Mining Time-delayed Gene Regulation Patterns from Gene Expression Data

Huang-Cheng Kuo and Pei-Cheng Tsai

***Abstract*—Discovered gene regulation networks are very helpful to predict unknown gene functions. The activating and deactivating relations between genes and genes are mined from microarray gene expression data. There are evidences showing that multiple time units delay exist in a gene regulation process. Association rule mining technique is very suitable for finding regulation relations among genes. However, current association rule mining techniques cannot handle temporally ordered transactions. We propose a modified association rule mining technique for efficiently discovering time-delayed regulation relationships among genes.**

By analyzing gene expression data, we can discover gene relations. Thus, we use modified association rule to mine gene regulation patterns. Our proposed method, BC3, is designed to mine time-delayed gene regulation patterns with length 3 from time series gene expression data. However, the front two items are regulators, and the last item is their affecting target. First we use Apriori to find frequent 2-itemset in order to figure backward to BL1. The Apriori mined the frequent 2-itemset in the same time point, so we make the L2 split to length one for having relation in the same time point. Then we combine BL1 with L1 to a new ordered-set BC2 with time-delayed relations. After pruning BC2 with the threshold, BL2 is derived. The results are worked out by BL2 joining itself to BC3, and sifting BL3 from BC3. We use yeast gene expression data to evaluate our method and analyze the results to show our work is efficient.

***Index Terms*—Association rule, Data mining, Genetic expression, Gene regulation**

I. Introduction

DNA microarray experiments measure the expression levels of genes during biological processes. A microarray is a chip with probes on glass or plastic, which is high-throughput technology for molecular biology [18]. DNA microarrays can be used to measure changes in expression levels. By analyzing gene expression data from microarray, we can uncover some relations between genes. These relations can be important when verifying them with biological evidences. The relations are found by computational algorithms without biological evidence. It may provide a direction for biologists to observe and discuss genetic relations. Gene regulation networks are widely used to predict unknown functions of genes. Thus, gene regulation networks help us confirm not only the interaction between genes, but also functions of genes. There is much formalism to describe a genetic regulatory system [9][14]. A Bayesian network is a common structure to model gene regulation network [19][6][25]. We use mining algorithms to find the time-delayed patterns of gene regulation.

A gene is up-regulated or down-regulated by proteins of other genes. When a gene is up-regulated, it takes time to transcribe the gene to mRNA, and translates the mRNA to protein for gene expression processes. Whereas, when a gene is down-regulated, it also takes time to decrease copies of mRNA's. The amount of time needed for a gene to response the regulation varies from gene to gene. Applying time-delayed property is important to model relations between regulations of genes [20][24][28].

Analyzing gene expression data can provide us information for recognizing the positive and negative regulations of genes. However, known methods for mining association rules are computationally expensive to find frequent patterns that include time-delayed items. We introduce a method to find time-delayed gene regulation patterns and obtain the patterns we desire for the co-regulation of genes. Since the method is tailored for this special application, thus it is faster than the Apriori algorithm.

II. Related Works

A. Association Rule

Association rule mining is a well studied method for discovering relations in large databases. Here we apply association rule mining to model gene relations [2][15]. Apriori is an algorithm for finding frequent itemsets [12][1][21]. Apriori is designed for databases containing transactions of items [10]. Association rule is of the form X => Y, where X and Y are sets of items. It means that when items of X are observed in a transaction, it is very likely to observe items of Y in the transaction. Support and confidence are two well-known interest measures of association rules. In this paper, association rule mining is applied to uncover gene networks.

The support of an itemset X is the number of transactions containing all items in X, denoted as sup(X). The support of an association rule X => Y is sup(X∪ Y). The confidence of the

Manuscript received February 29, 2012. This work was supported in part by the National Science Council under Grant NSC 99-2221-E-415-015.

Huang-Cheng Kuo is with the Department of Computer Science and Information Engineering, National Chiayi University, Chia-Yi City 600, Taiwan (corresponding author to provide phone: 886-5-271-7731; fax: 886-5-271-7741; e-mail: hckuo@mail.ncyu.edu.tw).

