

Inferring Gene Regulatory Relationships by Combining Target–Target Pattern Recognition and Regulator-Specific Motif Examination

Hairong Wei, Yiannis Kaznessis

Department of Chemical Engineering and Material Sciences, and Digital Technology Center, University of Minnesota, 421 Washington Avenue SE, Minneapolis, Minnesota 55455; telephone: 612-624-4197; fax: 612-626-7246; e-mail: yiannis@cems.umn.edu

Received 20 April 2004; accepted 5 August 2004

Published online 11 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/bit.20305

Abstract: Although microarray data have been successfully used for gene clustering and classification, the use of time series microarray data for constructing gene regulatory networks remains a particularly difficult task. The challenge lies in reliably inferring regulatory relationships from datasets that normally possess a large number of genes and a limited number of time points. In addition to the numerical challenge, the enormous complexity and dynamic properties of gene expression regulation also impede the progress of inferring gene regulatory relationships. Based on the accepted model of the relationship between regulator and target genes, we developed a new approach for inferring gene regulatory relationships by combining target–target pattern recognition and examination of regulator-specific binding sites in the promoter regions of putative target genes. Pattern recognition was accomplished in two steps: A first algorithm was used to search for the genes that share expression profile similarities with known target genes (KTGs) of each investigated regulator. The selected genes were further filtered by examining for the presence of regulator-specific binding sites in their promoter regions. As we implemented our approach to 18 yeast regulator genes and their known target genes, we discovered 267 new regulatory relationships, among which 15% are rediscovered, experimentally validated ones. Of the discovered target genes, 36.1% have the same or similar functions to a KTG of the regulator. An even larger number of inferred genes fall in the biological context and regulatory scope of their regulators. Since the regulatory relationships are inferred from pattern recognition between target–target genes, the method we present is especially suitable for inferring gene regulatory relationships in which there is a time delay between the expression of regulating and target genes. © 2004 Wiley Periodicals, Inc.

Keywords: pattern recognition; target pattern matching; gene networks

INTRODUCTION

One of the central problems in molecular biology is to understand gene regulatory networks and how they collectively operate with remarkable precision in response to environmental cues. In the last few years, there has been increasing interest in constructing gene regulatory networks using time series microarray data. Bayesian network algorithms (Friedman et al., 2000), smooth response surface algorithms (Xu et al., 2002) and inference algorithms (de la Fuente et al., 2002) have been developed to mine regulatory networks from gene expression profiles. The implementation of Bayesian networks is computationally hard because it requires learning the network while taking into account the temporal aspects of gene expression (Kwon et al., 2003). Some features of gene regulatory networks, for example a gene being potentially regulated by more than one gene, also pose great challenges to the Bayesian network approach. Smooth response surface algorithms are based on the ad hoc selection of a response surface function that bears little biological meaning (Xu et al., 2002). The inference method proposed by de la Fuente and co-workers (2002) requires quantitatively perturbing every gene's expression, which is not practically attainable. An additional potential problem for this approach is that it is not capable of distinguishing between direct and indirect regulation. There are also other methods attempting to infer gene regulatory relationships by examining whether a gene's expression is correlated to any other gene's expression. These approaches include correlation-based methods, mutual information theory, and edge detection (Filkov et al., 2002) as well as event methods (Kwon et al., 2003). These approaches are useful in finding gene pairs that share similarities, but have the following limitations when they are used to infer gene regulatory relationships: (1) Correlation or similarity between the expression profiles of two genes does not necessarily mean there is a regulatory relationship between them; (2) even if there is an actual relationship, it is not easy to tell

Correspondence to: Yiannis Kaznessis

Contract grant sponsors: Minnesota Supercomputing Institute (MSI); the Unisys Corporation; the University of Minnesota Biotechnology Institute

which one is the regulator and which one is the target; (3) these approaches can detect only limited regulatory relationships due to the complexity and multiple-layers of gene regulation as well as the potential resulting time-delay, which make regulatory relationship between a regulator (an activator or a repressor) and its target very elusive and difficult to recognize. For example, a target is potentially regulated by as many as dozens of genes (Costanzo et al., 2000; Friedman et al., 2000), which attenuates the correlation between a regulator and anyone of its target genes. In addition, the fact that a regulator controls its target genes periodically or during a specific stage but rarely all the time throughout the entire time course of the experiment reduces the reliability of these approaches, which favor global similarity over more localized similarities arising from conditional regulation. As reported, less than 20% of known regulatory pairs exhibit strong correlations in microarray datasets (Filkov et al., 2002), implying that no more than 20% regulatory relationships can be identified by correlation-based methods, even if the correlations between irrelevant genes can be successfully removed and regulators and targets can be told apart. Finally, selective translation of regulating mRNAs, post-translational modification of regulating proteins, and slow transcriptional assembly of target genes are very common in gene regulation (Avrava et al., 2003; Cosma et al., 1999; Mazumder et al., 2003; Nasmyth et al., 1990; Rosenfeld and Alon, 2003.). For example, a master regulator, GCN4p, of more than 500 target genes (Natarajan et al., 2001) in yeast is post-translationally regulated (Avrava et al., 2003).

Two very promising approaches for gene regulatory networks inference are the ones by Pilpel and co-workers (2001) and Ihmels and co-workers (2002), in which the analysis of gene expression profiles is coupled with promoter sequence analysis. In Pilpel's method, a database of known and putative regulatory motifs was established, and then used to screen genes. This information is then modeled to calculate the effect of each motif on gene expression. For each motif or combination thereof, a parameter called expression coherence score is calculated to measure the overall similarity of all genes containing that motif under several different conditions. A pair of motifs is considered "synergistic" if the expression coherence score of genes containing both motifs in their promoters is significantly greater than that of genes containing either motif. The signature algorithm proposed by Ihmels et al (2002) is similar to Pilpel's. The condition scores that measure the change in expression of input genes for each condition are calculated, and then used to select the genes possessing a significant and consistent change in expression under the specified condition. To infer regulatory relationships, a collection of the genes containing the relevant sequences in their upstream regions is used as an input. The signature algorithm singles out these genes that are co-regulated by the transcription factor associated with a particular regulatory motif. These two approaches are similar in that they both require large microarray datasets that are generated

under several conditions and the results are largely dependent on whether there are cohesive profiles of certain numbers of genes.

In this study, we develop a new method to infer gene regulatory relationships using one or a few known regulatory relationships and one or multiple time series microarray datasets. As compared to the methods developed by Pilpel et al. (2001) and Ihmels et al. (2002), ours is different in the following aspects. (1) Our method is designed to infer other target genes of a regulator when one or more target genes of this regulator are known, employing existing, experimentally validated knowledge of regulatory relationships. (2) The underlying principle is different. In our approach, the regulatory relationships are determined by a pattern recognition process, in which a gene is judged as a putative target gene of a given regulator when it partially shares similar expression profile patterns with at least one known target gene of that regulator, and the examination of the presence of regulator-specific binding sites. (3) Our approach starts with one to a few KTGs, and a single time series microarray dataset, and the results from inference can be summed up. Pilpel et al.'s and Ihmels et al.'s approaches look for the most significant modules that are active across several conditions, and thus require multiple microarray datasets from various conditions. In this sense, Pilpel et al.'s and Ihmels et al.'s approaches use a "top-down" strategy whereas our approach uses a "bottom-up" strategy. Therefore, the method we present in this article can serve as an additional, alternative approach for inferring gene regulatory relationships.

METHODS

According to the current model of transcription, the protein products of regulating genes known as activators or repressors, are composed of a DNA-binding domain, and can bind to specific DNA sequence elements in the promoter region of target genes. The bound regulatory proteins stimulate or repress transcription through either direct or through indirect interaction with basal transcriptional machinery. When a regulating gene indirectly interacts with the basal transcriptional machinery, one or a few intermediates, called co-activators or co-repressors, interplay between the activator/repressor and the basal transcriptional machinery or help activators and the transcriptional apparatus to navigate through the constraints of chromatin (Näär et al., 2001). Although co-regulators are important components of transcriptional regulation, the activators or repressors being anchored to the regulatory elements in the promoter regions seem to be the primary dictating factors of expression profiles. Co-regulators exert their influences to transcription in cooperation with the regulators. This assumption is, for example, supported by the sequential assembly of transcriptional machinery on the yeast HO promoter, in which SWI5p regulator binds to HO gene promoter first and then recruits other co-regulators and basal transcriptional machinery (Cosma et al., 1999).

Based on this model, we hypothesize that the expression profiles of genes under the control of the same activator or repressor often share similarities at specific stages or even a full time-course of a microarray experiment. With this hypothesis as a guiding tool we developed a new approach to infer gene regulatory networks in yeast. Our approach is illustrated in Figure 1. In each pattern recognition process, we chose a known regulator gene and its experimentally validated target genes. Using publicly available yeast time series microarray data (Cho et al., 1998), we extracted expression profile patterns of these known target genes. The patterns extracted from these KTGs were then used to query more putative target genes that share similar patterns as the KTGs. Dozens of genes were discovered each time. Falsely predicted target genes can be eliminated by examining the presence of regulator-specific binding sites in their promoter regions. Details of our approach and data are described below.

Selection of Regulating Genes and Their Target Genes

The regulating genes and their experimentally validated target genes were selected from the Yeast Proteome Data-

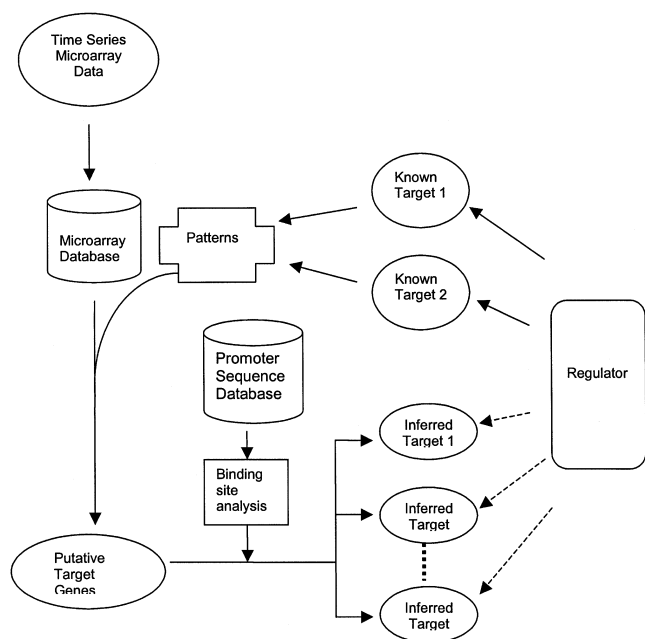


Figure 1. The flowchart of the approach described in this study. The time series microarray datasets from multiple environment conditions are cleaned and added to the database. A regulator and its known target genes can be obtained from previous publications, databases, experiments, or any other sources. Promoter sequences are stored in the same database with microarray data. Pattern selection serves as the first filter to remove the genes that do not share similarities in their expression profile patterns with the given target genes. Binding site examination provides the second filter at which the genes are filtered out that coincidentally share similar patterns with the known target, without having regulator-specific binding sites genes.

base (Costanzo et al., 2000), the TRANSFAC database (Matys et al., 2003) and previous publications (Lee et al., 2002; Simon et al., 2001). The criteria for selecting regulating genes are (1) the regulating genes must be known to activate or repress target genes by serving as an activator or a repressor, and have at least one experimentally validated KTG; (2) the regulator-specific binding motifs are available; (3) The expression levels of the regulating gene and its KTGs were not constant during the full time course of the microarray experiment. Specifically, if the ratio of the maximum expression value to the minimum expression value during the experimental period is smaller than 1.5 the genes are not considered. The genes that contain the regulator-specific binding site but were not experimentally proven to be target genes were not selected. Based on these criteria, we selected 18 regulators from more than 800 known regulators in yeast (Costanzo et al., 2000). Not all of these 800 regulators affect other genes' expression at the transcriptional level. Actually, most of them affect other genes expression at translational and/or post-translational levels. We eventually selected the 18 regulators that are well known to control gene expression at the transcriptional level.

This process renders our computational approach dependent on experimental evidence. We do believe however that employing experimental information in computational discovery is a sensible approach. As information on regulatory relationships becomes available we will conduct our analysis adding more relationships. In Table I, we present the 18 regulators used (first column).

When a regulator has more than one KTG, we compared the expression profiles of these KTGs and classified them in groups of similar patterns. From each group of KTGs, we arbitrarily chose a representative one for pattern recognition. Eventually, 38 known target genes were used for pattern recognition for all 18 regulators used. In the Appendix Table A1, we present these 38 known target genes used (second column).

Microarray Data

The approach we proposed in this study is applied to a public oligonucleotide GeneChip dataset produced for studying the yeast cell cycle (Cho et al., 1998). The data were collected over 17 time points for 6457 genes (including some test genes) on Affymetrix Ye6000 chip. The values of the fluorescence intensity were used for our study. The genes that have negative expression values or smaller positive values less than 30 at all time points were removed from our study. Considering some highly functional proteins may naturally express at low level, we did not impute any genes with expression levels between 30 and 150. Finally, 5016 genes were retained and imported into the database for studying. The resulting dataset was imported into the YCELLCYCLE database designed in Microsoft SQL Server 2000 at the Minnesota Supercomputing Institute.

