BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii101–ii107
doi:10.1093/bioinformatics/bti1117

Genes and Genomes

Computational discovery of transcriptional regulatory rules

Tho Hoan Pham[1,*], José Carlos Clemente[1], Kenji Satou[1,2], Tu Bao Ho[1,2]

[1]Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan and
[2]Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Science Plaza, 5-3 Yonban-cho, Chiyoda-ku, Tokyo 102-8666, Japan

ABSTRACT

Motivation: Even in a simple organism like yeast *Saccharomyces cerevisiae*, transcription is an extremely complex process. The expression of sets of genes can be turned on or off by the binding of specific transcription factors to the promoter regions of genes. Experimental and computational approaches have been proposed to establish mappings of DNA-binding locations of transcription factors. However, although location data obtained from experimental methods are noisy owing to imperfections in the measuring methods, computational approaches suffer from over-prediction problems owing to the short length of the sequence motifs bound by the transcription factors. Also, these interactions are usually environment-dependent: many regulators only bind to the promoter region of genes under specific environmental conditions. Even more, the presence of regulators at a promoter region indicates binding but not necessarily function: the regulator may act positively, negatively or not act at all. Therefore, identifying true and functional interactions between transcription factors and genes in specific environment conditions and describing the relationship between them are still open problems.

Results: We developed a method that combines expression data with genomic location information to discover (1) relevant transcription factors from the set of potential transcription factors of a target gene; and (2) the relationship between the expression behavior of a target gene and that of its relevant transcription factors. Our method is based on rule induction, a machine learning technique that can efficiently deal with noisy domains. When applied to genomic location data with a confidence criterion relaxed to P-value = 0.005, and three different expression datasets of yeast *S.cerevisiae*, we obtained a set of regulatory rules describing the relationship between the expression behavior of a specific target gene and that of its relevant transcription factors. The resulting rules provide strong evidence of true positive gene-regulator interactions, as well as of protein–protein interactions that could serve to identify transcription complexes.

Availability: Supplementary files are available from http://www.jaist.ac.jp/~h-pham/regulatory-rules

Contact: h-pham@jaist.ac.jp

1 INTRODUCTION

Even in a simple organism like *Saccharomyces cerevisiae* the mechanisms of gene transcriptional regulation are extremely complex and uncovering them is one of the key problems in computational biology. About 10% of genes in any organism can produce proteins having a transcriptional role. These proteins are called transcription factors or regulators, and their DNA-binding interactions make the set of downstream genes express.

Mapping of DNA-binding locations of transcription factors has been proposed both by experimental (Iyer *et al*., 2001; Ren *et al*., 2000; Lee *et al*., 2002; Lieb *et al*., 2001) and computational approaches (Roth *et al*., 1998; Liu *et al*., 2002; Matys *et al*., 2003; Timothy *et al*., 1995; Kellis *et al*., 2003). However, although genomic location data from experimental approaches is noisy owing to imperfect measuring methods (Lee *et al*., 2002), computational approaches suffer from over-prediction problems owing to the short length of the motifs bound by the transcription factors. Harbison *et al*. (2004) have constructed a map of yeast transcriptional regulatory code at different confidence levels by incorporating results from both kinds of methods. The frequency of false positives in genome-wide location data ranges from 6 to 10%, and about one-third of actual DNA-regulator interactions are not reported at the 0.001 P-value level (Lee *et al*., 2002). Nevertheless, increasing the P-value to include more true DNA-regulator interactions makes the rate of false positives increase. Also, interactions between regulators and DNA-binding sites are environment-dependent (Harbison *et al*., 2004): many regulators bind only to the promoter of certain genes under specific conditions. Even more, presence of regulators at a promoter region indicates binding but not necessarily function: the regulator may act positively, negatively or not act at all. Therefore, recognizing relevant regulators of a gene and describing how they regulate it under specific environmental conditions are still unanswered problems.

DNA-binding locations of transcription factors and gene-expression profiles are important information to reveal regulatory modules, which describe the regulation of a group of transcription factors on a set of genes. There are two main approaches to uncover regulatory modules: genes can be clustered into modules based on the similarity of their expression profiles, and then common binding sites of transcription factors in the promoter region of genes in each module can be found (Pilpel *et al*., 2001; Ihmels *et al*., 2002). Alternatively, we can group genes into modules that are commonly bound by a set of transcription factors and then validate the expression profiles to confirm these modules (Bar-Joseph *et al*., 2003; Pham *et al*., 2004). However, these methods do not mention under which environmental conditions the interactions take place, nor how transcription factors regulate genes in each module. Segal *et al*. (2003) introduced a probabilistic method for inferring not only regulatory modules, but also their regulatory program (describing the relationship between the expression of target genes and that of transcription factors). This approach uses the Expectation–Maximization (EM) algorithm to search the model (structure of regulatory modules and its parameters) with highest Bayesian score. The main drawback of

*To whom correspondence should be addressed.

