Mining Yeast Transcriptional Regulatory Modules from Factor DNA-Binding Sites and Gene Expression Data

Tho Hoan Pham[1] **Kenji Satou**[1,2] **Tu Bao Ho**[1]
h-pham@jaist.ac.jp ken@jaist.ac.jp bao@jaist.ac.jp

[1] Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
[2] Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST)

Abstract

In eukaryotes, gene expression is controlled by various transcription factors that bind to the promoter regions. Transcription factors may act positively, negatively or not at all. Different combinations of them may also activate or repress gene expression, and form regulatory networks of transcription. Uncovering such regulatory networks is a central challenge in genomic biology.

In this study, we first defined a new kind of motifs in regulatory networks, transcriptional regulatory modules (TRMs), with the form $factorset \rightarrow geneset$, which emphasizes the combinatorial gene control of the group of factors $factorset$ on the group of genes $geneset$. Second, we developed an efficient method based on a closed itemset mining technique for finding the two most informative kinds of TRMs, *closed inf-TRMs* and *closed sup-TRMs*, from factor DNA-binding sites and gene expression profiles data. The set of all closed inf-TRMs and closed sup-TRMs is often orders of magnitude smaller than the set of all TRMs but does not loss any information. When being applied to yeast data, our method produced results that are more compact, concise and comprehensive than those from previous studies to identify and interpret the transcriptional role of regulator combinations on sets of genes. **Availability:** Supplementary files: `http://www.jaist.ac.jp/~h-pham/regulation/`.

Keywords: regulatory network, factor DNA-binding sites, gene expression profiles, association rule mining, closed itemsets.

1 Introduction

In eukaryotes, gene expression is controlled by various transcription factors that bind to the promoter regions and can act in combination. With combinatorial control, a given transcription factor does not necessarily have a single, simply definable fucntion as commander of a particular battery of genes or specifier of a particular cell type. Rather, transcription factors can be likened to the words of a language: they are used with different meanings in a variety of contexts and rarely alone; it is the well-chosen combination that conveys the information that specifies a gene regulatory event [2]. Each cell is the product of specific gene expression programs that involve direct or indirect interactions between DNAs and transcription factors, which are in turn the products of gene expression in previous time course. Such genetic regulatory networks and mechanisms inside them have long been investigated. Traditionally, these studies required labor-intensive and gene-specific work. Recently, with the complete genome sequences of a number of organisms and the development of several high-throughput genomic technologies, such studies have shifted to a new level, whole-genomic scale [4].

Many studies have attempted to construct genetic regulatory networks based on datasets derived from the whole-genome methodologies. Such datasets are gene expression profiles and DNA-binding

locations of transcription factors. Different levels of success have been achieved by exploiting exclusively one of these databases [7, 10, 11]. The method GRAM was introduced in [3] that combines both databases and had some advantages over the previous ones.

The GRAM algorithm discovers gene modules that are a set of coexpressed genes to which the same set of transcription factors binds. Roughly, the algorithm GRAM scans all subsets of factors (factorsets) and analyzes the expression profiles of genes to which each factorset binds. If the expression profiles of these genes are significantly similar, they would be controlled by the factorset. However, scanning all factorsets in the subset space is an extremely demanding task. GRAM uses some heuristic rules to tackle this problem, but when the number of factors, genes and interactions between them are large, dealing with all the possible subsets becomes an infeasible task.

In this work, we start by defining a new kind of motifs in regulatory networks, transcriptional regulatory modules (TRMs) with the form $r = factorset \rightarrow geneset$, where *geneset* is a set of genes to which the group of factors $factorset$ binds and controls their expression. This rule means that factors in $factorset$ act in combination to transcriptionally regulate the expression of genes in *geneset*. We then introduced the two most informative kinds of TRMs, called *closed inf-TRMs* and *closed sup-TRMs*, both of which are biologically meaningful. The set of closed inf-TRMs and closed sup-TRMs is often orders of magnitude smaller than the set of all TRMs but represents the same knowledge.

