Mining Association Rules in Analysis of Transcription Factors Essential to Gene Expressions

Ruzhu Chen , Qiyou Jiang, Honglin Yuan and Le Gruenwald**
School of Computer Science,
The University of Oklahoma
Norman, OK 73019

ABSTRACT

Knowledge discovery from gene expression databases has become an important research area for biologists since the growing number of gene sequences was obtained. This paper studies the transcription factor(s) required for expression of the target genes using data mining association rule techniques. To apply the association rules to mine the transcription factors essential to certain gene expressions, we defined each type of tissues as a set of transactions or a dataset. Each dataset consists of transcription factors and the target genes. The Apriori mining algorithm was prototyped and the gene sequence data were tested. The results were obtained by pruning the itemsets before and after applying the Apriori algorithm, in which the false results were eliminated. The data items (transcription factors) obtained from this program were compared with those data obtained through experimental research. The comparison results indicated that it may be effective to apply data mining association rules to obtain transcription factors associated with gene expressions.

INTRODUCTION

Gene sequence databases have been updated frequently since the increased number of gene sequence data was obtained in recent years. The knowledge discovery from these databases has become a major research area (bioinformatics) in biology. Using computation techniques such as data mining to find the association relationship among these gene data is of great interest and challenging aspect. With experimental research methods such as gene knockout and microarray, biologists have been able to understand the regulation of certain gene expressions. However, experimental research is time consuming, and does not use the currently available gene sequence data. Using data mining techniques, especially association rule techniques, we may discover some regulation elements (transcription factors) that are essential to the expression of a gene in very short time. The goal of our research is to obtain the transcription factors required for the gene expression of a target gene in the databases. Transcription factors consist of a large number of proteins that were classified into different families.

This project reported the results and analysis of mining association rules in gene expression patterns. We studied the association rules between transcription factors and genes, designed and implemented an efficient algorithm of mining association rules for the analysis, finally generated datasets to test this algorithm.

RELATED RESEARCH WORK

1). Introduction to the study of gene sequence and expression

A biological process usually is a cascade of reactions involving more than one factor or protein participated. Gene expression in eukaryote, for instance, is triggered by a series of factors binding to the regulatory domains resulting in gene transcription. The breakthrough of modern molecular biology techniques enables us to obtain and understand the nature of biomolecules. Using automatic gene sequencing tools, scientists world-wide have been producing a large number of gene sequences, which are stored in a number of databases. However, the understanding of these large sequence data is fallen behind. The recently developed microarray technique has prompted molecular biologists to illustrate different genes expressed in the same and different tissues (Morishita *et al.,* 1999).

Knowledge discovery from these data is becoming an interested research area for biologists. For gene sequence analysis, as stated by Brazma (2000), we can study the similarities of gene expression profiles, classification of gene sequences into functional classes and the correlation between functional classes and expression levels. From the gene expression data sources, we may discover the association relationship among genes and thus resulting in understanding the regulation of gene expression. We may also obtain information of expression patterns of similar genes.

2). Current research in gene expression analysis using data mining

The sequence data provided in gene databases are individually stored but without any clustering and classification based on any category (Lawson, 1999).

** Contact Author: Le Gruenwald, ggruenwald@ou.edu

While the experimental molecular methodology continues to collect gene sequence and expression data, it also provides information of gene expression patterns (Lin *et al.*, 1998; Smith and Hager, 1998). However, this methodology is time consuming, expensive and is not taking the advantage of using the available large amount of data presented in various databases. Traditional data analysis techniques, such as pair-wise matching, EST (Expression Sequence Tags) and statistical analysis, may help to understand the similarity relationship among genes. Data mining as a new computational technique has been used increasingly in exploring information from gene sequence databases (Satou *et al.*,1997; Brazma *et al*., 1998; Morishita *et al*., 1999). Most of the work in mining the biological data focused on the analysis of biosequences using clustering, decision tree and visualization (Brazma, 1999). Using association rules, Nakaya *et al*. (2000) has successfully analyzed the association of oral glucose tolerance with the target genes. Satou *et al.*(1997) analyzed the association of protein structures and function using mining association rules. As proposed by Morishita (1999), the association rules can also be applied to study the association of transcription factors with their target gene. These results indicated that the application of data mining techniques is feasible and data mining will be among the leading analysis tools in gene analysis and bioinformatics.

