

Transcription factor binding element detection using functional clustering of mutant expression data

Gengxin Chen, Naoya Hata and Michael Q. Zhang*

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

Received October 9, 2003; Revised and Accepted March 30, 2004

ABSTRACT

As a powerful tool to reveal gene functions, gene mutation has been used extensively in molecular biology studies. With high throughput technologies, such as DNA microarray, genome-wide gene expression changes can be monitored in mutants. Here we present a simple approach to detect the transcription-factor-binding motif using microarray expression data from a mutant in which the relevant transcription factor is deleted. A core part of our approach is clustering of differentially expressed genes based on functional annotations, such as Gene Ontology (GO). We tested our method with eight microarray data sets from the Rosetta Compendium and were able to detect canonical binding motifs for at least four transcription factors. With the support of chromatin IP chip data, we also predict a possible variant of the Swi4 binding motif and recover a core motif for Arg80. Our approach should be readily applicable to microarray experiments using other types of molecular biology techniques, such as conditional knockout/overexpression or RNAi-mediated 'knockdown', to perturb the expression of a transcription factor. Functional clustering included in our approach may also provide new insights into the function of the relevant transcription factor.

INTRODUCTION

After the completion of human genome sequencing, the next great challenge is to understand the functions of all human genes. Elucidating the regulation of genes and eventually deciphering the entire genetic network will reveal the functions of genes during development processes and responses to environmental stimuli. This provides a deeper insight into the mechanisms of diseases and the identification of therapeutic targets. However, even in a relatively 'simple' organism, such as yeast, understanding the gene regulation network is still a formidable task.

The first step towards the goal of understanding gene regulation is to identify the regulator–target relations. Generations of biologists have made tremendous efforts to delineate such relations experimentally. Databases, such as

TRANSFAC (1) and SCPD (2), have been established to collect information from the literature about transcription factors (TFs), their target genes and binding sites. However, experimental identification of TF binding sites is slow and laborious. Computational methods have become increasingly important, especially after the emergence of high throughput technologies, such as DNA microarrays. In addition, large-scale projects such as the Saccharomyces Genome Deletions Consortium (3), Compendium of expression profiles (4) and Gene Ontology Consortium (5), also provide computational biologists with new opportunities to cross-check and integrate information from different sources to infer TF–target relations and determine the binding motifs of TFs.

A popular method to analyze microarray data at present is to cluster genes based on the similarity of their expression profiles (6,7). It has also been used to identify *cis*-regulatory elements (8,9). The rationale is that co-expressed genes are likely to be co-regulated and, therefore, may share common regulatory elements. A further development is to incorporate TF binding information and identify combinatorial regulation of TFs (Kato *et al.*, submitted for publication). Bussemaker *et al.* (10) took an alternative strategy to identify motifs correlated with gene expression by fitting a linear regression model. These methods successfully discovered some motifs corresponding to known binding sites and predicted some new motifs, but they did not directly reveal which TFs might bind to those sequence elements. More recently, Birnbaum *et al.* (11) and Zhu *et al.* (12) searched for TF/*cis*-element relationships by correlating the expression profile of a TF to composite or simple profiles of putative target genes. Methods described above achieve various degrees of success, however, each of them is limited in their effectiveness in some aspects. For example, a gene may be regulated by several TFs cooperatively or by different TFs under different conditions. Therefore, mRNA levels of genes regulated by the same TF may not be well correlated with each other or with the mRNA level of the regulating TF across all experimental conditions. In addition, a TF may require cofactors or be regulated post-translationally, thus the transcriptional activity of that TF, which is determined by the concentration of related functional proteins, is not always correlated with its mRNA level.

Mutants have been used to study gene function extensively in the history of genetics. Since the invention of DNA microarrays, this technology has been widely applied to large-scale experiments in which a TF is deleted or overexpressed via genetic manipulation, in order to identify the global target genes of the TF (see for example 13,14). However, a big

*To whom correspondence should be addressed. Tel: +1 516 367 8865; Fax: +1 516 367 8461; Email: mzhang@cshl.edu

Table 1. Known binding consensus sequences of the TFs under study

TF	SCPD	TRANSFAC	Other
Gcn4	TGANTN	RTGACTCATNS or ARTGACTCW	
Gln3	3 sites in record, consensus not constructed		TTNCTGATAAGG (34)
Mac1	GAGCAAA		
Mbp1	WCGCGW		ACGCGT (8)
Ste12	TGAAACA	ATGAAAC	
Swi4	CNCGAAA		CACGAAA or CGCGAAA (8)
Swi5	KGCTGR		RRCCAGCR (8)
Yap1	TTANTAA (as AP-1)	TGASTCAG or TGASTMA (AP-1 of multiple species pooled)	TTACTAA (25)

challenge to this type of method is the difficulty in distinguishing between direct and indirect effects of the genetic manipulation. When a TF is deleted or overexpressed, some of its affected target genes may be TFs as well, which in turn up-regulate or down-regulate target genes not directly regulated by the original TF. In this paper, we present a simple method to make use of TF mutant microarray data to discover the binding motif of the manipulated TF. To deal with the indirect effects mentioned previously, we use a functional clustering technique based on GO annotation of yeast genes (15). The rationale of our approach is that genes involved in the same process or pathway are more likely to be co-regulated (16). Functional clustering may not separate direct and indirect target genes completely, but if direct target genes of the TF are concentrated in one of the clusters and the motif sequence signal is sufficiently strong, our motif search algorithm will likely detect the binding motif of the TF. In fact, many known *cis*-regulatory elements are recovered by applying promoter motif search tools to functional groups of genes (17,18), such as MIPS gene categories. Although a similar GO-based method has been proposed by Pavlidis *et al.* (19), the effectiveness of this type of clustering for TF binding motif detection has not been assessed. A more recent publication of large-scale chromatin immunoprecipitation (ChIP) chip experiments (20) provides genome-wide *in vivo* TF binding information, which is very valuable for TF motif searches. We do not include those data in our primary analysis (see Discussion), but use them to validate the clusters and motifs we obtained. We tested our method with eight microarray datasets from the Rosetta Compendium (4), in which a TF with a known binding motif is deleted in each experiment. We were able to detect the canonical binding motif of the deleted TF in five experiments (four with high confidence), suggesting that our method may be able to predict candidate motifs for TFs whose binding sites are unknown and provide a guide for future experimental design. We believe that the method presented here is not limited to TF gene deletion, but should be applicable to other types of manipulations of TFs, such as overexpression or RNAi. We name our algorithm BEAUTI, for binding element analysis using TF intervention (available at <http://rulai.cshl.edu/tools/beauti/>).