Pei-Cheng Tsai is graduated from the Department of Computer Science and Information Engineering, National Chiayi University, Chia-Yi City 600, Taiwan. (e-mail: s095031@mail.ncyu.edu.tw).

DOI: 10.5176_2010-2283_2.1.130

association rule is an association rule X => Y is defined as sup(X ∪ Y)/sup(X), denoted as conf(X,Y). In our utilizing situation, items are expression level of genes with discretization. The discretized gene expression levels are represented with expression labels. We use H, M, and L for expression labels. Label H represents the relatively high gene expression level and label L represents the relatively low gene expression level. Label M is for genes without differential expression levels. The label M genes are not considered, since we are more interested in up-regulated or down-regulated genes. A sequentially numbered gene id is combined with the expression label. For example, for a microarray at a certain time point, H1 means that the expression level of gene 1 is high. Before introducing time-delay, an association rule {H1, L2}=>{L3} means that if expression level of gene 1 is high and expression level of gene 2 is low, then it is very likely the expression level of gene 3 is low. In other words, gene 3 is down-regulated.

B. Gene Regulation Network

We use microarray gene expression data to infer gene regulation relationships. A gene is transcribed into mRNA, and mRNA goes on to make protein. Protein/protein and protein/mRNA interactions affect the production of another protein which is translated from a certain mRNA. For an interaction, the concentration of the protein or mRNA affects the concentration of another. All the proteins and their corresponding genes and the interactions form a network. The edges of the network are interactions, and the nodes are proteins. The network is called gene expression network [4][16][5].

III. Mining the Patterns

A. Framework

Frequent itemsets discovered by current methods do not contain items from different transactions [29]. A series of microarray data, after discretization, is a series of transactions which are temporally ordered. Temporal patterns composed of some items in a transaction and other items in the following transaction are of interest to biologists. The temporal patterns can be further transformed using the association rule format: resulting in some item transactions occurring in later specified time units, and others immediately following the transactions. In this paper, we focus on finding time-delayed patterns which contain up to 2 items in a transaction and an item in the next transaction (next time unit). For example, after introducing time-delayed concept, an association rule {H1, L2}=>{L3} means that if we observe high expression of gene 1 and low expression of gene 2 at a certain time point, then it is very likely to observe low expression of gene 3 at a later time point.

The main idea of our method is to use backward frequent 1-itemset, denoted BL_1. We first find L_2, then find the backward frequent 1-itemset which contains only the items in L_2. The framework of the method is shown in Figure 1.

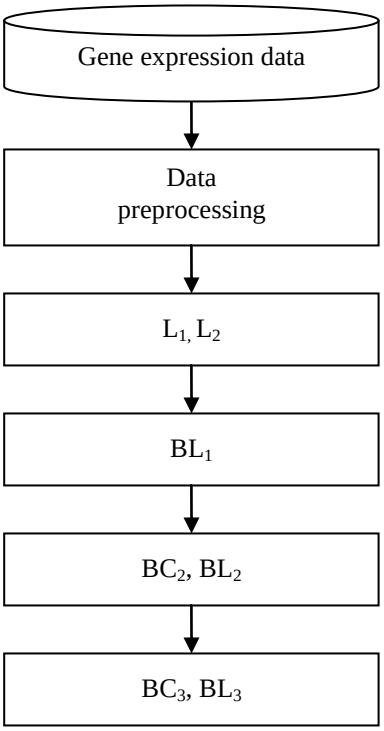

Fig. 1 Framework

B. From Gene Expression Data to Transactions

The form of original gene expression data we use are the experimental results by Campbell [8]. We consider two modes of processing: singular and fluctuation. In the singular mode, an expression profile of a microarray is correspondent to a transaction. In the fluctuation mode, the difference between two temporally adjacent microarray is considered.