Table I. Target genes inferred from regulators and known targets. More detailed information can be found in the Appendix.

Regulators	Inferred target genes
ABF1	YDL192W, YAL038W, YOL139C, YCR031C, YML012W, YNL135C, YLR293C, YNL031C, YBR011C, YOR099W, YCR012W, YBR093C, YBR106W, YPR165W, YNL301C, YHR010W, YBR031W, YDL130W, YIL018W, YNL162W, YDR012W, YJR145C, YLR448W, YBR084C-A, YNL302C, YDR025W, YML026C, YFL045C, YNL209W, YGL123W, YBR118W, YPR163C, YDR353W, YLR249W, YHL050C
ACE2	YER124C, YGL028C
CCBFSWI4SWI6	YNL289W
GCN4	YNL065W, YJL071W, YJL088W, YPR025C, YHR175W, YPR167C, YDR318W, YGL257C, YNL316C, YMR270C, YOR213C, YHL014C, YPR077C, YHR150W, YEL036C, YDL192W, YDL171C, YGL009C, YER091C, YDL232W, YER117W, YDR025W, YCR053W, YDR002W
GCR1	YBL092W, YBL087C, YBL027W, YBR084C-A, YBR181C, YBR189W, YBR118W, YBR191W
GZF3	YDR511W
HAP1	YJR121W, YEL011W, YOR032C, YOL111C, YKL085W, YFR033C, YLL013C, YDL210W, YLR327C, YKL067W, YIL100W, YKL066W
LEU3	YNL104C, YPR145W, YOR375C, YEL017C-A, YEL027W
MAC1	YPL270W, YLR051C, YOL137W
MCM1	YGR177C, YLR131C, YGR108W, YMR001C, YOL158C, YOR274W, YOL043C, YLR071C, YIL158W, YJL051W
MIG1	YML042W
MBP1	YNL233W, YGR041W, YLR103C, YPL241C, YPL256C, YBR161W, YLR381W, YPR135W, YOR033C, YPR175W, YBR278W, YOR144C, YJL196C, YCR065W, YLL022C, YBR215W, YIL026C, YCL024W, YDR503C, YDL003W, YOL090W, YKL165C, YDR097C, YDL127W, YNL082W, YNL102W, YBL035C, YBR088C, YDL103C, YKL113C, YER095W, YPL153C, YAR007C, YNL312W, YER070W, YGR152C, YNL206C, YHR154W, YAR008W, YJL074C, YFL008W, YGL207W, YKL042W, YLR313C, YLR389CYNL039W, YNL273W, YLR183C, YLR212C, YKR077W, YPR174C, YPL267W, YGR151C
PHO4	YBR249C, YAL003W, YAR071W, YBR106W, YBL087C, YKR057W, YDL136W, YLR354C, YLR249W, YAR075W
RAP1	YAL038W, YAL040C, 007C, L003W, 45W, L034W, L014W, YCR012W, YDR447C, YOR224C, YDL081C, YOL039W, YDL130W,

Table I. Continued

Regulators	Inferred target genes
	YIL018W, YDL082W, YBL087C, YBL027W, YDL075W, YMR142C, YER117W, YBR084C-A, YDR471W, YGR148C, YDR382W, YBL092W, YHR203C, YMR143W, YBL072C, YDR025W, YER102W, YBR048W, YGR118W, YER074W, YIL069C, YDR077W, YDL191W, YDL229W, YGL135W, YGL123W, YBR118W, YDR050C, YBR127C, YDL061C, YBR111C, YEL077C, YDL083C
ROX1	YJR047C, YJR047C, YDL070W, YKL096W, YDR225W, YBL051C, YBL079W, YLR373C, YBR095C, YBR123C, YBR156C, YAR035W, YAL053W, YBL095W
RPN4	YOR362C
SWI5	YBR158W, YLR295C, YKL185W, YOR264W, YNR067C, YDR512C, YDR516C, YOR317W, YCL040W, YOR155C, YNL173C, YDR296W, YDR194C, YPL097W, YNL078W, YLR093C, YIL136W, YDL179W, YDR079W, YOR158W, YJL078C, YOR153W, YKR093W, YGR044C, YDR487C, YOR018W, YDR041W, YMR191W, YMR261C, YDR513W, YFR003C, YNL046W, YGR048W, YOL129W, YFR003C, YMR243C, YPL158C, YOL048C, YML040W, YOR203W, YDR271C, YJL197W, YNL046W, YMR040W, YOR203W, YDR271C, YOL048C, YDR512C
YAP1	YJR148W, YDR304C, YPL091W, YGR234W, YBR054W, YLL049W

Pattern Extraction

Consider M identical genome-wide time series microarray experiments that contain N time points of G genes each. The average expression level μ_{ij} of each gene i ($i = 1, 2, \dots, G$) at time t_j ($j = 1, 2, \dots, N$) can be estimated by $\mu_{ij} = \sum_{m=1}^M e_{ijm}/M$ where e_{ijm} is the expression value of i^{th} gene at time point t_j on array m ($m = 1, 2, \dots, M$).

The profile P_i of each gene i can be expressed as follows:

$$P_i = \{\mu_i \in R^N: \mu_{i1} > 0, \mu_{i2} > 0, \mu_{i3} > 0, \dots, \mu_{ij} > 0, \dots, \mu_{iN} > 0\} \text{ where } i = 1, \dots, G$$

Consider w KTGs of a known regulator. The expression profile of the k^{th} known target gene T_k , is:

$$T_k = \{\mu_k \in R^N: \mu_{k1} > 0, \mu_{k2} > 0, \mu_{k3} > 0, \dots, \mu_{kj} > 0, \dots, \mu_{kN} > 0\} \text{ where } k = 1, \dots, w$$

To facilitate pattern extraction, the derivative profile of k^{th} known target gene, D_k , is expressed as:

$$D_k = \{d_k \in R^{N-1}: d_{k1}, d_{k2}, d_{k3}, \dots, d_{kj}, \dots, d_{k(N-1)}\}$$

where d_{kj} is calculated as

$$d_{kj} = (\mu_{k(j+1)} - \mu_{kj}) / (t_{(j+1)} - t_j)$$

Therefore, d_{kj} represents the trend of expression of k^{th} KTG from time t_j to t_{j+1} .

Pattern extraction and discovery was performed by two algorithms we developed: TREND-SELECT and SCALE-SELECT. The pseudocodes of TREND-SELECT and SCALE-SELECT algorithms are available at <http://www.cems.umn.edu/research/kaznessis/inference.html>. The details of TREND-SELECT and SCALE-SELECT algorithms are described below.

TREND-SELECT Algorithm

TREND-SELECT navigates through all time intervals sequentially to find out the ones in which all selected KTGs have the same trend. The input is the derivative profile D_k of w selected KTGs. Details of the procedure are listed below.

Step 1. Go through each time interval starting from t_1 towards t_{N-1} and check if all w KTGs used for inference have the same trend, namely increase, remain constant or decrease by using the derivative profile D_k .

Step 2. If all target genes consistently have the same trend in a time interval, the two time points of this time interval and the sign of the derivative of expression are saved. In addition, the average slope of all given KTGs in this time interval is calculated and stored.

Step 3. After the completion of all time intervals, the number of time intervals in which all KTGs have the same trend are counted. If the number is less than eight, input of new KTGs is then required. This number was chosen based on the investigation of the relationship between the number of time intervals, in which all KTGs have the same trend, and the number of genes that can be queried from the database, as shown in Figure 2. We tested 54 KTGs of 18 regulators. For each KTG, five independent queries were performed. In each query, the time intervals of certain number were randomly selected from all 16 time intervals. The gene number shown in Figure 2 is the average gene number from five independent queries. Although there is variation in the gene number at a certain number of time intervals, the queried gene number reduced to an average number of 55 when the number of time intervals increased to eight. Since a regulator in yeast may control up to several dozens of genes as shown in the YDP database (Costanzo et al., 2000), we thus chose eight time intervals as the criterion for pattern extraction.

Step 4. If all KTGs consistently have the same trend in more than eight time intervals, only eight time intervals are used. The selection of eight time intervals is based on the corresponding average slopes, regardless of the sign of

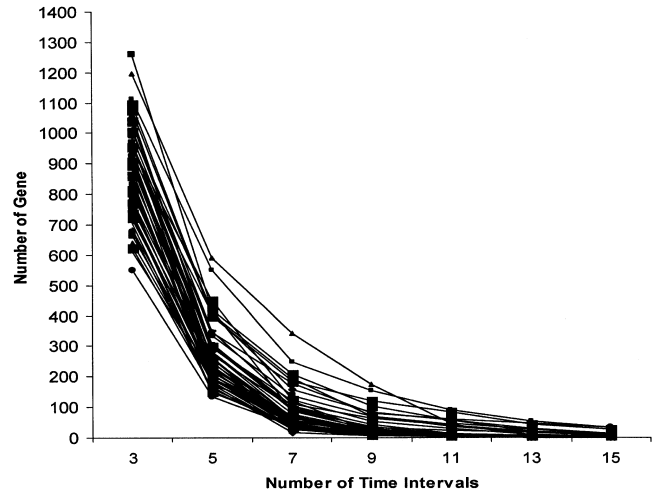


Figure 2. The number of genes queried from the database as a function of the number of time intervals from which the query constraints were extracted. 54 KTGs under the control of 18 regulators were used as templates for probing the relationships. For all 54 KTGs, the number of extracted genes decreases when the number of time intervals used for pattern extraction increases.

slope. As a result, the eight time intervals that are corresponding to the top eight average derivatives are chosen.

Step 5. All genes, $P_i = \{\mu_i \in R^N: \mu_{i1} > 0, \mu_{i2} > 0, \mu_{i3} > 0, \dots, \mu_{ij} > 0, \dots, \mu_{iN} > 0\}$ where $i = 1, \dots, G$, which have the same trends as KTGs in eight time intervals, are selected and used as an input for SCALE-SELECT.

SCALE-SELECT Algorithm

SCALE-SELECT is used to find two pairs of time points possessed by all KTGs, comprising of a peak and a trough. The ratio of the expression value at the peak to the expression value at the trough of all KTGs must be either the largest or second largest ratio among all possible ratios that can be generated from all peaks and troughs possessed by all given KTGs. The inputs of this algorithm include all candidate genes acquired from the TREND-SELECT algorithm, the KTG expression profile T_k , and the derivative profile D_k . Details of the algorithm are described below.

Step 1. Go through the time course starting from t_1 towards t_N and check whether a time point encountered is a peak or a trough for all KTGs by comparing the signs of two adjacent derivatives.

Step 2. If a time point is identified as a peak for all KTGs, this time point and the expression value of each KTG at this peak are saved. If a time point is identified as a trough for all KTGs, this time point and the expression values of all KTGs' at this trough are saved.

Step 3. Read w expression values of KTGs at a peak and w expression values of KTGs at a trough, calculate the ratio of each KTG's peak value to the trough value. Only the minimum of all w ratios, together with the time point of the peak and time point of the trough, is saved. This

process is repeated until all combinations between peaks and troughs are computed.

Step 4. To find the largest and the second largest ratios, sorting is performed for the ratios saved in Step 3. Whenever two ratios are switched in order, so too are the two peak time points and the two trough time points corresponding to these two ratios.

Step 5. The candidate genes from TREND-SELECT are selected using the largest and second largest ratios as two constraints. A gene is selected only if it has a larger ratio than either constraint.

IMPLEMENTATION OF ALGORITHMS

The TREND-SELECT and SCALE-SELECT algorithms were implemented in Microsoft SQL server 2000 with Transact-SQL (pseudocodes are presented in the appendix). Expression profiles of all yeast genes were stored in the database. The input was one to a few KTGs, and the expression profiles, and derivative profiles of all KTGs were automatically generated when algorithms were run. The output of TREND-SELECT and SCALE-SELECT were two query strings that were used to query the database for putative target genes.

Binding Motifs Search

After identifying putative target genes with pattern recognition we searched for regulator-specific binding elements in the promoter regions of queried putative genes. To decide how far upstream of transcription initiation sites we needed to check for the binding sites, we examined the distribution of 20 regulatory elements in the promoter regions of nearly 300 yeast genes and found that there is a sharp drop of regulatory elements at 1 kb upstream of transcriptional initiation sites. Specifically, 93% regulatory elements are located within this 1 kb region. It is known that most of the intergenic sequences between two yeast genes are around 500 bp (Zhu and Zhang, 1999), but at present, it is not clear whether the regulatory motifs within the adjacent upstream gene play a role in the downstream gene activation, in particular, when the upstream gene is not expressed. More specifically, overlapping of the promoter regions with upstream genes' coding regions is common for class III polymerase transcribed genes (Fantoni et al., 1994; Nakaar et al., 1994). However, this phenomenon is not often seen in class I and class II polymerase transcribed genes. Perhaps, the reason for this is that researchers often assume the motifs in the upstream gene's coding region are not likely to affect downstream gene's expression and usually just test the intergenic sequences (promoters) of downstream genes. Although there is still very limited bibliographical evidence regarding how long the promoters can be, recent emerging evidence indicates the motifs in the upstream genes can affect downstream gene expression. For example, yeast GST1 gene has a 1042 bp functional promoter region, within which there

are two other open reading frames located at -310 and -829, respectively. This is one example where the promoter overlaps with the coding regions (Tonozuka et al., 2001). In another example, ChIP genome-wide experiments indicate that yeast RAP1 regulator binds to the promoter regions of 182 target genes, but also to 23 genes' coding regions (Lieb et al., 2001).

