ARTICLE

BIOTECHNOLOGY and BIOENGINEERING

Computational Identification of Combinatorial Regulation and Transcription Factor Binding Sites

Taewoo Ryu,[1] Younghoon Kim,[1] Dae-Won Kim,[2] Doheon Lee[1]

[1]Department of BioSystems, Korea Advanced Institute of Science and Technology, 373-1, Guseong-dong, Yuseong-gu, Daejeon, 305-701, South Korea; telephone: +82-42-869-4316; fax: +82-42-869-4310; e-mail: dhlee@biosoft.kaist.ac.kr

[2]School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea

Received 12 June 2006; accepted 16 January 2007

Published online 24 January 2007 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bit.21354

ABSTRACT: A number of computational methods have been used to unravel the core mechanisms governing the regulation of gene expression, but these techniques examine only portions of the genetic regulatory mechanism. For example, some studies have failed to include the combined action of multiple transcription factors (TFs) or the importance of TF binding constraints (i.e., the binding position and orientation), while others have examined combinations of only 2 or 3 TFs. Thus, we sought to develop a new method for identifying regulatory modules in yeast, using an algorithm that includes all combinations of TFs plus a number of binding constraints when identifying target genes. We successfully developed a computational method for using microarray and TF–DNA interaction data to identify regulatory modules. All possible combinations of yeast TFs and various binding constraints were tested to identify regulatory modules. Within the identified modules, target genes were found to have common binding constraints such as fixed binding regions and orientations for each TF. Moreover, targets showed similar mRNA expression profiles and high functional coherence. Our novel approach, which accounts for both combined actions of TFs and their binding constraints, can be used to identify target genes and reliably predict regulatory modules over a broad range of functional categories. Complete results and additional information are available online at http://bisl.kaist.ac.kr/~dhlee/comModule/index.html.

Biotechnol. Bioeng. 2007;97: 1594–1602.

KEYWORDS: transcription factor; *Saccharomyces cerevisiae*; combinatorial regulation; *cis*-element

Introduction

Specific spatiotemporal expression of genes and proteins is essential in living organisms. Critical control of these activities is overseen by the transcription factors (TFs), which bind to specific DNA sequences and up- or down-regulate nearby genes. These TF-binding DNA sequences, called *cis*-elements or motifs, form the link between TFs and their target genes, and have been identified using a variety of experimental approaches. For example, conserved motifs have been identified by multiple alignment of coexpressed genes (Hughes et al., 2000; Tavazoie et al., 1999), and linear regression has been used to model the associations between mRNA expression levels and the abundance of specific motifs (Bussemaker et al., 2001; Conlon et al., 2003; Keles et al., 2002; Phuong et al., 2004). Microarray-based techniques such as ChIP-chip and damID have also been used for genome-wide in vivo mapping of TFs (Lee et al., 2002; van Steensel et al., 2001).

Researchers have also attempted to identify global regulatory networks, especially in yeast. These studies have involved the prediction of relationships between TFs and their target genes (Gao et al., 2004; Qian et al., 2003), as well as assessment of combinatorial TF regulation, in which multiple TFs cooperate to regulate gene sets in response to diverse environmental signals. Researchers have developed computational motif discovery algorithms capable of identifying TFs that act together in combinatorial regulation (Banerjee and Zhang, 2003; Beer and Tavazoie, 2004; Hvidsten et al., 2005; Pilpel et al., 2001; Segal et al., 2003). However, these methods have been limited by an inability to precisely link the TFs to their motifs and targets. Other methods have used raw ChIP-chip data to directly link TFs and target genes for the identification of TF regulatory

Correspondence to: D. Lee

Contract grant sponsor: Basic Research Program of KOSEF
Contract grant number: R01-2005-000-10266-0

networks (Bar-Joseph et al., 2003; Harbison et al., 2004; Kato et al., 2004; Yu et al., 2003).

As the physical interactions of TFs with other proteins such as polymerases and/or other TFs often require proper orientation on the DNA strand, recent studies have focused on identifying TF binding constraints and using them to select target genes. Beer and Tavazoie (2004) identified multiple constraints for specific functional groups of TFs using clustering-based motif discovery and Bayesian learning. However, they were unable to precisely infer the relationship between TFs and corresponding motifs, because computationally identified motifs were used. In addition, clustering cannot group all genes with the same motif, as noted by Bussemaker et al. (2001). These limitations would seem to suggest that the methods of Beer and Tavazoie (2004) are likely to miss potential modules. In another study, Zhu et al. (2005) sought to identify human regulatory modules by accounting for TF combinations and some binding constraints. However, the authors examined only two TFs in one module, and the utilized constraints were limited to the proximity of two *cis*-elements.