In singular mode, gene expression levels of the microarray at time point i are the members of a transaction. Then we discretize the numeric expression value into labels in order to fit association rule mining setting. A transaction ti = {p1, p2, …, pn | n is the number of genes}, where pj is a discretized expression label. In (1), average value and standard deviation of the expression level are computed. An expression is regarded as high if the value is one standard deviation higher than the average. An expression is regarded as low if the value is one standard deviation lower than the average.

$$p_j = \begin{cases} H_j, \text{if value} > \bar{x}_j + S. \\ L_j, \text{if value} < \bar{x}_j - S. \\ M_j, \text{otherwise.} \end{cases} \quad (1)$$

In fluctuation mode, we design another way to preprocess original dataset. A gene is transformed to an H label if the expression level of the gene increases. Otherwise, an L label is assigned to the gene. The formula is in (2).

$$p_j = \begin{cases} H_j, \text{if } x_{j,t} < x_{j,t+1}. \\ L_j, \text{if } x_{j,t} > x_{j,t+1}. \end{cases} \quad (2)$$

In the literature, investigation shows that several genes contribute to a disease. The expression level of a disease related gene is affected when expression levels of some other genes rise or fall [22]. So, instead of considering only a single microarray as a transaction, we would like to consider the fluctuation of genes also.

C. Time-delayed Patterns

For the time being, we consider only the patterns that two genes co-regulate another gene. The time delay is 1 unless otherwise specified. For the above example {H1, L2}=>{L3}, since L3 is at the later time point, in order to distinguish the time point, L3 is prefixed with "D", meaning time-delayed, and is renamed to DL3.

D. Using Apriori to Find L_1 and L_2

The way to find L_1 is the same as in the Apriori algorithm. Our proposed method counts the occurrence of every item in all transactions. An item is an expression label which is either H or L. The process goes on as Apriori until L_2 is obtained.

E. Finding Backward Frequent BL_1, BC_2 and BL_2

BL_1 is defined as the set of itemsets in which each itemset contains an item in L_2. Since the set is obtained backward from L_2, we name the set as BL_1.

Then BC_2= {<a, b> | a∈BL_1, b∈L_1, a≠ b} is defined. BC_2 is a set of ordered sets. The first element of an ordered set is an expression label in BL_1. The second element is from L_1.

After the ordered set with less than minimum supports are pruned from BC_2, BL_2 is obtained. While counting the support for a candidate ordered set, the first element is from a transaction, and the second element is from the temporally next transaction.

F. Finding BC_3 and BL_3

Perform self join on BL_2 to get BC_3. Since we consider patters that two genes co-regulate another gene with time delay, so order of the expression labels in BL_2 is important. Two ordered sets of BL_2 can join if they have the same second item. Patters are those ordered sets in BL_3, which is obtained by pruning itemsets whose supports are less than minimum support in BC_3.

We illustrate the mining process by a short example. Let BL_2 = {<a, d,>, <b, d>, <c, d>}. The BC_3 is <a, b, d>, <a, c, d>, and <b, c, d>. Same as the support counting for BC_2, the first two elements of an ordered set in BC_3 are from a same transaction, and the last element of the ordered set is from the next transaction.

G. Variant Patterns

The time delay can be specified other than 1 time unit. If Δt time delay is given by users, the patterns are defined as (3).

$$BL_3 = \{a, b\} \xrightarrow{\Delta t} \{c\} \quad (3)$$

Genes a and b co-regulate gene c and the effect appears Δt later. The computation for BC_2 and BC_3 are modified such that the last element of the itemset is from a transaction which is Δt time units later.

More complicated patterns can be specified as (4). Different time delays can be specified for gene *a* and gene *b*. We can combine the results for BL_3 from different lengths of time delay. Then, we can discover the patterns of regulators with different lengths of time delay to co-regulate the same target.

$$BL_3 = \{a, b\} \xrightarrow{\Delta t(a),\Delta t(b)} \{c\} \quad (4)$$

The algorithm is modified so that BC_3 is computed by joining $BL_2(\Delta t(a))$ and $BL_2(\Delta t(b))$. While the support counting for a candidate itemset in BC_3 has to consider 3 transactions together.