We then develope an efficient data mining method to discover all closed inf-TRMs and closed sup-TRMs from factor DNA-binding sites and gene expression profiles data. Our method is based on a closed itemsets mining, a powerful technique in data mining for finding association rules among sets of items from databases of transactions. The closed itemsets mining has a solid theoretical background [9] and has been intensively studied and successfully applied in data mining [6, 12].

Our method has been applied to yeast data for finding closed sup-TRMs and closed inf-TRMs. The analysis of result TRMs reveals some gene modules found by previous studies. Moreover, closed inf-TRMs and closed sup-TRMs, together with the measures *support* and *similar_ratio*, are more compact, concise and comprehensive to identify and interpret the transcriptional control of combinations of regulators.

2 Mining Frequent Itemsets and Closed Itemsets

Frequent itemsets mining is the most important and demanding task in many data mining applications [1]. Let $\mathcal{I} = \{a_1, \ldots, a_M\}$ be a finite set of items and $\mathcal{D}$ be a finite set of transactions (the dataset) where each transaction $t \in \mathcal{D}$ is a list of distinct items $t = \{x_0, \ldots, x_T\}, x_i \in \mathcal{I}$. An ordered sequence of n distinct items $I = \{i_0, i_1, \ldots, i_n\} | i_j \in \mathcal{I}$ is called an itemset of length n, or n-itemset. The number of transactions in the dataset including an itemset I is defined as the *support* of I, denoted by $supp(I)$. Given a threshold $MinSup$, an itemset is said to be frequent if its support is greater than or equal to $MinSup$, infrequent otherwise.

There are basically two kinds of algorithms for finding frequent itemsets. The first is Apriori algorithm [1] and its variants (see the work of Zaki and Hsiao [13] for an overview). They use the basic properties (Apriori properties) that all subsets of a frequent itemset are frequent and that all supersets of an infrequent itemset are infrequent in order to prune elements of the space of itemsets. These properties make it possible to effectively mine sparse datasets. However, with dense datasets, which contain strongly related transactions, it becomes much harder to mine since only a few itemsets can be pruned and the number of frequent itemsets grows very quickly while decreasing of $MinSup$ threshold. As a consequence, the mining task becomes rapidly intractable by these algorithms, which try to extract all the frequent itemsets.

The second type of algorithms, which finds frequent closed itemsets, can avoid the above mentioned problem. A closed itemset is described as a maximal set of items common to a set of transactions. In other words, an itemset I is a closed itemset if there exists no itemset I' such that $I' \supset I$ and

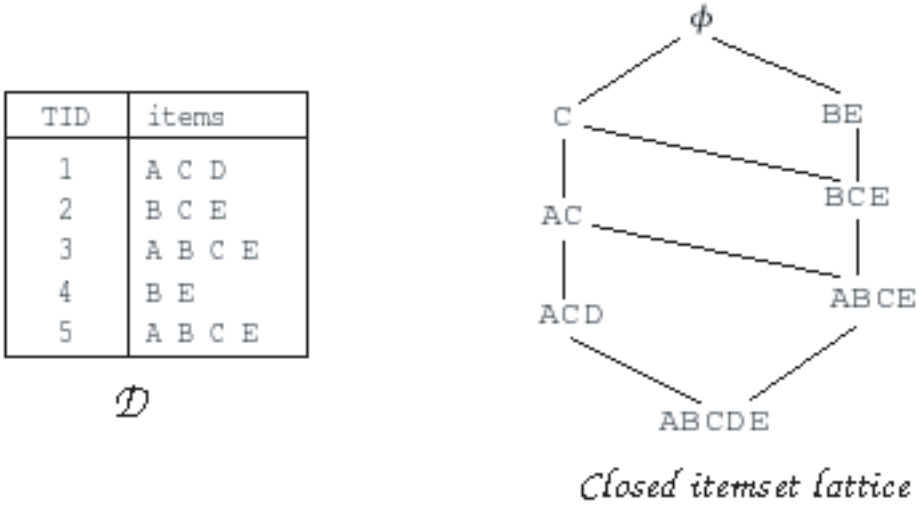

Figure 1: Closed itemsets from a database of transactions.