We herein report the development of a novel computational method that identifies regulatory modules by checking all possible combinations of TFs and examining 75 binding constraints for each TF. Our strategy integrates information regarding TF combinations, *cis*-elements, binding constraints and target genes into regulatory modules. TFs may both activate and repress target genes; our algorithm covers both cases because global expression patterns are analyzed under various experimental conditions, allowing identification of activated or repressed target genes. When this algorithm was tested in budding yeast using cell cycle data (Spellman et al., 1998) and environmental stress data (Gasch et al., 2000), we successfully identified a number of modules, most of which contained 2–5 TFs. Many genes with common motifs and constraints showed functional relatedness and high correlations in their microarray expression profiles. These results reveal that simultaneous consideration of TF combinations and binding constraints can yield significantly better identification of regulatory modules versus assessment based on each factor alone.

Methods

Data Preparation

Known TF binding sites were obtained from TRANSFAC 8.2 and ChIP-chip experiments. Among these, 98 TFs listed in the TRANSFAC and Lee et al. (2002) were used for the analysis. Microarray data from Gasch et al. (2000) and Spellman et al. (1998) were used for the yeast gene expression profiles. For the functional annotation of modules, MIPS FunCat Scheme 2.0 was downloaded from the MIPS website.

Motif Assignment

Chromosomal DNA sequences and gene annotation data were downloaded from the *Saccharomyces* Genome Database (Christie et al., 2004). For all yeast genes, upstream sequences from −1,000 to 100 relative to the translation start site were extracted. TF binding sites were assigned to each gene in position weight matrix (pwm) or consensus sequence format. For pwm, the search mechanism used in Kel et al. (2003) was applied with 80% similarity score threshold setting. For consensus sequence, an 85% similarity score threshold was applied.

Identifying Genetic Regulatory Modules

To define the TF binding constraints, which are crucial in our algorithm for module identification, we first divided each upstream sequence into overlapping windows, with denser overlaps near the translation start site (TSS). Starting from the +100 bp position, the first 50 bp (i.e., from +100 to +50) were set to the first window, and subsequent 50 bp windows were applied at 25 bp intervals. For example, the second window covered +75 to +25, the third +50 to 0, and so on until the final window covered −100 to −150. Beginning at −150, the window size was expanded to 100 bp and the interval was expanded to 50 bp until all sequences out to −1,000 bp were covered by overlapping windows. Each region was additionally constrained by an orientation constraint (i.e., $5' \rightarrow 3'$, $3' \rightarrow 5'$ or both directions). The 25 divided regions and 3 orientation constraints yielded 75 distinct constraints for each TF.

Our method has expanded previous research in the sense that it allows a large number of TF combinations and constraints in the regulatory module. The number of possible TFs in a module is unlimited as long as the modules meet the criterion as described below. For all combinations of TF pairs and constraints, those with more than three target genes and average pairwise Pearson correlation coefficients above 0.5 for the cell cycle data (Spellman et al., 1998) and 0.71 for the stress data (Gasch et al., 2000) (corresponding to the top 1 percentile in the distribution of all gene pairs in the microarray data) were selected as modules. We required pairs of TFs instead of single TFs because the latter did not yield significant results (see Table II), and we were interested in combinatorial regulation by multiple TFs. Once a module was initially established, it was then separately tested with each remaining TF in the dataset. Based on the assumption that combinatorial regulation with more TFs will lead to the more synchronized expression of target genes over the broad experimental conditions, each new module (containing the new TF) was approved if the average correlation of target genes was larger than that of the previous module. In this way, all possible combinations of TFs were examined, and constraints and targets were selected together. From the final result, redundant modules that share overlapping binding constraints with the same set of TFs and targets are

consolidated, and overall binding constraints are identified. Because our algorithm starts with TF pairs and tests additional TFs only when the module satisfies the standard, it does not search unnecessary combinations, which means the search space is reduced greatly and regulatory modules are found efficiently.

Finally, we used the functional annotations from the MIPS database to assess the functional coherence of the TFs and target genes in the discovered modules.

Determination of Statistical Significance of TF Binding

To obtain the significance of each TF combination, we first determined the binding probability for each TF. Hypergeometric distribution was used to calculate the probability of TF binding to the observed or greater number of genes by chance. The probability for each TF is given by:

$$\mathrm{HG}(t; G, g, T) = \frac{\binom{T}{t}\binom{G-T}{g-t}}{\binom{G}{g}}$$

where G is the total number of genes in the microarray, g is the number of target genes in the module, T is the total number of genes bound by each TF with the binding constraint identified in the module, and t is the number of genes bound by the TFs in the module. These probabilities were then multiplied for each TF combination; we assumed an independent relationship because each TF may bind DNA independently, as discussed below. A significance level of 0.01 was used as the cutoff, and the value was divided by the number of TFs, and then divided by the number of binding constraints in our experiment for Bonferroni correction. Thus, P-value lower than 1.36E-6 was considered significant.