3). Mining association rules and gene expression analysis

Using the association rule approach, we can analyze 1). The expression of one gene leads to the induction of a serial of target gene expressions. This expression pattern is denoted regulation of gene expression. The relationship between one gene and the other target genes can be viewed as an associative relation. 2). Several gene expressions lead to the expression of one target gene. Transcription factors and their target gene is one of many examples in this category (Morishita, 1999). 3). Gene expression leads to the induction of a new biological function (Nakaya *et al*., 2000).

Several algorithms of association rules have been proposed. Apriori (Agrawal and Skikant, 1994) constructs a candidate set of large itemsets, counts the number of occurrences of each candidate itemset, and then determines large itemsets based on a predetermined minimum support. Apriori requires c+1 passes to generate the large itemsets where c is the maximal cardinality of a large item set (Hidber, 1999). DHP (Park at el., 1995) employs a hash table, which is built in the previous pass, to test the eligibility of a *k*-itemset. DHP adds a k-itemset into the candidate set only if that k-itemset is hashed into a hash entry whose value is larger than or equal to the minimum transaction support required. The 2-pass algorithm by Savasere et al. (1999) partitions the database into blocks and then merges them together after the computation of large itemsets for each block. Carma (Continuous Association Rule Mining Algorithm), recently proposed by Hidber (1999), performs the computation of large itemsets on line, allowing end users to change support at any time and taking at most 2 scans. Nag et al. (1999) proposed an interesting method that uses a knowledge cache to shorten the query time significantly by remembering the results of previous queries and thus minimizing the size of the database that the user works with. Several other variants such as "look ahead", DIC (Dynamic Itemset Counting) (Aggarwal, 1998), constrained-based and multidimensional data mining (Han et al., 1999), and the one using random sampling to generate the large itemsets (Aggarwal, 1998) were also used by some researchers in many applications.

Many researchers have evaluated the performance of these approaches (Chen et al., 1999; Aggarwal, 1998). Basically, data mining is an application-dependent issue and different applications may require different mining techniques to cope with. To apply mining association rules in gene expression pattern analysis, we need to understand the properties of gene data. The data in a gene expression database are very large, and are divided into different tissues or organs. Most genes are duplicated in different tissues. To study the associations between genes, we need to eliminate these duplicated data since they are present in every tissue (we view all genes in a tissue as a dataset). Thus, to apply an existing algorithm to the gene analysis, we will need to modify it to filter the data. Also, we have to consider the negative implication of the results. Furthermore, unlike data mining in business applications, the size of a transaction in gene analysis is relatively small since the number of tissues (each tissue or organ is viewed as a set of transactions) present in an organism (e.g. human tissues) is limited. However, the number of items (genes) in one single transaction is very large. When we select an algorithm to facilitate this analysis, the number of passes is not a major factor to be considered. Therefore, in our study, we will use the algorithm Apriori as the basic algorithm with some necessary modifications.

In this study, we will analyze the relationship of transcription factors and their target genes, especially

those transcription factor(s) essential to the target gene expression. The association rules of transcription factors and their target genes were outlined by Morishita (1999), but no data mining association of the transcription factors has been studied. Our results will provide a computation approach to study the gene transcription regulation.

RESEARCH METHODOLOGY

1). Apriori algorithm

The Apriori algorithm (Agrawal and Srikant, 1994) was used as the fundamental algorithm in this study. Apriori is the most recognized algorithm in mining association rules. It has become the base algorithm for many newly developed algorithms, such as DHP (Park et al., 1995), Partitioning (Savasere et al., 1999), and DIC (Aggarwal, 1998). The algorithm gives more reliable results although it takes more run time. The reliability of the resulting relationship among genes is important and is what we are mostly interested in. As we stated in Section 2, the number of passes is not a major factor that we need to consider here because of the relatively small number of transactions. Most of the other algorithms we mentioned above focusing on decreasing the number of passes. Empirical evaluation shows the rationality of Apriori (Agrawal and Srikant, 1994).