$supp(I') = supp(I)$. For example, in the database $\mathcal{D}$ in Fig. 1, the itemset BCE is a closed itemset since it is the maximal set of items common to the transactions {2,3,5}. It is called a frequent closed itemset for $MinSup = 2$ as $supp(BCE) = 3 \geq MinSup$. The itemset BC is not a closed itemset since it is not a maximal group of items common to some transactions: all transactions including the items B and C also include the item E. All closed itemsets of a dataset form a lattice that is dually isomorphic to the Galois lattice [9]. In the figure, the lattice contains 8 closed itemsets. It is much smaller than the complete space of itemsets, which in this case includes up to 32 (5 items: 2^5) itemsets. The exact definition of closed itemsets and their useful properties have been described in the work of Pasquier *et al.* [9] and Zaki [12].

The set of closed itemsets is often much smaller than the set of all itemsets, but it presents exactly the same knowledge in a more succinct way. From the set of closed itemsets it is straightforward to derive both the identities and supports of all itemsets. Mining the frequent closed itemsets is thus semantically equivalent to mining all frequent itemsets, but with the great advantage that frequent closed itemsets are often orders of magnitude fewer than frequent ones. Using closed itemsets we implicitly benefit from data correlations which strongly reduce problem complexity [12].

Many algorithms for finding frequent closed itemsets have been developed such as CHARM [13], A-close [9], FPClose [6], etc. In our work, we used FPClose by Grahne and Zhu, implemented in C language.

3 Definition of Closed Sup-TRM and Closed Inf-TRM

Our purpose is to discover groups of factors where each group (or $factorset$) binds to a set of genes ($geneset$) and regulates their expression. In other words, we want to find all transcriptional regulatory modules (TRMs) having a form $factorset \rightarrow geneset$, where $geneset$ is a set of genes that are bound by $factorset$ and similar in their expression profiles.

However, many TRMs are not informative since we cannot infer the biological meaning from them. For example, from the database of transcription factor binding sites in Fig. 2 we cannot infer that $TF2\_TF4 \rightarrow G1\_G3\_G8$ means "$TF2\_TF4$ transcriptionally regulates $G1\_G3\_G8$", because in addition to $TF2$ and $TF4$, there is another factor ($TF5$) binding to all $G1, G3$ and $G8$. In other words, $TF2\_TF4$ is not a maximal set of factors commonly binding to {G1, G3, G8}. A TRM