In principle, motif searches with position specific weight matrices (Stormo, 2000) usually ignore dependencies between nucleotide positions in regulatory motifs, and use relatively more rigorous algorithms (Frith, 2002) as compared to searches with a consensus pattern. We used matrix searching for those motifs that have a matrix profile available, for the motifs bound by ABF1, GCN4, MCM1, PHO4, RAP1, and SWI5 regulators. We used consensus patterns, for searching all other motifs. We should note here that when using consensus sequences for all the searches, the results were not significantly different. Nonetheless, profile-based motif searching is more rigorous and should be used whenever possible.

The promoter sequences of yeast genes and the matrix profiles were obtained from the Promoter Database of *Saccharomyces cerevisiae*: SCPD (Zhu and Zhang, 1999). The regulator-specific motifs were obtained from several different sources including SCPD (Zhu and Zhang, 1999), TRANSFAC (Matys et al., 2003), and previous publications (Kellis et al., 2003; Lieb et al., 2001). Searching for the binding motifs present in the promoter region of all putative target genes is facilitated by the Searching Putative Regulatory Element (SPRE) tools available at the SCPD website (<http://cgsigma.cshl.org/jian/>). SPRE tools are capable of searching a motif in multiple sequences using either matrix profiles or consensus sequences.

RESULTS

In this study, we used 18 known regulatory yeast genes known to activate or repress gene expression by binding to the promoter regions of their target genes, and 38 of their known target genes (KTGs). All are presented in the Appendix Table A1. Using the criteria described in Methods these genes were chosen after searching the Yeast Proteome Database (Costanzo et al., 2000), the TRANSFAC database (Matys et al., 2003) and previous publications (Lee et al., 2002; Simon et al., 2001). The expression profiles for pattern recognition used in this study are from the public oligonucleotide GeneChip dataset produced for studying the yeast cell cycle (Cho et al., 1998). Using the expression profile patterns of the KTGs of each regulator gene as templates we discovered hundreds of putative target genes that share expression pattern similarities with KTGs. Since the number of genes is very large and the number of time points in the experimental expression profile is only 17, a lot of these putative genes will be false positives, i.e., they are not actually regulated by the regulator gene in question. We filtered the putative targets genes searching for regulator-specific binding sites in their

promoter regions. There were 267 genes discovered that share similar expression profiles with KTGs and regulator-specific binding sites. All the inferred target genes that are under the control of 18 yeast regulators are shown in Table I. We detail the results below.

Pattern Extraction and Database Queries

Our method does not require identity or even close similarity of expression profiles for the entire course of the experiment. As discussed in the sections describing the algorithms, if the selected KTGs have the same expression profile trends over eight time intervals, the patterns were extracted from these KTGs using a program written in Transact-SQL, in which the TREND-SELECT and SCALE-SELECT algorithms were implemented. The TREND-SELECT algorithm, which serves as the first filter, was designed to eliminate genes according to the expression profile trends (increasing, remaining constant, or decreasing) at eight time intervals. The SCALE-SELECT algorithm is capable of picking up one or two most significant features in the vertical dimension of expression profiles (magnitude of expression) and using them as additional constraints to the constraint string generated by the TREND-SELECT algorithm. The SCALE-SELECT algorithm does not impose any constraints on the absolute level of gene expression but instead to the shape of overall curves. The patterns as represented by the constraints generated by TREND-SELECT and SCALE-SELECT were then used for querying putative target genes from the database.

We use a few examples, shown in Figure 3, to illustrate how efficiently the TREND-SELECT and SCALE-SELECT algorithms find the genes sharing similar patterns. The expression profiles of four groups of KTGs that are under the control of four example regulators, ACE2, YAP1, GCN4 and MCM1, are shown in A1, A2, A3, and A4 of Figure 3, respectively. It is obvious that target genes under the control of the same regulator indeed share similar expression profiles, albeit not identical. The level of similarity is an adjustable degree of freedom for our algorithm, and we detail the parameters in the Methods section. To obtain the genes sharing similar patterns as the KTGs, we queried the microarray database using the patterns extracted from KTGs. Up to more than 100 putative target genes were obtained for each query. The expression profiles of those queried putative target genes are shown in B1, B2, B3, and B4 in Figure 3. These genes shown in B1-4 have indeed similar patterns with the KTGs. In Table II, the numbers of resulting putative target genes for each regulator gene are shown.

Examination of the Presence of Regulator-Specific Binding Sites

Due to the limited number of time points, and large gene space as well as the constraints being applied only to partial

time courses, the genes obtained were not necessarily target genes of the specific regulator, even if they coincidentally shared similar or even identical expression profiles with the real target genes. To eliminate these genes, we examined the presence of regulator-specific binding motifs in the promoter regions of all putative target genes. We found the average number of binding motifs being present in 1kb promoter regions of inferred genes to be almost double the number these motifs were present in the 1kb promoter regions of all genes of *S. cerevisiae*, indicating that the results are statistically significant (Table III).

We found that some inferred target genes share a highly homologous motif that is at most one-base different from the consensus sequences used. In the Appendix Table A1, we present the binding motifs for 18 regulators, and the motif locations in the promoter regions of all inferred genes that both share similar expression profile patterns with KTGs, and have regulator-specific binding sites. All motifs of regulator-specific binding sites can be present on both plus and minus strands, indicating that most motifs may function in an orientation-independent manner. A statistical analysis performed on these promoters suggests that about 44% of promoters have binding sites only on the plus strands, 23% of promoters only on the minus strands, and 33% of promoters on both strands. These results indicate that the binding motif may favor the plus strand over the minus strand. Although it is usually believed that most binding motifs can function in either orientation, only limited studies have been carried out with conclusive results. There is indeed evidence suggesting that regulatory elements can function either in an orientation-dependent (Ares et al., 1985; Natesan and Gilman, 1993; Ueda et al., 1994) or orientation-independent (Liu and Bateman, 1993; Wood et al., 1999; Wu and Lee, 1998) manner.

We then examined all KTGs that were used for pattern recognition and all KTGs rediscovered by the method (target genes discovered by our method that subsequent literature search reveals to be indeed target genes). It was found that the 38 KTGs used for pattern extraction have on average 2.1 binding sites present in their promoter regions and those newly inferred KTGs have on average 2.2 binding sites present in their promoter region. All inferred target genes contain two binding sites in their promoter regions.

For the previous example of four regulators, after examining the presence of regulator-specific binding sites, we obtained the genes whose expression profiles are shown in C1, C2, C3, and C4 in Figure 3.

Examination of Inferred Regulatory Relationships

Although the expression profiles of all KTGs of a given regulator need not be the same over the entire course of the experiment, there are time intervals during which genes share significant similarities in their expression profiles. In this study, we utilized this feature for inference and eventually, after using the binding site filter, we discovered

267 inferred regulatory relationships, listed in Table I. Although most of them are unknown, 40 (about 15%) regulatory relationships were rediscovered ones, i.e., ones that have already been proven experimentally. This rate of rediscovering experimentally validated regulatory rela-

tionships is a first indication of the strength of our approach and illustrates the value of using a subset of KTGs for predicting a subset of unknown target genes (UTGs).

To further analyze the inferred relationships and illustrate the strength of our method, we examined the function of

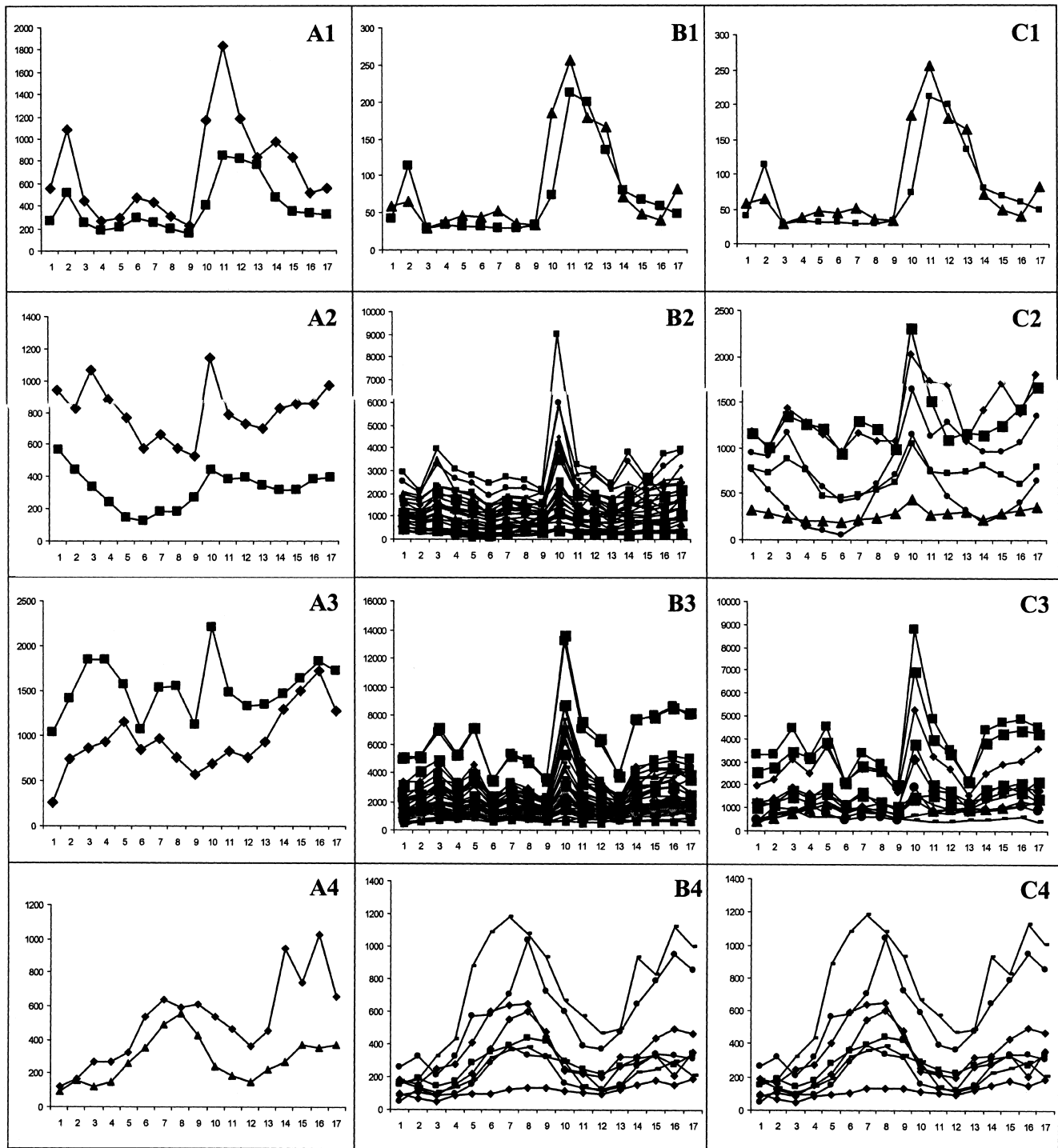


Figure 3. The expression profiles of the KTGs under the control of four example regulators, ACE2, YAP1, GCN4, MCM1, are shown as A1, A2, A3, and A4, respectively. The putative target genes queried from the database with the patterns extracted from the corresponding KTGs are shown in B1, B2, B3, and B4. The putative target genes that were kept, after searching for regulator-specific binding sites, are displayed in C1, C2, C3, and C4, respectively. Y-axis: Expression (intensity). X-axis: Time ($\times 10$ min).

Table II. The genes identified or rejected at each step.

Regulator	The number of genes containing motifs	KTGs used	Expression profile matching	Motif presence		Inferred target genes	
				Filtered	Filtered EVG	Total	Rediscovered EVG
ABF1	1391	2	104	71	0	35	3
ACE2	942	2	2	0	0	2	2
CCBF, SWI4, SWI6	305	1	1	0	0	1	1
GCN4	862	4	78	58	0	24	5
GCR1	3712	3	11	3	0	8	1
GZF3	1353	1	2	1	0	1	0
HAP1	1121	2	24	12	0	12	0
LEU3	45	3	135	130	0	5	2
MAC1	162	2	44	42	0	3	0
MCM1	921	2	28	19	0	10	2
MBP1	584	2	83	30	0	53	5
MIG	371	2	1	0	0	1	2
PHO4	375	2	129	117	0	10	0
RAP1	1085	3	118	74	0	46	8
ROX1	2953	2	17	29	0	13	1
RPN4	79	1	14	13	0	1	0
SWI5	1432	2	71	31	0	38	7
YAP1	760	2	26	20	0	6	1

Note: EVG: Experimentally validated gene.

each one of the inferred target genes to determine if there is any biological context for the discovered regulatory relationships. In particular, we thoroughly evaluated the inferred target genes under the control of ABF1 and MBP1 regulators and we report on them below. Ideally, we would present detailed descriptions of all the inferred regulatory relationships. This would, however, make this article too long and would distract from our method. The interested reader can use Table I and the more detailed Appendix Table A1 as a guide to search the yeast protein database for detailed descriptions and the biological context of all relationships discovered in this paper.