MATERIALS AND METHODS

Data source

We use yeast microarray expression data from the Rosetta Compendium (4). Among the 276 deletion mutants examined

by Rosetta, 40 deleted genes are identified as TFs or cofactors in the MIPS gene functional category 'transcriptional control' (<http://mips.gsf.de/proj/yeast/catalogues/funcat/>). Among those 40 genes, eight DNA-binding TFs are documented in SCPD (2). The consensus sequences of the binding sites are also verified in TRANSFAC (1) or the literature (Table 1). The deletion mutant expression profiles of those eight genes were selected to test our method for detecting the binding motif of each of the factors. In an attempt to predict novel binding motifs, we also applied our method to a few other TF deletion experiments in the Rosetta Compendium in which the binding sites of the relevant TFs are not well defined. TF binding (ChIP chip) data from Lee *et al.* (20) and Iyer *et al.* (21) were used to verify the motifs found with our method.

Significant gene and background gene selection

The Rosetta dataset provides the \log_{10} ratio (mutant versus wild-type or control) of the expression level for each gene on the arrays. A *P*-value calculated based on their error model is also available to indicate the likelihood of differential expression between the mutant and the control (4, Supplementary Material). We say that a gene is significant if its expression is up-regulated or down-regulated at least 1.5-fold in the mutant with a *P*-value <0.05. For the purpose of motif search, we also need specific background genes. For each TF, we select all the genes on the array with a *P*-value >0.5 and sort them by the \log_{10} ratio of expression, the 2000 genes with the least absolute \log_{10} ratio being selected as background genes.

Functional clustering based on GO annotation

We downloaded Biological Process Gene Ontology (5) from the GO Consortium website (<http://www.geneontology.org/>). The annotation of all yeast genes in GO terms in the Saccharomyces Genome Database (SGD) (15) is also available on the GO Consortium website. In GO relationships, one node may have more than one parent. Therefore, strictly speaking, the relationships between GO terms (nodes) form a directed acyclic graph (DAG). In addition, one gene may be assigned to more than one node because it may be involved in multiple processes/pathways.

Conceptually, we regard every node in the GO DAG as a potential cluster that contains all the significant genes assigned to this node and all its descendent (direct and indirect children) nodes. With a limited number of significant genes, the vast majority of these clusters are typically empty and we discard all the clusters containing less than five genes because a

sufficient number of genes are usually required for a successful motif search. The key to identify meaningful clusters is to assess their statistical significance. We use a hypergeometric distribution to calculate the P -value for each candidate cluster,

$$p_c = P(x \geq s_c) = \sum_{x=s_c}^{n_c} \frac{\binom{S}{x} \binom{N-S}{n_c-x}}{\binom{N}{n_c}}$$

where N is the total number of genes on the array, S is the number of significant genes obtained in our previous step, n_c is the number of genes on the array that is assigned to the current GO node c and all its descendents and s_c is the cluster size, i.e. the number of significant genes in or under node c . Only clusters with a P -value $< 5E-6$ are selected for further motif search. Very often clusters formed on a parent and a child node are identical. Because one gene can be assigned to different nodes, we occasionally obtained identical clusters from two 'unrelated' nodes. Whenever this happens, we collapse the identical clusters and obtain a final set of unique clusters.

To enhance the sensitivity to detect the binding site of different types of transcription factors (activator, repressor or those with both activator and repressor functions), we performed functional clustering on three sets of genes for each array: all significant genes, only up-regulated genes in the significant gene set (positive \log_{10} ratio) and only down-regulated genes in the significant gene set (negative \log_{10} ratio).

Motif search

An in-house word-counting based program (Hata *et al.*, submitted for publication) is used for motif searches. This program takes a set of foreground sequences and a set of background sequences and identifies the over-represented words in the foreground against the background. In this study, the promoter region is defined as the 700 bp sequence upstream of the translation start site (ATG) of a gene in the yeast genomic sequence. Previous studies showed that most transcription start sites in yeast are located close to the coding regions and the majority of mapped TF binding sites are located within 700 bp upstream of ATG (2). The sequences are cut from 1000 bp UTR5 sequences downloaded from the SGD. The promoters of the genes in each cluster are taken as the foreground and the promoters of the least differentially expressed 2000 genes are taken as the background. We selected a word length of seven for the motif search. With this word length and a relatively small foreground sequence set, the expected occurrence frequency of a word is normally smaller than five and very often smaller than one. Therefore, we used the Fisher's exact test instead of the χ^2 approximation to estimate the P -value of each word. A 7mer with a P -value $< 1E-4$ is empirically determined as significant and for each cluster, we rank the significant words by their P -value and pick the top two to pool into the final result as the candidate binding motifs for the deleted transcription factor. In addition, we require that a significant motif must be present in $>50\%$ of the promoters in the foreground set, to avoid very skewed distributions.

Random control

Two sets of random control experiments are conducted to verify the robustness of our algorithm and parameter selection. The first control experiment is designed to test the procedure from significant gene selection up to functional clustering, in which we shuffle the gene IDs in an array so that the gene expression values are dissociated from the functional annotations. We repeat the procedure on the shuffled data 500 times for eight TFs in our test set. The second control experiment is to test our motif search algorithm. In this experiment, we randomly select 5–50 foreground genes and 2000 background genes from all the genes on the array, then extract the promoter sequences and perform a motif search and filtering. This test procedure is repeated 2000 times in total. Such repetition times for the control tests are chosen as a compromise between the resolution of P -value estimation and computation time.

Motif search without functional clustering

For comparison, the motif search algorithm is applied (as previously described) to three sets of genes for each array: all significant genes, only up-regulated genes in the significant gene set (positive \log_{10} ratio) and only down-regulated genes in the significant gene set (negative \log_{10} ratio). The significant genes are not subdivided into clusters or, in other words, there is only one cluster for each case. We will refer to this test as NOGO in the following text.