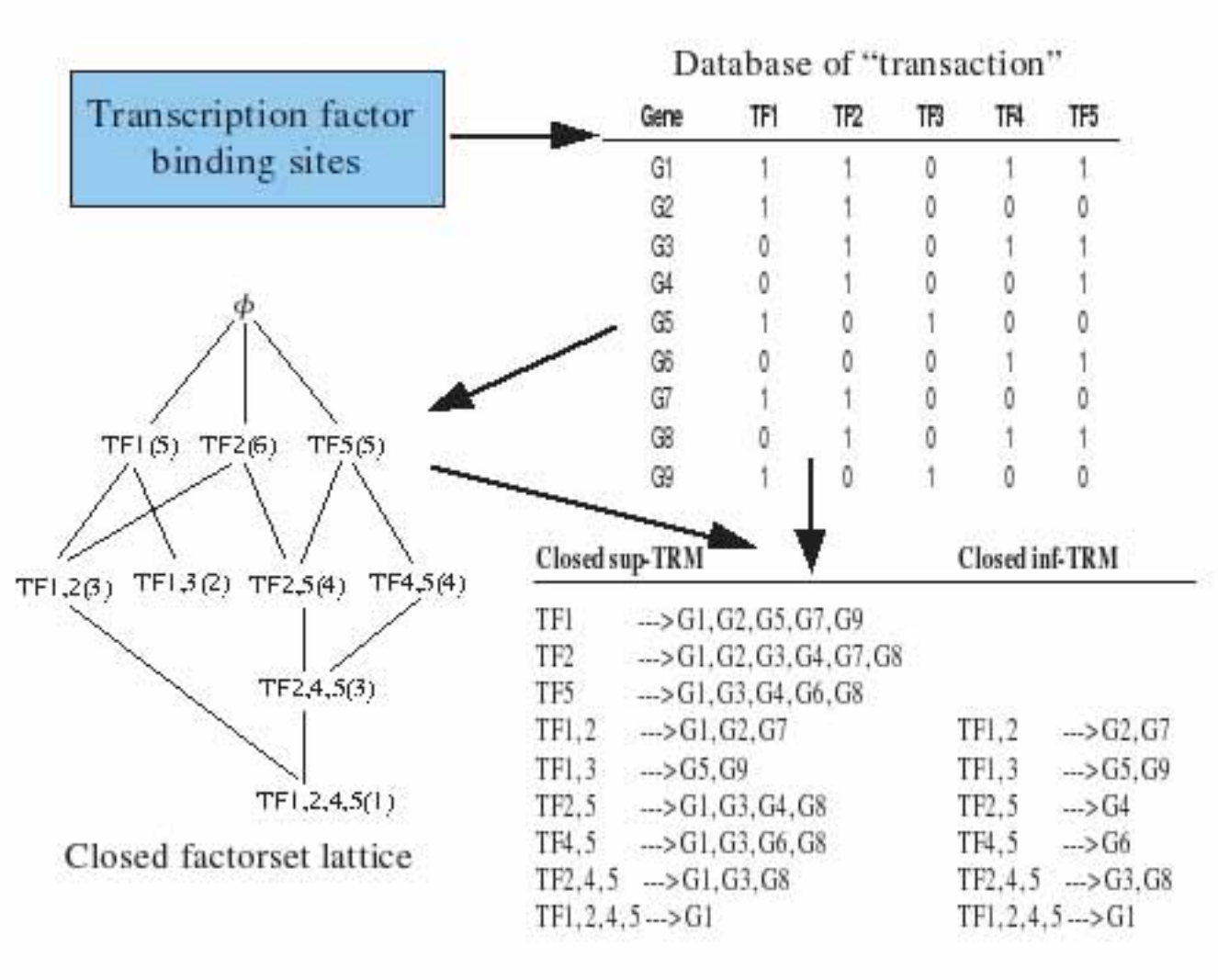

Figure 2: Closed sup-TRMs and closed inf-TRMs.

is informative only if its $factorset$ is a maximal set of factors (i.e.,*a closed factorset*) commonly binding to a set of genes. In this work, we focus only on informative TRMs, which have a form $closed\_factorset \rightarrow geneset$ (we will refer to this as *closed TRM*). As explained above, the set of closed factorsets is much smaller than the set of all factorsets, so the search space for closed TRMs is greatly reduced.

Moreover, we would like to find closed TRMs that emphasize the transcriptional role of their $closed\_factorset$. To do this, we can analyze the expression profiles of genes (*geneset*) that $closed\_factorset$ binds to. If they are similar we believe that *geneset* is controlled by $closed\_factorset$. Here, there are two strategies to group genes into *geneset* of a closed TRM. First we can set *geneset* to be the maximal set of genes to which factors in $closed\_factorset$ commonly bind (we refer to it as *sup_geneset* and the corresponding TRM as *closed sup-TRM*). Second we can set *geneset* to be only genes bound exactly by $closed\_factorset$ (we refer to it as $inf\_geneset$ and the corresponding TRM as *closed inf-TRM*). For example, in Fig. 2 the maximal set of genes that the closed factorset $TF2\_TF4\_TF5$ commonly binds to is $\{G1, G3, G8\}$, therefore $TF2\_TF4\_TF5 \rightarrow G1\_G3\_G8$ is a closed sup-TRM. There are only 2 genes $G3$ and $G8$ that the exact closed factorset $TF2\_TF4\_TF5$ binds to, therefore $TF2\_TF4\_TF5 \rightarrow G3\_G8$ is a closed inf-TRM.