Association rules are generated by putting the first two elements of a frequent itemset in BL_3 on the left hand side of a rule and the last item of the itemset on the right hand side of the rule. And then check confidence of the rule.

IV. Experimental Results

In this section, we describe the experiment to compare our method with Apriori and FP-growth algorithms. We use the dataset provided by Campbell [8]. There are serial 17 time points (0 to 160 minutes) and 6601 genes in the expression data. In the experiment, the computer has a Pentium D CPU 3.4GHz and 1GB RAM. The proposed method is implemented in Matlab.

A. Association Rules

Some association rules are listed in Tabl I using minimum support 0.7. We notice that in some rules, the same expression labels show on both sides of the rules. For example, in the second rule, gene 329 is on both sides.

TABLE I
ASSOCIATION RULES

L324,L329=>DL53,conf:(1)
L324,L329=>DL329,conf:(1)
L324,L329=>DL1282,conf:(1)
L324,L4245=>DL53,conf:(1)
L324,L4245=>DL329,conf:(1)
L1222,L1266=>DL324,conf:(1)
H3323,H3425=>DH3323,conf:(0.93)
L22,H3323=>DH3617,conf:(0.92)

For the fluctuation mode, with the same minimum support for mining, only 6 rules are obtained as list in Table II. No genes appear on both sides of a rule. The first rule shows that when the expression levels of the 2631st gene(YGR189C) and the 6486th gene(YPR159W/KRE6) are low, the 2722nd gene expression level will be high after a time unit delay. This might suggest that genes YGR189C and YPR159W/KRE6 might co-regulate gene YGR279C.

TABLE II
ASSOCIATION RULES FOR FLUCTUATION MODE

L2631,L6486=>DH2722,conf:(1)
H2754,H6584=>DL3390,conf:(1)
H1838,L2213=>DH5074,conf:(1)
H2754,H6584=>DL5139,conf:(1)
L3427,L3916=>DH5971,conf:(1)
L4633,L4858=>DL6486,conf:(1)

B. Transaction Preparation for Apriori and FP-Growth Algorithms

After the expression levels of a microarray are converted to expression labels, a transaction is constructed by two consecutive microarrays. We have to perform this preparation because regular Apriori and FP-growth algorithms cannot handle patterns across transactions. The expression labels of the second microarray are prefixed with "D". If two consecutive microarrays are {H1, L2, M3, L4} and {M1, H2, L3, L4}, then the transaction contains H1, L2, L4, DH2, DL3, and DL4. M labels are discarded. Obviously, the total number of items is double as the original discretized transactions.

For different length time delay, the time points of the transactions should be different accordingly.

C. Comparison to Apriori and FP-Growth

The minimum support and minimum confidence are set to 0.7 and 0.9 respectively for our experiments. Weka [23] is a known data mining tool. We compare the results and the time cost by directly using Apriori built in Weka with our proposed method. In the procedure of Apriori, the frequent patterns are searched until no candidate itemset occurs. We implement FP-growth algorithm with Matlab.

Our proposed method is designed to find patterns with co-regulated genes targeting a gene. We set BL_3 only to find the frequent patterns whose length is three and the time delays between the two regulators to the target gene are the same with a time unit. The experimental results show that Apriori costs more time than BL_3 because Apriori works out longer patterns. Furthermore, the target genes perform more than once compared with BL_3. As a result, unnecessary time is needed to perform the items in the same time point using Apriori.

We conduct experiments with different number of genes. Comparison of the running times for the three methods with singular mode data is shown in Table III. The Apriori algorithm spends more time than the FP-growth algorithm for number of genes from 1000 to 5000. Our proposed method performs best among the three methods.