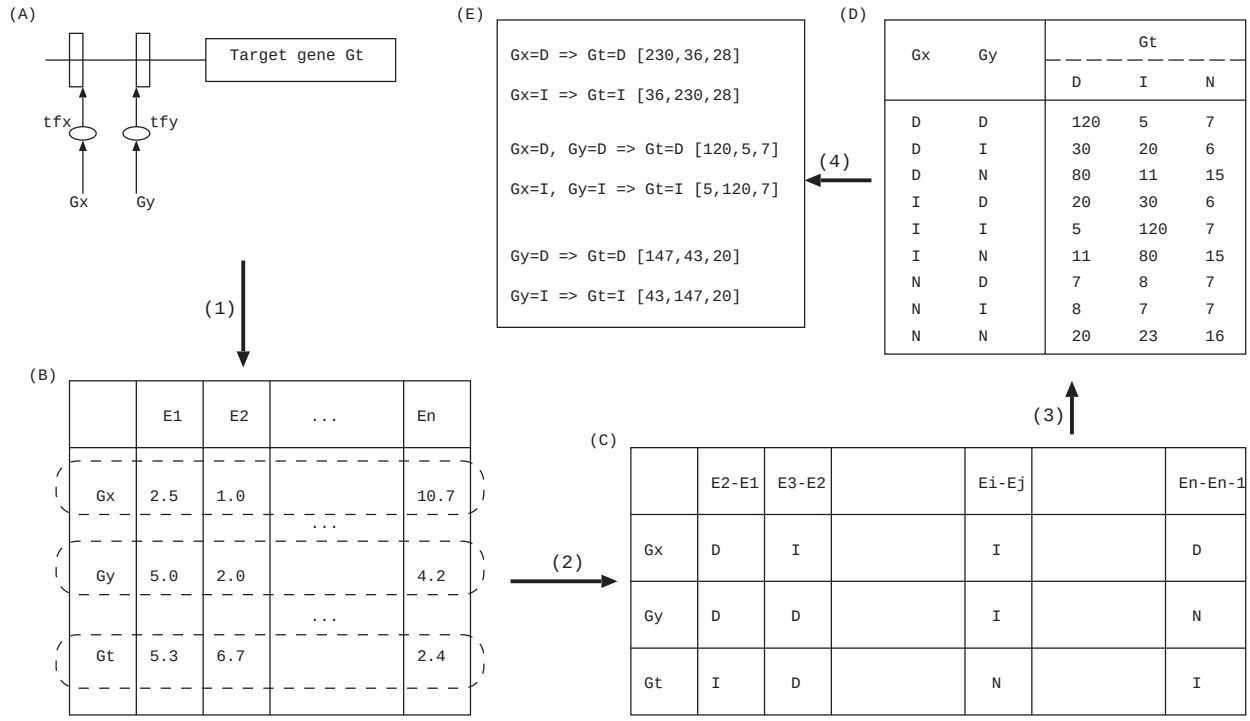

Fig. 1. Approach overview. (**A**) Genomic locations; (**B**) Gene-expression profiles; (**C**) Gene-expression behaviors; (**D**) Regulatory table; (**E**) Regulatory rules.

this method lies precisely in the EM algorithm, which is known to converge often to local optima, as well as to depend strongly on the initial models supplied.

In this paper, we propose a method that combines expression data with genomic location data to discover qualitative relationships between the expression of a gene (target gene) and that of its relevant transcription factors. We assume that transcription factors regulating the expression of a gene must bind to its promoter, and the expression of the target gene must be consistent, in a specific way, to the expression behavior of these transcription factors.

By analyzing data from microarray experiments, we can see how the expression of a gene changes related to modifications in the expression level of its transcription factors. We represent these expression behaviors by three states: downregulation (D), upregulation (I) and no change (N). We then develop a method based on rule induction (a machine learning technique that can efficiently deal with noisy domains), to discover consistent relationships between the expression behavior of target genes and the expression behavior of their transcription factors. Our method can find (1) relevant transcription factors from the set of potential transcription factors of a target gene; and (2) the relationship between the expression behavior of a target gene and the expression behavior of these relevant transcription factors.

When applied to genomic location data with a relaxed confidence criterion (P-value $= 0.005$) and three different expression datasets of yeast *S.cerevisiae*, our method produced a set of regulatory rules comprehensively describing the relationship between the expression behavior of a target gene and that of its relevant transcription factors. We could find the most frequent regulators occurring in regulatory rules under different conditions: response to environmental stress, response to DNA-damaging agents and during the cell cycle. We illustrate how the resulting regulatory rules provide strong evidence of true positive gene-regulator interactions, as well as evidences of protein–protein interactions that could serve to identify transcription complexes.

2 METHODS

2.1 Approach overview

In this work we find regulatory rules that relate the expression of a gene with that of its regulators. Given n potential transcription factors $tf_1, \ldots, tf_n$ binding to the promoter of a target gene G_t and assuming genes $G_1, \ldots, G_n$ are responsible for expressing these factors (Fig. 1A for an example with $n = 2$), we build a regulatory table as follows: we first determine the expression profiles of $G_1, \ldots, G_n$ and G_t (Fig. 1B) from the expression data. By comparing the results of pairs of experiments, we can determine if the expression of genes increased (I), decreased (D) or did not change (N) at the same time (Fig. 1C). With this information we can construct the regulatory table with instances of the form $(G_1 = v_1, \ldots, G_n = v_n, G_t = v_t, \text{count} = k)$, with $v_i = I, D$ or N (Fig. 1D). Section 2.2 provides more information on the regulatory tables. From the regulatory tables, we then apply the CN2-SD rule induction system (Section 2.3) to produce a set of regulatory rules (Fig. 1E).