The problem of mining association rules usually includes two steps (Chen et al., 1996): 1). Compute the large datasets with transaction support above the user defined one, and 2). Generate the association rules from the large datasets. The basic idea for Apriori is that any subset of a large dataset must be large. Thus, the candidate datasets having k items can be generated by joining large datasets having k-1 items, and deleting those that contain any subset that is not large.

2). Association studies of the transcription factors participating in gene expression patterns

Gene expression data are stored in the database grouped by the tissues or organs where they are present. A gene (an item) in the database is identified by its access number and name. A gene with the same access number can exist in different tissues or organs. In gene databases, a gene data consists of three attributes: a unique access number, name of the gene, and the sequence of the gene. To investigate the association of genes in the database, we need to understand the properties of gene expression. Most genes are expressed consistently in every tissue and organ, which are defined as housekeeping genes. Some genes are induced to express by other gene expressions or factors. The rules of transcription factors (x_1, x_2,…,x_n) and their target gene (y) are defined as follows:

1). If x_s (one or more) exists $\Rightarrow$ y exists in the dataset;
2). If one of the x_s does not exists $\Rightarrow$ y does not exist in the dataset.
3). An x_s exists $\Rightarrow$ different y exists, i.e., a transcription factor may participate in different target gene expressions.

Definition of dataset

To apply the association rules for mining the transcription factors of the target gene, we define each type of tissues as a set of transactions or a dataset (e.g. HL60, one of the blood tissues). In a dataset, each tissue sample that consists of many genes (transcription factors and target genes) is viewed as one transaction (Table 1).

Table 1. An example of datasets.

tissue 1 (trans 1)	tissue 2 (trans 2)	tissue 3 (trans 3)	…	…	…	tissue n (trans n)
Tf_{11}	Tf_{21}	Tf_{31}	…	…	…	Tf_{n1}
Tf_{12}	Tf_{22}	Tf_{32}	…	…	…	Tf_{n2}
Tf_{13}	Tf_{23}	Tf_{33}	…	…	…	Tf_{n3}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

(Note: The sizes of transactions are not necessarily the same for all transactions.)

3). Program design and implementation using the Apriori algorithm to analyze the association rules in gene data.

In this study, we mainly focused on the transcription associated gene expression patterns. Based on the rules of gene data described previously, our application program was designed to preprocess the data from the database prior to applying the Apriori algorithm as follows. First, the data items were read from the database and the datasets were generated by extracting all the transcription factors and the targets genes from the database. The datasets were then divided into two distinct groups of datasets. One group of itemsets consists of the corresponding target gene (Y), while the other group does not contain Y. The datasets with Y were proceeded to algorithm analysis. The primary results from the algorithm were then compared with the group of datasets without the target gene. If a primary result is in any dataset of this group, then this result is false and removed from the primary results. Finally, the expected results were generated and printed out to screen. Figure 1 shows the data flow in our program.

We wrote an analysis program prototyping the Apriori algorithm in the C++ language. Both candidate datasets (C) and large datasets (L) were

designed as linked lists. Since linked lists can freely grow and shrink based on the sizes of different problems, this program has the merit of generality**.**

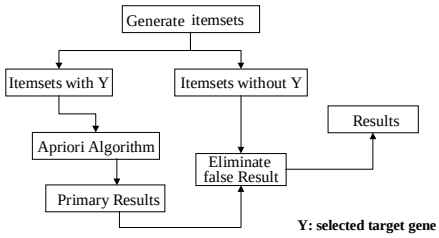

Figure 1. Flowchart of our program

RESEARCH RESULTS AND DISCUSSION

1). Data process and transcription factor extraction

To find the association relationship of transcription factors and their target genes, we used gene sequence data available in the *bodymap* database (Hishiki *et al.,* 2000). A total of 41 transcription factors are extracted from the tissues viewed as itemsets consisting of different items. To select target genes to be executed in our program, ten of genes were randomly selected from the database.