TABLE III
TIME COMPARISON FOR SINGULAR MODE

Methods \ Sizes	1000	2000	3000	4000	5000
Apriori	6s	24s	60s	318s	998s
FP-growth	3.77 s	11.99 s	25.08 s	60.74 s	149.61s
BL_3	0.14 s	0.29s	0.67s	1.36s	2.37s

When preparing transactions with fluctuation mode, the execution times for Apriori and FP-growth algorithms change a lot. FP-growth algorithm needs much more time than Apriori algorithm. The reason might be a long F-list generated by FP-growth algorithm.

TABLE IV
TIME COMPARISON FOR FLUCTUATION MODE

Methods \ Sizes	1000	2000	3000	4000	5000
Apriori	5s	19s	42s	95s	344s
FP-growth	7.32s	20.14s	61.39s	138.47s	271.13s
BL_3	0.1s	0.33s	0.73s	1.32s	2.04s

V. CONCLUSION

In this paper, a method with backward frequent itemsets to find time-delayed gene regulation patterns is proposed. We provide a different viewpoint from other methods. Our method applies results performed by Apriori to find possible gene regulation relationships. However, the BL_3 method discovered that two or more genes can affect a single gene, and BL3 also facilitates the process of finding regulation patterns. The execution time of our method is much shorter than Apriori and FP-growth. We can find the patterns more efficiently than common association rule mining algorithms.

In addition, no matter how many time units are delayed between regulators and their targets, the length of the time delay can be assigned manually by the user. The interactions between genes in any length of time delay in which you want to know can be found in a time series DNA microarray gene expression dataset by our method. We thought that the patterns we found might be used for gene function prediction.

In the future, the predicted gene regulation patterns should be verified by biological evidence and they should consistent with the yeast cell cycle phase information. We can apply our framework to a gene expression dataset further to test and verify if the results are helpful for predicting, or if they are applicable to other analyses. The time-delayed gene regulation patterns can be used for predicting protein-protein interaction. If there is such gene regulation between genes of two proteins, it is likely that the two proteins interact with each other. When integrated with amino acid patterns on the binding sites of proteins [13], better protein-protein interaction prediction can be achieved.

REFERENCES

[1] R. Agrawal and T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases," ACM SIGMOD International Conference on Management of Data, pp. 207-216, 1993.

[2] C. Becquet, S. Blachon, B. Jeudy, J. Boulicaut and O. Gandrillon, "Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data," Genome Biology, vol. 3, no. 12, research0067, pp. 1-16, 2002.

[3] C. Creighton and S. Hanash, "Mining gene expression databases for association rules," Bioinformatics, vol. 19, no. 1, pp. 79-86, 2003.