To test the significance of each module, we conducted random modeling in which sets of TFs, binding constraints and genes were randomly organized into modules, and tested each module for significance. First, we randomly selected a number of regulators (N) from 2 to 7, because the discovered modules all had between 2 and 7 TFs (Table IV). Then, N TFs and N binding constraints were randomly picked. Genes with promoters containing binding sites for the selected TFs and having the proper constraints were retrieved and average pairwise Pearson correlation coefficients were calculated from the generated modules. The entire module generation process was repeated 10,000 times for each microarray dataset, and we calculated the fraction of significant modules in which target genes showed correlations higher than the threshold value (e.g., 0.5 for the cell cycle data and 0.71 for the stress data).

Results

Identifying Regulatory Modules

Previously, researchers have used computational methods to identify novel transcription-related motifs, and have then compared these motifs with known TF binding sites to predict candidate regulators (Banerjee and Zhang, 2003; Beer and Tavazoie, 2004; Bussemaker et al., 2001; Hvidsten et al., 2005). However, these methods cannot precisely link the TFs to their motifs, and often do not account for DNA binding constraints vital to the proper regulation of TFs. We herein report a new computational method for identifying genetic regulatory modules using an algorithm that accounts for all TF combinations, as well as a number of binding constraints and target genes (Fig. 1). We utilized data from ChIP-chip experiments (Lee et al., 2002) and yeast TF binding site information from the TRANSFAC database

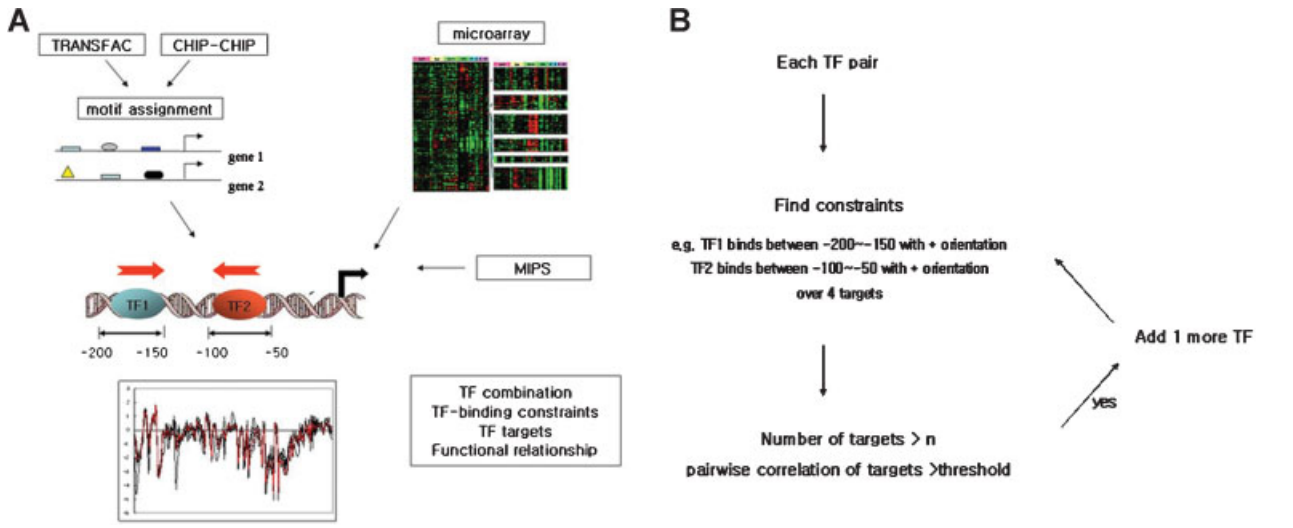


Figure 1. An overview of the method. **A**: Schematic diagram of method. **B**: Algorithm used for discovering combinatorial regulatory modules.

DOI 10.1002/bit

(Matys et al., 2003), and iteratively scanned the upstream sequences of all yeast genes for TF binding sites (see Methods Section and Fig. 1A).

As TFs must be properly oriented on the DNA strand in order to properly interact with the target gene sequences and other proteins such as polymerases and/or other TFs, we then sought to account for binding site constraints, including the position of the motif relative to the translation start site (ATG) and the orientation of the motif. We divided the upstream sequence of each gene into 25 overlapping regions, with denser overlaps near the translation start site (TSS). Each region could be defined as having one of three orientations, namely $5' \rightarrow 3'$, $3' \rightarrow 5'$, or both. The $5'' \rightarrow 3'$ orientation means that all target genes in the module have TF binding sites on the sense strand, while $3' \rightarrow 5'$ oriented regions have their binding sites on the antisense strand. The third category, "both orientations," means that target genes are not limited by orientation and the TFBS position is the only constraint.