RESULTS

Verification using TFs with a known binding motif

The eight yeast transcription factors used to test our method are: Gcn4, Gln3, Mac1, Mbp1, Ste12, Swi4, Swi5 and Yap1. Their binding motifs (consensus sequences) are shown in Table 1. The clustering and motif search results are summarized in Table 2. More detailed results, including GO terms associated with the clusters and ChIP verifications, are available in Supplementary Material Table S1. In most cases, we obtain ~ 100 significant genes that meet our criteria (fold change > 1.5 , P -value < 0.05). The mbp1 and swi4 arrays are notable exceptions (Table 2). With the mbp1 array, we find 11 significant genes. At the other extreme, we obtain 854 significant genes with the swi4 array. The proportion of up-regulated or down-regulated genes varies from array to array. For example, nearly all the significant genes on the gcn4 array are down-regulated, while almost all the significant genes on the mbp1 array are up-regulated. Among the significant genes from most of the arrays, a portion (up to nearly 50%) are annotated as 'biological_process unknown' (GO id: GO:0000004). We did not explicitly exclude those genes in our analysis, however, node GO:0000004 never reached our threshold for significant clusters. Therefore, these genes never entered our motif search step. For the background genes selected, their expression ratios were typically within 10% deviation from one.

With our criteria for significant clusters, we obtained no more than 19 clusters for each set of significant genes. When the number of significant genes was relatively small, we sometimes obtained no significant clusters for that set, e.g. the positive gene set of the ste12 array and the negative gene set of

Table 2. Summary of significant motifs found with each set of deletion mutant microarray data

Experiment	Motif	Best rank	Occurrence	
gcn4 significant genes: 108 (4+/104-) biological_process unknown: 26 (0+/26-) significant clusters: all(17), neg(16) significant motifs: 10	AAAAAAT, ATTTTTT	2	all(1), neg(1)	
	AAATTCC, GGAATTT	2	all(1), neg(1)	
	AAGCCAC, GTGGCTT	2	all(1), neg(1)	
	AATTCCG, CGGAATT	1	all(1), neg(1)	
	ATATATA, TATATAT	1	all(1), neg(1)	
	ATGACTC, GAGTCAT	1	all(3), neg(3)	
	CACGTGA, TCACGTG	1	all(4), neg(4)	
	GAGTCAC, GTGACTC	1	all(1), neg(1)	
	GGAGTCA, TGACTCC	2	all(1), neg(1)	
	TGACTCA, TGAGTCA	1	all(12), neg(11)	
	gln3 significant genes: 118 (83+/35-) biological_process unknown: 29 (21+/8-) significant clusters: all(13), pos(11), neg(3) significant motifs: 10	AAATTCC, GGAATTT	2	all(2), pos(2)
		AATTCCG, CGGAATT	1	all(2), pos(3)
		ACAGCGG, CCGCTGT	2	all(1), pos(1)
ACTGTGG, CCACAGT		1	all(1), neg(1)	
AGAAATA, TATTTCT		2	all(1)	
ATATATA, TATATAT		1	all(2), neg(2)	
ATGACTC, GAGTCAT ^a		1	all(4), pos(4)	
CACGTGA, TCACGTG		1	all(1), neg(3)	
CTATGTC, GACATAG		1	pos(1)	
TGACTCA, TGAGTCA ^a		1	all(5), pos(6)	
mac1 significant genes: 89 (46+/43-) biological_process unknown: 24 (17+/7-) significant clusters: all(11), pos(7), neg(5) significant motifs: 8		ATAAGGG, CCCTTAT	1	all(1), neg(1)
		ATGACTC, GAGTCAT ^a	1	all(2), neg(3)
		CAGGTGC, GCACCTG	2	pos(1)
	GAGCAAA, TTTGCTC	2	all(1), pos(1)	
	GCAAAAA, TTTTTCG	2	all(1)	
	GGGTGCA, TGCACCC	1	all(7), pos(7)	
	GGTGCAA, TTGCACC	2	all(1), pos(1)	
	TGACTCA, TGAGTCA ^a	1	all(2), neg(2)	
	mbp1 significant genes: 11 (10+/1-) biological_process unknown: 1 (1+/0-) significant clusters: all(2), pos(2) significant motifs: 3	AACGCGT, ACGCGTT	1	all(2), pos(2)
		ACGCGTA, TACGCGT	2	all(1), pos(1)
ACGCGTC, GACGCGT		2	all(1), pos(1)	
ste12 significant genes: 79 (26+/53-) biological_process unknown: 38 (13+/25-) significant clusters: all(3), neg(4) significant motifs: 2		ATGAAAC, GTTTCAT	2	all(3), neg(3)
	TGAAACA, TGTTTCA	1	all(3), neg(3)	
	swi4 significant genes: 854 (573+/281-) biological_process unknown: 341 (261+/80-) significant clusters: all(15), pos(19), neg(5) significant motifs: 11	AAATAGC, GCTATTT	1	neg(1)
		AAATTCC, GGAATTT	1	all(2), pos(2)
AAGCGAA, TTCGCTT		2	neg(1)	
AATTCCG, CGGAATT		2	all(1), pos(1)	
ACCGGCT, AGCCGGT		2	all(1)	
ATATATA, TATATAT		1	all(2), pos(3)	
ATGACTC, GAGTCAT ^a		2	all(1), pos(6)	
ATGCGAA, TTCGCAT		2	neg(1)	
CACGTGA, TCACGTG		1	all(1), pos(3)	
TATATAA, TTATATA		1	neg(1)	
TGACTCA, TGAGTCA ^a		1	all(9), pos(12)	
swi5 significant genes: 103 (81+/22-) biological_process unknown: 45 (36+/9-) significant clusters: all(7), pos(7) significant motifs: 4		AAGCCAC, GTGGCTT	2	all(2), pos(2)
		ATGACTC, GAGTCAT ^a	1	all(2), pos(2)
	CACGTGA, TCACGTG	1	all(4), pos(4)	
	TGACTCA, TGAGTCA ^a	1	all(3), pos(3)	
yap1 significant genes: 98 (72+/26-) biological_process unknown: 30 (20+/10-) significant clusters: all(10), pos(9), neg(3) significant motifs: 8	AAGCCAC, GTGGCTT	2	all(2), pos(2)	
	AATGACT, AGTCATT ^a	2	all(1), pos(1)	
	ATGACTC, GAGTCAT ^a	1	all(1), pos(2)	
	CACGTGA, TCACGTG	1	all(4), pos(4)	
	CAGGGTC, GACCCTG	1	all(2), pos(2)	
	GTGAATA, TATTCAC	1	all(1)	
	TGACTCA, TGAGTCA ^a	1	all(2), pos(3)	
	TTACTAA, TTAGTAA	1	neg(1)	

In column one, the experiment name represents the TF being deleted in the mutant. +, up-regulated genes; -, down-regulated genes. In the first column, all(*n*) means that *n* significant clusters are obtained when we perform functional clustering using all significant genes from that array. Similarly, pos(*n*) and neg(*n*) mean that *n* significant clusters are obtained when we use only up-regulated genes or only down-regulated genes, respectively. In column two, motifs are presented as pairs of reverse complement sequences. Motifs in bold are those matching the known consensus. Motifs in italic are those similar to the known consensus. Column three shows the best rank of the motif found in the clusters. Column four shows the number of clusters in which the motif was detected. The meanings of 'all', 'pos' and 'neg' are the same as those in column one.

