence motifs that may regulate gene expression. Consequently, we expanded the focus of phylogenetic footprinting from the conservation of contiguous sequences to higher-order features of TCR's, namely the spatial organization of individual binding sites. Since transcription factors participating in multifactorial regulation may require physical proximity among their binding sites, we searched for groups of conserved sequences that were more closely spaced in TCR's than expected. We refer to these spatially organized sequences as conserved word templates. As a proof of principle, we started with the simplest example of such templates: pairs of conserved 6-bp words. Conservation was assessed using the genome sequences of three additional *Saccharomyces* species, which were chosen to be sequenced in order to elucidate regulatory sequences conserved among these closely related species [25]. To exploit this comparative genome data, we have devised a method that systematically tested sequence pairs for joint conservation across genomes and close spacing within individual TCR's. Since genes regulated by the same set of transcription factors often display similar gene expression patterns in certain experimental conditions, we identified conserved word pair templates whose gene targets were associated with common changes in gene expression. We adopted a group-by-sequence approach to first identify genes that contained the word pair templates and then to test for significant associations with expression levels of the identified genes [26]. Significant associations between conserved word pair templates and specific gene expression changes, the prevalence of known transcription factor binding sites, and the enrichment for common functional roles among gene groups, suggest that conserved word pair templates comprise sequences important for multifactorial regulation in yeast.

2. CONSERVED WORD PAIR TEMPLATE ALGORITHM

2.1 Overview

We present a method to find conserved higher-order sequence templates from related *Saccharomyces* genomes. Our method incorporates sequential statistical tests, with each step focusing on a distinct property of conserved sequence templates. The simplest instances of sequence templates involve word pairs and their relative spacing. First, word pairs that show enriched conservation as a unit were identified using a chi-square test for independence. Next, the relative spacing of conserved word pairs was assessed using a permutation test. Finally, those conserved word pairs with close spacing were verified for functional importance by testing for gene expression differences between matching genes and the rest of the genome. The output for this algorithm is a P × C data matrix, whose entries correspond to the strength of association with differential gene expression, i.e. the K-S significance level (see §2.5). Note that P is the number of significant conserved word pairs, and C is the number of gene expression conditions.

2.2 Datasets

Whole-genome shotgun sequencing of *Saccharomyces bayanus*, *Saccharomyces mikatae*, and *Saccharomyces paradoxus* has been previously described [25]. All of these organisms are highly related to *Saccharomyces cerevisiae*, as they are grouped within the *sensu stricto* branch of the *Saccharomyces* genus [47]. Intergenic regions were aligned using CLUSTALW as described [25] and are available from the Saccharomyces Genome Database [43]. A total of 4101 CLUSTALW alignments were analyzed. These alignments were filtered for orthologs in at least 3 genomes.

Gene expression measurements were obtained from the Stanford Microarray Database [48] and Rosetta [34]. The main experimental types among the 342 conditions examined include diauxic shift [27], cell cycle [29] [30], environmental stress response [28], DNA damage [31] [32], low phosphate [33], cadmium (N. Ogawa and P. O. Brown, unpublished data), and inhibition of ergosterol biosynthesis [34]. This data has been log-transformed (base 2), and each experimental condition has been median normalized.

2.3 Dependent Conservation of Word Pairs

To assess whether two words were co-conserved in the same intergenic regions, a chi-square test of independence was systematically conducted for all possible words of length six. We defined a word to include a 6-bp sequence and its reverse complement. Define a transcription control region (TCR) for a gene as the 600 base pairs upstream of its translation start site. TCR's shared between divergently transcribed genes less than 600 bp long were only counted once. A word was labeled conserved in a TCR if all six bases were identical among three or more genomes in the CLUSTALW alignment. For each word pair (W, V) whose overlap was less than 4, a contingency table $\boldsymbol{C}_{wv}$ was constructed. In this table, $\boldsymbol{C}_{wv} = \#\,\mathrm{TCR}(\boldsymbol{I}_w \cap \boldsymbol{I}_v)$, where $\boldsymbol{I}_w$, $\boldsymbol{I}_v$ are indicator variables for the presence of each conserved word in a TCR. TCR's shared between divergently transcribed genes less than 600 bp long were only counted once. The expected counts $\boldsymbol{E}_{wv}$ were obtained from an independence assumption, *i.e.* the product of the individual word conservation probabilities, multiplied by the total number of TCR's. Thus the chi-square statistic with Yates continuity correction was computed according to the definition:

$$\chi^2_{wv} = \sum_{I_w=0}^{1} \sum_{I_v=0}^{1} \frac{(|C_{wv} - E_{wv}| - \frac{1}{2})^2}{E_{wv}} \qquad (1)$$

2.4 Spatial Proximity of Word Pairs

The second requirement for a conserved sequence template involved constraints on spatial arrangements between individual words. Any method that evaluates spacing distributions between word pairs must take into account positional biases that may be present for individual words (A. M. Moses, unpublished results). We used a permutation test to evaluate the significance of the average minimum distance, excluding overlaps, between conserved word pairs. By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions. Within any given TCR $\boldsymbol{t}$, define $\boldsymbol{p_t(W)} = \{\boldsymbol{p_t^1(W)}, \ldots, \boldsymbol{p_t^j(W)}\}$ as a vector of positions in *S. cerevisiae* where word $\boldsymbol{W}$ is conserved. Suppose that words $\boldsymbol{W}$ and $\boldsymbol{V}$ were jointly conserved in TCR's $\boldsymbol{T_1} \ldots \boldsymbol{T_N}$. Then the average minimum distance, $\overline{\boldsymbol{D}}$, can be computed as:

$$\overline{D}_{wv} = \frac{1}{T}\sum_{t=1}^{T}\min_{j,k}\left|p_t^j(W) - p_t^k(V)\right| \qquad (2)$$

We used a permutation test to generate an empirical null distribution of $\overline{\boldsymbol{D}}$ for all word pairs with $N \geq 10$. After randomly permuting the labels $\boldsymbol{t}$ for the position vectors of word V, a permutation test statistic, $\overline{\boldsymbol{D}}^*$, can be calculated as above. By repeating this resampling procedure R times, an empirical null distribution $\overline{\boldsymbol{D}}_{null} = \{\overline{\boldsymbol{D}}^{*1}, \ldots, \overline{\boldsymbol{D}}^{*R}\}$ can be obtained. The significance of the observed average minimum distance, $\overline{\boldsymbol{D}}$, in the N promoters was calculated as its quantile in the empirical null distribution $\overline{\boldsymbol{D}}_{null}$. We set an upper bound of $R = 10^6$, but stopped permutations early if 20 or more values in $\overline{\boldsymbol{D}}_{null}$ were found less than $\overline{\boldsymbol{D}}$.

