Finding associations between genes by time-series microarray sequential patterns analysis

Ho Jung Nam Doheon Lee
Department of BioSystems, Korea Advanced Institute Science and Technology (KAIST), Korea
Email : {hjnam, dhlee}@biosoft.kaist.ac.kr

ABSTRACT: Data mining techniques can be applied to identify patterns of interest in the gene expression data. One goal in mining gene expression data is to determine how the expression of any particular gene might affect the expression of other genes. To find relationships between different genes, association rules have been applied to gene expression data set [1]. A notable limitation of association rule mining method is that only the association in a single profile experiment can be detected. It cannot be used to find rules across different condition profiles or different time point profile experiments. However, with the appearance of time-series microarray data, it became possible to analyze the temporal relationship between genes. In this paper, we analyze the time-series microarray gene expression data to extract the sequential patterns which are similar to the association rules between genes among different time points in the yeast cell cycle. The sequential patterns found in our work can catch the associations between different genes which express or repress at diverse time points. We have applied sequential pattern mining method to time-series microarray gene expression data and discovered a number of sequential patterns from two groups of genes (test, control) and more sequential patterns have been discovered from test group (same GO term group) than from the control group (different GO term group). This result can be a support for the potential of sequential patterns which is capable of catching the biologically meaningful association between genes.

1 INTRODUCION

With a tremendous increase of microarray gene expression data, data mining techniques, techniques for extracting useful information from large databases, can be applied to identify patterns of interest in the gene expression data. One goal in mining gene expression data is to determine how the expression of any particular gene might affect the expression of other genes. To find relationships between different genes, association rules have been applied to gene expression data set [1]. An association rule has the form *LHS* => *RHS*, where *LHS* and *RHS* are sets of items, the *RHS* set being likely to occur whenever the *LHS* set occurs. An example of an association rule mined from expression data might be {*cancer*} => {*gene A* ↑, *gene B* ↓, *gene C* ↑}, meaning that, in most profile experiments with cancerous cell, gene A was measured as highly expressed, gene B was highly repressed, and gene C was highly expressed. A notable limitation of association rule mining method is that only the association in a single profile experiment can be detected. It cannot be used to find rules across different condition profiles or different time point profile experiments.

However, with the appearance of time-series microarray data, it became possible to analyze the temporal relationship between genes (e.g. before, after). In this paper, we analyze the time-series microarray gene expression data to extract the sequential patterns which are similar to the association rules between genes among different time points (profile experiments) in the yeast cell cycle. An example of such a pattern is that {*gene A*↑=> *gene B*↓=> *gene C*↑}, which means high expression level of gene A followed by strong repression of gene B and significant expression of gene C in turn. The general meaning of sequential patterns is significantly repeated patterns in almost every given time cycle. Coherently, sequential patterns which are discovered from time-series microarray gene expression data are also repeated patterns in given particular size of time cycle. The sequential patterns found in our work can catch the associations between different genes which express or repress at diverse time points.

Rest of this paper is organized as follows: Section 2 informs the related work, association rule mining method which is applied to gene expression data, Section 3 gives a basic review of sequential pattern mining, extending the concept as it could be applied to gene expression data. Section 4 gives experiments and its results. Discussion and future work of this work is given in section 5.

2 RELATED WORK

2.1 Mining gene expression for association rules

One widespread data mining technique for finding and describing relationships between different items in a large data set is to find for association rules in the data. An association rule has the form '*LHS(Left Hand Side)* => *RHS(Right Hand Side)*', where *LHS* and *RHS* are sets of items, the *RHS* set being likely to occur whenever the *LHS* set occurs. Association rules discovery method is widely used in the industry named 'market basket analysis'. In market basket analysis, an asscoation rule stands for a set of items that are likely to be purchased together; for example, the rule {bread} =>{milk, juice} would mean that whenever a shopper purchases bread, he or she is likely to purchse both milk and juice as well in the same transaction. Association rules mining method have been used as well to mine to gene expression data [1]. In the analysis of gene expression data, the items in an association rule can represent genes that are hignly expressed or highly repressed. For instance, an association rule mined from expression data might be {cancer} => {gene A ↑, gene B ↓, gene C ↑}, meaning that, in most profile experiments with

cancerous cell, gene A was measured as highly expressed, gene B was highly repressed, and gene C was highly expressed. *C.Creighton* group has discovered numerous rules in the compendium from Hughes *et al* [2] of expression profiles for 6316 transcripts corresponding to 300 diverse mutations and chemical treatments in yeast.