ABF1 Regulated Genes

ABF1p is known to positively regulate transcription of those genes that are involved in protein synthesis and transport, glycolysis, fermentation, and energy pathways as well as cell wall organization (Dolinski et al., 2002; Matys et al., 2003; Zhu and Zhang, 1999). For the inference scheme to be deemed reliable, inferred genes should likely encode proteins that are functionally similar to KTGs of ABF1 or should be involved in these processes.

When RPS28A that encodes a ribosome protein, and ACS1 that encodes an acetate-CoA ligase are used as KTGs for inference, we obtained 35 putative target genes (Table I). Among these, CRY1, RPS16A, RPL17A, RPS7B, RPL16B, RPL27, RPL5A, RPS18A, RPS18B, RPL19B, RPL2B, RPL2A, RPLA3, RPS31, and SUP44 encode proteins that are structural constituents of the ribosome and are involved in protein synthesis. YEF3, TEF1, TIF3, CDC33, and FPR1 are also involved in protein synthesis. In particular, YEF3

and TEF1 encode two proteins that are involved in translation elongation, and TIF3, like DED1, a KTG of ABF1 (Buchman and Kornberg, 1990), encodes a protein that is a translation initiation factor. CDC33 encodes a translation termination factor, and FPR1 encodes a protein that is involved in ribosome assembly.

Several inferred genes, GSP1, RHO1, ARF1, SSB1, and SSB2, have the same or similar functions as YPT1, a KTG

Table III. The average number of regulator-specific binding sites in the entire yeast genome and in the promoter regions of inferred target genes.

Regulators	Genome-wide Binding site(s) /gene	Inferred genes Binding site(s) /gene
ABF1	1.17	1.5
ACE2	1.15	2.5
CCBF, SWI4/6	1.05	1.0
GCN4	1.14	2.29
GCR1	2.0	2.5
GZF3	1.21	2.0
HAP1	1.19	2.16
LEU3	1.0	1.75
MAC1	1.03	1.5
MCM1	1.15	1.2
MIG1	1.04	2.0
MBP1	1.24	2.46
RAP1	1.12	2.26
PHO4	1.09	1.5
ROX1	1.58	4.38
RPN4	1.01	1.0
SWI5	1.23	3.29
YAP1	1.10	1.16
Average	1.2	2.0

of ABF1 (Halfter et al., 1989), which catalyzes the reaction ($\text{GTP} + \text{H}_2\text{O} \rightarrow \text{GDP} + \text{phosphate}$). RHO1, GSP1, and ARF1 catalyze the reaction ($\text{GTP} + \text{H}_2\text{O} \rightarrow \text{GDP} + \text{phosphate}$). SSB2 encodes an ATP phosphohydrolase, which catalyzes the reaction ($\text{ATP} + \text{H}_2\text{O} \rightarrow \text{ADP} + \text{phosphate}$) during protein synthesis. ARF1p, like YPT1p, is responsible for protein transport from ER to Golgi. RHO1p is involved in cell wall organization, and SSB1p and SSB2p are both involved in protein synthesis.

ACS1, the third KTG (Kratzer and Schuller, 1997) used for pattern extraction encodes an acetate-CoA ligase, catalyzes the reaction ($\text{ATP} + \text{acetate} + \text{CoA} = \text{AMP} + \text{diphosphate} + \text{acetyl-CoA}$). Like ACS1, two inferred genes, PGK1 and CDC19 which are KTGs (Svetlov and Cooper, 1995), encode proteins catalyzing two similar reactions ($\text{ATP} + 3\text{-phospho-D-glycerate} = \text{ADP} + 3\text{-phospho-D-glyceroyl phosphate}$) and ($\text{ATP} + 3\text{-phospho-D-glycerate} = \text{ADP} + 3\text{-phospho-D-glyceroyl phosphate}$), respectively. The inferred gene SEC53 encodes a phosphomannose mutase that catalyzes the reaction ($\text{D-mannose 1-phosphate} = \text{D-mannose 6-phosphate}$), and IPP1 encodes a diphosphate phosphohydrolase that catalyzes the reaction ($\text{diphosphate} + \text{H}_2\text{O} = 2 \text{ phosphate}$). PHO5p catalyzes the reaction: ($\text{orthophosphoric monoester} + \text{H}_2\text{O} = \text{an alcohol} + \text{phosphate}$), and PHO88p is involved in inorganic phosphate transport within or between cells.

KTR1p is an alpha-1,2-mannosyltransferase that is located on the Golgi apparatus involved in *N*-glycan processing, *O*-linked glycosylation, and cell wall mannoprotein biosynthesis. ERV25, a functionally unknown protein that serves as a unit in a protein complex, is involved in protein transport from endoplasmic reticulum to Golgi transport. Like YPT1, both SEC53 and ERV25 are involved in protein ER to Golgi transport. KTR1 is on Golgi.

There are only two genes, TRR1 and YHL050C, which we cannot explain because they do not fall in the function scope of ABF1 regulated genes or have similar functions with any KTGs. YHL050C encodes a helicase but the biological process it is involved in is not clear at present.

MBP1 Regulated Genes

The second regulator, MBP1p, is known to bind to the MCB element in the promoter regions of those genes that are involved in cell cycle control, DNA replication, and repair in budding yeast. Previous studies (Costanzo et al., 2000; Koch et al., 1993) have unambiguously demonstrated that MBP1p controls the transcription of CDC2, and CDC21 genes. When CDC2 and CDC21 were used for inference, 54 genes were obtained (shown in Table I). Among these 54 genes, 15 genes including POL1, POL12, POL30, CDC45, RNR1, ECO1, RAD27, RAD53, MSH2, MSH6, PMS1, RFA1, RFA2, DBP2, and CTF4 encode proteins that are involved in DNA replication and/or repair. Like SPT4, a KTG of MBP1 (Porter et al., 2002), four inferred genes, HCM1, SPT16, TOS4, and HPC2 encode transcription factors. TOS4p binds to a number of pro-

motor regions, in particular, the promoters of some genes involved in pheromone response and cell cycle. HPC2p is required for normal cell-cycle regulation of histone gene transcription. Six genes including SMC1, SMC3, IRR1, CTF3, CTF4, and MCD1 are functionally associated with the chromosome replication. SMC1p is a protein involved in mitotic chromosome segregation, and SMC3p, IRR1p, CTF3p, and CTF4p are chromatin cohesion factor. MCD1p is mitotic chromosome determinant and may function in chromosome morphogenesis from the S phase through mitosis. TUB4 is a KTG of MBP1 (Sobel and Snyder, 1995) that encodes gamma-tubulin and functions in attaching chromosomes to the mitotic apparatus. One inferred gene CIN2 encodes tubulin folding factor *c*. SPC42 is a KTG (Donaldson and Kilmartin, 1996) and encodes a structural constituent of cytoskeleton that is involved in spindle pole body duplication. TOF1 encodes topoisomerase I interacting factor 1 involving in DNA replication checkpoint DNA topological change. CLN1 and CLN2 have cyclin-dependent protein kinase regulator activity and KCC4p is a protein kinase located at the bud neck and plays a role in axial budding, and bud growth. BUD9p, RSR1p, and SPH1p have a function in bud-site selection, BNI4p localizes different enzymes to bud neck. Both HIF1p and TFC5p are functionally unknown nuclear proteins that interact with HAT11p, a histone acetyltransferase, presumably having a function in DNA replication. Finally, three inferred genes, RTT106p, RTT107p, and ELG1p are three regulators of Ty1 transposition, and may play a role in reducing DNA recombination and transposition during the cell cycle. There are seven genes including MCD4, CSH1, SEN34, STE23, ELO1, LPP1, and QRI1 that do not fall in the functional scope of MBP1 regulated genes or have same or similar functions to any KTGs of MBP1.

The analysis of inferred genes regulated by ABF1p and MBP1p clearly illustrates that most of the inferred genes fall reasonably well in the biological context and regulatory scope of the particular regulators. This scope could indeed be very wide with many functionally distinct protein groups being included.

To obtain an overall picture of all inferred target genes of 18 regulators, we examined the protein functions of all these genes, and found that about 41.7% of all inferred target genes encode proteins of the same or similar functions as the proteins encoded by their KTGs. Same function is defined based on the proteins being of the same type, for example, dehydrogenases, kinases, transcription initiation factors, transferases, etc. Functionally similar proteins refer to two proteins that catalyze similar reactions, for example, GTPase and ATPase. Without taking into account all proteins with unknown function, we were then able to calculate the percentage of the genes that have the same or similar function to a KTG and the genes with different functions from KTGs. As shown in Figure 4, among 222 genes that have a known function or are KTGs, about 18% genes are in average KTGs (previously validated by

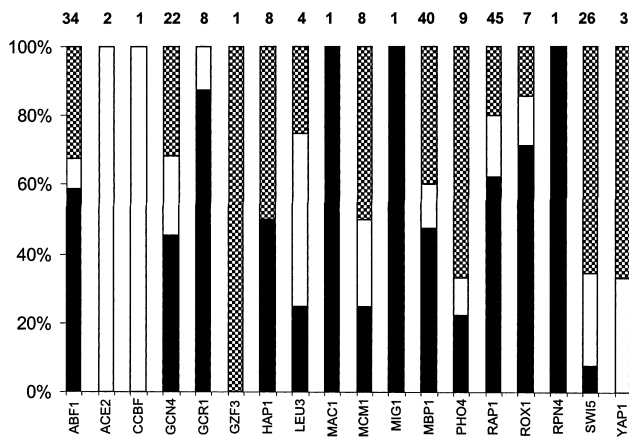


Figure 4. The percentage of rediscovered KTGs (white bars), the genes (black bars) that have same or similar functions to KTGs, and the genes (mosaic) that have different functions in all inferred genes that are functionally known genes. The number of all functionally known genes is shown at the top of each bar.

experiment), 36.1% genes have the same or similar function to a KTG, and 35% genes have different functions.

There is a small fraction of discovered relationships that cannot be explained at present because the inferred genes have yet to be studied experimentally. Either the protein functions or the biological processes they are involved in or both are not clear. For those inferred genes that encode proteins with different function from all KTGs, the relationships may be just a new type that has not been discovered experimentally. Of course, they may simply be false positives. Nonetheless, the number is small and these newly discovered relationships might enable refocusing of experiments and potentially lead to the discovery of new target gene types of known regulators.

The Role of Pattern Recognition and Motif Sequence Presence

The method we developed can be divided into two steps: pattern recognition and motif sequence examination. Both processes contribute greatly to the identification of regulatory relationships (Table II). For example, 1391 genes in yeast genome contain at least one ABF1p binding site in their 1 kb promoter regions. Expression profile matching identified 104 genes when 2 KTGs were used for making the inference. Of these, 69 genes were eliminated in the motif sequence examination. As a result, 35 genes are obtained as the target genes. On the other hand, 1356 genes containing ABF1 binding sites were ruled out in the two-step inference. These 1356 genes can be classified into two groups: those that were not controlled by the ABF1p under the experimental condition and those that were indeed regulated by ABF1 under the experiment condition but can not be inferred by the two KTGs used for inference or by the binding motif used. These genes can be inferred when other KTGs or binding sequences of ABF1p are

used for making an inference. In the 69 genes filtered by the motif presence examination, none was found to be a KTG. Therefore there are no false negative predictions, i.e., no genes determined experimentally to be target genes were filtered out. This was true for all the regulators used in this study.

The combined contribution of pattern recognition and motif examination to target gene inference is further illustrated in Table II where the numbers of genes discovered in each step are shown for all used regulator genes.

The Performance of the Method

We have demonstrated that the two algorithms developed can be used to discover other target genes under the control of the same given regulator. To further demonstrate that the results are statistically significant, we performed the following randomization test. We randomly assigned 352 genes as the target genes of the 18 regulators used in this study. Fourteen genes were assigned to ABF1, 17 to ACE2, 4 to CCBF, 48 to GCN4, 18 to GCR1, 4 to GZF1, 21 to HAP1, 7 to LEU3, 8 to MAC1, 23 MCM1, 27 to MIG1, 25 to MBP1, 10 to PHO4, 30 to RAP1, 20 to ROX1, 7 to RPN4, and 22 to SWI5, 46 to YAP1. The number assigned to each regulator is the same number of actual KTGs of this regulator. We then employed the algorithms to see how many putative target genes we could obtain. When this process was repeated for 1,000 and 10,000 times (changing the 352 genes each time), the results summa-

Table IV. Number of randomly discovered putative target genes (average after repeating 1000 and 10,000 times) and number of genes extracted using the developed algorithms.

