BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 22 2011, pages 3142–3148
doi:10.1093/bioinformatics/btr526

Gene expression

Advance Access publication September 16, 2011

Discovering relational-based association rules with multiple minimum supports on microarray datasets

Yu-Cheng Liu[1], Chun-Pei Cheng[1] and Vincent S. Tseng[1,2,*]

[1]Department of Computer Science and Information Engineering and [2]Institute of Medical Informatics, National Cheng Kung University, Taiwan, R.O.C.

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Association rule analysis methods are important techniques applied to gene expression data for finding expression relationships between genes. However, previous methods implicitly assume that all genes have similar importance, or they ignore the individual importance of each gene. The relation intensity between any two items has never been taken into consideration. Therefore, we proposed a technique named *REMMAR* (RElational-based Multiple Minimum supports Association Rules) algorithm to tackle this problem. This method adjusts the minimum relation support (*MRS*) for each gene pair depending on the regulatory relation intensity to discover more important association rules with stronger biological meaning.

Results: In the actual case study of this research, *REMMAR* utilized the *shortest distance* between any two genes in the *Saccharomyces cerevisiae* gene regulatory network (GRN) as the relation intensity to discover the association rules from two *S.cerevisiae* gene expression datasets. Under experimental evaluation, *REMMAR* can generate more rules with stronger relation intensity, and filter out rules without biological meaning in the protein–protein interaction network (PPIN). Furthermore, the proposed method has a higher *precision* (100%) than the *precision* of reference Apriori method (87.5%) for the discovered rules use a literature survey. Therefore, the proposed *REMMAR* algorithm can discover stronger association rules in biological relationships dissimilated by traditional methods to assist biologists in complicated genetic exploration.

Availability: The source code in Java and other materials used in this study are available at http://websystem.csie.ncku.edu.tw/REMMAR_Program.rar.

Contact: tsengsm@mail.ncku.edu.tw

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 18, 2011; revised on August 29, 2011; accepted on September 3, 2011

1 INTRODUCTION

Formerly, biologists were constrained to performing laboratory experiments to clarify gene expressions; however, manipulation of multiple genes is very costly, in both time and money. Microarray technology is a useful tool for gaining the expression value of many genes in the cell environment while simultaneously tackling the manipulation problem. At present, how to explain the cell's inner working with efficiency has become another challenge for the biologist. Data mining technologies are ideal for discovering information from large datasets. Thus, a growing number of researchers are applying data mining technologies for efficient elucidation of genetic information.

Recently, various types of analysis methods were proposed to analyze gene expression data, such as clustering analysis (Cheng and Church, 2000; Jiang *et al.*, 2004; Madeira and Oliveira, 2004; Pei *et al.*, 2003; Prelic *et al.*, 2006; Thalamuthu *et al.*, 2006), association rule analysis (Creighton and Hanash, 2003; Georgii *et al.*, 2005; Martinez *et al.*, 2008; McIntosh and Chawla, 2007), classification analysis (Bhasin and Raghava, 2004; Liu and Xu, 2009; Martella, 2009), among others. These methods equip biologists to extract biological knowledge quickly for different analysis purposes. Association rule analysis is used to find and describe relationships between genes. Every rule indicates whether a given gene is expressed (↑) or repressed (↓) to describe expression relationships in a cellular environment. Suppose that we discover the rule $\{gene\ x\uparrow\} => \{gene\ y\downarrow, gene\ z\uparrow\}$ from a gene expression dataset. This rule states that when *gene x* is expressed, *gene y* is repressed and *gene z* is expressed together in this dataset. Two important thresholds exist to measure the significance of a rule, *support* and *confidence*. The *support* of a rule is the frequency that *gene x*↑, *gene y*↓ and *gene z*↑ together in a sample. The *confidence* of a rule is the frequency that *gene y*↓, *gene z*↑ occurs when *gene x*↑ occurs (Creighton and Hanash, 2003).

Although various types of association rule analysis methods were proposed to analyze gene expression data (Creighton and Hanash, 2003; Georgii *et al.*, 2005; Martinez *et al.*, 2008; McIntosh and Chawla, 2007), not all of the proposed methods fit with the goal of our work in exploring the regulatory relations of genes from microarray datasets. Georgii *et al.* (2005) adopted the half-space technique to discover quantitative association rules on numeric microarray datasets without requiring a discretization process. However, their approach cannot mine a complete set of relevant rules when applying the method on microarray profile. Martinez *et al.* (2008) proposed the GenMiner method to mine association rules from a set of gene expression profiles and the publicly available gene ontology (GO) terms. However, the main purpose of their approach is different from ours since they were in an attempt to extract relationships between certain genes and its annotated GO terms. Finally, McIntosh and Chawla (2007) proposed an association rule mining algorithm called MaxConf, which was developed with a row-enumeration method. However, it did not consider the constraint

*To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

of *minimum support* threshold. Moreover, the results of applying the algorithm on microarray dataset can only show high *confidence* gene regulations. Nevertheless, MaxConf may fail in discovering the longer rules with the context containing many items. However, the longer rules are usually more important than the rules that are shorter in size in the real applications (Alves *et al.*, 2010). On the other hand, the rules with low *support* would also be mined using their algorithm. These rules are likely to be regarded as false positive as long as they appear infrequently over all samples. Differentiated with the previous works as described above, our method proposed in this article is effective for mining putative gene regulations from microarray profiles.