As shown in Figure 1b, we analyzed all possible TF pairs, accounting for 75 different binding constraints for each TF, and selected pairs with three or more target genes having a average correlation above 0.5 for the cell cycle data (Spellman et al., 1998) and 0.71 for the stress data (Gasch et al., 2000). This threshold corresponds to the top 1 percentile of all gene pairs in the microarray data, and is higher than that previously used in Banerjee and Zhang (2003) and Zhu et al. (2005). We then iteratively tested every other TF in the dataset, looking to see if the addition of the new TF improved the correlation of the existing module. Thus, our novel method was capable of considering all possible combination of TFs and selecting constraints and targets together. Moreover, because we used *cis*-elements with known binding TFs, this work allowed a more precise definition of TF-target gene relationships than that seen in the previous studies. Application of this method to yeast data allowed identification of 4,158 and 949 putative regulatory modules containing 163 and 120 target genes from the cell cycle and stress datasets, respectively (see Fig. 2 for representative examples). Comparison of our modules with annotated data from MIPS (Mewes et al., 2004) revealed that our modules showed high coherences in their functional categories.

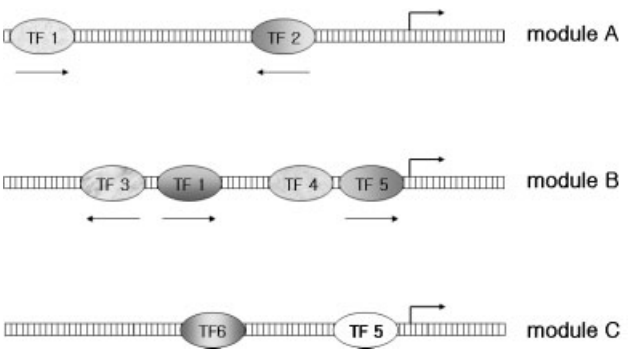


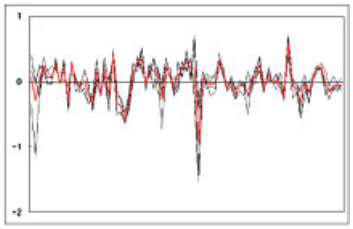
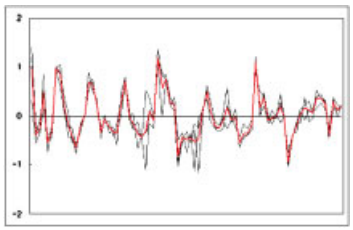
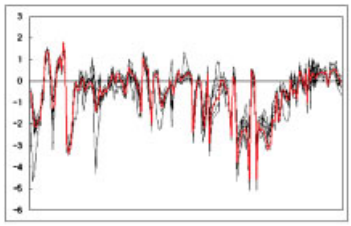
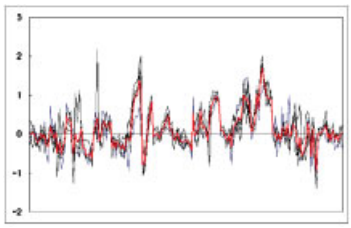
Figure 2. Characteristics of discovered modules. Modules may include 2 TFs (module **A**), or almost any other combination of TFs (module **B**), including the binding of 2 TFs in both orientations (module **C**). The specific binding positions for each TF and the orientations of the TF binding sites are also included in the modules.

For further study, we chose four identified modules of various sizes having high functional coherence (Table I), two from the cell cycle dataset and two from the yeast stress dataset. Within all of these modules, the target genes tended to show similar expression patterns. Moreover, the TFs and target genes showed functional relatedness. However, it should be noted that identification of combinatorial regulation does not necessarily imply that the TFs undergo physical interactions with each other (Harbison et al., 2004); the involved TFs may physically interact to co-regulate target gene expression, or they may bind separately to the same target under different conditions.

The first representative module was obtained from the cell cycle dataset (Spellman et al., 1998) and contained two TFs (CBF1 and MET31) known to play regulatory roles in the cell cycle. Previous studies by Kato et al. (2004) and Tavazoie et al. (1999) identified these TFs as being involved in combinatorial regulation. Our analysis indicated that CBF1 binds all of the identified target gene promoters between −400 and −250 from the translation start site (ATG); this is consistent with the CBF1 binding region reported in the TRANSFAC database (Matys et al., 2003) (see our website). The target genes in this module include four genes encoding L-asparaginase II (ASP3-1, ASP3-2, ASP3-3, and ASP3-4), which are spread out over thousands or tens of thousands of base pairs on chromosome 12. According to the MIPS categories, these genes encode proteins involved with aspartate metabolism, nitrogen and sulfur utilization, and stress responses. The two TFs in this module, CBF1 and MET31, also play roles in the metabolisms of amino acids, nitrogen and sulfur. The four target genes showed similar expression patterns, with an average pairwise correlation of 0.6. Thus, this module showed a high coherence in terms of functionality and expression.