^aMotifs matching the consensus Gcn4 binding site in experiments other than gcn4 deletion (see also Discussion).

the *swi5* array (Table 2). Although the final clusters are distinct, in the sense that they differ by at least one gene, they often overlap with each other. Some clusters are completely contained within another one. Since each cluster may represent a different functional group, we performed a motif search with each of them. As expected, many clusters generate identical or similar significant motifs.

An interesting observation is that even when we pooled up-regulated and down-regulated genes in the functional clustering, the resulting clusters tended to be homogeneous. In other words, all or the majority of the genes in a cluster tended to change their expression in the same direction, suggesting a correlation between gene function and expression regulation. This notion has already been demonstrated with other clustering methods [see for example figure 2 of Ashburner *et al.* (5)]. An example with a *mac1* array experiment is shown in Figure 1.

Among the eight TFs in the test set, our method reported between two and eleven candidate motifs for each (Table 2). Overall, in five of the eight test cases, at least one significant 7 nt word (motif) found with our method matched the known consensus binding sequence of the deleted TF in Table 1. These were *Gcn4*, *Mac1*, *Mbp1*, *Ste12* and *Yap1* (Table 2, bold motifs). Some of the matching motifs rank the best in several clusters. There are cases where several closely related words all match the consensus. In addition, in most cases where a canonical motif was found in a cluster, the promoters of the majority of genes in this cluster bound the corresponding transcriptional regulator based on data from Lee *et al.* (20)

(Supplementary Material Table S1). The only exception is *Mac1*. Although three of the five genes in the cluster 'iron transport' (GO:0006826) contain at least one copy of GAGCAA in their promoters, only *FRE1* is shown to bind *Mac1*. Therefore, this hit may be questionable (see Discussion).

On the other hand, although the canonical *Swi4* motif CRCGAA is not found in any cluster of the *Swi4* deletion

```

Input sig. genes: 89 (46+/43-)
Final significant clusters found: 11
Genes biological_process unknown: 24 (17+/7-)

Cluster relations:
Containment:

C1(++)
C3(--) > C4(--)
C9(+) > C6(+)
C10(+) > C2(++) > C5(++), C8(++), C11(--)
C11(--) > C7(--)

```

Figure 1. Relation and polarity of clusters obtained from *mac1* array experiment when all the significant genes are used. *C_n*, cluster label; >, the cluster on the left completely contains the cluster on the right; ++, all the genes in the cluster are up-regulated; +, at least 80% of the genes in the cluster are up-regulated; --, all the genes in the cluster are down-regulated; -, at least 80% of the genes in the cluster are down-regulated.

Table 3. Four pairs of histone genes in 'chromatin assembly/disassembly' (GO:0006333) cluster from an *swi4* experiment

Systematic name	Gene name	Expression		Lee <i>et al.</i> <i>Swi4</i> ChIP <i>P</i> -value	Iyer <i>et al.</i> <i>Swi4</i> target	Lee <i>et al.</i> <i>Mbp1</i> ChIP <i>P</i> -value	Iyer <i>et al.</i> <i>Mbp1</i> target	Position relative to ATG	
		log ratio	<i>P</i> -value					ATGCGAA	TTCGCAT
YBL002W	HTB2	-0.402	8.72E-05	1.1E-03 ^a		3.9E-03 ^a	Yes		-397
YBL003C	HTA2	-0.39	4.47E-05					-309	
YBR009C	HHF1	-0.403	9.76E-05	2.3E-02 ^a		9.3E-02		-375	-299
YBR010W	HHT1	-0.198	1.00E-02					-354	-278
YDR224C	HTB1	-0.25	2.73E-03	5.6E-06 ^b		6.0E-02			
YDR225W	HTA1	-0.444	1.67E-05						
YNL030W	HHF2	-0.387	1.17E-04	4.5E-01	Yes	1.0E+00	Yes	-383	-316
YNL031C	HHT2	-0.185	1.83E-03					-367	-300

^a*P* < 0.05.
^b*P* < 0.001.

Table 4. Seven genes in 'regulation of CDK activity' (GO:0000079) cluster from an *swi4* experiment

Systematic name	Gene name	Expression		Lee <i>et al.</i> <i>Swi4</i> ChIP <i>P</i> -value	Iyer <i>et al.</i> <i>Swi4</i> target	Lee <i>et al.</i> <i>Mbp1</i> ChIP <i>P</i> -value	Iyer <i>et al.</i> <i>Mbp1</i> target	Position relative to ATG	
		log ratio	<i>P</i> -value					ATGCGAA	TTCGCAT
YBL056W	PTC3	-0.193	3.89E-02	6.6E-01		5.2E-01			-536
YDL155W	CLB3	-0.207	2.13E-02	5.1E-01		7.8E-01		-225	-52
YGR108W	CLB1	-0.674	8.69E-06	7.8E-02	Yes	1.1E-01	Yes		
YGR109C	CLB6	-0.642	1.31E-03	5.7E-05 ^b	Yes	1.1E-08 ^b	Yes		
YMR199W	CLN1	-0.42	6.48E-05	1.2E-06 ^b	Yes	1.5E-04 ^b			-185
YPL256C	CLN2	-0.253	1.11E-02	5.0E-03 ^a	Yes	4.3E-02 ^a			-80
YPR119W	CLB2	-0.517	1.49E-05	4.9E-05 ^b	Yes	5.1E-03 ^a		-326	

^a*P* < 0.05.
^b*P* < 0.001.