However, the traditional association rule analysis implicitly assumes that all items have similar importance without considering their significance in the data. Regardless, this is often not the case in real-life applications. Therefore, a number of researchers have proposed weight-based (Cai *et al.*, 1998; Ramkumar *et al.*, 1998; Tao *et al.*, 2003; Tseng *et al.*, 2010; Wang *et al.*, 2000; Yun and Leggett, 2005) or multiple support-based (Liu *et al.*, 1999; Su *et al.*, 2008) association rule mining to engage this problem in a transaction database. Nevertheless, traditionally weighted or multiple support-based association rule mining techniques only take into account the importance of each item. In a number of real-life case, the degree of importance is dependent on relationships between items. Both techniques never consider the relation intensity between any two items. For this reason, we propose the *REMMAR* (*RElational-based Multiple Minimum supports Association Rules*) algorithm, which adjusts the *minimum relation support* (*MRS*) for each paired item dependent on the relation intensity, to discover the rules that have stronger biological meaning. In a real case study, we used the *shortest distance* between any two genes in the gene regulatory network (GRN) as the relation importance. Two real *S.cerevisiae* microarray datasets were used to evaluate the comparison results between the traditional association rule mining method and the proposed algorithm.

The remainder of this article is organized as follows: Section 2 provides a brief review of traditional association rule mining and presents the proposed method; Section 3 consists of the application of the approach to *S.cerevisiae* datasets, to study the significance of the discovered rules. Finally, in Section 4, we present a conclusion regarding our findings.

2 METHODS

In this section, we first describe the reference Apriori algorithm and research using the Apriori framework to discover gene interaction rules from gene expression data. Before we utilize the *REMMAR* algorithm to discover association rules, microarray data must undergo a transformation into transactional data format. Thereafter, the problem in the research must be defined. Finally, Section 2.3 shows the proposed *REMMAR* algorithm in detail.

2.1 Traditional association rule mining

Agrawal *et al.* (1993, 1994) first proposed association rule analysis. It was used as the Apriori algorithm (Agrawal *et al.*, 1993; Agrawal and Srikant, 1994) to analyze databases, to discover valuable relationships between items. The *support* of an itemset is defined as the frequency that occurs in all transactions. If the *support* of an itemset was not less than the user-specified *minimum support*, it would be recognized as a *large itemset* (*frequent itemset*), because strong relationships exist between items in this itemset.

	G1	G2	G3	G4	G5
Sample 1	-0.21	0.04	0.02	-0.01	-0.23
Sample 2	0.02	-0.07	0.22	0.01	0.07
Sample 3	0.03	-0.25	0.01	-0.08	0.33
Sample 4	0.08	0.05	-0.03	0.03	-0.06

	G1↑	G2↑	G3↑	G4↑	G5↑	G1↓	G2↓	G3↓	G4↓	G5↓
Sample 1	0	0	0	0	0	**1**	0	0	0	**1**
Sample 2	0	0	**1**	0	0	0	0	0	0	0
Sample 3	0	0	0	0	**1**	0	**1**	0	0	0
Sample 4	0	0	0	0	0	0	0	0	0	0

Fig. 1. Example of transforming gene expression data into transaction data format.

The main concept of all Apriori-based methods is that any subitemsets of a *large itemset* is surely also a *large itemset*. In other words, an itemset is not required to be revised if it exceeds *minimum support* as long as any one of subitemsets of the itemset is not a *large itemset*. For example, if a *large itemset* generated by an Apriori-based algorithm contains k items, all of the subitemsets involving 1 to $k-1$ items are certainly *large itemsets*. In contrast, with a brute force method, an itemset needs to be verified even if it contains an item whose frequency does not exceed the *minimum support*. Taken together, using an Apriori-based method can reduce the overhead in examining whether the produced itemsets are large ones. For this reason, the Apriori-based methods are more efficient since the verification of ineligible itemsets is not required. When all *large itemsets* are identified, any *large itemset* possessing more than one item can be divided into two itemsets, S_x and S_y. An association rule is described as $S_x => S_y$, where S_x and S_y are the left-hand side (LHS) and right-hand side (RHS) of this rule, respectively. This rule means that S_y can possibly occur where S_x exists. The *confidence* of a candidate rule is the frequency that S_y occurs when S_x exists. If the *confidence* of this candidate rule is not less than a user-specified *minimum confidence*, it would be defined as an association rule in this data.