2.2 Regulatory tables

Given a gene and a set of its potential transcription factors, the regulatory table of this gene (the target gene) is a contingency table describing the relation between the expression behavior of the gene and its regulators. If a gene has n potential regulators, its regulatory table consists of at most 3^n rows, since the expression behavior of each regulator has three states: I (upregulation), D (downregulation) and N (no change). For each set $(G_1, \ldots, G_n, G_t)$ of regulators and target gene, we then study their expression profiles. Every experiment is compared against all others to determine if the expression of a gene increased ($e_x - e_y > T$), decreased ($e_x - e_y < -T$) or did not

change ($|e_x - e_y| \leq T$). Section 2.6 describes how the threshold T is determined. Fig. 1D is an example of regulatory table of a gene with two potential regulators.

These regulatory tables have two important characteristics. First, they contain noise from three different sources: (1) Imperfect measurement methods to collect gene-expression data, (2) uncertainty of interactions between transcription factors and the target gene (as explained in Section 1) and (3) method to obtain the threshold value T. Owing to these factors, the expression behavior of potential regulators turns out to be often inconsistent with the expression behavior of the target genes. To alleviate this problem, we use the counts for each state of the expression behavior of the gene (Fig. 1D). A second important characteristic of the regulatory tables is that they are sometimes incomplete. Since we construct them from expression data, some combinations of expression behavior of the set of regulators may have never happened under any conditions, or occurred with very low frequency as a result of noise.

Even though regulatory tables can be incomplete and noisy, consistent relationships between expression behavior of genes can nevertheless be uncovered from them. These relationships are represented in the form of a rule $G_{i_1} = v_{i_1}, \ldots, G_{i_k} = v_{i_k} \rightarrow G_t = v_t$, which takes account only of transcription factors $G_{i_1}, \ldots, G_{i_k}$ relevant for the expression behavior of the target gene G_t, and that ignores other non-relevant factors. In the following subsections, we will present a machine learning technique, rule induction, to efficiently discover such kinds of rules from regulatory tables.

2.3 Descriptive rule induction by CN2-SD

Rule induction from examples is a machine learning technique successfully used as a support tool for knowledge acquisition and prediction. The induced rules are usually expressed as condition $\rightarrow$ class, where condition and class are logic expressions of the form (variable$_1$ = value$_1 \wedge \cdots \wedge$ variable$_k$ = value$_k$). There are three kinds of rule inducting algorithms: covering, decision tree-based and association rule-based. Covering algorithms make use of a 'separate-and-conquer' strategy over the search space to learn a rule set (Furnkranz, 1999). This strategy searches for a rule that explains (covers) a part of its training instances, separates (or reassigns with lower weight) these examples and recursively conquers the remaining examples by learning more rules until no examples remain. Decision tree-based algorithms use a 'divide-and-conquer' strategy (Quinland, 1986, 1987). Decision trees can be easily turned into a rule set by generating one rule for each path from the root to a leaf. Finally, association rule-based algorithms use an 'exhaustive search' strategy by exploring almost the whole search space (Liu *et al*., 1998). The basic idea is to use an association rule algorithm to gather all rules that predict the class attribute and also pass a minimum quality criterion.

By implementation, the divide-and-conquer strategy (in decision tree-based algorithms) is restricted to learn non-overlapping rules only. The exhaustive strategy (in association rule-based algorithms) has the problem of producing many redundant rules. Separate-and-conquer algorithms can partially avoid these disadvantages (Furnkranz, 1999; Lavrac *et al*., 2004).

CN2 (Clark and Nibblet, 1989; Clark and Boswell, 1991) is a rule induction system implementing the separate-and-conquer strategy. It was originally designed to solve classification and prediction tasks. CN2 can induce a set of independent rules, where each rule describes a specific subgroup of instances. However, owing to the maner in which CN2 iteratively removes examples, only the first few induced rules are usually of interest. Subsequently induced rules are obtained from biased example subsets, i.e. subsets including only positive examples not covered by previously induced rules. This is not suitable for description tasks (discovering individual rules describing interesting patterns, as presented in the work here), where desired rules may cover overlapped instances. CN2-SD (Lavrac *et al*., 2004), a modification of CN2 for subgroup discovery, solves this problem and will be therefore used in the rest of this paper.

CN2-SD generalizes the covering algorithm by introducing example weights. Initially, all examples have a weight of 1.0. However, the weights of examples covered by a rule will not be set to 0 (as in CN2), but instead will be reduced by a certain factor. The resulting number of rules is typically higher than with CN2, since most examples will be covered by more than one rule. CN2-SD has, therefore, two complementary advantages: it can learn better local patterns since the influence of previously covered patterns is reduced, but not completely ignored; and, it can produce a better classifier by combining the evidence of more induced rules.