experiment, two similar motifs, ATGCGAA and AAGCGAA, are found in the 'chromatin assembly/disassembly' (GO:0006333) cluster and 'regulation of CDK activity' (GO:0000079) cluster, respectively (Tables 3 and 4). The eight genes in Table 3 are four pairs of divergent histone genes. Each pair of genes shares a common promoter region of ~650–820 bp long. ATGCGAA is the second ranked motif following the TATA box (TATATAA) in the histone cluster (Supplementary Material Table S1). In fact, if we allow one degenerate letter in the motif, ABGCGAA becomes the most significant motif and each of the four histone gene promoters contains one to three copies of this motif. A striking feature is that the occurrences of this motif are highly localized around the mid-point of the promoter region (~300–400 bp from the translation start site ATG). Upstream activation (UAS) elements with consensus GCGAAAAANTNNGAAC have been experimentally identified within the promoters of these histone genes (22) and additional putative UAS elements were found by a Gibbs sampling motif search algorithm (23). The occurrences of the motif ATGCGAA found with our method all overlap with the previously found UAS elements. As Zhang (23) pointed out, the similarity of the Swi4 motif to part of this histone UAS element might suggest the involvement of Swi4 in the regulation of histone genes. The fact that all these eight histone genes are repressed after Swi4 deletion supports this notion. ChIP chip data also appear to support this, although the discrepancy between Lee *et al.* (20) and Iyer *et al.* (21) may suggest a relatively weak binding affinity.

In the 'regulation of CDK activity' (GO:0000079) cluster (Table 4), AAGCGAA is the second ranked motif (Supplementary Material Table S1). The top motif AAATAGC matches a sub-string of some URS1H elements in SCPD (2) and the 'repressor of CAR1 expression' binding sites (related to URS1 elements) in TRANSFAC (1). However, AAATAGC is not the core but at the flank of the binding matrix of the 'repressor of CAR1 expression' in TRANSFAC. So, it may not be related to the URS1 element. We do not find a match to this motif in CompareACE (<http://atlas.med.harvard.edu/>) either. Therefore, its identity or plausibility is unclear to us. Among the genes in this cluster, CLN1 and CLN2 are known Swi4 targets documented in SCPD. An intriguing one is CLB2, a known Mcm1 + SFF target. It was shown to be a target of Swi4 by both Lee *et al.* (20) and Iyer *et al.* (21) and yet does not have a single copy of a relatively relaxed CNCGAAA motif in its 700 bp promoter region. It is possible that the binding site for Swi4 is beyond 700 bp upstream of the translation start site, since the ChIP chip experiments used the whole intergenic sequences. However, it is also possible that AAGCGAA is a new variant of the binding motif for Swi4 detected by our method, as is ATGCGAA in the histone cluster discussed previously. In fact, if we look more closely, the putative motif site in the CLB2 promoter is actually AAGCGAAA, with only one mismatch to the canonical Swi4 motif. We will further discuss the implications of this in the next section.

Random control tests

In our first random control test, we shuffled the gene IDs in an array to disrupt the association between gene expression values and the functional annotations. Only eight significant clusters were obtained in 12 000 trials on eight arrays with the

same parameters as in the previous experiments, suggesting that the probability of obtaining a significant cluster purely by chance is ~0.00067. In our second random control test, we performed a motif search on different sizes of randomly picked foreground gene sets against 2000 random background genes. On average, 0.13 motifs per trial were found and passed our filtering procedure from a total of 2000 trials. Therefore, the clusters and motifs found with our method are likely to be biologically meaningful.

These results indicate that our method is capable of detecting the binding elements of the relevant TFs from the deletion mutant microarray expression data without using any *a priori* knowledge about the motif sequence patterns. Even when ChIP data are not available, there is a good chance that the candidate motif list generated by our method includes the true binding motif of the relevant TF.

Comparison to motif search without functional clustering (NOGO)

When we apply the same filtering criteria for significant motifs as previously described, i.e. saving only the top two ranked motifs and requiring the motif to be present in >50% of the promoters in the foreground set, we only obtained significant motifs for four of the eight tested TFs (Gcn4, Gln3, Mbp1 and Ste12). Among those, the motif for Gln3 is a TA repeat. For the remaining three TFs, at least one significant motif matched the corresponding known consensus. This may not be a fair comparison, however, because we typically obtained several clusters (up to 19 in the case of swi4) with functional clustering and each cluster may generate two significant candidate motifs. Another issue is that functional clustering divides the significant genes into smaller clusters. There is a greater chance of a small cluster satisfying the 50% presence criterion.

For a more convincing comparison, we therefore relaxed the filtering criteria in NOGO such that up to 40 top ranked motifs were saved (motif *P*-value cut-off still applied) and the 50% presence criterion was dropped. The detailed results are shown in Supplementary Material Table S2. With such relaxed criteria, NOGO still failed to find a motif that matched the known consensus in the mac1 and yap1 experiments. Neither did it find the two possible motif variants for Swi4. On the other hand, NOGO succeeded in finding motifs matching the known consensus in the swi5 experiment. However, those motifs were only present in ~20–30% of the foreground promoters.

Application to less known TFs

To test our method on less studied TFs, we applied the same procedure to a few TFs documented in TRANSFAC but without well-defined binding motifs, including Arg80, Cin5 (Yap4), Ppr1 and Oaf1 (Yaf1). Cin5 has one artificial binding site TTACTAA in TRANSFAC. The other three TFs have one relatively long binding sequence (16–24 bp) each, but the core motifs are unknown. We did not obtain any significant functional clusters for Rosetta ppr1 and oaf1 deletion experiments. In both cases, the numbers of significant genes were small (18 for the ppr1 and 4 for the oaf1 experiment). A significant motif TGACTIONA was detected in a few clusters in the cin5 experiment, similar to its family member Yap1.

However, no genes in these clusters were shown to bind Cin5 according to the data of Lee *et al.* (20).