2). Data output from our program

Our program reads a given target gene access number from the input file. After execution, the output is the expected transcription factor gene access numbers. We used a small value of minimum support since the target gene is only present in a few datasets. The results obtained with minimum support = 0.25 were summarized in table 2. When using other large minimum supports to analyze the data, the expected results were not found.

Table 2. Results of our program corresponding to the target genes.

Y (input)	X_s (results)
V00491	M84810
U11276	X51345, X51346, U15410
M13577	X62829, X53145, X51346
X04803	L42856
V00497	X53145,X74070
X00373	X53145,X74070
S54005	X53145, X62585
J03040	L19067, M42856, M62831
M31520	X62585, X53145
M17987	M62831, X62585, X53145

4). Result analysis

The output results from our program were matched to the corresponding names of genes by searching the database (gene bank) using the gene access numbers. To determine whether these results were positively associated with the corresponding transcription factors, we compared the results with the available experimental research data. Table 3 concluded the association rules of the transcription factors and their target genes. It shows that, for each target gene, there is an association rule between the target gene and transcription factors obtained from our program. Four of the results were verified by the experimental research results. Due to the incomplete data of transcription factors in the database (more than 100 transcription factors were reported in human genome), our results were unable to include all the essential transcription factors associated with the target gene. Also, more than half of our results are still waiting for confirmation because of lack of complete experimental data.

We used known transcription factors in the database to study the target genes which have association relationships with the transcription factors. The results from our program showed that data mining in analysis of gene expression patterns can be an efficient way to discover the relationships among gene sequence data. The data mining results provide two ways for biologists in research on gene expression patterns. First, for a known gene, all its associated genes can be extracted from the databases by data mining. It saves tremendous time for scientists because they only need to identify (or evaluate) the candidate associated genes instead of going through the whole time-consuming, untargeted experimental research. Second, when gene is unknown (X and Y), the method gives a number of result sets which provide information for new association research and thus may lead to a breakthrough in finding new expression patterns.

Conclusion

We have studied the association rules between a given transcription factors gene and its target genes using the Apriori algorithm. Our program obtained the positive associations among genes, although some results may still contained non-associated transcription factors. It takes much less time comparing with the experimental methods. These results implied that applying the Apriori mining technique in gene analysis is feasible. The Apriori algorithm is effective in discovering gene expression patterns comparing to the experimental method.

To obtain more accurate results in gene data mining, we still need to improve the implementation so that the non-associated target genes can be eliminated.

Also, the time required for computation of large transactions in our program should be minimized.

Table 3. Determination of association rules between transcription factors and target genes using data mining and experimental research methods.

Transcription factors (Y)	Output from program (Xs)	Association determined by experimental methods
Human alpha-globin	α-globin factor CP2	Yes (chae *et al.*,1999)
hNKR-P1a protein (NKR-P1A)	Transcription factor jun-B, Transcription factor jun-D, Transcription factor NTAFx	Yes (Persico *et al., 1995*) Yes (Persico *et al., 1995*) Yes (Amasaki *et al.*, 1998)
Myelin basic Protein (MBP)	Transcription factor ETR 103, Transcription factor jun-B Transcription factor jun-D,	Yes (Yamaguchi *et al.*, 1998 and Shimizu *et al.*, 1992) Yes (Persico *et al., 1995*) Yes (Persico *et al., 1995*)
Ubiquitin	Transcription factor SIII p18	Yes (Garrett *et al.*, 1995)
haptoglobin α 1S (Hpa 1S)	Transcription factor jun-B, Transcription factor BTF-3	? ?
Human myoglobin gene	Transcription factor jun-B, Transcription factor BTF-3	? ?
Thymosin beta-10	Transcription factor jun-B, Transcription elongation factor TFIIS-1	? ?
Osteonectin	NF-kappa-B transcription factor p65, Transcription elongation factor TFIIS-1, Transcription factor ETR103	? ? ?
Ribosomal protein S24	Transcription elongation factor TFIIS-1, Transcription factor jun-B	? ?
Beta-2-microglobulin	Transcription factor ETR103, Transcription elongation factor TFIIS-1, Transcription factor jun-B	? ? ?