For description tasks, besides the weighted covering method, CN2-SD needs also a rule evaluation heuristic that favors rules with higher generality (Pham *et al*., 2005). In the rest of this paper, we will use a weighted relative accuracy heuristic [Equation (1)]. During the weighted covering strategy tends to find rules that explain overlapped subgroups of instances in the search space, the weighted relative accuracy heuristic produces highly general rules that express the knowledge contained in one specific subgroup.

$$h_{\mathrm{WRA}}(\text{condition} \rightarrow \text{class}) = \frac{p(\text{condition})}{p(\text{class}|\text{condition}) - p(\text{class})} \qquad (1)$$

2.4 Filtering regulatory rules

By using a weighted covering strategy CN2-SD can restrict the redundancy of learned rules and guarantee the scanning of the whole search space. However, uninteresting rules are still produced. In addition to the significance test, which ensures that the distribution of examples among classes covered by a rule is significantly different to the distribution obtained by random assignment (Clark and Nibblet, 1989), we use two other heuristics to filter out undesired rules. Given a regulatory rule r, $TF_1 = v_1, \ldots, TF_n = v_m \rightarrow$ Target_gene $= v[n_D, n_I, n_N]$, where $v_i (i = 1, \ldots, m)$ and v are expression behavior values ($\{D, I, N\}$) of genes or transcription factors; and $[n_D, n_I, n_N]$ are the class distribution of examples covered by r.

Removing trivial rules and irrelevant conditions in a rule. r is called a trivial regulatory rule if the predictive value of Target_gene is N (no change). This rule can be interpreted as: 'there is no relationship between the target gene and its transcription factors'. Since these kind of rules are trivial, they are removed from the learned rule set.

If there is any transcription factor in the condition part of a rule appearing with value N (no change), this transcription factor has no role in regulating the expression of the target gene. We also remove these irrelevant factors in the condition part of the rules, and update the class distribution for the new rules.

Removing inconsistent regulatory rules. The consistence (cons) of a non-trivial regulatory rule r is defined as

$$\mathrm{cons}(r) = \frac{n_p}{n_D + n_I} \times \frac{n_p}{n_D + n_I + n_N} \qquad (2)$$

where n_p is equal to n_D if r is a classification prediction rule for Target_gene belonging to class D, and equal to n_I if r is a classification prediction rule for Target_gene belonging to class I. Consistence takes into account two factors: a confidence-without-noise $n_p/(n_D + n_I)$ and confidence-with-noise $n_p/(n_D + n_I + n_N)$, where n_N is the parameter representing noise in microarray data. Clearly, $0 \leq \mathrm{cons}(r) \leq 1$, and the higher the value of cons(r), the higher the confidence that regulatory rule r is true.

2.5 Datasets

In our experiments we used genomic location data (as described in Harbison *et al*., 2004) as a source for potential gene–transcription factor interactions. This dataset contains interactions between 106 transcription factors and ~6200 genes of yeast *S.cerevisiae*, with a relaxed binding criterion of confidence P-value ≤ 0.005 (in order to increase the number of true and functional interactions that can be found), and conserved in at least one other yeast species. Three expression datasets (Gasch *et al*., 2000, 2001; Spellman *et al*., 1998) are also used to analyze the expression behavior of target genes as well as transcription factors response to environmental stresses, response to DNA-damaging agents and during the cell cycle. The number of experiments of these three datasets is 172, 52 and 77.

Table 1. Summary of produced regulatory rules

	Response to environmental changes[a]	Response to DNA damage[b]	Cell cycle[c]	All
Number of rules	2438	1974	506	3707
Number of genes	1002	889	288	1336
Number of interactions found	2206	1938	580	3033
Number of interactions found with P-value ≤ 0.001	1518 (68.8%)	1350 (69.7%)	401 (69.1%)	2103 (69.3%)
Number with p-value ≤ 0.001, no interaction found	475/1993 (23.8%)	557/1927 (28.9%)	389/790 (48.2%)	455/2558 (17.8%)
Ten most frequent regulators	RAP1, ABF1, STE12, FHL1, REB1, NRG1, HSF1, SWI6, UME6, CBF1	RAP1, FHL1, HSF1, GCN4, ABF1, STE12, CIN5, MSN4, CBF1, MBP1	SWI6, STE12, SWI4, MBP1, DIG1, MSN4, PHD1, FKH1, FKH2, ABF1	RAP1, ABF1, STE12, REB1, GCN4, HSF1, NRG1, CBF1, SWI6, FHL1

[a]Gasch *et al*. (2000).
[b]Gasch *et al*. (2001).
[c]Spellman *et al*. (1998).