To further limit the search space of candidate rules, they looked only for rules where either the LHS or the RHS set of the rule LHS => RHS contain only one item. They found the rules under 10% minimum support for frequent itemsets and 80% minimum confidence for association rule. After cursory analysis, some of rules reveals numerous associations between certain genes, many of which make sense biologically, others suggesting new hypotheses that may warrant further investigation.

3 METHODS

3.1 Sequential patterns

A classic example of a sequential pattern is that customers typically rent "Star Wars", then "Empire Strikes Back", and then "Return of the Jedi". Note that these rentals need not be consecutive. Customers who rent some other videos in between also support this sequential pattern.

Problem Statement All the transaction of a customer can together be viewed as a sequence, where each transaction corresponds to a set of items, and the list of transaction, ordered by increasing transaction-time, correspond to a sequence. Such a sequence is called a customer-sequence. Formally, the transactions of a customer ordered by increasing transaction-time, T1, T2, ..., Tn. The set of items in Ti be denoted by itemset(Ti). The customer-sequence for this customer is the sequence <itemset(T1), itemset(T2) ... itemset(Tn)>. A customer supports a sequence s if s is contained in the customer-sequence for this customer. The support for a sequence is defined as the fraction of total customers who support this sequence. Given a database D of customer transactions, the problem of mining sequential patterns is to find the maximal sequences among all sequences that have a certain user-specified minimum support. Each such maximal sequence represents a sequential pattern.

Example Consider the database shown in table 1. (This database has been sorted on customer-id and transaction-time.) Table 2 shows this database expressed as a set of customer sequences [3].

With minimum support set to 25%, i.e., a minimum support of 2 customers, two sequences: < (30) (90)> and < (30) (40 70)> are maximal shown in Table 3 among those satisfying the support constraint, and are the desired sequential patterns. The sequential pattern < (30) (90)> is supported by customer 1 and 4. Customer 4 buys item (40 70) in between items 30 and 90, but supports the pattern < (30) (90)> since we are looking for patterns that are not necessarily contiguous. The sequential pattern < (30) (40 70)> is supported by customer 2 and 4. Customer 2 buys 60 along with 40 and 70, but supports this pattern since (40 70) is a subset of (40 60 70).

Customer ID	Transaction time	Item purchased
1	6/25/1993	30
1	6/30/1993	90
2	6/10/1993	10, 20
2	6/15/1993	30
2	6/20/1993	40, 60, 70
3	6/25/1993	30, 50, 70
4	6/25/1993	30
4	6/30/1993	40, 70
4	7/25/1993	90
5	6/12/1993	90

Table 1: Database Sorted by Customer Id and Transaction Time

Customer Id	Customer Sequence
1	<(30)(90)>
2	<(10 20)(30)(40 60 70)>
3	<(30 50 70)>
4	<(30)(40 70)(90)>
5	<(90)>

Table 2: Customer-Sequence Version of the Database

Sequential Patterns with support > 25%
<(30)(90)>
<(30)(40 70)>

Table 3: The answer set with support 25%

3.2 In case of time-series microarray data

In the perspective of market basket analysis, a gene expression profile can be regarded as a single transaction and each transcript or protein can be thought of as an item. Nevertheless, in market basket analysis, any particular item is either purchased or not purchased in a transaction, while in an expression profile each transcript or protein is assigned a real value that indicates the relative abundance of that transcript or protein in the profiled sample. To apply sequential pattern mining method to time-series gene expression data, binning step is needed at first for each measured value as being up (highly expressed), down (highly repressed), or neither up nor down. After finishing the binning step, gene expression profiles have to be converted into 3-dimensional space which correspond with {customer-id, transaction time, sequence of items purchased}. In this paper, we mapped the data into {a set of experiment profile per one cycle, transaction time, binned gene expression} dimensional space.