Regulators	Randomization	Randomization	This implementation
	test I 1000 times	test II 10,000 times	
ABF1	0.314	0.35	3
ACE2	0.006	0.009	2
CCBF, SWI4, SWI6	0.002	0.001	1
GCN4_1	0.337	0.325	3
GCN4_1	0.729	0.682	2
GCR1	0.15	0.106	1
GZF3	0.01	0.086	0
HAP1	0.182	0.158	0
LEU3_1	0.024	0.027	1
LEU3_2	0.26	0.241	1
MAC1	0.129	0.131	0
MCM1	0.004	0.005	2
MIG1	0.018	0.016	0
MBP1	0.4	0.399	5
PHO4	0.24	0.241	2
RAP1	0.67	0.653	8
ROX1	0.076	0.075	1
RPN4	0.035	0.025	0
SWI5	0.23	0.269	7
YAP1	0.425	0.415	1
Sum	4.241	4.214	40

rized in Table IV were obtained, in terms of the average number of randomly assigned putative target genes that can be extracted using the algorithms.

The results of this randomization test indicate that only about four known target genes can be inferred randomly. In our implementation, 40 known target genes were identified. We conducted a paired student's *t* test between the data from 1000 runs and the real KTGs captured, and calculated $t = 3.71 > t_{0.01}(19) = 2.86$. For the student's *t* test between the data from 10,000 runs, and the number of KTGs captured, $t = 3.72 > t_{0.01}(19) = 2.86$. Thus, we can say with 99% confidence that the algorithms discover the true KTGs 9.5 times more frequently than a random process does. More importantly, the maximum number of all KTGs captured during 1000 runs and 10,000 runs are 9 and 12, respectively, far less than the discovered 40.

The same randomization test was performed with the functions of inferred target genes examined and compared to that of assigned putative KTGs. The results indicate that only 12.2% of all randomly inferred target genes have the same or similar function to known target genes. In our implementation, 36.1% inferred target genes have same or similar function to known target genes, indicating that the frequent discovery of functionally similar genes is not accidental.

DISCUSSION

We proposed that genes under the control of the same activator or repressor gene often share similar expression profiles at specific stages or even through the full time-course of a microarray experiment. We developed an algorithm to infer the target genes of known regulators at a genome-wide scale. There are several distinct features in our approach.

1. A gene is identified as a target gene of a specified regulator when it shows some similarity to at least one of the experimentally validated target genes of that regulator and contains regulator specific binding sites in its promoter region. This distinguishes ours from other

approaches that attempt to identify the regulatory relationships by comparing regulating genes and target genes. Even with significant time delays between the regulatory and target genes expression, because of transcriptional regulation, translational control, and post-translational modifications, our approach allows the regulatory relationships to be inferred. When regulators and their targets have dissimilar profiles other approaches will have difficulty in inferring these target genes. For example, the expression profiles of regulators HAP1 and MBP1 and their target genes, (CYC1 and CYC7 for HAP1, and CDC2 and CDC21 for MBP1), are shown in Figure 5. The differences are apparent between the expression profiles of the two regulators and their experimentally validated target genes.

2. Pattern searching is not full time-course based. We did not use distance-based clustering methods to find the genes sharing similar patterns because these methods classify two genes into the same cluster only when they have the same expression profile over the whole time course. The genes that have the same expression profile only over a shorter period are not likely to be classified together. Actually, a regulating gene often controls its target genes only occasionally, which results in its target genes sharing the same profile over a certain time period rather than the full time course of the experiment.
3. Our approach relies on the existing pool of knowledge of gene regulatory networks built by life scientists. Since experimentally validated relationships are used as inputs, our approach fully utilizes the available resources. In addition, the inferred relationships are those that are located in the neighborhood of existing networks and thus can stimulate life scientists to test them. Importantly, since discovery of gene regulatory relationships via experimental means is still time-consuming and tedious, the approach we describe can rapidly generate a large number of gene regulatory relationships, enable focusing of experiments, and potentially accelerate the assembly of gene regulatory networks.
4. The entire process can be repeatedly executed with different KTGs and different binding motifs, and the

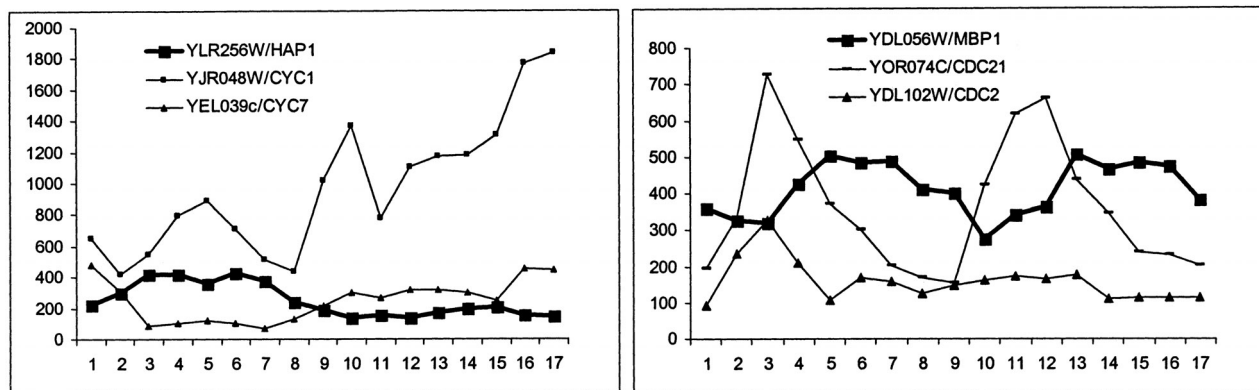


Figure 5. Differences in the expression profiles of regulators and target genes.

results from different inputs can be compared to draw firm conclusions, making it particularly useful and convenient method for life scientists.

Increased performance can potentially be achieved by selecting more KTGs and using more efficient motif searching algorithms. It is preferable that multiple KTGs are used, since the patterns extracted are more likely to represent the signature of the regulator. Also, other recently developed motif searching algorithms (Brazma et al., 1998; Solan et al., 2004) may discover more putative target genes from the candidate genes.

The efficiency of our approach can be greatly improved by automating the whole process. We have developed the Transact-SQL program so that the examination of the presence of regulator-specific binding sites in the promoter regions can be automatically executed. As a result, only the name of known target genes and the consensus sequences are needed as an input. Then the Transact-SQL program automatically extracts the patterns from the given KTGs, queries the microarray database, and examines the presence of regulator-specific binding sites. During this process, most intermediate results are saved in temporary tables and are finally discarded when the session is over. The important results, such as the putative target genes that share the same patterns with the given target genes, and the inferred genes that not only share the same patterns with given target genes but also contain the regulator-specific binding sites in their promoter regions, as well as the details of binding sites number and position are saved in the normal tables. We calculated the CPU time and found that it usually takes 200–300 ms to parse and compile the program and 100–150 ms to execute the compiled program on the SQL 2000 Server. It thus becomes a straightforward exercise to infer up to thousands of regulatory relationships in just one day, if the microarray data, known regulator and target genes, and regulator-specific binding sites are provided. Moreover, further improvements in predictive ability can result by implementing the algorithms in an iterative fashion: the discovered genes could be added to the initial KTGs and then the database could be searched anew. Changing the intervals required for sim-

ilarity in each iteration could result in “fainter” targets being discovered.

Since different approaches are radically different in design and implementation as well as data preparation, it is difficult to compare the efficiency of our approach with most of the others. Nonetheless, careful examination reveals that although our approach can detect 40 experimentally validated relationships from the 267 inferred relationships, no known regulatory relationship were recovered out of top 40 discovered relationships by the smooth response surface algorithm (Xu et al., 2002). Moreover, the calculations of the smooth response surface algorithm are particularly computationally intensive.

It is not possible to conduct a thorough sensitivity and specificity analysis for the method we developed, since the numbers of known regulators and known target genes is rather small, and to our knowledge there is no regulator with all of its target genes completely discovered. Nonetheless, by carefully examining the discovered target genes and the discarded ones, our confidence to the proposed scheme increases, mainly because we can explain the biological context of the regulatory relationships. Moreover, as shown in Table II, none of the filtered out genes, after searching for regulator-specific binding sites was found to be an experimentally validated gene.

Currently, one limitation for this approach is that the binding information of most regulators is still not known. Recently emerging techniques look promising in generating this knowledge (Lee et al., 2002; Lieb et al., 2001; Liu et al., 2002), and a number of computational approaches (Sinha and Tompa, 2002, Thijs et al., 2001; van Helden et al., 1998), together with comparative genomics (Kellis et al., 2003), have been employed to accomplish this goal.

In conclusion, the algorithm we developed is particularly suitable for inferring gene regulatory relationships in cases where a regulating gene and its targets genes do not share similar expression profiles because of complicated, multiple layers of gene regulation and time delay. It is also valuable in that it provides a framework for mining gene regulatory relationships using the microarray data in conjunction with several publicly available databases and existing experimental knowledge.

APPENDIX

Table A1. Target genes inferred from regulators and known targets. The first column shows the 18 regulator genes used in this study. The second shows the genes known to be target genes of the regulators. The third column shows the number of binding sites in the promoter region of KTGs. The fourth column shows the discovered target genes (the fifth is the same information with the alternative gene names). The sixth through eighth columns provide more detailed information on the regular-specific binding sites.

Regulators	Matrices / Consensus sequences		Inferred target genes		Binding sites	Position in promoters	
	Known target genes	Number of binding sites					
	Matrix (cut_off value = 0.8)		YDL192W ‡	ARF1 ‡	TCACATATAACG	+	–232 ~ –221
	YGR118W (RPS28A)	1 (–)	YAL038W *	CDC19 *	CGTCGGTTGTGA	–	–837 ~ –826
	YAL054C (ACS1)	3 (–+–)			TCTCTCGCAACG	+	–608 ~ –597
			YOL139C	CDC33	CGTTTAGAGTAA	–	–248 ~ –237
			YCR031C ‡	CRY1 ‡	TCACTGTGACCG	+	–627 ~ –616

Table A1. Continued

Regulators	Matrices / Consensus sequences		Inferred target genes	Binding sites	Position in promoters		
	Known target genes	Number of binding sites					
ABF1			YML012W	ERV25	TCATCAACAACG	+	-421 ~ -410
			YNL135C	FPR1	TCATTGACCACG	+	-911 ~ -900
					CGTCAGTATTGA	-	-486 ~ -475
			YLR293C	GSP1	TCGCTCTGGACG	+	-223 ~ -212
			YNL031C	HHT2	CGTGTTCAAGTGT	-	-905 ~ -894
			YBR011C	IPP1	TCCCCACCCACG	+	-447 ~ -436
					TCATTAGATACG	+	-185 ~ -174
			YOR099W	KTR1	TCACATAATACG	+	-463 ~ -452
					TCACGTGCAACG	+	-442 ~ -431
			YCR012W *	PGK1 *	TCATAAAGCACG	+	-552 ~ -541
					CGTCGCTCGTGA	-	-514 ~ -503
			YBR093C	PHO5	CCTCTTTAGTGA	-	-867 ~ -856
					TCAGTCCCCACG	+	-610 ~ -599
			YBR106W	PHO88	TCAAGTTGAACG	+	-830 ~ -819
					TCACTTTACACG	+	-253 ~ -242
			YPR165W †	RHO1 †	CGTGCAGCGTGA	-	-377 ~ -366
			YNL301C †	RP28B †	ACACCAAACACG	+	-187 ~ -176
			YHR010W †	RPL27 †	TCACTAGTAAAG	+	-131 ~ -120
			YBR031W *	RPL2A *	TCACCCTCAACG	+	-533 ~ -522
					CGTTCAATGGGA	-	-489 ~ -478
					TCACGTACACG	+	-181 ~ -170
			YDL130W †	RPLA3 †	TCACTATATACG	+	-921 ~ -910
					CTTTCAAAGTGA	-	-486 ~ -475
			YIL018W †	RPL5A †	TCAAAAAATACG	+	-927 ~ -916
					TCACAAACAACG	+	-613 ~ -602
			YNL162W †	RPL41A †	TCCCTAGTAAACG	+	-738 ~ -727
			YDR012W †	RPL2B †	CGTTAGTTTTGA	-	-327 ~ -316
					CGTTAGTTTTGA	-	-327 ~ -316
			YJR145C †	RPS7B †	TCACTTAGTACG	+	-201 ~ -190
			YLR448W †	RPL16B †	TCACTTTGTACG	+	-819 ~ -808
			YBR084C-A †	RPL19B †	CGTTCTCGGTGA	-	-360 ~ -349
			YNL302C †	RPS16A †	GGTGTGTTGGTGA	-	-214 ~ -203
			YDR025W †	RPS18A †	CGTTTCCGGTGA	-	-991 ~ -980
			YML026C †	RPS18B †	TCACAATGAAAG	+	-467 ~ -456
		YFL045C	SEC53	TGACCATGAACG	+	-492 ~ -481	
				TCATGAGACACG	+	-228 ~ -217	
		YNL209W †	SSB2 †	TCACGATAACCG	+	-772 ~ -761	
				TCATTCGAAACG	+	-225 ~ -214	
		YGL123W †	SUP44 †	TCATCATTAACG	+	-294 ~ -283	
		YBR118W †	TEF2 †	GGTTGTTTGTGA	-	-392 ~ -381	
				TCACCAAAAAAAG	+	-344 ~ -333	
		YPR163C †	TIF3 †	TCACGCACCACG	+	-199 ~ -188	
		YDR353W	TRR1	CGTTTGATTTGA	-	-702 ~ -691	
				TCACTGTGAACG	+	-309 ~ -298	
		YLR249W	YEF3	TCACTTCAACC	+	-318 ~ -307	
				TCACTTAGCACG	+	-262 ~ -251	
		YHL050C		TCAAGAGTAACG	+	-132 ~ -121	
ACE2		GCTGGT	YER124C *	DSE1 *	GCTGGT	+	-446 ~ -441
	YLR286C (CTS1)	3 (+ + -)			GCTGGT	-	-398 ~ -393
	YHR143W (DSE2)	3 (+ - -)			GCTGGT	+	-253 ~ -248
			YGL028C *	SCW11 *	GCTGGT	+	-924 ~ -919
				GCTGGT	-	-533 ~ -528	
CCBF SWI4 SWI6		CGCGAAA	YNL289W *	PCL1 *	CGCGAAA	+	-303 ~ -297
	YPL256C (CLN2)	1 (+)				-	-294 ~ -288
		Matrix (cut_off value = 0.9)	YNL065W †	AQR1 *	TGACTC	+	-551 ~ -546
	YIL116W (HIS5)	3 (- - +)			TGACTC	+	-451 ~ -446