In 2003, Creighton and Hanash (2003) used the Apriori method to discover numerous gene interaction rules from gene expression data, many of which are biologically sensible. Every rule can indicate whether a given gene is expressed (↑) or repressed (↓), to describe the expression relationship in a cellular environment. Suppose that the user discovers the rule $\{gene\ x\uparrow\} => \{gene\ y\downarrow, gene\ z\uparrow\}$ from the gene expression dataset. This rule states that if *gene x* is expressed, then *gene y* is repressed and *gene z* is expressed together in this dataset.

2.2 Gene expression data transformation and basic definitions of our method

2.2.1 Transformation of gene expression data As with the model Creighton and Hanash (2003) proposed, which consists of applying the association-based method on gene expression data analysis, each sample can be recognized as a transaction. Each expression value in the data can be transformed as *up* (↑; expressed; readings are >0.2 for the log base 10 of fold change 1.58 as upper bound), *down* (↓; repressed; readings are lower than -0.2 for the log base 10 of fold change -1.58 as lower bound) or *normal* (neither expressed nor repressed). For the threshold value of ± 0.2, it is set in a reasonable range (fold change from 1.5 to 2.0) based on most microarray analysis studies in identifying the differentially expressed probes. The *up* or *down* notation suffices in displaying whether a gene is significant and has the probability to influence other genes. For this reason, only the expression value for each gene can be transformed as *up* or *down* in transactional data format, as shown in Figure 1. After gene expression data transformation, *gene x*↑ and *gene x*↓ will be defined in different items. Thus, the number of items would be increased when either decreasing the upper bound or increasing lower bound, whereas that of items would be decreased when either increasing upper bound or decreasing lower bound.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Fig. 2. An example of GRN.

	G1	G2	G3	G4	G5
G1	-	3	1	2	1
G2	-	-	3	1	2
G3	-	-	-	2	1
G4	-	-	-	-	1
G5	-	-	-	-	-

Fig. 3. An example of gene *shortest distance* matrix retrieved from the GRN of Figure 2.

2.2.2 Initial minimum support The *support* of an itemset S_x is the frequency of S_x occurrence in all transactions (Samples). The *initial minimum support* (*IMS*) is a user-specified baseline threshold to judge the *support* of each item. Items greater than, or equal to, *IMS* are called *large* 1-*itemsets*.

2.2.3 Initial minimum confidence The *confidence* of an association rule $S_x => S_y$ is the frequency of S_y occurrence when S_x occurs. The *initial minimum confidence* (*IMC*) is a user-specified baseline threshold to judge the *confidence* of each rule.

2.2.4 Minimum relation support In the real case study of this research, we used the *shortest distance* between any two genes in the GRN as the relational importance. As an example, Figure 2 is a GRN. For any pair of neighboring genes (physical interaction) in the network, the distance between them will be defined as 1. Hence, the *shortest distance* between G2 and G4 is 1, and that between G1 and G2 is 3. In theory, two genes are more likely to involve the same pathways or biological functions if they have a shorter distance. Figure 3 is the gene *shortest distance* matrix retrieved from the GRN of Figure 2. The *maximum shortest distance* (*Max_Dist*), *minimum shortest distance* (*Min_Dist*) and *average shortest distance* (*Avg_Dist*) of this matrix are 3, 1 and 1.7, respectively. Suppose the user-specified *IMS* is 20%, and the *shortest distance* between G2 and G4 is 1 [$Dist$(G2↑,G4↑) = $Dist$(G2↑,G4↓) = $Dist$(G2↓,G4↑) = $Dist$(G2↓,G4↓) = 1]. Depending on Equation (1), G2 and G4 can get the *MRS* to 17.55% [MRS(G2↑,G4↑) = MRS(G2↑,G4↓) = MRS(G2↓,G4↑) = MRS(G2↓, G4↓) = 17.55%], because G2 and G4 have a stronger relation intensity (with smaller *shortest distance*) than others in the GRN. Furthermore, if the itemset S has more than two items, based on the second equation, the *MRS* of S will be the maximum *MRS* of each item-pair in S. A large itemset would be generated as long as the *support* value of the itemset (contains at least two items) is not less than its corresponding *MRS*.

$$\mathrm{MRS}(x,y)=\begin{cases}\mathrm{IMS}+\left(\mathrm{IMS}^{*}\left(\frac{\mathrm{Dist}(x,y)-\mathrm{Avg\_Dist}}{\mathrm{Max\_Dist-Min\_Dist}}\right)^{2}\right) & (\mathrm{Dist}(x,y)-\mathrm{Avg\_Dist})>0\\ \mathrm{IMS}-\left(\mathrm{IMS}^{*}\left(\frac{\mathrm{Dist}(x,y)-\mathrm{Avg\_Dist}}{\mathrm{Max\_Dist-Min\_Dist}}\right)^{2}\right) & \text{Otherwise}\end{cases} \quad (1)$$

$$\mathrm{MRS}(S)=\mathrm{Max}\{\mathrm{MRS}(x,y)|x,y\in S \text{ and } x\neq y, \text{ where } S \text{ is an itemset containing more than } \textit{one} \text{ item}\} \quad (2)$$