Correction for multiple testing involved control of the proportion of false positives using a False Discovery Rate method [1]. This method has increased power over Bonferroni-type methods. Permutation quantiles for all N word pairs tested were sorted in non-decreasing order: $\boldsymbol{q_1} \leq \ldots \leq \boldsymbol{q_N}$. Let $\boldsymbol{k} = \max\left(\boldsymbol{i} : \boldsymbol{q_i} < \frac{0.05\boldsymbol{i}}{\boldsymbol{N}}\right)$. Then the first $\boldsymbol{k}$ word pairs in the ordering had a corrected significance level of $q < 0.05$, *i.e.* the rate of false positives is approximately 5%.

2.5 Association between Template-Specified Gene Groups and Gene Expression Changes

So far we have identified word pairs with two properties: dependent conservation and spatial proximity among all TCR's in the whole genome. These word pairs can be viewed as sequence-based rules for selecting a subset of genes based on the conservation of an element of TCR architecture. In this stage, we would like to evaluate the transcriptional information associated with these rules by assaying for gene expression changes among genes that match these sequence constraints.

For each gene expression condition $\boldsymbol{c}$ in our dataset, $\boldsymbol{c} \in \{1, \ldots, 342\}$, we tested the null hypothesis that a gene subset $\boldsymbol{G}_{\mathrm{wv}} \subseteq \boldsymbol{G}$ selected by a conserved word pair (w, v) had the same distribution of gene expression ratios ($\boldsymbol{E}_{\mathrm{wv}}^{\mathrm{c}}$) as the entire genome ($\boldsymbol{E}^c$). The alternate hypothesis stated that the two gene expression distributions were significantly different. Any gene was an element of $\boldsymbol{G}_{\mathrm{o}}$ if its corresponding TCR conserved both sequences in the word pair. Since the size N_{o} of gene subsets may be small and the distributions may not be normally distributed, we used the nonparametric Kolmogorov-Smirnov (K-S) test. The test statistic $\boldsymbol{K}$ compares the cumulative distribution functions $\boldsymbol{F}_{\mathrm{wv}}^{\boldsymbol{c}}$ and $\boldsymbol{F}^{\boldsymbol{c}}$ corresponding to $\boldsymbol{E}_{\mathrm{wv}}^{c}$ and $\boldsymbol{E}^c$ by the formula $\boldsymbol{K} = \max_{\boldsymbol{x}}\left|\boldsymbol{Fwv}^{\boldsymbol{c}}(\boldsymbol{x}) - \boldsymbol{F}^{\boldsymbol{c}}(\boldsymbol{x})\right|$. The significance level of an observed value $\boldsymbol{K}^*$ can be obtained using a numerical approximation [51].

A gene subset determined by a word pair was deemed to have significantly different expression if its K-S p-value was less than a certain threshold. To correct for multiple testing, this threshold was established by controlling the False Discovery Rate. The significance levels $\boldsymbol{p_i}$ from each K-S test were ordered in ascending order. Let $\boldsymbol{N}$ represent the total number of K-S tests performed, i.e. the number of jointly conserved, closely spaced word pairs times the number of gene expression experiments). If $\boldsymbol{k}$ was the largest $\boldsymbol{i}$ such that $\boldsymbol{p_i} < \boldsymbol{i}\alpha / \boldsymbol{N}$, then the first $\boldsymbol{k}$ word pairs in the ordering were deemed to have a significance level of $p < \alpha$.

We ensured that the K-S p-value for the conserved word pair subset $\boldsymbol{G}_{\mathrm{o}}$ was more significant than subsets $\boldsymbol{G_w}$ or $\boldsymbol{G_v}$ comprised of only one conserved word by computing $\boldsymbol{K}$ for $\boldsymbol{E}_w^{\mathrm{c}}$ vs. $\boldsymbol{E}_v^c$, as well as for $\boldsymbol{E}_w^{\mathrm{c}}$ vs. $\boldsymbol{E}^c$. The marginal improvement of the joint word pair was defined as: $\boldsymbol{K}$ ($\boldsymbol{F}_{\mathrm{o}}^{\boldsymbol{c}}$ vs. $\boldsymbol{F}^{\boldsymbol{c}}$) – max($\boldsymbol{K}$ ($\boldsymbol{F}_{\mathrm{w}}^{\boldsymbol{c}}$ vs. $\boldsymbol{F}^{\boldsymbol{c}}$), $\boldsymbol{K}$ ($\boldsymbol{F}_{\mathrm{v}}^{\boldsymbol{c}}$ vs. $\boldsymbol{F}^{\boldsymbol{c}}$)).

3. RESULTS

3.1 Identification of conserved word pair templates

We initialized our word list using all 2080 words of length six, treating a given word and its reverse complement as identical. For each TCR (consisting up to 600 bp upstream of an open reading frame), a word was labeled conserved if all six bases were identical in at least three of the four *Saccharomyces* genomes, based on the CLUSTALW alignment of that TCR. To systematically test whether words were conserved more often in the same intergenic regions of the *Saccharomyces* genomes than expected by independent conservation, a chi-square test was performed on all possible pairwise combinations of words (see §2.3). Pairs of words that overlapped each other by more than three nucleotides were excluded. A significant proportion of word pairs showed dependent conservation: among the 2.16 million word pairs tested, 8452 of them (~0.4%) had conservation c2 scores greater than 31.1. This threshold corresponds to a probability of 0.05 for obtaining one or more false positives after a Bonferroni correction for multiple testing.