Table A1. Continued

Regulators	Matrices / Consensus sequences		Inferred target genes	Binding sites		Position in promoters	
	Known target genes	Number of binding sites					
GCN4	YOR202W(HIS3)	2 (-+)	YJL071W *	ARG2 *	TAGTCA	-	-113 ~ -108
			YJL088W *	ARG3 *	TGACTC	+	-939 ~ -934
	YPR025C ‡	CCL1 ‡	TGACTC	+	-132 ~ -127		
			AAGTCA	-	-960 ~ -955		
			TGACTC	+	-488 ~ -483		
			TGACTG	+	-725 ~ -720		
			TGACTG	+	-640 ~ -635		
			TGACTT	+	-583 ~ -578		
			TGACTG	+	-578 ~ -573		
			TGACTT	+	-391 ~ -386		
			TGACTC	+	-148 ~ -143		
			YPR167C *	MET16 *	TGACTC	+	-148 ~ -143
			YDR318W *	MCM21	CAGTCA	-	-762 ~ -757
	YGL257C ‡	MNT2 ‡	CAGTCA	-	-302 ~ -297		
			TGACTA	+	-991 ~ -986		
	YNL316C ‡	PHA2 ‡	TGACTC	+	-921 ~ -916		
			TGACTT	+	-454 ~ -449		
	YMR270C ‡	RRN9 ‡	TGACTC	+	-26 ~ -21		
			TGACTG	+	-713 ~ -708		
	YOR213C ‡	SAS5 ‡	TGACTC	+	-530 ~ -525		
			CAGTCA	-	-304 ~ -299		
	YHL014C	YLF2	AAGTCA	-	-164 ~ -159		
			TGACTC	+	-789 ~ -784		
YPR077C		TGACTT	+	-838 ~ -833			
YHR150W		TGACTG	+	-654 ~ -649			
		AAGTCA	-	-634 ~ -629			
		GAGTCA	-	-83 ~ -78			
	Matrix (cut_off = 0.90)						
GCR1	YER055C (HIS1)	1 (+)	YEL036C ‡	ANP1 ‡	TGACTG	+	-418 ~ -413
	YCL030C (HIS4)	3 (+ + +)	YDL192W	ARF1	AAGTCA	-	-592 ~ -587
	YDL171C *	GLT1 *	AAGTCA	-	-439 ~ -434		
			AAGTCA	-	-341 ~ -336		
			TGACTC	+	-572 ~ -567		
	YGL009C ‡	LEU1 ‡	AAGTCA	-	-455 ~ -450		
			TAGTCA	-	-365 ~ -360		
	YER091C ‡	MET6 ‡	AAGTCA	-	-883 ~ -878		
			TGACTC	+	-301 ~ -296		
	YER091C ‡	MET6 ‡	CAGTCA	-	-76 ~ -71		
			CAGTCA	-	-668 ~ -663		
	YDL232W	OST4	TGACTT	+	-491 ~ -486		
			GAGTCA	-	-459 ~ -454		
	YER117W	RPL17B	GAGTCA	-	-354 ~ -349		
			CAGTCA	-	-345 ~ -340		
	YDR025W	RPS18A	TAGTCA	-	-298 ~ -293		
			GAGTCA	-	-298 ~ -293		
	YCR053W *	THR4 *	AAGTCA	-	-603 ~ -598		
			TGACTC	+	-37 ~ -32		
	YDR002W ‡	YRB1 ‡	TGACTC	+	-547 ~ -542		
			GAGTCA	-	-879 ~ -874		
	YDR002W ‡	YRB1 ‡	GAGTC	-	-225 ~ -220		
			TGACTC	+	-415 ~ -410		
YDR002W ‡	YRB1 ‡	GAGTCA	-	-179 ~ -174			
		TAGTCA	-	-126 ~ -121			
YDR002W ‡	YRB1 ‡	TGACTA	+	-844 ~ -839			
		TGACTC	+	-577 ~ -572			
YDR002W ‡	YRB1 ‡	CAGTCA	-	-475 ~ -470			
	CTTCC						
GCR1	YAL038W (CDC19)	6 (5+, 1 -)	YBL092W ‡	RPL32 ‡	CTTCC	+	-890 ~ -886
	YCR012W (PGK1)	5 (3+, 2 -)	YBL087C ‡	RPL17A ‡	CTTCC	-	-108 ~ -104
	YOL086C (ADH1)	4 (- + - +)			CTTCC	+	-969 ~ -965
					CTTCC	-	-872 ~ -868
				CTTCC	-	-782 ~ -778	

Table A1. Continued

Regulators	Matrices / Consensus sequences		Inferred target genes	Binding sites	Position in promoters
	Known target genes	Number of binding sites			
			YBL027W ‡ RPL19A ‡	CTTCC	+ -702 ~ -698
			YBR084C-A ‡ RPL19B ‡	CTTCC	- -338 ~ -334
				CTTCC	+ -328 ~ -324
				CTTCC	- -156 ~ -152
			YBR181C ‡ RPS10A ‡	CTTCC	+ -147 ~ -143
				CTTCC	- -842 ~ -838
				CTTCC	+ -790 ~ -786
				CTTCC	+ -550 ~ -546
				CTTCC	- -49 ~ -45
			YBR189W ‡ SUP46 ‡	CTTCC	+ -731 ~ -727
				CTTCC	- -347 ~ -343
			YBR118W * TEF2 *	CTTCC	+ -952 ~ -948
				CTTCC	+ -519 ~ -515
			YBR191W ‡ URP1 ‡	CTTCC	+ -802 ~ -798
				CTTCC	- -190 ~ -186
		GATAAC			
GZF3	YDL210W (UGA4)	2 (+ -)	YDR511W ACN9	GATAAC	+ -838 ~ -833
				GATAAC	+ -365 ~ -360
		CGGNNNTANCGG			
HAP1	YJR048W (CYC1)	1 (-)	YJR121W ATP2	CGCGAATCACGG	- -404 ~ -393
	YEL039C (CYC7)	1 (FUP -)	YEL011W GLC3	CGCTATCATCGC	- -657 ~ -646
			YOR032C ‡ HMS1 ‡	CGCTTTTAACGG	- -453 ~ -442
			YOL111C HRE212	CGCGGTGTGCGC	- -930 ~ -919
				CGCACACCGCGC	+ -929 ~ -918
				CGCCCGCACCGC	+ -626 ~ -615
			YKL085W ‡ MDH1 ‡	CGGGGAAGGCGG	- -544 ~ -533
				CGGATTTGGCGG	- -504 ~ -493
				CGCGCCAAGCGG	+ -234 ~ -223
				CGCCGGGCTCGC	+ -569 ~ -558
				CGCCAAATCCGC	+ -503 ~ -492
			YFR033C ‡ QCR6 ‡	CGCAAAGTGCGG	+ -834 ~ -823
				CGCCTTCTGCGC	- -552 ~ -541
				CGCGCCGCACGG	+ -512 ~ -501
				CGCTTTGGACGG	- -322 ~ -311
				CGGATGTACCGG	- -290 ~ -279
				CGGGCATATCGG	+ -225 ~ -214
			YLL013C PUF3	CGGTCTGATCGG	- -551 ~ -540
				CGCTATCTACGG	+ -433 ~ -422
			YDL210W ‡ UGA4 ‡	CGCAAGCCGCGG	- -490 ~ -479
				CGGATAACGCGG	+ -254 ~ -243
			YLR327C	CGGATTGTGCGG	+ -381 ~ -370
			YKL067W YNK1	CGCGTAATTCGC	+ -214 ~ -203
			YIL100W	CGCTAGCTCCGC	- -463 ~ -452
				CGCAGGGAACGG	- -346 ~ -335
			YKL066W	CGCGTAATTCGC	+ -683 ~ -672
		CGGGACCGG			
LEU3	YGL009C (LEU1)	1 (+)	YNL104C * LEU4 *	CGGGACCGG	+ -449 ~ -441
		CGGWWCCGG or GCCGTSKACGGC	YPR145W ‡ ASN1 ‡	GCCGTGGACGGC	+ -331 ~ -320
			YOR375C * GDH1 *	CGGTTCCGG	- -404 ~ -396
	YLR355C (ILV5)	2 (- +)		CGGAACCGG	+ -403 ~ -395
	YGL009C (LEU1)	2 (- +)	YEL017C-A PMP2	CGGATCCGG	- -972 ~ -964
			YEL027W	GCCGTCTACGGC	+ -278 ~ -267
				GCCGTCTACGGC	+ -268 ~ -257
		TTTGCTCA			
MAC1	YPR124W (CTR1)	2 (+ -)	YPL270W ‡ MDL2 ‡	TTTGCTCA	+ -955 ~ -948
	YFR055W	1 (+)	YLR051C	TTTGCTCA	- -273 ~ -266
			YOL137W	TTTGCTCA	+ -876 ~ -869

Table A1. Continued

Matrices / Consensus sequences							
Regulators	Known target genes	Number of binding sites	Inferred target genes		Binding sites	Position in promoters	
Matrix (cut_off = 0.83)							
MCM1	YDR146C (SWI5)	1 (+)	YGR177C	ATF2	CCTATTTTGG	-	-233 ~ -224
	YPL119W (CLB2)	1 (+)	YLR131C	ACE2	CCCAATAAGG	+	-748 ~ -739
			YGR108W *	CLB1 *	CCCAAAGAGG	+	-616 ~ -607
			YMR001C *	CDC5 *	CCTAAAAGGG	-	-519 ~ -510
			YOL158C ¥	ENB1 ¥	CCAAATAAGG	+	-336 ~ -327
			YOR274W	MOD5	CCTTATTGGC	-	-661 ~ -652
					CCGGTTTAGG	+	-323 ~ -314
			YOL043C	NTG2	CCTAACTTGG	-	-124 ~ -115
			YLR071C ¥	RGR1 ¥	CCCTGATAGG	+	-819 ~ -810
					CCTGATAGGG	-	-818 ~ -809
			YIL158W		CCTTTTCGGG	-	-161 ~ -152
			YJL051W		CCTTTTTGGG	-	-255 ~ -246
	GCGGGG						
MIG1	YDR516C (EM12)	1 (+)	YML042W ¥	CAT2	GCGGGG	-	-254 ~ -249
	YKR075C	4 (- + - +)		YOL158C	GCGGGG	-	-238 ~ -233
ACGCGTNA							
MBP1	YOR074C (CDC21)	2 (+ -)	YNL233W	BN14	ACGCGTCT	-	-291 ~ -284
	YDL102W (CDC2)	4 (+ + - -)			ACGCGTCA	+	-289 ~ -282
			YGR041W	BUD9	TCGCGTGT	+	-720 ~ -713
					TCGCGTCA	-	-701 ~ -694
			YLR103C	CDC45	ACGCGTAT	+	-180 ~ -173
					ACGCGTTA	-	-152 ~ -145
					ACGCGTTT	+	-150 ~ -143
			YPL241C	CIN2	ACGCGTCT	+	-85 ~ -78
					TCGCGTTT	-	-64 ~ -57
			YPL256C ¥	CLN2 ¥	TCGCGTCA	+	-965 ~ -958
			YBR161W	CSH1	TCGCGTAT	+	-731 ~ -724
					TCGCGTAT	-	-504 ~ -497
			YLR381W	CTF3	TCGCGTAT	-	-111 ~ -104
			YPR135W	CTF4	TCGCGTTA	-	-130 ~ -123
					ACGCGTAA	+	-123 ~ -116
			YOR033C ¥	DHS1 ¥	TCGCGTTT	+	-310 ~ -303
					ACGCGTAA	-	-256 ~ -249
					ACGCGTCT	+	-254 ~ -247
			YPR175W ¥	DPB2 ¥	ACGCGTTT	-	-409 ~ -402
					ACGCGTGT	+	-407 ~ -400
					ACGCGTCA	-	-127 ~ -120
					ACGCGTTA	+	-125 ~ -118
					TCGCGTGA	+	-113 ~ -106
			YBR278W	DPB3	TCGCGTAA	-	-79 ~ -72
			YOR144C	ELG1	ACGCGTTT	-	-149 ~ -142
					ACGCGTAA	+	-147 ~ -140
			YJL196C	ELO1	ACGCGTGA	+	-357 ~ -350
			YCR065W ¥	HCM1 ¥	TCGCGTCT	-	-382 ~ -375
					ACGCGTTA	-	-319 ~ -312
					ACGCGTTA	+	-317 ~ -310
					ACGCGTTT	-	-271 ~ -264
					ACGCGTTT	-	-514 ~ -507
			YLL022C	HIF1	ACGCGTCA	+	-512 ~ -505
				TCGCGTTT	+	-88 ~ -81	
				TCGCGTAT	+	-250 ~ -243	
		YBR215W ¥	HPC2 ¥	ACGCGTGA	+	-135 ~ -128	
		YIL026C	IRR1	ACGCGTCA	-	-130 ~ -123	
				ACGCGTAA	+	-128 ~ -121	
				ACGCGTTT	-	-168 ~ -161	
		YCL024W ¥	KCC4 ¥	ACGCGTAA	+	-166 ~ -159	
				ACGCGTGA	-	-162 ~ -155	
		YDR503C	LPP1	ACGCGTAA	+	-160 ~ -153	