2.2.5 Minimum relation confidence As Creighton and Hanash (2003) proposed, this study focused on the rule $S_x => S_y$, where S_x has a single item, as to limit the search space of candidate rules. In the *Minimum relation confidence* (*MRC*), we also used the *shortest distance* between any two genes in the GRN as the relation importance. Based on Equations (3) and (4), the *MRC* of rule $S_x => S_y$ will be the maximum *MRC* of each item-pair between S_x and S_y. Suppose the user-specified *IMC* is 45% and the $MRC(\{G2↑\} => \{G4↑, G5↓\}) = \mathrm{Max}\{MRC(G2↑, G4↑), MRC(G2↑, G5↓)\} = \mathrm{Max}\{39.4875, 46.0125\} = 46.0125$. Therefore, in this example, if the *confidence* value of the rule ($\{G2↑\} => \{G4↑, G5↓\}$) is not less than its corresponding *MRC* (46.0125), the rule would be an eligible association rule.

$$\mathrm{MRC}(S_x \Rightarrow S_y)=\mathrm{Max}\{\mathrm{MRC}(x,y)|x\in S_x \text{ and } y\in S_y, \text{ where } S_x \text{ and } S_y \text{ are itemsets}\} \quad (3)$$

$$\mathrm{MRC}(x,y)=\begin{cases}\mathrm{IMC}+\left(\mathrm{IMC}^{*}\left(\frac{\mathrm{Dist}(x,y)-\mathrm{Avg\_Dist}}{\mathrm{Max\_Dist-Min\_Dist}}\right)^{2}\right) & (\mathrm{Dist}(x,y)-\mathrm{Avg\_Dist})>0\\ \mathrm{IMC}-\left(\mathrm{IMC}^{*}\left(\frac{\mathrm{Dist}(x,y)-\mathrm{Avg\_Dist}}{\mathrm{Max\_Dist-Min\_Dist}}\right)^{2}\right) & \text{Otherwise}\end{cases} \quad (4)$$

2.3 Relational-based multiple minimum supports association rules mining

The *REMMAR* algorithm was proposed to discover gene interaction rules from the gene expression database on relation intensity. As stated above in Section 2.1, *REMMAR* utilizes the main concept of Apriori to improve the efficiency of mining association rules compared with the brute force method. The *REMMAR* algorithm is shown below:

Algorithm *REMMAR*

1. for($k=1; L_{k-1}\neq\phi; k++$) do
2. if ($k=1$) then
3. for all transactions t in database D do
4. L_1={c in C_1|c.count>=$c.IMS$}
5. end
6. else
7. C_k=*candidate-gen*(L_{k-1})
8. for all transactions t in D do
9. C_t=*subset*(C_k,t)
10. for all candidates c in C_t do
11. c.count++
12. end
13. end
14. end
15. L_k={c in C_k|c.count>=$c.MRS$}
16. end
17. Answer=$U_k L_k$

The main process of the *REMMAR* algorithm is described as follows:

(1) Count the *support* of each item in Database D (all samples). If the *support* of the item is greater than, or equal to, the *IMS*, it is called *large* 1-*itemsets* in L_1 (the set that includes all the *large* 1-*itemsets*).

(2) Utilize the *candidate-gen* function to generate *candidate k-itemsets* in C_k (the set that includes all the *candidate k-itemsets*) based on *large* $k-1$-*itemsets* in L_{k-1}.

(3) Utilize the *subset* function to count the *support* of each *candidate k-itemset* in D.

(4) If the *support* of the *candidate k-itemset* in C_k is greater than, or equal to, the *MRS* of this itemset, it is called *large k-itemsets* in L_k.

(5) Increment k by 1 and repeat the procedure from Stage 2 to Stage 4 until *candidate itemsets* can no longer be generated.

After the process of *REMMAR*, the algorithm can generate all *large itemsets* that satisfy the *MRS*. In the next phase, for each set of *large itemset* L and each possible subset A of L, they can generate the association rule $A \rightarrow (L–A)$ that satisfies the *MRC* themselves.