Next, we selected word pairs that displayed closer physical spacing in intergenic regions than expected by chance. As a metric for the closeness in relative spacing between word pairs, the average minimum distance between two words in *S. cerevisiae*, $\overline{\quad}$, was calculated based on the genes whose TCR's conserved both words. If two non-overlapping words were closely spaced in all TCR's, we should find $\overline{\boldsymbol{D}}$ to be smaller than expected by chance. This spacing was assessed using a permutation test by selecting the set of genes that contained a conserved word pair and then randomizing the assignment of one of the words to the genes containing that word (see §2.4). By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions.

After correcting for multiple testing, a total of 1117 out of 8452 word pairs (~13%) had significantly small values (FDR $q < 0.05$) for $\overline{\boldsymbol{D}}$ (Figure 1). As a negative control, we also assayed a sample of word pairs that did not show dependent conservation (conservation $\chi^2 < 1$), yet were jointly conserved in at least 10 TCR's. Only 161 out of 42801 (~0.4%) random word pairs with non-dependent conservation ($\chi^2 < 1$) showed significantly small values for $\overline{\boldsymbol{D}}$. Figure 2 illustrates the distributions of $\overline{\boldsymbol{D}}$ for conserved word pair templates, jointly conserved word pairs, and randomly conserved word pairs. The medians of these distance distributions were 100 nucleotides, 116.5 nucleotides and 132 nucleotides, respectively. Notably, the median $\overline{\boldsymbol{D}}$ for template pairs was significantly smaller ($p < 0.05$) than the median $\overline{\boldsymbol{D}}$ for randomly conserved pairs. These results indicate that many of the word pairs that were conserved in the same intergenic regions of multiple *Saccharomyces* genomes also exhibited closer spacing in TCR's.

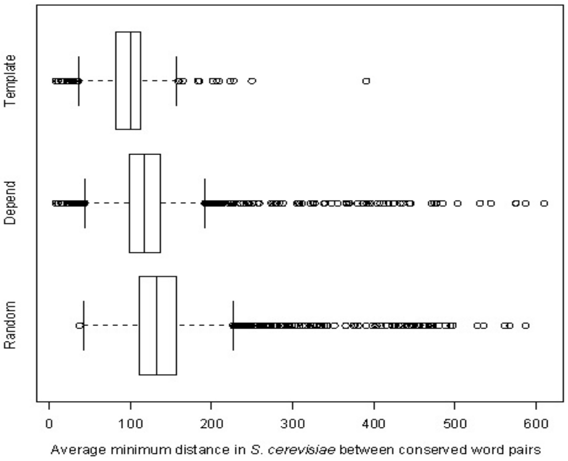

Figure 1) Word pairs in conserved word pair templates are closely spaced in *S. cerevisiae*

Template denotes closely spaced and jointly conserved word pairs ($\chi^2 > 31.1$, spacing $q < 0.05$, $N = 1117$). *Depend* denotes dependently conserved word pairs ($\chi^2 > 31.1$, $N = 8452$) and includes all of the word pairs in the template category. *Random* denotes a sample of randomly conserved word pairs ($\chi^2 < 1$, $N = 4667$). For each category, the distribution of average minimum distances is represented by a box-and-whisker plot.

3.2 Conserved word pair templates were significantly associated with gene expression

Our method identified conserved word pair templates that were statistically significant with respect to both co-conservation in multiple genomes and close spacing in *S. cerevisiae* TCR's. To evaluate the regulatory information in these templates, we assessed the statistical association between gene groups that shared a template and changes in gene expression. Similar to other group-by-sequence approaches for finding regulatory sequences, we expect that gene subsets defined by common TCR sequence rules should have gene expression patterns that are similar under conditions where the transcription factors are active, yet are different from the average expression of genes in the genome [26].

To assess the association between conserved word pair templates and differentially expressed genes, we identified gene subsets that contain both conserved words in the template within their TCR's and observed their expression patterns in S. cerevisiae in publicly available datasets ([27] to [34], see §2.5). We then conducted Kolmogorov-Smirnov (K-S) tests to evaluate for differential gene expression between each gene subset and the whole genome. A P × C matrix was computed: each conserved word pair in P was assigned a K-S p-value for each experimental condition observed in C. (see §2.5). Entries in this matrix (K-S p-values) were filtered out if the K-S *p*-value: (1) did not meet the threshold for multiple testing; or (2) was less than 10 times more significant than the K-S *p*-value for a gene subset associated with either word alone (see §2.5). The latter criterion discounts gene expression changes that are due predominantly to the action of a single transcription factor.

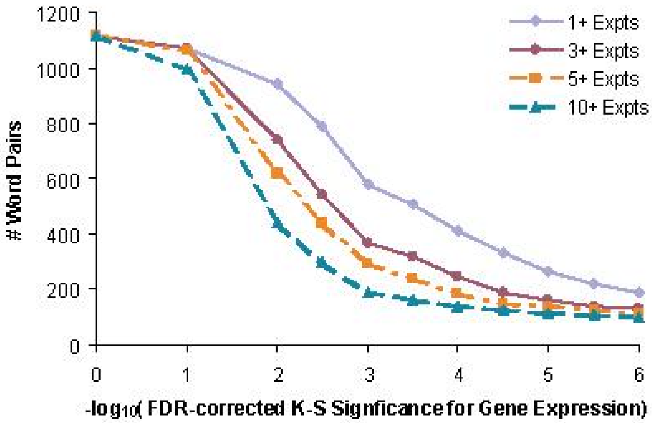

Figure 2) Total number of conserved word pair template associations at different K-S significance values

The horizontal axis shows different multiple testing-corrected significance levels for the K-S test (see §2.5). The number of closely spaced word pairs meeting this cutoff for different minimum numbers of expression conditions is shown on the vertical axis. Word pairs were also filtered for an improvement of 10× over the K-S significance from any single word.