References

Aggarwal, Charu, Yu, Philip: *Bulletin of the IEEE Technical Committee on Data Engineering*, Vol 21, No.1, Page 23-31, March 1998.

Agrawal R., and Srikant, R.: *Proc. 20th Int'l Conf. Very Large Data Bases*, Sept. 1994, pp.478-499.

Amasaki Y ; Masuda ES ; Imamura R ; Arai K ; Arai N: *JOURNAL OF IMMUNOLOGY* 1998, 160 (5): 2324-33.

Brazma, Alvis, 2000: *In*: http://industry.ebi.ac.uk/ ~brazma/dm.html. [visited on March 05, 2000].

Brazma, A., I. Jonassen, I.Eidhammer, D. Gilbert. 1998: *Journal of Computational Biology*, Vol 5(2), pp. 277-303.

Brazma, A. 1999: In: *Bioinformer* (http: //bioinformer.ebi.ac.uk/newsletter/archives/4/lead_article.html). (visited on March 05, 2000).

Chae J.H., Lee Y. H. and Kim C. G., 1999: *Biochem. Biophys. Res. Commun*. 263:580-583.

Chen, Ming-Syan, Han, Jiawei, and Yu, Philip: *IEEE Transactions on knowledge and Data Engineering*, Vol.8, No. 6, December 1996, pp.866-883.

Garrett KP ; Aso T ; Bradsher JN ; Foundling SI ; Lane WS ; Conaway RC ; Conaway JW: *Proc Natl Acad Sci U S A* 1995, 92 (16): 7172-6.

Han, Jiawei, Lakshmanan, Laks V.S., and Ng, Raymond T.: *Computer*, August 1999, pp 46-50.

Hidber, Christian: *ACM SIGMOD Conference on Management of Data*, May 1999.

Hishiki, J. Sese, A. Nakaya and S. Morishita, 2000: In*: http://bodymap.ims.u-tokyo.ac.jp.* (visited Feb 28,2000).

Lawson, D.: *PARASITOLOGY*, vol. 118, S15-S18, 1999.

Lin Richard Z, Chen Jin, Hu, Zhuo-Wei, and Hoffman Brian B, 1998: Jounal of Biological Chemistry, vol. 273 (45), pp30033-30038.

Morishita, Shinichi , Hishiki, eruyoshi and Okubo, Kousaku : *Proceedings of 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (DMKD), pages 21-25, June 1999.

Morishita, Shinichi and Jun Sese, 2000: In: *Proc. of ACM SIGACT-SIGMOD-SIGART Symposium on Database Systems* (PODS),2000.

Nag, Biswadeep, Deshpande, Prasad, DeWitt, David: *ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, Aug. 1999.

Nakaya, Akihiro, Hishigaki, Harutsugu and Morishita, Shinichi: In Proc. of Pacific Symposium on Biocomputing, pages 367-379, January 4-9, 2000.

Park, J.-S., Chen, M.-S., and Yu P.S.: Proc. ACM SIGMOD, May 1995, pp.175-186.

Persico AM ; Schindler CW ; Zaczek R ; Brannock MT ; Uhl GR : *Synapse* 1995, 19 (3): 212-27

Satou K ; Ono T ; Yamamura Y ; Furuichi E ; Kuhara S ,1997: Takagi T: *Ismb* (ISMB) ; 5: 254-7

Smith, C.L. and Hager,GL, 1997: *Journal of Biological Chemistry* 272(44), pp27493-27497.

Shimizu N ; Ohta M ; Fujiwara C ; Sagara J ; Mochizuki N ; Oda T ; Utiyama H: *JOURNAL OF BIOCHEMISTRY* 1992, 111 (2): 272-7.

Thornburn, A. W., 2000: http://molbio.med.utah. edu/thorburn /thorburn.html (visited on 3-7-2000).

Yamaguchi Y ; Nishio H ; Kasahara T ; Ackerman SJ; Koyanagi H ; Suda T: *Leukemia* 1998, 12 (9): 1430-9