Following the proposal of Agrawal *et al.* (1994), *REMMAR* have two important functions similar to other Apriori-based algorithms; the *candidate-gen* function and the *subset* function. The *candidate-gen* function is used to generate the *candidate k-itemsets* in C_k based on the *large* $k-1$*itemsets* in L_{k-1}.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

candidate-gen

insert into C_k
select p.item$_1$, p.item$_2$, …, p.item$_{k-1}$, q.item$_{k-1}$
from L_{k-1} p, L_{k-1} q
where p.item$_1$=q.item$_1$, …, p.item$_{k-2}$=q.item$_{k-2}$,
p.item$_{k-1}$ <q.item$_{k-1}$

The *subset* function will utilize the *candidate itemsets* generated from the *candidate-gen* function, to count the *support* of each *candidate k-itemset*. If the *candidate itemsets* from C_k are the subset of a transaction t, it will be defined as C_t (*candidate itemsets* in transaction t). The *support* of itemsets from C_t will then be incremented by 1. Each transaction t will be scanned once to measure the *support* of C_k dependent on the *Subset* function when each level of C_k has been generated.

Subset

1. Subset_collection=ϕ
2. for each itemset c in C_k
3. if c is subset of t
4. Subset_collection= Subset_collection Uc
5. return Subset_collection

3 RESULTS AND DISCUSSIONS

This section compares the association rules discovered from the proposed *REMMAR* algorithm to the reference Apriori method, and whether the former is superior at satisfying the biological meaning. The first part presents a brief introduction of the gene regulatory (GR) data, the protein–protein interaction (PPI) data and two yeast-related datasets that were utilized. The second part offers a comparison of the significance of rules *REMMAR* discovered, against of reference Apriori method, to verify with the PPI network. The third part displays the discovered rules that are verified with biological literatures.

3.1 Datasets

The GR and PPI data of *S.cerevisiae* were obtained from the *Saccharomyces* Genome Database (Cherry *et al.*, 1998). The GRN is composed of 5331 genes interacting with one another via 143 668 transcriptional regulation interactions. The protein–protein interaction network (PPIN) comprises 5603 proteins interacting with one another via 85 622 protein physical interactions.

To the best of our knowledge, the microarray profile can be conducted in the three major experiments involving temporal, duplicate and perturbation (McIntosh and Chawla, 2007). In this study, we applied our proposed method on two microarrays involving individual datasets performed by Gasch *et al.* (2000) and Brem *et al.* (2002). The former is a *S.cerevisiae* gene expression dataset on different stress in perturbation experiment, which contains a collection of 173 different stress conditions and a selection of 2993 genes. The latter is a *S.cerevisiae* gene expression dataset in duplicate on yeast segregation experiment. The dataset contains 6229 genes and 40 segregates with dye swap are included.

3.2 Evaluation with PPIN

In the real case study, the *shortest distance* between any two genes in the GRN was used as the relation importance to discover association rules. The distances of proteins that have physical interactions are defined as 1 in PPIN. Moreover, the *average shortest distance* (*ASD*) between any two proteins is 2.771 in PPIN. This section shows that the rules discovered with *REMMAR* have stronger biological meanings (has the shorter *ASD* in the PPIN) than those discovered with the reference Apriori algorithm. The focus on rules was on position where the LHS possesses only a single item, as Creighton and Hanash (2003) proposed, to limit the search space of candidate rules in this study. Therefore, if the rules represent their items with $\{gene\ x\uparrow\} => \{gene\ y\downarrow, gene\ z\uparrow\}$, the rule length (RL) is defined as 3 (RL3). In addition, if the *shortest distance* between *gene x* and *gene y* in PPIN is 2, then the *shortest distance* between *gene x* and *gene z* in PPIN is 4. Based on Equation (5), the *ASD* of this rule in PPIN is 3.

$$\text{Average\_Shortest\_Distance}(S_x \Rightarrow S_y) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \text{distance}(x, y_i) \,|\, x \in S_x \text{ and } y \in S_y, \right. \quad (5)$$
$$\left. \text{where } S_x \text{ and } S_y \text{ are itemsets, } n \text{ is the number } of \text{ items in } S_y \right\}$$

We conducted a novel method named *REMMAR*, which automatically adjusted two arguments involving *IMC* and *IMS*. To prove that our improvement is significant, we set *IMC* of our method as 80%, which is same as the *minimum confidence* value introduced in the primitive Apriori algorithm of a previous approach (Creighton and Hanash, 2003). On the other hand, we empirically set the values of *IMS* and *minimum support* since the aim of this research is to manually evaluate the mined rules by either *REMMAR* or Apriori with literature. By decreasing the values (lower threasholds) of *IMS* for *REMMAR* and *minimum support* for Apriori, a large number of satisfied rules will be mined. This arises a big problem that comprehensive evaluations of these massive rules would be beyond our ability. On the contrary, few rules will be mined when we increased the parameter values (higher threasholds). In facing this trade-off, in this study, we set both the *IMS* value for *REMMAR* and the *minimum support* value for Apriori as 60% in the first dataset (Gasch *et al.*, 2000) and 92% in the second dataset (Brem *et al.*, 2002).