Figure 2 displays the number of conserved word pair templates that were significantly associated with gene expression changes, for varying significance levels of the K-S test, which have been corrected for multiple testing (see §2.5). Each line indicates the number of gene subsets that were significant in a different minimum number of experimental conditions. Several hundred closely spaced word pairs were significantly associated with differential gene expression. For example, 293 word pairs met an FDR-corrected significance threshold of $p < 10^{-3}$ for 5 or more experimental conditions.

3.3 Many identified sequences represent known transcription factor binding sites

Conserved word pair templates that were most strongly associated with gene expression changes also agreed with prior experiments on transcription factors [35]. In all analyses described below, we used a set of 339 word pairs that had significant associations with gene expression changes at an FDR-corrected multiple testing threshold of 0.005 for 10 or more experiments. For visualization purposes, we organized the P × C matrix by hierarchically clustering the K-S p-values for the 339 word pairs.

Hierarchical clustering of this output matrix identified groups of word pairs with similar K-S p-values in specific subsets of experimental conditions (Figure 4). In many cases, the word pairs that clustered together also comprised overlapping hexamer sequences, suggesting that some of the hexamers in different pairs may represent a larger, somewhat variable sequence (Table 1). For example, group #13 in Figure 4 includes 8 word pairs. In each of these word pairs, one of the component words – such as TCACGT, GCACGT or CACGTG – matched part of the Cbf1p or Pho4p binding sites. The other component word in each pair – such as AACTGT, ACTGTG, CTGTGG, TGTGGC or GTGGCT – represented part of the known Met31/32p binding site (AAACTGTGG). Therefore, genes whose TCR's contained any word pair within this group likely contained a conserved Cbf1p or Pho4p binding site, along with a conserved Met31/32p binding site, and the distances between the conserved sites in these genes were also smaller than expected by chance. These results agree with the known interaction of Cbf1p and Met31/32p for the regulation of genes involved in sulfur utilization (see Discussion).

Table 1 shows a partial list of the 13 most significant groups of consensus sequences, which were assembled by joining adjacent word pairs in the clustered output matrix with overlapping sequences. Many of these consensus sequences matched transcription factor binding sites that had been biochemically verified. Several pairs of transcription factors, denoted by boldface in Table 1, were not previously known to participate in multifactorial regulation. Three of these pairs included new putative transcription factor binding sites. In group 8, the word ACAGCC is found in a template with the GATA motif. In group 9, the word CGGGCC is found in a template with the binding site for the stress-induced transcription factor, Msn2/4p. In group 2, one of the words in each word pair was the binding site for Swi4/6p, which regulates the expression of cell-cycle dependent genes. The other word in each pair was an invariant CGCCAA, which is highly similar to, though distinct from, the characterized Swi4/6 binding site CRCGAAA [35].

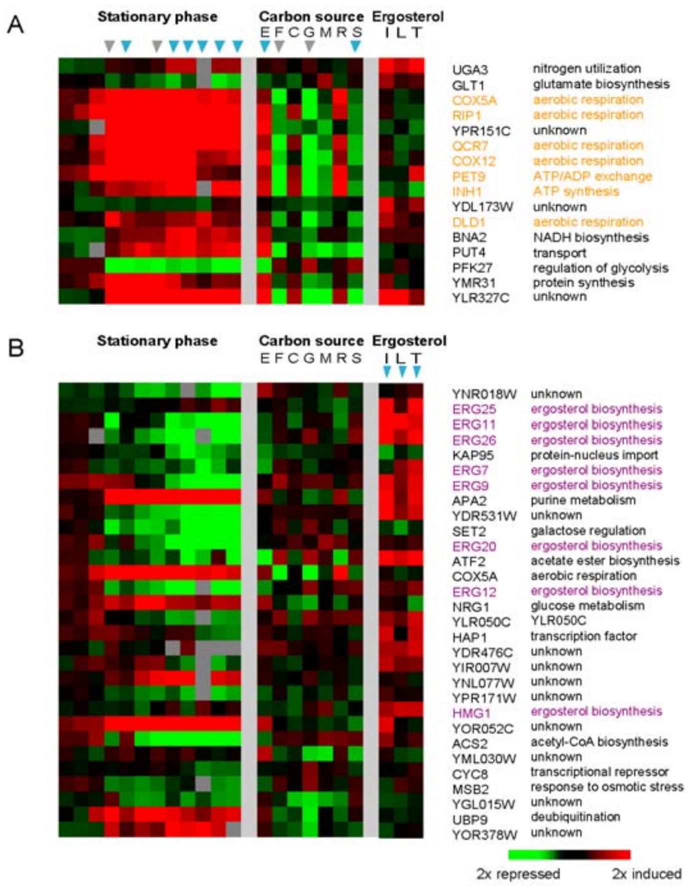

Figure 3) Multifactorial regulation of Hap1p target genes

Gene expression patterns are shown for genes whose TCR's contain binding sites for: (A) Hap1p (CCGATA) and Hap2/3/4/5p (CCAATC); or (B) Hap1p (CCGATA) and Ecm22p/Upc2p (TCGTTT). The genes are listed in ascending order of minimum distance between the two conserved words in the corresponding TCR of *S. cerevisiae*. Each row represents a given gene's expression pattern under the conditions shown in each column: progression into stationary phase (2 h, 4 h, 8 h, 12 h, 1 day, 2 days, 3 days, 5 days, 7 days, 13 days, 22 days, 28 days of growth) [28]; steady-state growth on different carbon sources: ethanol (E), fructose (F), galactose (C), glucose (G), mannose (M), raffinose (R) and sucrose (S) [28]; and growth in the presence of drugs that inhibit ergosterol biosynthesis: itraconazole (I), lovastatin (L) and terbinafine (T) [34]. A red color indicates that the gene's expression was induced under those conditions, while a green color indicates that the gene was repressed under those conditions; black indicates no detectible change in expression, and grey indicates missing data. Gene names highlighted in orange (A) or in purple (B) correspond to genes whose products are involved in respiration and ergosterol biosynthesis, respectively. Arrows above the columns indicate conditions in which the displayed gene groups show significant gene expression changes according to the Kolmogorov-Smirnov test, after False Discovery Rate correction for multiple testing at a *p*-value of 0.005 (blue) or 0.01 (grey).