Table A1. Continued

Regulators	Matrices / Consensus sequences		Inferred target genes	Binding sites	Position in promoters	
	Known target genes	Number of binding sites				
				ACGCGTCA	+ -906 ~ -899	
				ACGCGTTT	- -374 ~ -367	
			YDL003W	MCD1	ACGCGTTT	+ -372 ~ -365
				TCGCGTTT	+ -334 ~ -327	
				TCGCGTCT	+ -278 ~ -271	
				TCGCGTTT	- -152 ~ -145	
			YOL090W ‡	MSH2 ‡	TCGCGTTT	- -119 ~ -112
			YKL165C	MCD4	ACGCGTTT	- -198 ~ -191
			YDR097C ‡	MSH6 ‡	ACGCGTGA	+ -196 ~ -189
				ACGCGTAT	- -178 ~ -171	
				ACGCGTCT	+ -176 ~ -169	
				TCGCGTCT	- -579 ~ -572	
			YDL127W ‡	PCL2 ‡	TCGCGTCA	- -396 ~ -389
				TCGCGTAA	+ -378 ~ -371	
				ACGCGTCT	- -35 ~ -28	
			YNL082W ‡	PMS1 ‡	ACGCGTCT	+ -33 ~ -26
				ACGCGTTT	- -210 ~ -203	
			YNL102W *	POL1 *	ACGCGTTA	+ -208 ~ -201
				ACGCGTTT	- -175 ~ -168	
				TCGCGTGT	+ -168 ~ -161	
			YBL035C ‡	POL12 ‡	ACGCGTCT	- -224 ~ -217
				ACGCGTAA	+ -222 ~ -215	
				ACGCGTCA	- -201 ~ -194	
				ACGCGTCT	+ -199 ~ -192	
			YBR088C ‡	POL30 ‡	TCGCGTTT	- -204 ~ -197
				ACGCGTTT	- -197 ~ -190	
				ACGCGTAA	+ -195 ~ -188	
			YDL103C	QRI1	ACGCGTAA	- -132 ~ -125
				ACGCGTAA	+ -130 ~ -123	
			YKL113C *	RAD27 *	ACGCGTCT	+ -838 ~ -831
				ACGCGTTA	- -182 ~ -175	
				ACGCGTTT	+ -180 ~ -173	
				TCGCGTTT	- -135 ~ -128	
			YER095W ‡	RAD51 ‡	ACGCGTAA	+ -128 ~ -121
			YPL153C ‡	RAD53 ‡	ACGCGTCA	+ -201 ~ -194
				ACGCGTCA	- -261 ~ -254	
			YAR007C	RFA1	ACGCGTAA	+ -259 ~ -252
				ACGCGTGA	- -167 ~ -160	
				ACGCGTAA	+ -165 ~ -158	
				ACGCGTCT	- -137 ~ -130	
			YNL312W	RFA2	ACGCGTGA	+ -135 ~ -128
				ACGCGTTT	- -122 ~ -115	
				ACGCGTTA	+ -120 ~ -113	
			YER070W *	RNR1 *	ACGCGTTT	- -110 ~ -103
				ACGCGTAA	- -494 ~ -487	
				ACGCGTTT	+ -492 ~ -485	
				ACGCGTTT	- -444 ~ -437	
				ACGCGTAA	+ -442 ~ -435	
				ACGCGTTT	- -373 ~ -366	
				ACGCGTCA	+ -371 ~ -364	
			YGR152C ‡	RSR1 ‡	ACGCGTAA	+ -306 ~ -299
				TCGCGTCA	+ -293 ~ -286	
			YNL206C	RTT106	TCGCGTTT	- -260 ~ -253
			YHR154W	RTT107	ACGCGTCA	- -132 ~ -125
				ACGCGTGA	- -189 ~ -182	
				ACGCGTTA	+ -187 ~ -180	
			YAR008W	SEN34	TCGCGTCA	+ -123 ~ -116
				TCGCGTGA	- -592 ~ -585	
				ACGCGTGA	- -218 ~ -211	
				ACGCGTCT	+ -216 ~ -209	
				ACGCGTAA	- -188 ~ -181	
			YJL074C ‡	SMC3 ‡	ACGCGTGA	+ -186 ~ -179

Table A1. Continued

Regulators	Matrices / Consensus sequences		Inferred target genes	Binding sites	Position in promoters	
	Known target genes	Number of binding sites				
				GGACCCA	+ -713 ~ -707	
				GGACCCA	+ -668 ~ -662	
			YCR012W *	PGK1 *	AAACCCA	+ -474 ~ -468
			YDR447C †	RP51B †	GCACCCC	+ -972 ~ -705
			YOR224C	RPB8	AAACCCA	+ -712 ~ -706
			YDL081C †	RPLA1 †	TGGGTGA	- -974 ~ -968
			YOL039W †	RPLA2 †	TGGGTTC	- -955 ~ -949
					TGGGTCT	- -763 ~ -757
			YDL130W †	RPLA3 †	TCACCCA	+ -768 ~ -762
					AAACCCA	+ -276 ~ -270
			YIL018W †	RPL5A †	GTACCCA	+ -976 ~ -970
					GAACCCA	+ -416 ~ -410
					AAACCCA	+ -936 ~ -390
					AAACCCA	+ -372 ~ -366
					TGGGTGG	- -914 ~ -188
			YDL082W †	RPL13A †	TGGGTGT	- -277 ~ -928
			YBL087C †	RPL17A †	ACACCCC	+ -339 ~ -333
					AAACCCA	+ -318 ~ -312
					AAACCCA	+ -190 ~ -184
			YBL027W *	RPL19A	AAACCCA	+ -388 ~ -382
			YDL075W †	RPL43A †	GTACCCA	+ -939 ~ -933
					TGGGTGT	- -807 ~ -999
			YMR142C †	RPL13B †	TGGGTGC	- -418 ~ -412
					GCACCCA	+ -333 ~ -327
			YER117W †	RPL17B †	GAACCCA	+ -372 ~ -366
			YBR084C-A *	RPL19B *	CCACCCA	+ -780 ~ -774
					ACACCCA	+ -434 ~ -428
			YDR471W †	RPL27B †	ACACCCT	+ -668 ~ -662
					ATACCCA	+ -639 ~ -633
					ACACCCA	+ -352 ~ -346
					CCACCCA	+ -337 ~ -934
					TGGGTGA	- -310 ~ -304
			YGR148C †	RPL30B †	AAACCCA	+ -745 ~ -739
					TGGGTGG	- -442 ~ -436
			YDR382W *	RPL45 *	GAACCCA	+ -933 ~ -387
			YBL092W †	RPL32 †	TGGGTGT	- -784 ~ -778
			YHR203C †	RPS7A †	TGGGTTT	- -975 ~ -969
					ATACCCA	+ -953 ~ -947
					GCACCCA	+ -810 ~ -804
					TCACCCA	+ -470 ~ -464
			YMR143W †	RPS16A †	TGGGTGC	- -749 ~ -743
					GCACCCA	+ -664 ~ -658
			YBL072C †	RPS8A †	CCACCCA	+ -729 ~ -723
					ACACCCT	+ -680 ~ -674
					ACACCCT	+ -668 ~ -662
			YDR025W †	RPS11A †	ACACCCA	+ -384 ~ -378
					TGGGTGG	- -242 ~ -236
					GGACCCA	+ -917 ~ -920
			YER102W †	RPS8B †	AAACCCA	+ -750 ~ -744
					ATACCCA	+ -357 ~ -936
					TGGGTCT	- -142 ~ -136
			YBR048W †	RPS18B †	GCACCCA	+ -306 ~ -930
					AAACCCA	+ -290 ~ -284
			YGR118W †	RPS28A †	ACACCCA	+ -387 ~ -939
					ACACCCA	+ -370 ~ -364
					ACACCCA	+ -224 ~ -218
			YER074W †	RPS24EA †	TGGGTGA	- -803 ~ -977
					ACACCCA	+ -338 ~ -332
					CCACCCA	+ -129 ~ -123
			YIL069C †	RPS24EB †	GTACCCA	+ -179 ~ -173
			YDR077W	SED1	ACACCCA	+ -906 ~ -990
			YDL191W †	SOS1 †	GCACCCC	+ -520 ~ -514

Table A1. Continued

Regulators	Matrices / Consensus sequences		Inferred target genes	Binding sites	Position in promoters
	Known target genes	Number of binding sites			
				ATTGTT	+ -875 ~ -870
				ATTGTT	- -347 ~ -342
				AACAAT	- -93 ~ -88
				AACCAT	- -70 ~ -65
				AACAAG	- -19 ~ -14
			YBR095C RXT2	AACAAG	+ -369 ~ -364
				ATTGTT	- -170 ~ -165
				AACAAT	+ -34 ~ -29
			YBR123C ‡ TFC1 ‡	ATTGTT	- -667 ~ -662
			YBR156C SLI15	AACAAG	- -890 ~ -885
				AACAAT	- -497 ~ -492
				AACCAT	- -246 ~ -241
			YAR035W ‡ YAT1 ‡	AACAAT	- -466 ~ -461
				AACCAT	- -350 ~ -345
			YAL053W	AACAAG	+ -807 ~ -802
				ATTGTT	- -332 ~ -327
				AACAAT	- -97 ~ -92
			YBL095W	AACAAG	+ -920 ~ -915
				ATTGTT	- -914 ~ -909
				AACCAT	- -811 ~ -806
				AACAAG	- -741 ~ -736
				AACAAT	- -472 ~ -467
				AACAAG	- -311 ~ -306
				AACAAG	- -64 ~ -59
		GGTGGCAAA			
RPN4	YGL048C (RPT6)	1 (+)	YOR362C PRE10 ‡	GGTGGCAAA	+ -128 ~ -120
		Matrix (cut_off = 0.85)			
SWI5	YKL164C (PIR1)	2 (++ -)	YBR158W * AMN1 *	TGCTGC	+ -777 ~ -772
	YKL163W (PIR3)	2 (+-)		TGCTGG	+ -703 ~ -698
				TGCTGG	+ -468 ~ -463
			YLR295C ATP14	TCAGCA	+ -541 ~ -536
			YKL185W ASH1 *	CCAGCA	- -639 ~ -634
				CCAGCA	- -463 ~ -458
			YOR264W DSE3 *	TGCTGG	- -415 ~ -410
				CCAGCG	+ -340 ~ -335
				CCAGCC	- -314 ~ -309
				ACAGCA	- -291 ~ -286
				CCAGCA	- -262 ~ -257
			YNR067C DSE4	TCAGCA	- -485 ~ -480
				TGCTGG	- -400 ~ -395
				CCAGCA	+ -282 ~ -277
				AGCTGG	- -194 ~ -189
			YDR512C EMI1	TGCTGA	+ -762 ~ -757
				TGCTGC	+ -124 ~ -119
			YDR516C EMI2	TCAGCA	+ -691 ~ -686
				CCAGCC	- -527 ~ -522
			YOR317W FAA1	CCAGCT	- -624 ~ -619
				ACAGCA	- -293 ~ -288
			YCL040W GLK1	GCAGCA	- -659 ~ -654
				CCAGCA	- -630 ~ -625
				TGCTGC	- -549 ~ -544
				ACAGCA	+ -450 ~ -445
			YOR155C ISN1	ACAGCA	- -843 ~ -838
				CCAGCG	- -585 ~ -580
			YNL173C MDG1	CCAGCT	- -805 ~ -800
				TCAGCA	- -470 ~ -465
				GGCTGG	- -274 ~ -269
			YDR296W ‡ MHR1 ‡	AGCTGG	+ -929 ~ -924
				CCAGCA	+ -899 ~ -894
				CCAGCA	- -740 ~ -735
			YDR194C MSS116	AGCTGG	- -933 ~ -928