Table 1 indicates the resultant rule numbers under different methods (Apriori and *REMMAR*) and different RLs in the first dataset. Here 'Common' represents the intersection of the results produced by Apriori and *REMMAR* algorithms. 'Apriori (no common)' and '*REMMAR* (no common)' represent the respective results without the common intersection. In other words, all rules within the respective results are unique. From RL2 to RL7, both Apriori and *REMMAR* generated 42 739 (including common) and 53 616 (including common) rules, respectively. In all, 37 771 of the discovered rules appear in both generated rule sets. From RL2 to RL7, only a difference of 4968 and 15 845 rules exists. Furthermore, Table 2 indicates the *ASD* under different methods and different RLs in the first dataset. In this table, *REMMAR* (no common) can discover shorter *ASD* rules in PPIN under any RL in contrast with Apriori (no common). The *ASD* (2.173) of a total number of association rules mined by *REMMAR* (no common) is better than

Table 1. Rule number comparison under different methods and RL in the first dataset

	RL2	RL3	RL4	RL5	RL6	RL7	Total
Apriori (no common)	242	668	1813	1679	522	44	4968
Common	3307	11 519	14 438	7062	1372	73	37 771
REMMAR (no common)	281	3476	6078	4611	1253	146	15 845

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

that (2.288) of the Apriori (no common). Therefore, in the result of this dataset, *REMMAR* can evidently mine not only more rules, but also rules of more importance in real biological significance than traditional association rule mining methods.

Table 3 indicates the result rule numbers under different methods (Apriori and *REMMAR*) and different RLs in the second dataset. From RL2 to RL14, Apriori and *REMMAR* generated 613 799 (including common) and 541 165 (including common) rules, respectively. However, 541 133 rules are shared in both generated rule sets. Hence, from RL2 to RL14, only 72 666 and 32 rules are different. Moreover, Table 4 indicates the *ASD* under different methods and different RLs in the second dataset. In this table, *REMMAR* (no common) can discover shorter *ASD* rules in PPIN under any RL than Apriori (no common). The *ASD* (2.467) of a total number of association rules mined by *REMMAR* (no common) is better than that (2.873) of the Apriori (no common). The *ASD* 2.873 of Apriori (no common) is inferior to the *ASD* 2.771 between any two proteins in PPIN. Therefore, in the result of this dataset, *REMMAR* was evidently not generating many rules that have no real biological meanings.

3.3 Evaluation with literature

After evaluating the correlation of associated rules with *ASD* in PPIN, *REMMAR*, proposed in this study, has a lower *ASD* compared with the Apriori method. Genes involved in association rules have a high probability for representing an interaction at protein level, which includes three main types: positive regulation, negative regulation or physical interaction. However, the PPIN does not contain this information. This study applied *REMMAR* on two yeast-related datasets, composed of differentially expressed values of probes. Upregulation and downregulation of a gene in different conditions are represented as two items within a transaction database (see Section 2).

Both Apriori and *REMMAR* algorithms generated numerous rules from these two datasets, as shown in Tables 1 and 3. Verifying all these rules with literature is a difficult task. Therefore, if our identified rules have physical interactions in PPIN ($ASD = 1$), the rules whose regulations involve positive or negative regulations would be further verified with the literature. We showed the results in Table 5. In this table, we want to test whether these association rules are highly informative, especially in the biological domain. In previous studies, an increasing amount of literature has reported enormous critical regulations for gene pairs in certain conditions. These published papers were manually collected as ground truth for validating rules identified from these two datasets. In the Supplementary Material, the rules in each dataset in the beginning will be disassembled into a length of two as relations for verifying with the literature. For example, {YLL039C_Up} => {YDR074W_Up, YHR089C_Down} will be disassembled into {YLL039C_Up} => {YDR074W_Up} and {YLL039C_Up} => {YHR089C_Down} two disassembled relations. If prior studies have reported a disassembled relation, a positive serial number of corresponding papers will be assigned in its entry in the References column. On the contrary, the serial numbers in negative represent an opposite way between the relations and the statement of literature. For instance, for the first dataset results, we illustrate two relations: {YJL164C_Up} => {YPL203W_Up} and {YPL203W_Up} => {YJL164C_Up}. The former states that the overexpressed gene YPL203W is dependent on the upregulation of gene YJL164C, and

Table 2. *Average shortest distance* of rules in PPIN under different methods and RL in the first dataset

	RL2	RL3	RL4	RL5	RL6	RL7	Total
Apriori (no common)	2.384	2.292	2.311	2.271	2.222	2.178	2.288
Common	2.271	2.181	2.149	2.112	2.087	2.089	2.160
REMMAR (no common)	2.256	2.218	2.176	2.149	2.110	2.102	2.173

Table 5. Rule number comparison under different methods and RLs in the first and second datasets

	First dataset					Second dataset	
	RL2	RL3	RL4	RL5	Total	RL2	Total
Apriori (no common)	8	0	0	0	8	0	0
Common	104	34	8	2	148	2	2
REMMAR (no common)	14	14	5	0	33	2	2

Table 3. Rule number comparison under different methods and RLs in the second dataset