The second module, which was also obtained from the cell cycle dataset, consisted of four TFs (MCM1, STE12, MBP1, and PDR3) and three target genes (YRF1-3, YRF1-6, and YRF1-7). MCM1 and STE12 were previously shown to be involved in combinatorial regulation of cell cycle control (Harbison et al., 2004; Pilpel et al., 2001; Spellman et al., 1998). MBP1 also plays a cell cycle-related role (Kato et al., 2004; Yu et al., 2003), and was recently shown to be involved in combinatorial regulation with MCM1 (Nagamine et al., 2005). PDR3 has not been previously associated with combinatorial regulation or cell cycle control. The binding regions identified in our analysis for MCM1 and PDR3, −500 to −350 and −100 to −25, respectively, overlap with the known binding regions in the TRANSFAC database. The three target genes, YRF1-3, YRF1-6, and YRF1-7, encode members of the Y'-helicase protein 1 family, which fall into the MIPS categories of DNA synthesis and replication during the cell cycle. The expression levels of these three genes fluctuate together during the cell cycle (Table I),

Table I. Examples of discovered module.

TF (constraints)	Expression pattern	Target genes (correlation)	Module function
CBF1 (−400:−250/↔) MET31 (−1,000:−850/+)		ASP3-1 ASP3-2 ASP3-3 ASP3-4 (0.6)	Amino acid metabolism Nitrogen and sulfur metabolism Stress response
MCM1 (−500:−350/−) STE12 (−400:75/+) MBP1 (−125:−75/+) PDR3 (−100:−25/+)		YRF1-3 YRF1-6 YRF1-7 (0.79)	DNA processing Stress response Cell cycle
FKH1 (50:100/+) RAP1 (−450:−350/+)		RPL19B ENP1 RPL23B RPL26B RPL42B RPL39 RPS4A (0.878)	Ribosome biogenesis Cell cycle DNA processing RNA processing
MSN2 (−400:−250/−) CAT8 (−150:−75/−)		PAU1 PAU3 PAU6 YIL176C YLL064C (0.724)	Stress response

TF combinations with binding constraints are listed on the left side of the graph. The "+" designates the 5′ → 3′ direction, the "−" designates the 3′ → 5′ direction and the "↔" designates both orientations. Expression profiles for the target genes in the modules are displayed, with average target expression indicated with a red line. The y-axis is the log2 ratio of the expression value and the x-axis represents the experimental conditions of the microarray data. Target genes, their average pairwise correlation values and representative module functions are listed on the right side of the graph.

yielding a high correlation of 0.79. No previous work has shown a direct connection between the TFs and target genes in this module. However, MCM1 and STE12 have been shown to regulate other yeast helicases (Davis et al., 1992; Fitch et al., 2003), suggesting that these TFs may regulate the transcription of YRF1-3, YRF1-6, and YRF1-7 in vivo. The binding constraints identified for STE12 and MBP1 in our analysis did not match those in the TRANSFAC database. However, this does not necessarily indicate that these are false positive results, because the TF binding site databases are unlikely to contain all possible binding sites for all TFs. Future work will be required to confirm the binding sites newly identified in this work.

The third representative module was obtained from the yeast stress data (Gasch et al., 2000) and contained two TFs, FKH1 and RAP1, and seven target genes encoding seven ribosomal or ribosome-related proteins. FKH1 and RAP1 were previously shown to cooperate to regulate the ribosomal protein, RPL31B (Hvidsten et al., 2005). According to our analysis, RAP1 binds between −450 and −350, in the + orientation. This binding region overlaps with the known sites in the TRANSFAC database. Six of the seven target genes are ribosomal proteins, while the other target, ENP1 is involved in 20S pre-rRNA processing (Lai et al., 2005), giving this module a high functional coherence and a correlation of 0.878. FKH1 is involved in the stress

DOI 10.1002/bit

Table II. Evaluation of the proposed method.