Table A1. Continued

Regulators	Matrices / Consensus sequences		Inferred target genes	Binding sites		Position in promoters
	Known target genes	Number of binding sites				
				TGCTGG	+	-717 ~ -712
				TGCTGA	+	-661 ~ -656
				CGCTGG	+	-317 ~ -312
				GGCTGG	+	-222 ~ -217
			YPL097W	MSY1	+	-508 ~ -503
				CCAGCA	+	-409 ~ -404
				TGCTGT	-	-259 ~ -254
				ACAGCA	+	-30 ~ -25
			YNL078W *	NIS1 *	-	-985 ~ -980
				CCAGCT	+	-783 ~ -778
				TGCTGG	-	-757 ~ -752
				TGCTGA	+	-438 ~ -433
				GGCTGG	+	-355 ~ -350
				ACAGCA	+	-273 ~ -268
				TGCTGC	-	-256 ~ -251
			YLR093C	NYV1	+	-488 ~ -483
				ACAGCA	+	-478 ~ -473
			YIL136W	OM45	-	-573 ~ -568
				CCAGCC	+	-561 ~ -556
				CCAGCT	-	-424 ~ -419
				AGCTGG	-	-422 ~ -417
				ACAGCA	+	-292 ~ -287
			YDL179W *	PCL9 *	-	-324 ~ -319
				CCAGCG	-	-298 ~ -293
				TGCTGC	-	-298 ~ -293
				CCAGCT	+	-280 ~ -275
				ACAGCA	-	-37 ~ -32
			YDR079W	PET100	-	-774 ~ -769
				CCAGCT	-	-669 ~ -664
				TGCTGA	-	-73 ~ -68
			YOR158W	PET123	+	-984 ~ -979
				TGCTGC	-	-756 ~ -751
				CCAGCA	+	-633 ~ -628
				TGCTGC	-	-547 ~ -542
				TGCTGA	+	-306 ~ -301
				TGCTGA	+	-41 ~ -36
			YJL078C	PRY3	+	-485 ~ -480
				TGCTGG	+	-476 ~ -471
				GGCTGG	+	-466 ~ -461
				CCAGCA	+	-212 ~ -207
				TGCTGC	-	-54 ~ -49
			YOR153W	PDR5	+	-798 ~ -793
				TCAGCA	-	-398 ~ -393
			YKR093W	PTR2	-	-584 ~ -579
				TGCTGC	+	-85 ~ -80
			YGR044C *	RME1 *	+	-666 ~ -661
				CCAGCA	-	-334 ~ -329
				TGCTGG	-	-286 ~ -281
				TCAGCA	+	-9 ~ -4
			YDR487C	RIB3	-	-889 ~ -884
				ACAGCA	-	-763 ~ -758
				CGCTGG	-	-348 ~ -343
			YOR018W	ROD1	+	-172 ~ -167
			YDR041W	RSM10	+	-395 ~ -390
				ACAGCA	-	-178 ~ -173
			YMR191W	SPG5	-	-959 ~ -954
				GGCTGG	+	-482 ~ -477
				GGCTGG	+	-478 ~ -473
				TCAGCA	+	-453 ~ -448
				TGCTGG	-	-341 ~ -336
				AGCTGG	+	-57 ~ -52
			YMR261C	TPS3	+	-502 ~ -497
				CCAGCC	+	-502 ~ -497
				TCAGCA	-	-212 ~ -207

Table A1. Continued

Regulators	Matrices / Consensus sequences		Inferred target genes	Binding sites	Position in promoters		
	Known target genes	Number of binding sites					
YAPI			YDR513W	TTR1	GCAGCA	-	-205 ~ -200
					CCAGCT	-	-950 ~ -945
					GCAGCA	-	-905 ~ -900
			YFR003C ‡	YPI1 ‡	ACCAGC	-	-590 ~ -585
						+	-697 ~ -692
			YNL046W *		CCAGCC	590	-691 ~ -686
					GCAGCA	~ -	-535 ~ -530
					TGCTGC	585-	-532 ~ -527
			YJL197W	UBP12	TGCTGG	+	-48 ~ -43
					CGCTGG	+	-169 ~ -164
					GCAGCA	-	-780 ~ -785
					GGCTGG	+	-607 ~ -602
					TGCTGT	+	-588 ~ -583
			YNL046W *		TGGTGG	+	-841 ~ -836
					TGGTGC	+	-702 ~ -697
					TGGTGC	+	-423 ~ -418
					TGGTGG	+	-218 ~ -213
			YMR040W		CCACCA	-	-944 ~ -939
					CCACCA	-	-908 ~ -903
					CCACCA	-	-343 ~ -338
					TGGTGT	+	-301 ~ -296
			YOR203W		TGGTGA	+	-918 ~ -913
			YDR271C		TCACCA	-	-863 ~ -858
					TGGTGA	+	-425 ~ -420
					ACACCA	-	-287 ~ -282
					CCACCA	-	-284 ~ -279
			YOL048C		GGGTGG	+	-923 ~ -918
		TGGTGT	+	-268 ~ -263			
		TGGTGT	+	-155 ~ -150			
YDR512C		TGCTGA	+	-305 ~ -300			
	YGR209C (TRX2)	1 (-)	YJR148W	BAT2	TTACTAA	+	-864 ~ -858
	YNL134C	2 (- +)	YDR304C	CYP5	TTACTAA	+	-798 ~ -792
			YPL091W *	GLR1 *	TTACTAA	-	-191 ~ -185
			YGR234W	YHB1	TTACTAA	+	-407 ~ -401
					TTACTAA	+	-397 ~ -391
			YBR054W	YRO2	TTACTAA	-	-545 ~ -539
			YLL049W		TTACTAA	-	-312 ~ -306

Note: The genes marked with * are the experimentally proven target genes of the given regulators. The genes marked with ‡ are those that encode the protein of the same category as known target gene does. A plus sign means the binding site is present on the sense strand while a minus sign indicates that the binding site is present on the minus strand.

We extend our thanks to MSI staff Richard Flesvig and Haoyu Yu for their help during the development of this work.

References

- Ares M Jr, Mangin M, Weiner AM. 1985. Orientation-dependent transcriptional activator upstream of a human U2 snRNA gene. *Mol Cell Biol* 5(7):1560-1570.
- Avrava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 100:3889-3894.
- Brazma A, Jonassen I, Vilo J, Ukkonen E. 1998. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res* 8:1202-1215.
- Buchman AR, Kornberg RD. 1990. A yeast ARS-binding protein activates transcription synergistically in combination with other weak activating factors. *Mol Cell Biol* 10(3):887-897.
- Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* 2:65-73.
- Cosma MP, Tanaka T, Nasmyth K. 1999. Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle-and developmentally regulated promoter. *Cell* 97:299-311.
- Costanzo MC, Hogan JD, Cusick ME, Davis BP, Fancher AM, Hodges PE, Kondu P, Lengieza C, Lew-Smith JE, Lingner C, Roberg-Perez KJ, Tillberg M, Brooks JE, Garrels JI. 2000. The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): Comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res* 28(1):73-76.
- de la Fuente A, Brazhnik P, Mendes P. 2002. Linking the genes: Inferring quantitative gene networks from microarray data. *Trends Gen* 18(8):395-398.
- Dolinski K, Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Sethuraman A, Theesfeld CL, Binkley G, Lane C, Schroeder M, Dong S, Weng S, Andrada R, Botstein D, Cherry JM. 2002. *Saccharomyces* Genome Database. Available at <http://www.yeastgenome.org/>.

- Donaldson AD, Kilmartin JV. 1996. Spc42p: A phosphorylated component of the *S. cerevisiae* spindle pole body (SPB) with an essential function during SPB duplication. *J Cell Biol* 132(5):887–901.
- Fantoni A, Dare AO, Tschudi C. 1994. RNA polymerase III-mediated transcription of trypanosome U2 small nuclear RNA gene is controlled by both intragenic and extragenic regulatory elements. *Mol Cell Biol* 14(3):2021–2028.
- Filkov V, Skienan S, Zhi J. 2002. Analysis techniques for microarray time-series data. *J Comp Bio* 9(2):317–330.
- Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian network to analyze expression data. In Proceedings of the Fourth Annual Conference on Research in Computational Molecular Biology (RECOMB) '00. Tokyo: RECOMB. p 127–135.
- Frith MC, Spouge JL, Hansen U, Weng Z. 2002. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res* 30(14):3214–3224.
- Halfter H, Muller U, Winnacker EL, Gallwitz D. 1989. Isolation and DNA-binding characteristics of a protein involved in transcription activation of two divergently transcribed, essential yeast genes. *EMBO J* 8(10):3029–3037.
- Hope IA, Struhl K. 1986. Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. *Cell* 46:885–894.
- Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. 2002. Revealing modular organization in the yeast transcriptional network. *Nature Genetics* 31:370–377.
- Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–254.
- Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K. 1993. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science* 261(5128):1543–1544.
- Kratzer S, Schuller HJ. 1997. Transcriptional control of the yeast acetyl-CoA synthetase gene, ACS1, by the positive regulators CAT8 and ADR1 and the pleiotropic repressor UME6. *Mol Microbiol* 26(4):631–641.
- Kwon AT, Hoos HH, Ng R. 2003. Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics* 19:905–912.
- Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298(5594):799–804.
- Lieb J, Liu X, Botstein D, Brown PO. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genet* 28:327–334.
- Liu F, Bateman E. 1993. An upstream promoter element of the *Acanthamoeba castellanii* TBP gene binds a DNA sequence specific transcription activating protein, TPBF. *Nucleic Acids Res* 21(18):4321–4329.
- Liu X, Clarke ND. 2002. Rationalization of gene regulation by Eukaryotic transcription factor: Calculation of regulatory region occupancy from predicted binding affinities. *J Mol Biol* 323:1–8.
- Liu XS, Brutlag DL, Liu JS. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotech* 20(8):835–839.
- Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 31(1):374–378.
- Mazumder B, Seshadri V, Fox PL. 2003. Translational control by the 3'-UTR: the ends specify the means. *Trends Biochem Sci* 28(2):91–98.
- Näär AM, Lemon BD, Tjian R. 2001. Transcriptional coactivator complexes. *Ann Rev Biochem* 70:475–501.
- Nakaar V, Dare AO, Hong D, Ullu E, Tschudi C. 1994. Upstream tRNA genes are essential for expression of small nuclear and cytoplasmic RNA genes in trypanosomes. *Mol Cell Biol* 14(10):6736–6724.
- Nasmyth K, Adolf G, Lydall D, Seddon A. 1990. The identification of a second cell cycle control on the HO promoter in yeast: Cell cycle regulation of SWI5 nuclear entry. *Cell* 62:631–647.
- Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, Hinnebusch AG, Marton MJ. 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.
- Natesan S, Gilman M. 1993. DNA bending and orientation-dependent function of YY1 in the c-fos promoter. *Genes Dev* 7:2497–2503.
- Pilpel Y, Sudarsanam P, Church GM. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* 29(10):153–159.
- Porter SE, Washburn TM, Chang M, Jaehning JA. 2002. The yeast paf1-rRNA polymerase II complex is required for full expression of a subset of cell cycle-regulated genes. *Eukaryot Cell* 1(5):830–842.
- Rosenfeld R, Alon U. 2003. Response delays and the structure of transcription networks. *J Mol Biol* 329:645–654.
- Sinha S, Tompa M. 2002. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 30:5549–5560.
- Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA. 2001. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* 106(6):697–708.
- Sobel SG, Snyder M. 1995. A highly divergent gamma-tubulin gene is essential for cell growth and proper microtubule organization in *Saccharomyces cerevisiae*. *J Cell Biol* 131(6 Part 2):1775–1788.
- Solan Z, Horn D, Ruppin E, Edelman S, Lapidot M, Kaplan S, Garten Y, Pilpel Y. 2004. Motif extraction from promoter regions of *S. cerevisiae*. Available at <http://neuron.tau.ac.il/~horn/publications/csb4.pdf>.
- Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16:16–23.
- Svetlov VV, Cooper TG. 1995. Review: Compilation and characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. *Yeast* 11(15):1439–1484.
- Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze B, Moreau Y. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17:1113–1122.
- Tonozuka H, Wang J, Mitsui K, Saito T, Hamada Y, Tsurugi K. 2001. Analysis of the upstream regulatory region of the GST1 gene required for its Oscillatory expression. *J Biochem* 130:589–595.
- Ueda T, Wang Z, Pham N, Messing J. 1994. Identification of a transcriptional activator-binding element in the 27-kilodalton zein promoter, the -300 element. *Mol Cell Biol* 14(7):4350–4359.
- van Helden J, André B, Collado-vides J. 1998. Extracting regulatory site from upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 281:827–842.
- Wood TI, Griffith KL, Fawcett WP, Jair KW, Schneider TD, Wolf RE Jr. 1999. Interdependence of the position and orientation of SoxS binding sites in the transcriptional activation of the class I subset of *Escherichia coli* superoxide-inducible promoters. *Mol Microbiol* 34(3):414–430.
- Wu F, Lee AS. 1998. Identification of AP-2 as an interactive target of Rb and a regulator of the G1/S control element of the hamster histone H3.2 promoter. *Nucleic Acids Res* 26(21):4837–4845.
- Xu H, Wu P, Wu CFJ, Tidwell C, Wang Y. 2002. A smooth response surface algorithm for constructing a gene regulatory network. *Physiol Genomics* 11:11–20.
- Zhu J, Zhang MQ. 1999. SCPD: A promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* 15(7–8):607–611.