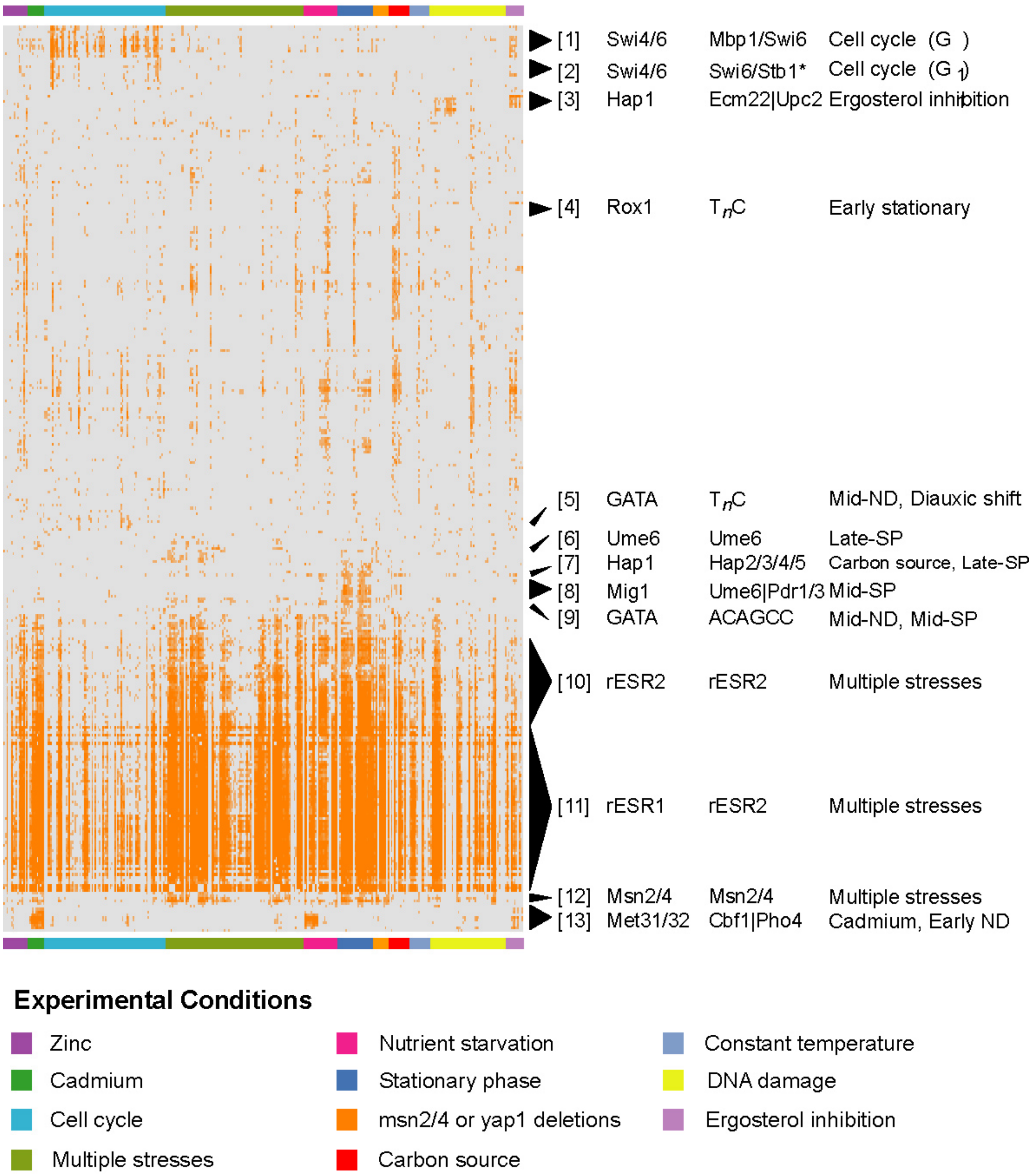

Figure 4) Specific patterns of gene expression changes associated with templates

The *P* × *C* matrix of K-S *p*-values was hierarchically clustered by rows and visualized with TreeView (http://rana.lbl.gov). Each row corresponds to a conserved word pair template, and each column represents a single gene expression experiment. The experimental conditions are indicated by the color bar above and below the figure, according to the key shown below. The value in each cell corresponds to the K-S *p*-value of gene expression changes in each condition (column) for a group of genes that contain the conserved word pair template (row) in their TCR's. An orange color denotes a K-S *p*-value below the FDR critical value of 0.005 for multiple testing, while grey represents values that were not significant. Word pairs that failed to meet a False Discovery Rate critical value of 0.005 for multiple testing in 10 or more experiments are not shown. Some of the most significant conserved word pair associations are labeled and annotated in Tables 1 and 2. Abbreviations for experimental conditions include: ND (nitrogen depletion), SP (stationary phase).

Table 1) Gene expression associations for most significant groups of word pairs

The output ***P*** × ***C*** matrix of word pairs (***P***) that were significantly associated ($p < 0.005$) with at least 10 or more environmental conditions (***C***) was ordered using hierarchical clustering. Numbers correspond to groups of overlapping word pairs indicated in Figure 4. Boldface denotes sequence pairs whose involvement in multifactorial regulation has not been previously reported. Consensus sequences were assembled from groups of word pairs that were found in adjacent rows in the ordering. Residues are shown in bold if it is contained in at least two hexamers. Numbers denote the groups that are indicated in Figure 4. IUPAC codes used: K (G or T); M (A or C); R (A or G); S (C or G); W (A or T).