2.6 Assigning expression behavior labels

We compare the expression values e_i and e_j of a gene between any two microarray experiments i and j to determine its expression behavior. If $e_j - e_i > T$, the expression behavior of the gene is upregulated (I) from experiment i to experiment j; if $e_j - e_i < T$, the expression behavior of the gene is downregulated (D); otherwise it is unchanged (N). When the threshold value T is large, our system will produce regulatory rules with high confidence-without-noise but low confidence-with-noise (Section 2.4). These regulatory rules are often true positives although by using a high threshold we are also discarding some relevant regulatory rules. Inversely, if the value of T is small, our system will produce many irrelevant regulatory rules owing to the noise in microarray data. To determine a reasonable threshold T for a microarray dataset, we first set an initial value T_0 large enough, then apply our method to find a set of true positive regulatory rules. This set is considered as previously known regulatory rules (since the set often includes true positives). We then tune the parameter T to get the highest value of the average measure cons [Equation (2)] over all the rules in this set. By using this method, we obtained threshold values T for the data (1.3) and data (0.75) of Gasch *et al*. (2000, 2001) and data (1.0) of Spellman *et al*. (1998).

3 RESULTS

Results were obtained by using the datasets for gene-expression profiles response to environmental conditions (Gasch *et al*., 2000, 2001; Spellman *et al*., 1998), with T threshold values calculated as in Section 2.6 (1.3, 0.75 and 1.0). These datasets represent the gene-expression profiles response to environmental stresses, response to DNA-damaging agents and during the cell cycle. Genomic location data of yeast *S.cerevisiae* with binding criterion relaxed to P-value ≤ 0.005 and conserved in at least one other yeast (Harbison *et al*., 2004) was used to determine potential transcription factors of a gene. We removed genes that are bound by no regulator or where 95% of the total number of expression behaviors were N(no change). There are 1800, 2133 and 1172 remaining genes that are bound by at least one transcription factor and significantly expressed, i.e. they have a number of expression behaviors of classes D or I >5% of their total number of behaviors. For each gene in these sets and each expression dataset, we constructed a table (Fig. 1), obtaining a total of 1800, 2133 and 1172 regulatory tables. The algorithm CN2-SD (Lavrac *et al*., 2004) with WRA heuristic [Equation (1)] was then applied to find all regulatory rules from these tables. Finally, we filtered out trivial rules, trivial conditions in rules (Section 2.4), regulatory rules covering few examples and rules with consistence <0.3.

We found 3707 regulatory rules for predicting 1336 target genes to be D and the same number of rules for prediction being I. Since we analyze any pair of experiments without considering their time order, with each regulatory rule for predicting the target gene to be D there is an equivalent regulatory rule for predicting it to be I, where variables in the condition part of the rule received the opposite values ($I \leftrightarrow D$). For example, rules $Gx = D, Gy = D \rightarrow Gt = I$ and $Gx = I, Gy = I \rightarrow Gt = D$ are equivalent. For simplicity, we will refer only to regulatory rules for predicting target gene Gt belonging to class D as the representative ones.

Table 1 shows the number of regulatory rules, number of genes controlled by these rules and the 10 most frequent transcription factors found from these three expression datasets. We found 1002 genes appearing in 2438 regulatory rules in response to environmental stresses; 889 genes appearing in 1974 rules in response to DNA-damaging agents; 288 genes appearing in 506 rules related to the cell cycle; and a total of 1336 genes in 3707 rules in all three kinds of environments. We also found that the most frequent transcription factors occurring in regulatory rules in response to environmental stresses (RAP1, ABF1, STE12, FHL1, REB1, etc.) and in response to DNA-damaging agents (RAP1, FHL1, HSF1, GCN4, ABF1, etc.) are quite similar and agree with the function they have been annotated with in Gene Ontology (Harris *et al*., 2004), whereas the most frequent transcription factors occurring in regulatory rules from the cell-cycle dataset (SWI6, STE12, SWI4, MBP1, DIG1, etc.) have functions previously reported to control the cell cycle during growth (Harris *et al*., 2004). The full set of regulatory rules can be obtained on-line (files **regulatory-rules.txt*).

It should be noticed how one gene is often regulated by one or more transcription factors depending on environmental conditions. For example, gene YPR145W (ASN1) is regulated by different subsets of regulators (Table 2) under environmental stresses. The transcription factors that most influenced the expression of YPR145W response to environmental changes are STE12, with regulatory rule YPR145W = D ← STE12 = D covering 2703 instances and consistence 0.82; and DAL82, which negatively regulates YPR145W. The transcription factors GCN4 and GLN3, when acting together

Table 2. Examples of regulatory rules (boldface indicates consistency ≥0.3)

Regulatory rules	Response environment changes		Response DNA-damage		Cell cycle	
	Class distribution	Consistence	Class distribution	Consistence	Class distribution	Consistence
YPR145W = D ← STE12 = D	[2198, 8, 479]	**0.82**	[36, 11, 130]	0.16	[0, 0, 102]	0.00
YPR145W = D ← GCN4 = I, STE12 = D	[300, 0, 33]	**0.90**	[0, 0, 0]	0.00	[0, 0, 5]	0.00
YPR145W = D ← DAL82 = I	[820, 56, 435]	**0.59**	[19, 42, 155]	0.03	[0, 1, 6]	0.00
YPR145W = D ← DAL82 = I, GCN4 = I	[256, 15, 95]	**0.66**	[0, 1, 3]	0.00	[0, 0, 0]	0.00
YPR145W = D ← DAL82 = I, GLN3 = I	[356, 11, 92]	**0.75**	[6, 2, 14]	0.2	[0, 0, 0]	0.00
YBR067C = D ← ASH1 = D	[1077, 111, 1102]	**0.43**	[111, 180, 303]	0.07	[408, 171, 696]	0.23
YBR067C = D ← ASH1 = D, HSF1 = I	[124, 0, 8]	**0.94**	[0, 7, 6]	0.00	[0, 3, 2]	0.00
YBR067C = D ← HSF1 = D, NRG1 = D	[74, 301, 228]	0.02	[24, 0, 6]	**0.80**	[0, 11, 16]	0.00
YBR067C = D ← HSF1 = D	[139, 350, 360]	0.05	[51, 6, 41]	**0.47**	[26, 1, 51]	**0.32**