The most interesting experiment was with *arg80*. Arg80 escaped our initial screening for TFs with known binding sites because its binding site documented in SCPD is referred to as ARC (ARginine Control) element without an associated TF name and in older literature Arg80 is referred to as ArgRI. In TRANSFAC, there is only one Arg80 binding sequence (23 bp long) documented for human. Before our awareness of the known Arg80 binding motif, we applied our method to the Rosetta *arg80* deletion experiment and found a significant motif CACTTAA (or TTAAGTG) in the 'arginine biosynthesis' (GO:0006526) cluster second to the top motif TGACTCA. All five genes in this cluster (ARG5,6, ARG3, ARG1, ARG8 and CPA1) bind Arg80 in the ChIP chip experiment of Lee *et al.* (20) and the expression of all of them was up-regulated in the *arg80* deletion experiment, consistent with the known function of Arg80 as a repressor of arginine syntheses. Based on these lines of evidence, we predicted that CACTTAA might be the core motif of the yeast Arg80 binding site. A further literature search confirmed this conclusion. Motif TTAAGTG is very similar to part of the consensus of the ARC element defined by Crabeel *et al.* (24). Particularly, TAA is one of the most conserved cores of ARC elements. It is also interesting that the top motif found in this cluster was TGACTCA, the binding motif of Gcn4, and most of the genes in this cluster were shown to bind Gcn4 with ChIP chip data. Gcn4 is known to be a regulator of a large number of amino acid biosynthetic genes (14). Therefore, the motifs found in this 'arginine biosynthesis' cluster are consistent with the knowledge of a hierarchical regulation scheme in which these genes are regulated by a general amino acid control factor (Gcn4) and an amino acid-specific factor (reviewed in 24). This result again demonstrates that our method is capable of discovering potentially new TF binding motifs when sufficient information is available in the data sources.

DISCUSSION

In the three experiments with *gcn4*, *mbp1* and *ste12*, the correct motifs found by our method were very strong: they were detected in several clusters with top ranks. These motifs could also be detected without functional clustering. The *yap1* experiment demonstrated the advantage of our method. The canonical Yap1 binding motif, TTACTAA, is detected in one 'negative' cluster consisting of five genes on GO node 'oxygen and reactive oxygen species metabolism', consistent with the previous finding of Yap1 function in the oxidative stress response (25). This motif could not be detected in the NOGO test even when we relax the filtering criteria in the motif search. The reason may be that the majority of the significant genes do not have Yap1 canonical binding sites in their promoters. Some of them may be Yap1 targets using degenerate/variant motifs, but more likely, many of them change their expression due to indirect effects of YAP1 deletion.

It is interesting to see that the canonical Gcn4 motif was found in several other experiments besides the GCN4 mutant, including GLN3, MAC1, SWI4, SWI5 and YAP1 (Table 2, marked ^a). Since Gcn4 was suggested to be a master regulator of gene expression in response to cellular stresses (14), it is

possible that those mutations may have triggered some compensatory responses involving GCN4. The *yap1* experiment is complicated by the fact that Yap1 was shown to bind to the Gcn4 site less optimally than to its canonical site (25). However, ChIP chip data do not support the binding of Yap1 to the Gcn4 site because positive Yap1 binding is only seen in the promoters of a small fraction (<20%) of genes in clusters where a significant Gcn4 motif is found (Supplementary Material Table S1). Consistent with the known function of Gcn4 as an activator responding to amino acid starvation and other cellular stresses in yeast (14), the vast majority of the significant genes in the GCN4 mutant were down-regulated and most of the significant clusters obtained with our method were on GO nodes related to amino acid metabolism or biosynthesis (Supplementary Material Table S1). The Gcn4 binding motif is detected in almost all of the above clusters, suggesting that these genes are likely to be activated by Gcn4 in the wild type even under non-starved conditions.

Several factors may have contributed to the failure of our approach to detect expected TF binding motifs. First, the experimental conditions may not be appropriate for the TF to manifest its function. For example, Gln3 is known to activate genes involved in the usage of poor nitrogen sources and those genes are repressed when readily used nitrogen sources are available (26). Therefore, it is possible that under normal culture conditions, as in the Rosetta experiments, the target genes of Gln3 are repressed in wild-type cells and the deletion of GLN3 would have no effect. In fact, among the 118 significant genes in the *gln3* experiment, only two gene promoters were shown to bind Gln3 in the ChIP chip data, even at a very loose *P*-value cut-off (0.05). Second, our knowledge of gene functions, as represented in the GO annotation, is far from complete. This prevents us from a successful functional clustering in some cases. For example, a NOGO motif search detected a Swi5 motif in the down-regulated gene set, but we did not obtain any significant functional clusters with the same set of genes. Among the 22 genes in this set, 7 (32%) are annotated as 'biological_process unknown' (GO:0000004). Even those genes with some known functions may still have other functions that have not been annotated. Third, functional redundancy of some TFs may have reduced the effect of gene deletion on the direct targets, therefore reducing the signals in the expression data. This could be one reason why we did not detect the canonical binding motif for SCB (Swi4) because it is well known that Mbp1 and Swi4 have overlapping functions. In addition, expression profiling using non-synchronized yeast cell populations may also have reduced the TF deletion effect on some cell cycle-related target genes. One example was reported by Koch *et al.* (27): the mRNA levels of some Mbp1 target genes in the MBP1 null mutant were intermediate between the peaks and troughs observed in wild-type cells during the cell cycle, possibly because the Mbp1–Swi6 complex could be an activator or repressor depending on the phase of the cell cycle. Therefore, the average mRNA levels for some cell cycle-related Mbp1 targets may be similar in the non-synchronized cell populations of both mutant and wild-type. Similar effects could be relevant to the Swi4 null mutant as well.

The *mac1* experiment may be a special case deserving more discussion. Our method reported the canonical Mac1 binding