			Most significant word pair in consensus group			
	Conserved Word Pairs (Consensus of overlapping words)	Known transcription factors or motifs	Conservation (χ^2, *p*-val via Bonferroni)	Avg. min dist $\overline{\boldsymbol{D}}$	# TCR	Expression conditions with significant gene subsets (FDR significance)
1	RCGAAA, RACGCG,	Swi4/6, Swi6/Mbp1	83.0 (7×10^{-13})	68.1	36	Cell cycle, G1 phase (10^{-6})
2*	**CACGAAA, CGCCAA**	**Swi4/6, Stb1 (putative)**	**55.6 (2×10^{-7})**	**78.2**	**25**	**Cell cycle, G1 phase (10^{-4})**
3*	**CCGATA, TC[GT]TTT**	**Hap1, Ecm22 | Upc2**	**36.2 (0.004)**	**81.7**	**30**	**Ergosterol inhibition (10^{-4}) MMS (DNA damage) (10^{-3})**
4*	**GATAAG, TTCTTT**	**GATA, T_nC**	**36.0 (0.005)**	**100.5**	**88**	**Nitrogen depletion 8h (10^{-5})**
5	GGCTAA CGGCGG	Ume6, Ume6	179.2 (2×10^{-34})	81.9	15	Late stationary phase (10^{-4})
6	CCGATA, CCAATC	Hap1, Hap2/3/4/5	35.0 (0.007)	88.6	16	Stationary phase (10^{-4}) Ethanol (10^{-4})
7	ACCCCA, CCGCCG	Mig1, Ume6 | Pdr1/3	66.7 (7×10^{-10})	70.5	16	Stationary phase (10^{-6}) Ethanol (10^{-5})
8*	**GATAAG, ACAGCC**	**GATA, Novel**	**39.5 (0.004)**	**75.5**	**21**	**Nitrogen depletion 8,12h (10^{-5}) Stationary phase 10h-2d (10^{-4})**
9*	**AAGGGG, CGGGCC**	**Msn2/4, Novel**	**33.4 (0.016)**	**79.6**	**14**	**Elutriation 2d, 4d, 6d (10^{-5}) Stationary phase 10h-3d (0.005)**
10	ANTGAAA, GAAAAWT	rESR2 (Overlap)	96.9 (2×10^{-16})	96.8	68	Repressed in multiple environmental stresses (10^{-6})
11	G[AC]GATGAG TGAAAATTTT	rESR1 motif, rESR2 motif	240.6 (10^{-49})	41.6	183	Repressed in multiple environmental stresses (10^{-6})
12	AWAAGG, AGGGG	Msn2/4 (Overlap)	94.7 (5×10^{-16})	99.0	29	Multiple stresses (10^{-3})
13	ACTGTGGC, [GT]CACGTG	Met31/32, Cbf1 | Pho4	47.5 (2×10^{-5})	43.5	22	Amino acid starv. (10^{-6}) Nitrogen depletion (10^{-6}) Cadmium (10^{-6})

Recent chromatin immunoprecipitation experiments suggested that this sequence may represent the binding site for Stb1, a transcription factor that binds Swi6 in vitro and is implicated in cell cycle regulation [36]. This sequence was found in several genes adjacent to intergenic regions bound by Stb1 *in vivo* [37].

Some groups of genes with shared word pair templates were enriched for known targets of transcription factors. Genes with a conserved half-site for the Hap1p transcription factor, as well as a conserved Hap2/3/4/5p binding site, in their TCR's were significantly associated with induction in late stationary phase (Figure 3A). In addition, many of these genes were more highly expressed in growth medium containing ethanol, relative to other carbon sources (Figure 3A). Many of these genes encode aerobic respiration enzymes, which are required for the switch from fermentation to respiration [38] [39]. Indeed, both the Hap1p transcription factor and the Hap2/3/4/5p transcription factor complex are known to regulate the expression of these genes in response to heme and/or oxygen availability and carbon source, respectively. By contrast, gene groups with both a conserved Hap1p binding site and a conserved Ecm22p or Upc2p binding site in their TCR's were only significantly associated with induction in the presence of a drug that inhibited ergosterol biosynthesis (Figure 3B). This group of 30 genes contained 8 ergosterol biosynthesis genes; this proportion represented an enrichment compared to the rest of the genome. The transcription factors Ecm22p and Upc2p have been shown to induce the expression of ergosterol biosynthesis genes in response to low intracellular concentrations of ergosterol, while Hap1p is known to regulate the expression of these genes according to the availability of heme and oxygen which are required for the pathway (see Discussion) [40]. Note that the gene groups shown in Figure 3 and Figure 3 showed significant gene expression changes in different sets of environmental conditions.

4. DISCUSSION

This work describes two principles for analyzing combinations of regulatory sequences. First, sequence conservation among closely related yeast species was used to find sequences that were more likely to be functionally important. Secondly, a template approach that considered joint positional distributions of word pairs increased the specificity of gene expression predictions using sequence-based rules. We have demonstrated that higher-order sequence features within TCR's were conserved across multiple *Saccharomyces* genomes. Closely spaced and jointly conserved word pairs were also more likely to be associated with gene expression changes. A large proportion of words contained in templates matched known transcription factor binding sites, and some of the uncharacterized words may represent novel regulatory sequences. In many cases, associations between templates and gene expression changes were significant in conditions when the corresponding transcription factors are known to be active. In addition, groups of genes that co-conserved both words in a template often were enriched for common functional roles. These results suggest that conserved word pair templates, which were discovered strictly based on higher-order properties of sequence conservation, also carry biological relevance.