Table 3. Genes regulated by MBP1 and SWI6

Rule	Env.$^+$	GO terms for target gene
MBP1 = D,SWI6 = D → YBR070C = D	b,c	Nuclear envelope–endoplasmic reticulum network
MBP1 = D SWI6 = I → YDR263C (DIN7) = D	a	DNA repair, mitochondrion
MBP1 = I SWI6 = I → YDR507C (GIN4) = D	c	Protein amino acid phosphorylation, protein kinase activity, bud neck
MBP1 = D SWI6 = I → YGL178W (MPT5) = D	a	Cell wall organization and biogenesis, mRNA binding,cytoplasm
MBP1 = I SWI6 = D → YGR109C (CLB6) = D	c	G1/S transition of mitotic cell cycle, cyclin-dependent protein kinase regulator activity
MBP1 = I SWI6 = D → YGR152C (RSR1) = D	a,c	Bipolar bud site selection, GTPase activity, plasma membrane
MBP1 = D SWI6 = I → YGR180C (RNR4) = D	b	DNA replication, ribonucleoside-diphosphate reductase activity, cytoplasm
MBP1 = D SWI6 = D → YJL187C (SWE1) = D	c	G_2/M transition of mitotic cell cycle, protein kinase activity, nucleus
MBP1 = D SWI6 = D → YKL008C (LAC1) = D	a,c	Ceramide biosynthesis, sphingosine *N*-acyltransferase activity, endoplasmic reticulum
MBP1 = I SWI6 = D → YMR179W (SPT21) = D	c	Regulation of transcription from Pol II promoter,nucleus
MBP1 = I SWI6 = D → YMR307W (GAS1) = D	a	Cell wall organization and biogenesis,'1,3-beta-glucanosyltransferase activity', mitochondrion
MBP1 = D SWI6 = D → YNR009W = D	a,c	Unknown, cytoplasm
MBP1 = I SWI6 = D → YNR009W = D	a	Unknown, cytoplasm
MBP1 = I SWI6 = D → YPL127C (HHO1) = D	c	'Regulation of transcription, DNA-dependent', DNA binding, nucleus
MBP1 = D SWI6 = D → YPR075C (OPY2) = D	c	Cell cycle arrest in response to pheromone, cytoplasm
MBP1 = I SWI6 = D → YPR120C (CLB5) = D	c	G1/S transition of mitotic cell cycle, cyclin-dependent, protein kinase regular activity, nucleus

$^+$Env. = a, b, c ≡ regulatory rule is activated in response to environmental stresses, in response to DNA-damaging agents or during cell cycle, respectively.

with STE12 or DAL82, can increase the activation ability of these factors. Conversely, a transcription factor (independently or cooperatively with others) can regulate many different genes at the same time. For example, MBP1 interacts with SWI6 in different ways to regulate the activity of 15 different genes (Table 3).

4 DISCUSSION

4.1 Relevant interactions from genomic locations data

We analyzed relevant interactions between 94 transcription factors and 1336 genes occurring in 3707 regulatory rules found by our method. We found 3033 relevant interactions among them (Table 1), 2103 (69.3%) of which have been reported in the genomic locations data with P-value ≤ 0.001 (Lee *et al.*, 2002). Therefore, 31.7% of the relevant interactions found in regulatory rules are from potential ones in the genomic locations with $0.001 \leq P\text{-value} \leq 0.005$. This result agrees with the work of Lee *et al.* (2002), where it was reported that about one-third of actual DNA-regulator interactions in genomic locations data are missed at P-value = 0.001. Details of interactions in regulatory rules and in genomic locations data for the 1336 genes can also be obtained from the complementary on-line material. Out of the 2558 interactions, 455 interactions, involving the 1336 genes (Table 1) and all interactions involving other genes from this genomic locations data were not found in regulatory rules. The reasons are: (1) the genomic locations data contain substantial noise; (2) in our experiments we only considered three kinds of environmental conditions, whereas the genomic locations data contain potential interactions that do not actually take place under the conditions we chose; and (3) many physically binding interactions are too weak or do not translate into a real function.