	RL2	RL3	RL4	RL5	RL6	RL7	RL8	RL9	RL10	RL11	RL12	RL13	RL14	Total
Apriori (no common)	140	1059	3620	7895	12 366	14 651	13 568	9999	5820	2574	804	156	14	72 666
Common	1524	11 001	40 200	87 520	124 086	121 121	84 976	44 739	18 400	5940	1404	208	14	541 133
REMMAR (no common)	32	0	0	0	0	0	0	0	0	0	0	0	0	32

Table 4. *Average shortest distance* of rules in PPIN under different methods and RLs in the second dataset

	RL2	RL3	RL4	RL5	RL6	RL7	RL8	RL9	RL10	RL11	RL12	RL13	RL14	Total
Apriori (no common)	2.636	2.741	2.791	2.834	2.865	2.886	2.896	2.897	2.890	2.878	2.867	2.856	2.846	2.873
Common	2.694	2.724	2.744	2.760	2.774	2.787	2.799	2.808	2.816	2.821	2.825	2.829	2.835	2.780
REMMAR (no common)	2.467	0	0	0	0	0	0	0	0	0	0	0	0	2.467

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

vice versa for the latter. In terms of this bidirectional regulation, these two genes may have a high cooperation or a similar function in a bioprocess. Interestingly, Toda *et al.* (1987) had identified that these two candidates can encode catalytic subunits of cAMP-dependent protein kinases, and be encoded by the *TPK* genes in 1987. A similar result showed that these two genes play critical roles for the maintenance of iron levels (Robertson *et al.*, 2000). For biologists, the most important issue is how these relations provide valuable information. The *REMMAR* method identified a novel relation between a pair of genes; {YCR057C_Down} => {YGL171W_Down}. To date, the relation has not been reported. To check the reliability of the novel regulation relation '{YCR057C_Down} => {YGL171W_Down}' discovered in this study, we get insights into the biological functions of these two genes through a literature survey. Both proteins, YCR057C and YGL171W, have been reported in the assembly of the 90S preribosomal particle to drive the process of ribosome biosynthesis (Dosil and Bustelo, 2004; Grandi *et al.*, 2002). Moreover, it has also been reported that they played functional roles involved in rRNA biogenesis in the nucleolus (Dragon *et al.*, 2002; Venema and Tollervey, 1999). However, to the best of our knowledge, it has yet been reported that both proteins directly regulate each other even though they would colocalize with ribosomes and nucleolus. Therefore, we can provide biologists with this kind of gene regulation relation with high reliance in validation. Moreover, it has been reported that YCR057C could predominantly accumulate in a complex with YJL069C (Dosil and Bustelo, 2004). This relation can also be found in our results. Hence, this study provides a novel relation for gene pairs among a total number of 34 disassembled relations. All remaining relations in the two datasets coincided with previous studies in certain conditions of yeast species. Therefore, when comparing to the reference algorithm in this study, the advantages of applying *REMMAR* to automatically adjust *MRS* yield a *precision* of 100% (33/33, without considering the novel one), rather than Apriori's 87.5% (7/8). We can find an incorrect disassembled relation, {YPL154C_Up} => {YMR174C_Up}, which does not correspond to the contents of previous literature because Pai3p (YMR174C) was reported as a specific Pep4p (YPL154C) inhibitor (Phylip *et al.*, 2001). However, since a pair of genes with long distance may involve in different pathways or play entirely different roles in a cell, we only verified the part of both methods that identified distinctive rules, whose *ASD* values are equal to 1 in PPIN. If the *ASD* of an identified rule is >1, this rule would not be considered for verification in this study. Definitely, there exist putative gene regulations that have yet been identified. Besides, all the rules and disassembled relations evaluated with literature are unique, as shown in the Supplementary Material. In fact, we found 33 verified disassembled relations that did not appear in the results by the reference Apriori algorithm. The results of this study suggest that the *REMMAR* method has a significant improvement, which can identify more association genes from microarray data.

4 CONCLUSION

In this study, we proposed the *REMMAR* algorithm to tackle the inefficiency of previous methods, which never took into account the importance of each item. Moreover, for discovering more meaningful rules in biology, the proposed *REMMAR* would adjust the *MRS* for each individual gene pair according to the intensities of two neighboring genes in the PPIN. Not only did we improve the traditional association algorithm with a multiple support mechanism called *REMMAR*, but it was also applied on two yeast datasets for mining association genes in different conditions. The *shortest distance* between any two genes in the GRN was used as the relation importance in this study. In the first part of the result evaluation, the proposed method not only generated more rules with stronger relation intensity, but also filtered rules without biological meaning in the PPIN. In the second part of the evaluation, the part of the discovered rules was verified with biological literature. The proposed method has a *precision* of 100%, whereas the reference Apriori method exhibits a *precision* of 87.5%. Therefore, the proposed method is more effective in helping biologists explore the relationships of biological activities.