TF	Binding constraints	Number of target genes	Average correlation
CBF1	Presence	3,856	0.03
	−400:−250/↔	878	0.027
MET31	Presence	235	0.026
	−1,000:−850/+	28	0.04
CBF1—MET31	Presence	220	0.025
	−400:−250/↔ and −1,000:−850/+	**4**	**0.6**
MCM1	Presence	4,220	0.03
	−500:−350/−	499	0.034
STE12	Presence	6,005	0.031
	−400:75/+	5,597	0.031
MBP1	Presence	2,125	0.028
	−125:−75/+	152	0.035
PDR3	Presence	1,689	0.029
	−100:−25/+	57	0.03
MCM1-STE12-MBP1-PDR3	Presence	455	0.027
	−500:−350/− and −400:75/+ and −125:−75/+ and −100:−25/+	**3**	**0.79**
FKH1	Presence	3,805	0.022
	50:100/+	199	0.027
RAP1	Presence	936	0.017
	−450:−350/+	108	0.053
FKH1—RAP1	Presence	720	0.019
	50:100/+ and −450:−350/+	**7**	**0.878**
MSN2	Presence	3,317	0.022
	−400:−250/−	332	0.031
CAT8	Presence	1,319	0.022
	−150:−75/−	41	0.017
MSN2—CAT8	Presence	880	0.026
	−400:−250/− and −150:−75/−	**5**	**0.724**

Comparison of the four modules shown in Table 1 (bold) versus several other modules with different constraints shows that both combinatorial TF regulation and binding constraints contribute to specific target selection. 'Presence' indicates that at least one TF binding site was identified upstream of the target gene. Modules with different constraints have large numbers of target genes and low average Pearson correlation scores.

response (Shapira et al., 2004) and RAP1 is known to bind upstream of ribosomal protein genes (Klein and Struhl, 1994; Miyoshi et al., 2003; Moehle and Hinnebusch, 1991). Ribosomal protein expression has been shown to change during multiple stress responses (Gasch et al., 2000), indicating the functional relatedness of this module.

The fourth representative module, also obtained from the stress data, consisted of two TFs (MSN2 and CAT8) and five target genes (PAU1, PAU3, PAU6, YIL176C, and YLL064C). Of these, MSN2, PAU1, PAU3, and PAU6 fall into the MIPS category of stress response. Although CAT8 is not contained within the stress response MIPS category, it has been reported to play a role in the response to nutrient stress (Tachibana et al., 2005), giving this module a high functional coherence. MSN2 and CAT8 are known to be involved in gene activation or repression under a variety of growth conditions (Kim and Iyer, 2004), but this is the first indication that they may be involved in combinatorial regulation. The binding interval identified for MSN2, between −400 and −250, is consistent with that in the TRANSFAC database. The target genes in this module included three PAU-protein family members (PAU1, PAU3, and PAU6) and two novel genes (YIL176C and YLL064C) that show strong sequence similarity to members of the Srp1/Tip1p family of cold and heat shock-induced PAU-family mannoproteins (Bourdineaud, 2000; Kwast et al., 2002). Thus, all target genes in this module are members of the PAU family and are thought to be closely related to the stress response, yielding a high correlation value of 0.724. One of the targets, PAU6, was previously shown to be regulated by MSN2 (Bruckmann et al., 2004), providing additional evidence for the veracity of the identified regulatory module.

In sum, our novel method allowed identification of many putative regulatory modules, and analysis of four representative modules identified both known and novel TF combinations and TF-target interactions. The identified binding constraints of many TFs overlapped with the known constraints in spite of a lack of sufficient information. Our results also revealed high target gene correlation and functional coherence within each module. A list of all discovered modules and additional related data are available on our website (http://bisl.kaist.ac.kr/∼dhlee/comModule/index.html).

Evaluation of the Discovered Modules

To evaluate the effectiveness of our method, we repeated our analysis with individual constrained TFs or unconstrained TF combinations (Table II). When we constrained our analysis for the presence of at least one TF binding site

Table III. Significance of combinatorial TF binding.

TF combination	Binding constraints	*P*-value
CBF1—MET31	−400:−250/↔ and −1,000:−850/+	1.45E-13
MCM1-STE12-MBP1-PDR3	−500:−350/− and −400:75/+ and −125:−75/+ and −100:−25/+	4.64E-15
FKH1-RAP1	50:100/+ and −450:−350/+	2.01E-22
MSN2—CAT8	−400:−250/− and −150:−75/−	3.03E-17

To evaluate the statistical significance of our results, we calculated *P* values for each combination listed in Table 1. All TF combinations placed within modules show *P* values <1.36E-6, which is the threshold for significance of the Bonferroni-corrected *P* values.

upstream of a target gene, we obtained modules having hundreds or thousands of target genes despite the use of stringent motif assignment thresholds designed to reduce inclusion of false binding sites. In addition, the pairwise Pearson correlation coefficients of these target genes were generally low. For example, 3,805 yeast genes were identified as having FKH1 binding sites, and these genes showed an average pairwise correlation of 0.022. Similarly, the use of unconstrained TF combinations in the analysis yielded large numbers of target genes having low correlations. Using the binding constraints for a single TF also failed to identify significant modules; the use of additional constraints reduced the number of targets, but the average correlation was low. These results collectively show that only modules identified using TF combinations and appropriate binding constraints showed high target gene correlation, indicating that these factors are vital to the proper identification of functional binding sites and target genes.