4.2 Regulatory modules and transcription complexes

We used a clustering method based on a closed itemset lattice, [Pham *et al.* (2004)] to group genes regulated by a common subset of transcription factors. We define a regulatory module as a system including three components: a set of genes, a set of regulators and a

Table 4. Description of some regulatory modules

Regulators	Roles of regulators	Number of genes	Significantly shared GO terms
RAP1	Transcriptional silencing of HML and HMR loci, activation of ribisomal glycolytic enzymes, etc.	111	(71/111) protein biosynthesis ($P = 4.99e-41$) (21/111) ribosome biogenesis ($P = 1.37e-10$)
ABF1	Chromatin-reorganizing activity involved in transcriptional activation, gene silencing, and DNA replication and repair	132	(27/132) ribisome biogenesis ($P = 4.74e-14$) (24/132) RNA processing ($P = 1.29e-08$)
STE12	Activates genes involved in mating or pseudohyphal/invasive growth pathways	78	(14/78) conjugation with cellular fusion ($P = 3.87e-12$) (13/78) response to abiotic stimulus ($P = 3.69e-7$)
FHL1	Similarity to DNA-binding domain of *Drosophila* forkhead, required for rRNA processing	82	(66/82) protein biosynthesis ($P = 8.41e-49$) (14/82) ribosomal subunit assembly ($P = 1.27e-15$)
HSF1	Heat shock transcription factor, activates multiple genes in response to hyperthermia	99	(17/99) protein folding ($P = 3.57e-18$) (17/99) response to stress ($P = 6.95e-06$)
SWI6	G_1/S transition, meiotic gene-expression localization regulated by phosphorylation	67	(16/67) development ($P = 6.62e-6$), (8/67) regulation of cell cycle ($P = 2.25e-5$)
MBP1	Involved in regulation of cell-cycle progression from G_1 to S phase	37	(6/37) DNA replication (1.35e − 5), (10/37) DNA metabolism (0.00023) (5/37) DNA repair (0.0006)
MBP1 and SWI6	Complex regulating transcription at the G_1/S transition	15	(6/15) regulation of cell cycle ($P = 1.24e-7$), (3/15) regulation of cyclin dependent protein kinase activity ($P = 3.91e-6$)
SWI4	Involved in cell-cycle dependent gene expression	36	(6/36) regulation of cell cycle ($P = 3.52e-5$), (4/36) G_1/S transition of mitotic cell cycle ($P = 8.71e-5$), (3/36) G2/M transition of mitotic cell cycle ($P = 0.00059$)
GCN4	Amino acid biosynthetic genes with respect to amino acid starvation	77	(21/77) amino acid and derivative metabolism ($P = 6.86e-16$)

set of regulatory rules between them. Since a gene can be regulated by different sets of regulators with different regulatory rules, it can belong to multiple regulatory modules. Genes in each module often have similar or related functions that agree with the role of their transcription factors. We used GO Term Finder to search for significantly shared GO terms directly or indirectly associated with the genes in each regulatory module. To determine significant terms, the algorithm examines a group of genes to find GO terms to which a high proportion of the genes are associated, as compared with the number of times the term is associated with other genes. Table 4 describes regulatory modules including some of the most frequent regulators. The modules obtained in this work are more complete than those in previous studies (Pilpel *et al*., 2001; Ihmels *et al*., 2002; Bar-Joseph *et al*., 2003; Segal *et al*., 2003; Pham *et al*., 2004), in addition to what genes and regulators compose each module, we also describe the regulatory relationship between them in the form of a rule under certain environmental conditions.

Table 3 shows a detailed description of one of the regulatory modules appearing in Table 4. This module consists of 15 distinct genes commonly regulated by MBP1 and SWI6. These two regulators have been reported to form a complex involved in regulation of cell-cycle progression (Koch *et al*., 1993; Bruin *et al*., 2004). Out of 15 genes in this module, 6 of them (GIN4, MPT5, CLB6, SWE1, OPY2 and CLB5) are related to the cell-cycle regulation ($P = 1.24 \times 10^{-7}$), as annotated in Gene Ontology.

This example suggests a possible use of our method to predict transcription complexes. We consider all regulatory modules with more than two regulators and containing at least five genes. Regulators that coactivate or corepress a specific set of genes are candidates to form transcription complexes. Table 5 shows candidate complexes that regulate $\geq$7 or more genes. For example, SWI4 and SWI6 coregulate 16 distinct genes; TEC1 and STE12, 15; INO2 and INO4, 11; HAP2 and HAP4, 10; FKH1 and FKH2, 7; and SWI4, SWI6 and STE12, 4. These coregulators have been also previously confirmed to interact in order to regulate genes (Table 4 for references). There are some other pairs of coregulators regulating $\geq$7 genes in the resulting rules for which we could not find any evidence in the BIND database (Bader *et al*., 2003). For example, FHL1 and RAP1 coregulate (almost always positively) up to 65 distinct genes, with most of them being related to the process 'protein biosynthesis'. Until experimental confirmation, we suggest that these pairs of coregulators could be new transcription complexes. The complete list of coregulators can be found on-line.