Funding: National Science Council, Taiwan, R.O.C. (grant no. NSC 99-311-B-006-003 and NSC 100-2627-B-006-022).

Conflict of Interest: none declared.

REFERENCES

Agrawal,R. *et al.* (1993) Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM Press, Washington, D.C., USA, pp. 207–216.

Agrawal,R. and Srikant,R. (1994) Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*. Morgan Kaufmann, Santiago de Chile, Chile, pp. 487–499.

Alves,R. *et al.* (2010) Gene association analysis: a survey of frequent pattern mining from gene expression data. *Brief. Bioinformatics*, **11**, 210–224.

Bhasin,M. and Raghava,G.P.S. (2004) SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics*, **20**, 421–423.

Brem,R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Cai,C.H. *et al.* (1998) Mining association rules with weighted items. In *Proceedings of the International Database Engineering and Applications Symposium*. IEEE Computer Society, Cardiff, Wales, UK, pp. 68–77.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California, USA, pp. 93–103.

Cherry,J.M. *et al.* (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.

Creighton,C. and Hanash,S. (2003) Mining gene expression databases for association rules. *Bioinformatics*, **19**, 79–86.

Dosil,M. and Bustelo,X.R. (2004) Functional characterization of Pwp2, a WD family protein essential for the assembly of the 90 S pre-ribosomal particle. *J. Biol. Chem.*, **279**, 37385–37397.

Dragon,F. *et al.* (2002) A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature*, **417**, 967–970.

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

Georgii,E. *et al.* (2005) Analyzing microarray data using quantitative association rules. *Bioinformatics*, **21**, 123–129.

Grandi,P. *et al.* (2002) 90S pre-ribosomes include the 35S pre-rRNA, the U3 snoRNP, and 40S subunit processing factors but predominantly lack 60S synthesis factors. *Mol. Cell*, **10**, 105–115.

Jiang,D. *et al.* (2004) Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.*, **16**, 1370–1386.

Liu,B. *et al.* (1999) Mining association rules with multiple minimum supports. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Diego, CA, USA, pp. 337-341.

Liu,K.H. and Xu,C.G. (2009) A genetic programming-based approach to the classification of multiclass microarray datasets. *Bioinformatics*, **25**, 331–337.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE Trans. Comput. Biol. Bioinformatics*, **1**, 24–45.

Martella,F. (2009) Classification of microarray data with factor mixture models. *Bioinformatics*, **22**, 202–208.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Martinez,R. *et al.* (2008) GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. *Bioinformatics*, **24**, 2643–2644.

McIntosh,T. and Chawla,S. (2007) High confidence rule mining for microarray analysis. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **4**, 611–623.

Pei,J. *et al.* (2003) MaPle: a fast algorithm for maximal pattern-based clustering. In *Proceedings of the 3rd IEEE International Conference on Data Mining*. IEEE Computer Society, Melbourne, Florida, USA, pp. 259–266.

Phylip,L.H., *et al.* (2001) The potency and specificity of the interaction between the IA3 inhibitor and its target aspartic proteinase from Saccharomyces cerevisiae. *J. Biol. Chem.*, **276**, 2023–2030.

Prelic,A. *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Ramkumar,G.D. *et al.* (1998) Weighted association rules: model and algorithm. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Robertson,L.S., *et al.* (2000) The yeast A kinases differentially regulate iron uptake and respiratory function. *Proc. Natl Acad. Sci. USA*, **97**, 5984–5988.

Su,J.H. *et al.* (2008) Effective ranking and recommendation on web page retrieval by integrating association mining and Pagerank. In *Proceedings of the Workshop on Optimization-Based Data Mining and Web Intelligence*. IEEE Computer Society, Sydney, NSW, Australia, pp. 455–458.

Tao,F. *et al.* (2003) Weighted association rule mining using weighted support and significance framework. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Washington, DC, USA, pp. 661–666.

Thalamuthu,A. *et al.* (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–2412.

Toda,T. *et al.* (1987) Three different genes in S. cerevisiae encode the catalytic subunits of the cAMP-dependent protein kinase. *Cell*, **50**, 277–287.

Tseng,V. *et al.* (2010) UP-Growth: an efficient algorithm for high utility itemsets Mining. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Washington, DC, USA, pp. 253–262.

Venema,J. and Tollervey,D. (1999) Ribosome synthesis in Saccharomyces cerevisiae. *Annu. Rev. Genet.*, **33**, 261–311.

Wang,W. *et al.* (2000) Efficient mining of weighted association rules. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Boston, MA, USA, pp. 270–274.

Yun,U. and Leggett,J.J. (2005) WFIM: weighted itemset mining with a weight range and a minimum weight. In *Proceedings of the SIAM International Data Mining Conference*. Society of Industrial and Applied Mathematics, Newport Beach, California, USA, pp. 270–274.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016