Conserved word pair templates may be classified under several distinct classes of regulatory elements in TCR's. One possible interpretation of templates is that closely spaced sequence pairs may promote direct or indirect interactions between transcription factors by increasing the local concentrations of the individual factors. For example, the proximity of Cbf1p and Met31/32p binding sites may promote interaction between these factors in recruiting their common transcriptional activators, Met4 and Met28. Experimental studies on the TCR's of *MET3* and *MET28* have demonstrated that the binding of Cbf1p enhances the DNA binding affinity of Met31/32p [41]. Indeed, biochemical experiments suggest that all of these proteins interact at the TCR's of some sulfur utilization genes [41].

Another possible regulatory scheme for conserved, closely-spaced word pairs is that individual sequences found in templates may correspond to binding sites for transcription factors that bind independently under the same or separate conditions. The Hap1p and Hap2/3/4/5p transcription factors, whose binding sites were identified in a template, represent an example of multifactorial regulation in response to different environmental stimuli [42]. In some cases, templates could discern genes that shared binding sites for one transcription factor, but were differentially expressed under certain sets of conditions. Genes that conserved both Hap1p and Hap2/3/4/5p binding sites in their TCR's included genes encoding mitochondrial enzymes, as well as respiration proteins, that showed significant induction during growth in stationary phase and ethanol (Figure 3A). By contrast, genes that conserved both the Hap1p and Upc2p/Ecm22p binding sites in their TCR's were enriched for genes encoding ergosterol biosynthesis enzymes. Unlike the genes encoding mitochondrial enzymes, these genes showed no expression changes in response to different carbon sources, yet they were significantly induced under treatment with drugs that inhibit ergosterol biosynthesis: itraconazole, lovastatin and terbinafine (Figure 3B) [34]. A biochemical link between these two enzyme categories may explain their common regulation via Hap1p: the protein products of these genes all require the cofactor heme, whose intracellular levels are sensed by Hap1p [43] [44]. Our results suggest that Hap1p controls the expression of all of these genes in response to heme and/or oxygen levels. The expression of genes encoding mitochondrial and respiration enzymes may be controlled by Hap2/3/4p in response to nonfermentable carbon sources, whereas the expression of ergosterol biosynthesis genes may be regulated by Ecm22p and/or Upc2p in response to ergosterol levels. Whether these factors act cooperatively with, or independently of, Hap1p will require further biochemical investigation to elucidate.

Close spacing between word pairs may be important for reasons other than the promotion of transcription factor interactions. Different regions of TCR's at varying windows away from translation start sites may be more competent at recruiting or inhibiting RNA polymerase. These differences may be influenced by nucleosome accessibility, chromatin structure, or DNA physical properties, which can be correlated with local A/T content (see [45] for references). Notably, we have also found that the relative proportions of A and T nucleotides vary considerably within the 200 bp closest to translation start sites (A. M. Moses, M. B. Eisen and Audrey Gasch, unpublished results). Low-complexity words that contained 4 or more A's or T's could be found in many templates (denoted by TnC in Figure 4 and Table 1); these words may serve as surrogates for a distance window from translation start. Transcription factor binding sites that are closely spaced to these low-complexity words may be found in more transcriptionally competent regions of TCR's.

Since transcription factor binding sites often contain degenerate positions that reflect specificity, a key limitation of our approach is the use of exact words [11]. The known binding sites listed in Table 1 correspond to transcription factors with high sequence specificities. Since other known binding sites are poorly modeled by exact words (in that they bind sequences with relaxed specificity at certain positions in their binding sites), our method has failed to include them in conserved word pair templates. In addition, our method currently requires sequence identity for a word to be labeled as conserved. This strict requirement omits binding sites that may retain their function, despite mutations in degenerate positions that may have little impact on transcription factor binding.

The consideration of joint conservation and close spacing has provided insights into how TCR organization may influence the multifactorial regulation of gene expression in *Saccharomyces cerevisiae*. These criteria were motivated by experimental studies on the positional organization of individual binding sites within TCR's, with the hypothesis that this underlying architecture would be functionally conserved. Even more complicated higher-order sequence rules are apparent in the organization of cis-regulatory modules in *Drosophila melanogaster* [46]. Nevertheless, a common organizational theme of the TCR's in both of these organisms is the importance of relative spacing between transcription factor binding sites. The discovery of additional principles for TCR organization will further advance our understanding of how regulatory information is encoded in genome sequences.

5. ACKNOWLEDGEMENTS

We thank Audrey Gasch and Justin Fay for their critical reading of the manuscript and insightful suggestions during the development of this work; Peter Bickel and John Storey for their statistical advice; and Nabuo Ogawa and Patrick Brown for sharing gene expression data on cadmium conditions prior to publication. D.Y.C. is a Howard Hughes Medical Institute Predoctoral Fellow, and M.B.E. is a Pew Scholar in the Biomedical Sciences. This work was conducted under the US Department of Energy contract No. ED-AC03-76SF00098.

6. REFERENCES

[1] Composite regulatory elements: structure, function and classification. [http://www.gene-regulation.com/ pub/databases/transcompel/compel.html]

[2] Wolberger C: Multiprotein-DNA complexes in transcriptional regulation. *Annu Rev Biophys Biomol Struct* 1999, 28:29-56.

[3] Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA: A compilation of composite regulatory elements affecting gene-transcription in vertebrates. *Nucleic Acids Res* 1995, 23:4097-4103.

[4] Gasch AP: The environmental stress response: a common yeast response to diverse environmental stresses. In: *Yeast Stress Responses* Edited by S Hohmann, WH Mager, vol. 1. pp. 11-70. Berlin: Springer; 2003: 11-70.

[5] Mead J, Bruning AR, Gill MK, Steiner AM, Acton TB, Vershon AK: Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast. *Mol Cell Biol* 2002, 22:4607-4621.