5 CONCLUSION

Data of DNA–transcription factor interactions from experimental and computational methods is often noisy and contains information about physically binding interactions, although not necessarily functional ones. By combining this data with expression profiles data, our rule induction method can discover relevant transcription factors for a given target gene, as well as the relationship between the expression behavior of the target gene and that of its relevant regulators. By using a relaxed confidence value we were able to uncover interactions

Table 5. Candidate complexes

Candidate complex	Number of genes	External evidences
FHL1 and RAP1	65	
DIG1 and STE12	21	Olson *et al*. (2000), BIND Id: 130453
SWI4 and SWI6	16	Siegmund and Nasmyth (1996), BIND Id: 24482
MBP1 and SWI6	15	Siegmund and Nasmyth (1996), BIND Id:24484
TEC1 and STE12	15	Kim *et al*. (2004)
CAD1 and YAP7	14	
TYE7 and CBF1	13	
MSN2 and MSN4	12	
YAP1 and YAP7	11	
DAL82 and GLN3	11	
INO2 and INO4	11	Wagner *et al*. (2001), BIND Id: 126362
HAP2 and HAP4	10	McNabb *et al*. (1997), BIND Id: 170195
CBF1 and GCN4	9	
CBF1 and ABF1	9	
PHD1 and NRG1	8	
SOK2 and CIN5	8	
CIN5 and NRG1	8	
MBP1 and STE12	8	
TEC1 and DIG1	7	
STB1 and MBP1	7	
FKH2 and FKH1	7	Hollenhorst *et al*. (2000), BIND id: 172668
CAD1 and YAP1	7	
YAP7 and GCN4	7	
TEC1 and DIC1 and STE12	4	
SWI4 and SWI6 and STE12	4	Breeden and Nasmyth (1987)

usually missed in other studies owing to an excessively strict *P*-value. The use of expression profiles obtained under three different environments allowed us to establish not only if an interaction takes place, but also if it is functionally active and under what conditions it would happen.

Our method also provides evidence of transcription factors that commonly regulate different groups of genes. This result could be used to identify potential transcription complexes, and we present examples of previously not reported complexes for which strong evidence was found.

ACKNOWLEDGEMENTS

We would like to thank Prof. Nada Lavrac and Dr Branko Kavsek for their help and comments. This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) 'Genome Information Science' from the Ministry of Education, Culture, Sports, Science and Technology of Japan; BIRD of Japan Science and Technology Agency (JST); and the COE project 'Knowledge Creation from Data Mining' (JCP KS 1) from Japan Advanced Institute of Science and Technology.

Conflict of Interest: none declared.

REFERENCES

Bader,G.D. *et al*. (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res.*, **31**, 248–250.

Bar-Joseph,Z. *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

Bruin,R.A. *et al*. (2004) Cln3 activates g1-specific transcription via phosphorylation of the SBF bound repressor Whi5. *Cell*, **117**, 887–898.

Clark,P. and Boswell,R. (1991) Rule induction with CN2: Some recent improvements. In *5th European Working Sessions on Learning*, Porto, Portugal, pp 151–163.

Clark,P. and Nibblet,T. (1989) The CN2 induction algorithm. *Machine Learning*, **3**, 261–283.

Furnkranz,J. (1999) Separate-and-conquer rule learning. *Artif. Intel. Rev.*, **13**(1), 03–54.

Gasch,A. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Gasch,A. *et al*. (2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell*, **12**, 2987–3003.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of an eukaryotic genome. *Nature*, **431**, 99–104.

Harris,M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, 258–261.

Ihmels,J. *et al*. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.

Iyer,V.R. *et al*. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Koch,C. *et al*. (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science*, **261**, 1551–1557.

Lavrac,N. *et al*. (2004) Subgroup discovery with CN2-SD. *J. Machine Learning Res.*, **5**, 153–188.

Lee,T.I., *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Lieb,J.D. *et al*. (2001) Promoter-specific binding of Rap1 revealed by genomewide maps of protein-DNA association. *Nat. Genet.*, **28**, 327–334.

Liu,B., Hsu,W. and Ma,Y. (1998) Integrating classification and association rule mining. In *4th Interl Conference on Knowledge Discovery and Data Mining*.

Liu,X.S. *et al*. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–8399.

Matys,V. *et al*. (2003) TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Pham,T.H. *et al*. (2004) Mining yeast transcriptional regulatory modules from factor DNA-binding sites and gene expression data. *Genome Informatics*, **15**, 287–295.

Pham,T.H., Ho,T.B., Satou,K. and Clemente,J.C. (2005) Rule evaluation heuristics for knowledge discovery. In *Workshop on Knowledge Discovery and Data Management in Biomedical Science (KDDMBS), 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-05)*, Vietnam, pp 29–44.

Pilpel,Y. *et al*. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Quinland,J. (1986) Induction of decision trees. *Machine Learning*, **1**, 81–106.

Quinland,J. (1987) Generating production rules from decision trees. In *10th International Joint Conference on Artificial Intelligence (IJCAI-87)*, pp 304–307.

Ren,B. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Roth,F.P. *et al*. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 907–908.

Segal,E. *et al*. (2003) Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genet.*, **34**, 166–176.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Timothy,L., Bailey, and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. In *Proceedings of the 3rd International Conference on Intelligent Systems Molecular Biology*, Menlo Park CA, USA, pp 21–29.