To confirm the significance of our results, *P* values of combinatorial binding were calculated for the discovered modules (Table III). The probability of TFs randomly binding to the same or greater number of genes than observed was calculated using hypergeometrical distribution. The resulting *P* values were lower than the Bonferroni-corrected significance level, 1.36E-6. The *P* values of each module may be accessed through our website.

The significance of modules was further tested by random modeling. To ensure that the discovered modules were more significant than those obtained by random chance, 10,000 modules were randomly generated for each microarray dataset (see Methods Section) and the average correlation of target genes in each module was compared to the abovementioned threshold. We obtained 1 significant module from each dataset, indicating that there is a <0.0001 probability that the utilized TFs, constraints and target genes could be organized into a single significant module by random chance.

Table IV. The statistics of discovered modules.

Data	Number of TFs within the module	Number of modules	Percentage (%)
Spellman et al. (1998)	2	324	7.79
	3	1,241	29.85
	4	1,411	33.93
	5	893	21.48
	6	267	6.42
	7	22	0.53
Gasch et al. (2000)	2	284	29.92
	3	472	49.74
	4	149	15.70
	5	44	4.64

Regulatory modules were analyzed based on the number of TFs within modules. Most of the discovered modules from either dataset consisted of 2–5 TFs.

TF Pairs in the Discovered Modules

To examine how many TFs are required for combinatorial regulation, we analyzed TF combinations from all discovered modules. Table IV shows the distribution of regulatory modules containing different numbers of TFs. Our finding that the majority of modules required 2–5 TFs in both datasets is consistent with the reports of Hvidsten et al. (2005). Other groups also obtained modules containing only a single TF. However, as our findings and the work of Hvidsten et al. (2005) suggest that modules with single TFs and high target gene correlation are rare, and our research was focused on combinatorial regulation, we did not examine single-TF modules in the present work.

Discussion and Conclusion

The identification of regulatory modules is an early step toward elucidating whole-cell signaling networks. As a relatively small number of TFs are responsible for controlling the entirety of gene expression, it is likely that they act in various combinations to provide a high level of functional diversity. Here, we analyzed these combinations, along with binding constraints and target genes, in an effort to better understand the precise mechanisms of TF regulation.

Our novel method for identifying regulatory modules is more accurate than the previously reported methods because we used TFs with known binding sites, providing a firm link between TFs, *cis*-element and targets. In addition, we considered the position and orientation of the binding sites, and searched all possible TF-binding site combinations. The modules identified using our method could not be identified using either parameter alone. Our TF combinations might represent physical interactions between TFs, competitive DNA binding, or conditional DNA binding. Future work with larger pools of TF interaction data will be required to distinguish among these possibilities.

DOI 10.1002/bit

Even though we herein present representative modules with some discussion of previous reports and functional annotation, further experimental evaluation will be required to confirm these findings. In future studies, the TF sets may be functionally analyzed by their co-expression along with appropriate target genes and/or by protein–protein interaction assays (e.g., co-immunoprecipitation) under different conditions. In addition, promoter deletion assays or perturbation of TF expression levels in a reporter system could be used to further confirm the relationship between a given TF and its target(s).

We selected budding yeast as a model organism, based on the availability of microarray and binding site information. However, our method should be easily applicable to higher eukaryotes and other experimental conditions, provided that sufficient binding site information is available. In the future, it should be possible to add more complex features of genetic regulatory networks to this analysis, such as activation, inhibition, feedback loops and regulatory/signaling proteins.

This work was supported by grant R01-2005-000-10266-0 from the Basic Research Program of KOSEF. We would like to thank CHUNG MoonSoul Center for BioInformation and BioElectronics and the IBM-SUR program for providing research and computing facilities.

References

Banerjee N, Zhang MQ. 2003. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. Nucleic Acids Res 31:7024–7031.

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi N, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. 2003. Computational discovery of gene modules and regulatory networks. Nat Biotechnol 21:1337–1342.

Beer MA, Tavazoie S. 2004. Predicting gene expression from sequence. Cell 117:185–198.

Bourdineaud JP. 2000. At acidic pH, the diminished hypoxic expression of the SRP1/TIR1 yeast gene depends on the GPA2-cAMP and HOG pathways. Res Microbiol 151:43–52.