The reason for clarifying these two kinds of TRMs, closed sup-TRMs and closed inf-TRMs, is that

the expression of some genes may be significantly changed (or may not) when one or more additional factors bind to their promoter. Closed inf-TRMs are useful to identify the transcriptional role of their $factorset$ without the impact from other factors, while closed sup-TRMs can include genes that are bound by additional factors other than their $factorset$ and these additional factors may have no transcriptional role. Furthermore, taking account both closed inf-TRMs and closed sup-TRMs is also useful to identify the regulators of not only a group of genes, but also an individual gene.

We also defined two measures for a TRM r: $support(r)$ and $similar\_ratio(r)$, where $support(r)$ is the number of genes in its *geneset* ($inf\_geneset$ with inf-TRM or $sup\_geneset$ with sup-TRM), and $similar\_ratio(r)$ is the rate of genes in its *geneset* whose expression profiles are significantly similar (see Section 4). Both $support(r)$ and $similar\_ratio(r)$ are important evidences to infer if r is a real transcriptional regulatory module or not, which emphasizes that $factorset$ transcriptionally cis-regulates *geneset*.

4 Mining Closed Sup-TRMs and Closed inf-TRMs

Our method for mining closed sup-TRMs and closed inf-TRMs is based on the search on the space of closed factorsets and consists of two phases. The first phase is to find the list (or lattice) of all closed factorsets with *supp* greater than a threshold $MinSup$ from the database of transcription factor binding sites (Fig. 2). We generate a database of "transactions" where each transaction corresponds to a gene and contains a list of transcription factors that bind to it. We then used a software library (FPClose) provided by Grahne and Zhu [6] to find all closed factorsets with $supp \geq MinSup$ in this transaction database.

The second phase concerns how to choose genes included in the *geneset* of a TRM regulated by each $closed\_factorset$. As explained above, there are two ways to generate respectively a closed sup-TRM and a closed inf-TRM. If we get *geneset* to be the maximal set of genes to which the common factors in the $closed\_factorset$ bind, we will produce a candidate closed sup-TRM with the *support* equal to the number of these genes. If we get *geneset* to be only genes to which factors in the $closed\_factorset$ and only these factors bind, we will produce a candidate closed inf-TRM with the *support* equal to the number of these genes. Our method takes account of both kinds of TRMs (Fig. 2). If genes in the *geneset* of a candidate TRM have significantly similar expression profiles, the TRM $r = closed\_factorset \rightarrow geneset$ will be produced.

How can we determine if a group of genes has significantly similar expression profiles? Let $E_1 = (e_{11}, e_{12}, \ldots, e_{1m}), E_2 = (e_{21}, e_{22}, \ldots, e_{2m}), \ldots, E_n = (e_{n1}, e_{n2}, \ldots, e_{nm})$ be expression vectors of n genes $E = (E_1, E_2, \ldots, E_n)$ under m experiments (after standardized as described in Section 5); some e_{ij} may be *null*; we define $E_{avr} = (a_1, a_2, \ldots, a_m)$ as the average expression profile (or the expression center) of the group of these genes ($a_j = average_{i=1,\ldots,n}(e_{ij} | e_{ij} \neq null)$. The distance between a gene E_i and the average expression profile (expression center) is defined as follows:

$$distance(E_i, E_{avr}) = average_{j=1,\ldots,m}(|e_{ij} - a_j| : e_{ij} \neq null)$$

As in the work of Bar-Joseph *et al.* [3], we determine a suitable distance threshold T_k to infer if the expression profiles of genes in a *k-geneset* are significantly similar ($P < 0.05$) based on randomization tests. Randomization tests have been extensively used in computational biology and provide good results. We select at random k genes, compute E_{avr} for this set, and determine the distance d of the 5% closest genes that were not included in the random sampled set. This process is repeated many times (it is actually performed as a pre-processing step, for different possible sizes of k), and we set the threshold T_k to be the median d obtained in these randomization tests.