motif GAGCAA (CuRE, copper-response element) in the cluster 'iron transport' (GO:0006826), second to the most significant motif TGCACCC (Supplementary Material Table S1). However, a close examination of this cluster raises questions about the validity of this result. First, among the five genes (FET5, FRE2, FRE1, FET3 and ENB1) in this cluster, only FRE1, a known Mac1 target (28,29), showed significant promoter binding in the ChIP chip data. In fact, only between two and five promoters among the 89 significant genes in the MAC1 mutant experiment were shown to bind Mac1 in ChIP chip data at *P*-value cut-offs of 0.001 and 0.05, respectively. Second, it is known that two copies of CuRE in the promoter are necessary for efficient activation of downstream gene transcription (TRANSFAC database). Among the known Mac1 target genes, CuREs tend to be close to each other in the promoter. However, the FET5 promoter contains only one CuRE and the two CuREs in the promoter of ENB1 are >350 bp apart, with one very far from the translation start site. Therefore, these two genes may not be true Mac1 targets (unless other variant or degenerate CuREs exist). On the other hand, the top motif, TGCACCC, perfectly matches the core of the RCS1 (AFT1) motif consensus in the TRANSFAC database. RCS1 is known to be involved in high affinity iron ion transport (SGD annotation), which is consistent with the 'iron transport' cluster. In fact, TGCACCC is detected as the top motif in all positive clusters in the mac1 experiment (Supplementary Material Table S1) and we see a 26% increase in RCS1 mRNA level in MAC1 mutant versus wild-type (ratio = 1.26, *P*-value = 0.07), consistent with the role of RCS1 as a transcriptional activator (29). Although the fold change and *P*-value does not reach our criteria for significant genes, it may be biologically significant, as it is well known that a small fluctuation in TF expression may have a big impact on downstream genes. Therefore, we think a more plausible interpretation is that many of the significant genes in the mac1 experiment are due to increased expression of RCS1, an indirect effect of MAC1 deletion, and that the detection of a CuRE in our analysis may be due to the coupling between iron and copper transport. Alternatively, the detection of a CuRE in our analysis could be an artifact. It is reported that the CuRE is strongly bound by Mac1 only under copper starvation (28). Therefore, it is also possible that under the yeast culture conditions used in the Rosetta Compendium, Mac1 was not active in the wild-type. In this case, deletion of MAC1 will not change expression of most of the Mac1 targets. This echoes our point in the previous paragraph, the experimental conditions are critical in the study of TF functions.

Although we do not detect the published SCB (Swi4) binding motif in the SWI4 deletion mutant experiment, our method identified two very similar motifs in the histone cluster and the CDK regulation cluster. An interesting point with these two motifs is that they are in the form A?GCGAA, which is somewhat similar to both the canonical Swi4 binding motif CGCGAAA and the Mbp1 binding motif ACGCGT. As can be seen in Tables 3 and 4, the promoters of many genes in these two clusters were shown to bind Mbp1 in the ChIP chip experiments of Lee *et al.* (20) and/or Iyer *et al.* (21). Some of these genes do not have a single copy of even a degenerate Mbp1 site, ACGCGN. Although a motif like A?GCGAA may have a lower affinity for Swi4 and Mbp1 than their canonical binding motifs, it may provide a mechanism for cross-talk

between the two pathways, as it is known that Swi4 and Mbp1 overlap functionally. A relatively low affinity for this motif may be compensated for by multiple occurrences of this motif in the promoters or via cooperation with other factors, like those of histone genes. It is also worth noting that in a recent study by Liu *et al.* (30), who applied their motif discovery algorithm (MDscan) to earlier published ChIP chip data, the top ranked motifs reported for Swi4 were ACGCGAA and AACGCGA, resembling the motif we found.

One difference between our method and many early microarray data analysis methods is that our method attempts to combine the information from genome-wide expression data and known gene functions, while others mostly use functional information *ad hoc* to confirm or interpret the resulting clusters. Another feature of our approach is that our motif search algorithm uses the promoters of a set of non-significant control genes as background instead of sequences based on a random model. That may have enhanced the sensitivity of our motif search because we at least partially corrected the bias in the sequence word distribution. As a consequence, our motif search algorithm does not seem to be severely affected by simple repeats [e.g. poly(A), poly(T) and dinucleotide repeats] in promoter sequences while some other motif search methods often need to mask these simple repeats before searching.

A recent study by Wang *et al.* (31) extended the REDUCE algorithm (10) and applied it to a dataset consisting of more than 500 microarrays, including the Rosetta Compendium, in an attempt at systematically reconstructing transcription networks. REDUCE is a powerful algorithm for motif detection, as demonstrated by Bussemaker (10) and more recently by van Steensel (32). Wang *et al.* (31) successfully rediscovered the known motifs of several TFs in corresponding TF perturbation experiments. However, their method appeared to be susceptible to indirect effects of TF perturbations. They reported TGACTCA as the motif for Yap1 and TGCACCC as the candidate motif for Mac1. In contrast, our method successfully detected the canonical Yap1 motif TTACTAA and arguably detected the CuRE GAGCAA for Mac1. Therefore, we believe our method is to some extent complementary to theirs. It also reveals that conclusions based solely on a single TF perturbation expression study may not be reliable. Other sources of information, such as ChIP data or multiple expression arrays with different types of perturbations on a TF, are needed to verify the results and reach a sensible conclusion.

As discussed previously, the effectiveness of our method relies on the level of present knowledge about gene functions. The eight TFs in our test experiments are relatively well studied. A key question is, did we merely recover TF/target gene information already in the functional annotation? This question is critical if one wants to extrapolate the performance of our method to other less studied TFs. We manually checked the evidence codes and references used for GO annotation in SGD for a few small clusters in the mbp1 and yap1 experiments. For the putative target genes in those clusters, none of the references involved direct binding assays of TF/*cis*-elements. Instead, most of the references involved phenotypic studies. A few annotations are linked to review papers or with the evidence code IEA (inferred from electronic annotation), which may include some information from

binding assays. Therefore, we believe that our method may have inferred TF/*cis*-element relations based on mRNA level changes and promoter sequences, combined with functional information mostly obtained from phenotypic studies. As an example, MBP1 is annotated in SGD as involved in 'DNA replication' (GO:0006260). One of the significant clusters found with our method in the MBP1 mutant experiment is on node 'DNA repair' (GO:0006281) and every gene in that cluster contains at least one copy of ACGCGT in its promoter. This suggests that Mbp1 may also be involved in DNA repair and that our method may be able to reveal new functional relationships between genes. The effort of GO annotation in SGD is still ongoing. The current annotation probably reflects only a subset of our knowledge about all yeast genes in the literature. When more functional annotation information is available, our method should become more effective.

In this study, we used ChIP chip data to verify the motifs found with our method. Of course, when ChIP chip data are available for a TF, it may be more desirable to use these data directly with a method such as MDscan (30) or that of Kato *et al.* (submitted for publication) to detect the binding motif of that TF. However, even after a large-scale study such as that of Lee *et al.* (20), ChIP chip data are still not available for many TFs in yeast. ChIP chip data for other species are far less common. Therefore, methods such as ours are valuable in detecting relevant TF binding motifs when only TF intervention data are available. As we have shown in the last section, our method is able to provide a promising candidate motif list without using ChIP chip data. In addition, the functional clustering algorithm we implemented is not limited to motif finding. It can be applied to lists of candidate genes obtained with other methods, for example, with genes that are differentially expressed between two tissue samples or genes significantly bound by a TF in ChIP experiments. This approach may reveal possible new functional relations among genes or provide new insights into the function of the TF in question.