Bruckmann A, Steensma HY, de Mattos MJT, van Heusden GPH. 2004. Regulation of transcription by Saccharomyces cerevisiae 14-3-3 proteins. Biochem J 382:867–875.

Bussemaker HJ, Li H, Siggia ED. 2001. Regulatory element detection using correlation with expression. Nat Genet 27:167–171.

Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, et al. 2004. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. Nucleic Acids Res 32:D311–D314.

Conlon EM, Liu XS, Lieb JD, Liu JS. 2003. Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci 100:3339–3344.

Davis JL, Kunisawa R, Thorner J. 1992. A Presumptive Helicase (MOTi Gene Product) Affects gene expression and is required for viability in the yeast Saccharomyces cerevisiae. Mol Cell Biol 12:1879–1892.

Fitch MJ, Donato JJ, Tye BK. 2003. Mcm7, a subunit of the presumptive MCM helicase, modulates its own expression in conjunction with Mcm1. J Biol Chem 278:25408–25416.

Gao F, Foat BC, Bussemaker HJ. 2004. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics 5:31.

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

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature 431:99–104.

Hughes JD, Estep PW, Tavazoie S, Church GM. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol 296:1205–1214.

Hvidsten TR, Wilczynski B, Kryshtafovych A, Tiuryn J, Komorowski J, Fidelis K. 2005. Discovering regulatory binding-site modules using rule-based learning. Genome Res 15:856–866.

Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ. 2004. Identifying combinatorial regulation of transcription factors and binding motifs. Genome Biol 5:R56.

Kel AE, Gobling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31:3576–3579.

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

Kim J, Iyer VR. 2004. Global role of TATA box-binding protein recruitment to promoters in mediating gene expression profiles. Mol Cell Biol 24:8104–8112.

Klein C, Struhl K. 1994. Protein kinase A mediates growth-regulated expression of yeast ribosomal protein genes by modulating RAP1 transcriptional activity. Mol Cell Biol 14:1920–1928.

Kwast KE, Lai LC, Menda N, James DT, III, Aref S, Burke PV. 2002. Genomic analyses of anaerobically induced genes Saccharomyces cerevisiae: Functional roles of Rox1 and other factors in mediating the anoxic response. J Bacteriol 184:250–265.

Lai LC, Kosorukoff AL, Burke PV, Kwast KE. 2005. Dynamical remodeling of the transcriptome during short-term anaerobiosis in Saccharomyces cerevisiae: Differential response and role of Msn2 and/or Msn4 and other factors in galactose and glucose media. Mol Cell Biol 25:4075–4091.

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

Matys V, Fricke E, Geffers R, Gobling E, Haurock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31:374–378.

Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V, et al. 2004. MIPS: Analysis and annotation of proteins from whole genomes. Nucleic Acids Res 32:D41–44.

Miyoshi K, Shirai C, Mizuta K. 2003. Transcription of genes encoding transacting factors required for rRNA maturation/ribosomal subunit assembly is coordinately regulated with ribosomal protein genes and involves Rap1 in Saccharomyces cerevisiae. Nucleic Acids Res 31:1969–1973.

Moehle CM, Hinnebusch AG. 1991. Association of RAPI binding sites with stringent control of ribosomal protein gene transcription in Saccharomyces cerevisiae. Mol Cell Biol 11:2723–2735.

Nagamine N, Kawada Y, Sakakibara Y. 2005. Identifying cooperative transcriptional regulations using protein-protein interactions. Nucleic Acids Res 33:4828–4837.

Phuong TM, Lee D, Lee KH. 2004. Regression trees for regulatory element identification. Bioinformatics 20:750–757.

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

Qian J, Lin J, Luscombe NM, Yu H, Gerstein M. 2003. Prediction of regulatory networks: Genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19:1917–1926.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. 2003. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34:166–176.

Shapira M, Segal E, Botstein D. 2004. Disruption of yeast forkhead-associated cell cycle transcription by oxidative stress. Mol Biol Cell 15:5659–5669.

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

Tachibana C, Yoo JY, Tagne JB, Kacherovsky N, Lee TI, Young ET. 2005. Combined global localization analysis and transcriptome data identify genes that are directly coregulated by Adr1 and Cat8. Mol Cell Biol 25:2138–2146.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 1999. Systematic determination of genetic network architecture. Nat Genet 22:281–285.

van Steensel B, Delrow J, Henikoff S. 2001. Chromatin profiling using targeted DNA adenine methyltransferase. Nat Genet 27:304–308.

Yu H, Luscombe NM, Qian J, Gerstein M. 2003. Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet 19:422–427.

Zhu Z, Shendure J, Church GM. 2005. Discovering functional transcription-factor combinations in the human cell cycle. Genome Res 15:848–855.

DOI 10.1002/bit