Genes in the *k-geneset* of a TRM r are said to have significantly similar expression patterns ($P < 0.05$) if their expression vectors are all in the "sphere" centered at E_{avr} with radius T_k. Unfortunately, most TRMs do not satisfy this condition due to experimental errors in the expression profiles as well

as in factor DNA-binding locations. Hence we introduce the algorithm REVISION (Table 1) to find the subset of genes in *k-geneset* whose expression profiles are significantly similar. The idea of this algorithm is to remove outliers (one outlier for each loop), recalculate the expression center, and set new significantly similar threshold T_k. After removing outliers, we have *k′-geneset* whose expression profiles are significantly similar. We defined the *similar_ratio* of the TRM r as the ratio $\frac{k'}{k}$.

Table 1: REVISION algorithm.

Input	*k-geneset*; $T[2,\ldots,n]$: $T[k]$ - threshold to infer *k-geneset* to be significantly similar (see the text); $E_1,\ldots,E_k$: $E_i = (e_{ij})_{j=1,\ldots,m}$ - expression profile;
Output	*k′-geneset*: subset of genes whose expression profiles are significantly similar.

```
1)  do
2)      E_aver = expression_center(E_1,...,E_k); //see the text
3)      for i = 1 to k d_i = distance(Ei, E_aver); //see the text
4)      j = max_{i=1,...,k}(d_i);
5)      if (d_j > T_k)
6)              Report E_j as an outlier;
7)              remove E_j from the list E_1,...,E_k;
8)  while (d_j > T_k)
9)  Report E_1,...,E_k are significantly similar;
```

In summary, our method can discover the two most informative kinds of TRMs: closed sup-TRMs and closed inf-TRMs. *Support* and *similar_ratio* measures of each TRM are important evidences to infer if the TRM is a real transcriptional regulatory module or not.

5 Datasets

The data of factor DNA-binding sites is from the work of Lee *et al.* [8]. This data (updated December 5, 2003) presents profiles for location analysis experiments of 113 factors. A confidence value (p-value) for each factor DNA-binding interaction is calculated by using an error model [8]. From this data, we extracted a database of "transactions", where each transaction has an unique gene identifier (`geneid`) and contains a set of factors that binds to its promoter with the confidence less than a prespecified threshold (0.001). We excluded all transactions that contain no factors. The number of remaining transactions in the database is 2363 (equal to the number of yeast genes that were bound by at least one of 113 transcriptional factors).

We used the ExpressDB from the work of Aach *et al.* [5] for gene expression data in our work. This data included 17.5 million pieces of data reported by 11 studies with three different kinds of high-throughput RNA assays and under 213 conditions. The data has been standardized as Estimated Relative Abundances (ERAs). We then normalized ERAs in the interval [0,1] by a simple linear transformation.

6 Results and Discussion

Our method was applied on the yeast data of factor DNA-binding sites and ExpressDB (see Section 5). It produced 405 candidate closed sup-TRMs and 157 candidate closed inf-TRMs with *support* greater than 5. Among these candidates, there are 141 closed sup-TRMs and 40 closed inf-TRMs with *similar_ratio* greater than 0.5 (see Files "sup_TRMs.htm" and "inf_TRMs.htm" respectively. All

files mentioned in this section are available at "`http://www.jaist.ac.jp/~h-pham/regulation/`"). There are 13 overlapped TRMs among them. Therefore we have 168 most informative TRMs in total. We named each found TRM by its regulators ($factorset$). Table 2 shows an example of closed sup-TRM regulated by factorset $HAP2\_HAP3\_HAP5$. It contains 5 genes, in which 4 genes have significantly similar expression profile. The last one YHR051W (marked by a symbol $\star$) was considered as an outlier.

Table 2: An example of closed sup-TRM.