[6] Bhoite LT, Allen JM, Garcia E, Thomas LR, Gregory ID, Voth WP, Whelihan K, Rolfes RJ, Stillman DJ: Mutations in the Pho2 (Bas2) transcription factor that differentially affect activation with its partner proteins Bas1, Pho4, and Swi5. *J Biol Chem* 2002, 277:37612-37618.

[7] Verger A, Duterque-Coquillaud M: When Ets transcription factors meet their partners. *Bioessays* 2002, 24:362-370.

[8] Ambrosetti DC, Basilico C, Dailey L: Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol Cell Biol* 1997, 17:6321-6329.

[9] Ludwig MZ, Patel NH, Kreitman M: Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* 1998, 125:949-958.

[10] Liu Z, Little JW: The spacing between binding sites controls the mode of cooperative DNA-protein interactions: implications for evolution of regulatory circuitry. *J Mol Biol* 1998, 278:331-338.

[11] Stormo GD: DNA binding sites: representation and discovery. *Bioinformatics* 2000, 16:16-23.

[12] Duret L, Bucher P: Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* 1997, 7:399-406.

[13] Pennacchio LA, Rubin EM: Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* 2001, 2:100-109.

[14] Blanchette M, Tompa M: Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*. 2002, 12:739-748.

[15] Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* 2002, 12:832-839.

[16] Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* 2000, 10:577-586.

[17] Pilpel Y, Sudarsanam P, Church GM: Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 2001, 29:153-159.

[18] Bussemaker HJ, Li H, Siggia ED: Regulatory element detection using correlation with expression. *Nat. Genet.* 2001, 27:167-171.

[19] Keles S, van der Laan M, Eisen MB: Identification of regulatory elements using a feature selection method. *Bioinformatics* 2002, 18:1167-1175.

[20] Wang W, Cherry JM, Botstein D, Li H: A systematic approach to reconstructing transcription networks in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* 2002, 99:16893-16898.

[21] Wagner A: Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* 1999, 15:776-784.

[22] Klingenhoff A, Frech K, Quandt K, Werner T: Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* 1999, 15:180-186.

[23] Pavlidis P, Furey TS, Liberto M, Haussler D, Grundy WN: Promoter region-based classification of genes. *Proceedings of the Pacific Symposium on Biocomputing* 2001, 6:151-164.

[24] Kel-Margoulis OV, Ivanova TG, Wingender E, Kel AE: Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac Symp Biocomput* 2002, 7:187-198.

[25] Kamvysselis M, Patterson N, Birren B, Berger B, Lander ES: Whole-genome comparative annotation and regulatory motif discovery in multiple yeast species. *Proceedings of the 7th International Conference on Research in Computational Molecular Biology 2003*, 7.

[26] Chiang DY, Brown PO, Eisen MB: Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics* 2001, 17 Suppl 1:S49-S55.

[27] DeRisi JL, Iyer VR, Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997, 278:680-686.

[28] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 2000, 11:4241-4257.

[29] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998, 9:3273-3297.

[30] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, et al: A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* 1998, 2:65-73.

[31] Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol Biol Cell* 2001, 12:2987-3003.

[32] Lee SE, Pellecioli A, Demeter J, Vaze MP, Gasch AP, Malkova A, Brown PO, Botstein D, Stearns T, Foiani M, et al: In: *Biological Responses to DNA Damage*, vol. 65. pp. 303-314. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2000: 303-314.

[33] Ogawa N, DeRisi J, Brown PO: New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. *Mol. Biol. Cell* 2000, 11:4309-4321.

[34] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai HY, He YDD, et al: Functional discovery via a compendium of expression profiles. *Cell* 2000, 102:109-126.

[35] Zhu J, Zhang MQ: SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* 1999, 15:607-611.

[36] Ho Y, Costanzo M, Moore L, Kobayashi R, Andrews BJ: Regulation of transcription at the Saccharomyces cerevisiae start transition by Stb1, a Swi6-binding protein. *Mol Cell Biol* 1999, 19:5267-5278.

[37] Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al: Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002, 298:799-804.

[38] Gancedo JM: Yeast carbon catabolite repression. *Microbiol Mol Biol Rev* 1998, 62:334-361.

[39] Kwast KE, Burke PV, Poyton RO: Oxygen sensing and the transcriptional regulation of oxygen-responsive genes in yeast. *J Exp Biol* 1998, 201:1177-1195.

[40] Vik A, Rine J: Upc2p and Ecm22p, dual regulators of sterol biosynthesis in Saccharomyces cerevisiae. *Mol Cell Biol* 2001, 21:6395-6405.

[41] Blaiseau PL, Thomas D: Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA. *EMBO J* 1998, 17:6327-6336.

[42] Guarente L, Lalonde B, Gifford P, Alani E: Distinctly regulated tandem upstream activation sites mediate catabolite repression of the CYC1 gene of *S. cerevisiae*. *Cell* 1984, 36:503-511.

[43] Burke PV, Poyton RO: Structure/function of oxygen-regulated isoforms in cytochrome c oxidase. *J Exp Biol* 1998, 201:1163-1175.

[44] Lees ND, Skaggs B, Kirsch DR, Bard M: Cloning of the late genes in the ergosterol biosynthetic pathway of Saccharomyces cerevisiae--a review. *Lipids* 1995, 30:221-226.

[45] Liao GC, Rehm EJ, Rubin GM: Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 2000, 97:3347-3351.

[46] Berman BP, Nibu Y, Pfeiffer BD, Tomancek P, Celniker SE, Levine M, Rubin GM, Eisen MB: Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* 2002, 99:757-762.

[47] Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M: Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res*. 2001, 11:1175-1186.

[48] Saccharomyces Genome Database. [http://genome-www.stanford.edu/Saccharomyces/]

[49] Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, et al: The Stanford Microarray Database. *Nucleic Acids Res* 2001, 29:152-155.

[50] Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 1995, 57:289-300.

[51] Press WH, Teukolsky SA, Vertterling WT, Flannery BP: *Numerical Recipes in C*, Second Edition. Cambridge: Cambridge University Press; 1992.