[4] J.K. Choi, U. Yu, O.J. Yoo and S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer," Bioinformatics, vol. 21, no. 24, pp. 4348-4355, 2005.
[5] I. Chaturvedi1 and J.C. Rajapakse, "Fusion of gene regulatory and protein interaction networks using skip-chain models," Springer Berlin / Heidelberg, Pattern Recognition in Bioinformatics, Third IAPR International Conference, vol. 5265, pp. 214-224, 2008.
[6] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," Journal of Computational Molecular Biology, vol. 7, no. 3-4, pp. 601-620, 2000.
[7] Genes & Gene Expression, http://www.biochemweb.org/genes.shtml
[8] Genome-Wide Analysis of Cell Cycle-Dependent Transcription, http://genomics.stanford.edu/yeast_cell_cycle/cellcycle.html
[9] H.D. Jong, "Modeling and simulation of genetic regulatory systems: a literature review," Journal of Computational Biology, vol. 9, no. 1, pp. 67-103, 2002.
[10] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. 2nd ed, Morgan Kaufmann publications, 2006.
[11] K. Koch, S. Schonauer, I. Jansen, J.v.d. Bussche and T. Burzykowski, "Finding clusters of positive and negative coregulated genes in gene expression data," IEEE International Conference on Bioinformatics and Bioengineering, pp. 93-99, 2007.
[12] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: a recent overview," GESTS International Transactions on Computer Science and Engineering, vol. 32, no. 1, pp. 71-82, 2006.
[13] Huang-Cheng Kuo, Jung-Chang Lin, Ping-Lin Ong, Jen-Peng Huang, "Discovering amino acid patterns on binding sites in protein complexes," Bioinformation, vol. 6, no. 1, pp. 10-14, 2011.
[14] P. Li, C. Zhang, E.J. Perkins, P. Gong, and Y. Deng, "Comparison of probabilistic boolean network and dynamic bayesian network approaches for inferring gene regulatory networks," BMC Bioinformatics, Vol. 8 (Suppl 7), pp.13-20, 2007.
[15] F.J. Lopez, A. Blanco, F. Garcia1, C. Cano1 and A. Marin, "Fuzzy association rules for biological data analysis: a case study on yeast," BMC Bioinformatics, vol. 9, pp.107-124, 2008.
[16] T.I. Lee, et al, "Transcriptional regulatory networks in saccharomyces cerevisiae," Science 298, 799, 2002.
[17] P.C.H. Ma and K.C.C. Chan, "Inferring gene regulatory networks from expression data by discovering fuzzy dependency relationships," IEEE Transactions on Fuzzy Systems, vol. 16, no. 2, pp. 455-465, 2008.
[18] P.A. Pevsner, et al, "Improved chips for sequencing by hybridization," Journal of Biomolecular Structure and Dynamics, vol. 9, no. 2, pp. 399-410, 1991.
[19] R. Ram, M. Chetty and T.I. Dix, "Causal modeling of gene regulatory network," IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, pp. 1-8, 2006.
[20] L. Sacchi, R. Bellazzi, R. Porreca, C. Larizza and P. Magni, "Precedence temporal networks from gene expression data," IEEE Symposium on Computer-Based Medical Systems, pp. 109-114, 2005.
[21] A. Tuzhilin and G. Adomavicius, "Handling very large numbers of association rules in the analysis of microarray data," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 396-404, 2002.
[22] H. Wang, J. Pei and P.S. Yu, "Pattern-based similarity search for microarray aata," ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 814-819, 2005.
[23] Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
[24] Z. Xing and D. Wu, "Modeling multiple time units delayed gene regulatory network using dynamic bayesian network," IEEE International Conference on Data Mining – Workshops, pp. 190-195, 2006.
[25] F. Yavari, F. Towhidkhah, Sh. Gharibzadeh, A.R. Khanteymoori and M.M. Homayounpour, "Modeling large-scale gene regulatory networks using gene ontology-based clustering and dynamic bayesian networks," International Conference on Bioinformatics and Biomedical Engineering, pp. 297-300, 2008.
[26] Y. Zhao, G. Wang, Y. Yin and G. Yu, "Mining positive and negative co-regulation patterns from microarray data," IEEE Symposium on BioInformatics and BioEngineering, pp. 86-93, 2006.
[27] S. Zainudin and S. Deris, "Towards evaluation of inferred gene network," International Conference on Computational Science and its Applications, pp. 57-64, 2007.
[28] Guoren Wang, Linjun Yin, Yuhai Zhao, Keming Mao, "Efficiently mining time-delayed gene expression patterns," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 40, no. 2, pp. 400-411, 2010.
[29] Jiong Yang, Wei Wang, Philip S Yu, "Mining asynchronous periodic patterns in time series data," IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3, pp. 613-628, 2000.

Huang-Cheng Kuo was born in Taiwan in 1964. He got a masters degree and a Ph.D. degree from Department of Computer Engineering and Science at Case Western Reserve University, Ohio, USA. He had been with Southern Taiwan University for three years. He is currently an assistant professor with the Department of Computer Science and Information Engineering at National Chiayi University, Taiwan. His research interest is in applying data mining methods to biological data.

Pei-Cheng Tsai was born in Taiwan in 1980. He got a masters degree from Department of Computer Science and Information Engineering at National Chiayi University, Taiwan. He is currently looking for a job in computer science area. His research interest is in applying data mining methods to biological data.