#module: 10			
Regs: HAP2_HAP3_HAP5	*Support:* 5	*Similar_ratio:* 0.80	
0.048	YPL207W	similarity to hypthetical proteins from A. fulgidus, M.thermoau totrophicum and M. jannaschii	
0.050	YLR220W	involved in calcium regulation	
0.053	YLL027W	mitochondrial protein required for normal iron metabolism	
0.059	YER174C	member of the subfamily of yeast glutaredoxins (Grx3, GRX4, and Grx5)	
$\star$0.105	YHR051W	cytochrome-c oxidase subunit VI	

There are 13 TRMs (see File "overlapped_TRMs.htm") overlapped between closed inf-TRMs and closed sup-TRMs. Taking account together both closed sup-TRMs and closed inf-TRMs will help us to understand more exactly the transcriptional role of regulators. For example, Closed inf-TRM No. 9 and Closed sup-TRM No. 10 both have the same factorset $HAP2\_HAP3\_HAP5$. The former contains 4 genes $YPL207W, YLR220W, YLL027W$ and $YER174C$ to which the exact 3-factorset $HAP2\_HAP3\_HAP5$ bind. All these 4 genes have significantly similar expression profiles. In the latter, in addition to 4 above genes, it includes one more gene $YHR051W$. However, this gene has been considered as an outlier because its expression profile is different from that of the remaining genes in the module. When looking at factors that bind to this gene, we found that, in addition to 3 factors $HAP2, HAP3$ and $HAP5$, there are 2 other factors $HAP4$ and $ABF1$ binding to it. Therefore, we strongly believe that these two factors make $YHR051W$ expressed so differently from the others. This example proves that the distinction between two kinds of TRMs (sup-TRMs and inf-TRMs) is necessary and useful to identify the transcriptional regulators of not only a group of genes but also of an individual gene.

The 27 remaining closed inf-TRMs are those not found in the list of closed sup-TRMs. This proves that when one or more additional factors bind to a gene, its expression may be changed. For example, in Closed inf-TRM No. 36 regulated by $NRG1\_CIN5\_YAP6$ (see File "inf_TRMs.htm"), there are only 6 genes that this 3-factorset binds exactly to, and all these 6 genes are significantly similar in their expression profiles. But there are up to 24 genes that share these 3 factors (see Module 319 in File "sup_TRMs_revision.htm"), and their expression profiles are not similar, since in addition to these 3 factors mentioned above, there are some other factors that also bind to some genes in the module and make them expressed so differently. Therefore closed inf-TRMs are useful to identify a group of genes transcriptionally regulated by a group of factors by avoiding the impact from other factors.

Many TRMs (both closed sup-TRMs and closed inf-TRMs) found by our method include genes whose function are related and consistent with the regulators' known roles. For example, Closed sup-TRM No. 30 regulated by $HIR1\_HIR2$ contains 7 genes, 6 of them have the function concerning "histone". Closed sup-TRM No. 98 and Closed inf-TRM No. 34 both regulated by the same factorset $PDR1\_FHL1\_RAP1$, include genes with the function mainly concerning "ribosomal protein", etc. We used *GO Term Finder* tool (`http://www.yeastgenome.org/help/goTermFinder.html`) to search for significant shared GO terms that are directly or indirectly associated with the genes in each TRM. To determine significance, the algorithm examines the group of genes to find GO terms to which a

high proportion of the genes are associated compared to the number of times that term is associated with other genes in the genome. As a result, 29 of 40 closed inf-TRMs and 81 of 141 closed sup-TRMs have significant ontology terms other than "biological_process unknown", "molecular_function unknown" and "cellular_component unknown". Therefore, TRMs found by our method are biologically meaningful. They will be useful to infer the function of uncharacterized genes.

Our method identifies not only biologically related sets of genes, but also some factors that are interacting to regulate the genes; for example, $HAP2\_HAP3\_HAP5$ (Module 10 in the list of closed sup-TRMs), $HAP2\_HAP4$ (Module 27), $HAP2\_HAP3$ (Module 28), $HIR1\_HIR2$ (Module 30), $MBP1\_SWI4$, $SWI6\_SWI4$, and other interactions can be found in some modules. Some of these interactions have been confirmed by previous studies (collected in the work of Bar-Joseph *et al.* [3]).