Our method at present is still rather simple and we can foresee several improvements. For example, in our motif search we now only count exact word matches without tolerating variants. The sensitivity may be improved if we allow degenerate motifs. The challenge, however, is that the top ranks in such a search will be dominated by variants of the strongest motif. Thus, better filtering methods will be required. Another issue is the *P* value cut-off in the functional clustering. Currently we determined the *P* value cut-off empirically as a relatively stringent 5E-6 in order to address the multiple test problem. While our random control test suggested that this cut-off should keep the random hits at a fairly low level, this fixed cut-off may not be optimal. It may be too stringent when the number of significant genes is small and may be too liberal when the number of significant genes is large. An adaptive cut-off based on a false discovery rate control (33) may be more desirable. In addition, we currently examine the distribution of the motif hits among the genes in a cluster and the locations of the matching sites in the promoters empirically and *ad hoc*. A more formal method of handling these aspects may further enhance the specificity of our prediction.

As discussed previously, expression array data of a TF intervention may not always contain enough information for

motif detection. A similar problem may exist for ChIP chip experiments as well. In some preliminary tests using the Gln3 and Mac1 ChIP chip data from Lee *et al.* (20) and an extended version of MDscan (30), we were unable to detect any motifs resembling the published motifs for Gln3 or Mac1. The reason may be similar to that for the Rosetta deletion experiment, these TFs are not active under the culture conditions of the experiments and few real targets are bound by the TF studied. Moreover, some binding detected by a ChIP chip may not be specific or functional. Thus, the signal may be too weak to be detected by a program like MDscan. With current technologies, both expression array data and ChIP chip data contain a significant amount of noise. However, they may reflect different aspects of the same biological process. Methods that integrate the information from expression arrays and ChIP chips are definitely worth more investigation. Additional improvement may be gained with further integration of other sources of information, including, but not limited to, biological knowledge in the literature and databases. Our study is a first step in that direction. Even with our current method, the success rate of our results is nevertheless encouraging. It is our belief that with the rapid accumulation of biological data, our approach, with further improvements, will be valuable for identifying TF/target relationships and for deciphering genetic networks of yeast and other living organisms.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Xianghong Zhou for her discussions on handling gene ontology data and Xiaole Liu for her help with MDscan. This study was supported by NIH grant GM60513.

REFERENCES

1. Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I. and Schacherer, F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
2. Zhu, J. and Zhang, M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.
3. Winzler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
4. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
5. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
6. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
7. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.
8. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998)

- Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
9. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
 10. Bussemaker, H.J., Li, H. and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nature Genet.*, **27**, 167–171.
 11. Birnbaum, K., Benfey, P.N. and Shasha, D.E. (2001) cis element/transcription factor analysis (cis/TF): a method for discovering transcription factor/cis element relationships. *Genome Res.*, **11**, 1567–1573.
 12. Zhu, Z., Pilpel, Y. and Church, G.M. (2002) Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol.*, **318**, 71–81.
 13. DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
 14. Natarajan, K., Meyer, M.R., Jackson, B.M., Slade, D., Roberts, C., Hinnebusch, A.G. and Marton, M.J. (2001) Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol. Cell. Biol.*, **21**, 4347–4368.
 15. Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., Sherlock, G. et al. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
 16. Zhou, X., Kao, M.C. and Wong, W.H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.
 17. 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.
 18. Zhu, J. and Zhang, M.Q. (2000) Cluster, function and promoter: analysis of yeast expression array. *Pac. Symp. Biocomput.*, 479–490.
 19. Pavlidis, P., Lewis, D.P. and Noble, W.S. (2002) Exploring gene expression data with class scores. *Pac. Symp. Biocomput.*, 474–485.
 20. 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.
 21. Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M. and Brown, P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
 22. Osley, M.A. (1991) The regulation of histone synthesis in the cell cycle. *Annu. Rev. Biochem.*, **60**, 827–861.
 23. Zhang, M.Q. (1999) Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.*, **23**, 233–250.
 24. Crabeel, M., de Rijcke, M., Seneca, S., Heimberg, H., Pfeiffer, I. and Matisova, A. (1995) Further definition of the sequence and position requirements of the arginine control element that mediates repression and induction by arginine in *Saccharomyces cerevisiae*. *Yeast*, **11**, 1367–1380.
 25. Fernandes, L., Rodrigues-Pousada, C. and Struhl, K. (1997) Yap, a novel family of eight bZIP proteins in *Saccharomyces cerevisiae* with distinct biological functions. *Mol. Cell. Biol.*, **17**, 6982–6993.
 26. Cunningham, T.S., Svetlov, V.V., Rai, R., Smart, W. and Cooper, T.G. (1996) G1n3p is capable of binding to UAS(NTR) elements and activating transcription in *Saccharomyces cerevisiae*. *J. Bacteriol.*, **178**, 3470–3479.
 27. Koch, C., Moll, T., Neuberg, M., Ahorn, H. and Nasmyth, K. (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science*, **261**, 1551–1557.
 28. Labbe, S., Zhu, Z. and Thiele, D.J. (1997) Copper-specific transcriptional repression of yeast genes encoding critical components in the copper transport pathway. *J. Biol. Chem.*, **272**, 15951–15958.
 29. Georgatsou, E. and Alexandraki, D. (1999) Regulated expression of the *Saccharomyces cerevisiae* Fre1p/Fre2p Fe/Cu reductase related genes. *Yeast*, **15**, 573–584.
 30. Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
 31. Wang, W., Cherry, J.M., Botstein, D. and Li, H. (2002) A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **99**, 16893–16898.
 32. van Steensel, B., Delrow, J. and Bussemaker, H.J. (2003) Genomewide analysis of *Drosophila* GAGA factor target genes reveals context-dependent DNA binding. *Proc. Natl Acad. Sci. USA*, **100**, 2580–2585.
 33. Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
 34. Bysani, N., Daugherty, J.R. and Cooper, T.G. (1991) Saturation mutagenesis of the UASNTR (GATAA) responsible for nitrogen catabolite repression-sensitive transcriptional activation of the allantoin pathway genes in *Saccharomyces cerevisiae*. *J. Bacteriol.*, **173**, 4977–4982.