Of 168 TRMs, 26 are very similar to some gene modules previously found by GRAM in the work of Bar-Joseph *et al.* [3], although their genes are not exactly same (see File "comparison.htm"). Some TRMs are overlapped with gene modules from their work. There are some differences between our method and GRAM. First, our method finds candidate TRMs based on the search on the closed factorset space that is much smaller than the space of all closed factorsets but does not loss any information. Second, our method for discovering closed sup-TRMs and inf-TRMs emphasizes the transcriptional control of combinations of regulators. This is the reason why TRMs generated by our method are different from gene modules generated by GRAM [3]. For example, Gene Module No. 52 from the results of GRAM regulated by $FKH1\_FKH2$ is not found by our method, because the factorset $FKH1\_FKH2$ binds up to 13 genes and the expression profiles of these genes are very different (see Closed sup-TRM No. 233 in File "sup_TRMs_revision.htm" for the revision process of our method).

7 Conclusion and Future Work

In this paper, we have defined two relevant kinds of TRMs: closed inf-TRMs and closed sup-TRMs that are compact, concise and comprehensive to identify and interprete transcriptional activity of combinations of regulators. We then developed an efficient data mining method to discover closed inf-TRMs and closed sup-TRMs from factor DNA-binding sites and gene expression data. Our method is based on a closed itemset mining that has a solid theoretical background and takes account only of relevant combinations of factors.

Our method has been applied to yeast data to find transcriptional regulatory modules (TRMs). The results are consistent with those previouly found by other methods. Moreover, TRMs found by our method are more concise and comprehensive to identify and interprete transcriptional role of regulators. In this work, we used the data of factor DNA-binding sites proved by Lee *et al.* [8], which were harvested from a microarray method. Each gene-factor interaction was assigned with a confidence value. In future work, we will apply our method to factor DNA-binding sites data that will be computationally predicted (i.e., from the TRANSFAC database).

Acknowledgments

This work was partly supported by a 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; and the COE project JCP KS1 from Japan Advanced Institute of Science and Technology. The first author has been supported by a Vietnamese government scholarship from the Ministry of Education and Training of Vietnam. We thank Jose C. Clemente and three anonymous reviewers for their criticism and suggestions of the reading of the manuscript.

References

[1] Agrawal, R., Imielinski, T., and Swami, A., Mining association rules between sets of items in large databases, *Proc. 1993 ACM SIGMOD Inter. Conference on Management of Data*, 207–216, 1993.

[2] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P., *Molecular Biology of the Cell, Fourth Edition*, Garland Science, Taylor&Francis Group, 2002.

[3] Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A., and Gifford, D.K., Computational discovery of gene modules, regulatory networks, *Nature Biotechnology*, 21:1337–1342, 2003.

[4] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I., The transcriptional program of sporulation in budding yeast, *Science*, 282(5389):699–705, 1998.

[5] Church, G.M., Aach, J., and Rindone, W., Systematic management and analysis of yeast gene expression data, *Genome Res.*, 10(4):431–445, 2000.

[6] Grahne, G. and Zhu, J., Efficiently using prefix-trees in mining frequent itemsets, *Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, Melbourne, FL, November, 2003.

[7] Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N., Revealing modular organization in the yeast transcriptional network, *Nature Genetics*, 31(4):370–377 2002.

[8] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., and Young, R.A., Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, 298:799–804, 2002.

[9] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L., Efficient mining of association rules using closed itemset lattices, *Information Systems, Elsevier Science*, 24(1):25–46, 1999.

[10] Pilpel, Y., Sudarsanam, P., and Church, G.M., Identifying regulatory networks by combinatorial analysis of promoter elements, *Nature Genetics*, 29(2):153–159, 2001.

[11] Qian, J., Dolled-Filhart, M., Lin, J., Yu, H., and Gerstein, M., Beyond synexpression relationships: Local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions, *J. Mol. Biol.*, 314(5):1053–1066, 2001.

[12] Zaki, M., Generating non-redundant association rules, *Data Mining and Knowledge Discovery: An International Journal*, 2004.

[13] Zaki, M. and Hsiao, C-J., CHARM: An efficient algorithm for closed itemset mining, *2nd SIAM Inter. Conference on Data Mining*, 2002.