4 EXPERIMENT

4.1 Data sets

We used *Saccharomyces cerevisiae* cell cycle alpha factor arrest synchronization data [4]. The time-series microarray consists of 18 time points and an interval between the sampling time points is 7 minutes. To show the significance of the sequential patterns which are extracted from the

Examples of discovered sequential patterns

[YCR093W_Down] => [YDR303C_Up]
[YDR217C_Up] => [YBR160W_Up]
[YBR160W_Up] => [YDR303C_Up][YOL139C_Up]
[YDR217C_Up] => [YDR303C_Up][YOL139C_Up]
[YDR363W_Down] => [YDR303C_Up][YOL139C_Up]
[YDR303C_Up][YIL131C_Up] => [YBR160W_Up]
[YNL068C_Up] => [YDR303C_Up][YOL139C_Up]
[YNL068C_Up] => [YIL131C_Up][YOL113W_Up]
[YKL203C_Down][YNL068C_Up] => [YDR303C_Up]
[YBR160W_Up] => [YDR303C_Up][YIL131C_Up][YOL139C_Up]
[YDR217C_Up] => [YDR303C_Up][YIL131C_Up][YOL139C_Up]
[YDR363W_Down] => [YDR303C_Up][YIL131C_Up][YOL139C_Up]
[YDR303C_Up][YIL131C_Up] => [YDR303C_Up][YOL139C_Up]
[YDR303C_Up][YIL131C_Up] => [YIL131C_Up][YOL113W_Up]
[YNL068C_Up] => [YDR303C_Up][YIL131C_Up][YOL139C_Up]
[YDR303C_Up][YIL131C_Up] => [YDR303C_Up][YIL131C_Up][YOL139C_Up]

Table 4: Discovered sequential patterns (Cycle size 5, Shifting +2)

Group	Test group (same GO term)					Control group (different GO term)				
Shifting size	+0	+1	+2	+3	+4	+0	+1	+2	+3	+4
Cycle size 2	0	-	-	-	-	0	-	-	-	-
Cycle size 3	0	0	-	-	-	0	0	-	-	-
Cycle size 4	0	1	1	-	-	2	1	0	-	-
Cycle size 5	0	0	16	4	-	0	0	4	1	-
Cycle size 6	67	141	68	100	130	0	0	0	22	12

Table 5: The number of sequential patterns discovered from time-series microarray

time-series microarray, we selected two groups of genes. Among three categories of Gene ontology, *molecular function*, *biological process*, *cellular component*, experiment work showed that *biological process* agrees best with the hypothesis that similar expressions indicate a functional relation [6]. In this reason, two groups of genes are selected out under the criterion whether the genes have same biological process GO term or not. A test group of genes assigned Gene ontology (GO) term [5] as *'regulation cell cycle'*. There are 24 distinct genes which have GO term as *'regulation cell cycle'* and having no missing value in their microarray expression data. A control group of genes no pair of which share the same GO term also consists of 24 genes to normalize the number of genes between two groups. To maximize the correctness, we choose each group 24 genes which have no missing value in expression data. Under the hypothesis that the genes which share same GO term could have functional relation, if sequential patterns have potential to indicate the functional association among different genes, then the sequential patterns can be more frequently found from the group which have same GO term than the group which have no same pair of GO term. To extract sequential patterns from the given data set, standard cycle size and a shifting size are important features. For example, discovered sequential patterns from stocks data could be different whether the cycle size is a day or a week or a month or a year and could be also different whether shifting April ~ July to May ~ August. However, in case of expression level in cell cycle data, there is no known proper cycle size to catch the sequential patterns of gene expression. For this reason, we tested with the various cycle size, from 2 to 6 and tested with different shifting size (from 0 to given cycle size -1)

4.2 Results

We used IBM Intelligence Miner to find sequential patterns from time-series microarray gene expression data. We specified the minimum support value for sequential patterns 90%. We ran the application on a desktop computer with an Intel Pentium 4 2.4G processor. Examples of discovered sequential patterns are shown in table 4. 16 distinct sequential patterns have been found under the condition 5 time point cycle and 2 time points shifting. [YCR093W_Down] => [YDR303C_Up], one of the discovered rules, means strong repression of ORF YCR093W followed by highly expression of ORF YDR303C can be found in every 5 sampling time points cycle. As you can see in table 5, discovered sequential patterns which are discovered under the condition 2, 3

time points cycle, there is no detected sequential pattern. In table 5, comparing the number of sequential patterns between two groups (test, control), we can significantly notify that more sequential patterns have been discovered from test group (same GO term group) than from the control group (different GO term group). This result can be a support for the potential of sequential patterns which is capable of catching the biologically meaningful association between genes.

All sequential patterns found in our work is available at http://biosoft.kaist.ac.kr/~hjnam/sequential_patterns.zip.

5 DISCUSSION

We have applied sequential pattern mining method to time-series microarray gene expression data and discovered a number of sequential patterns from two groups of genes. The sequential patterns that we have discovered definitely represent a small part of all of the possible patterns among genes. The rest of sequential patterns could be found by using different minimum support value or different time interval period or using other data sets. Sequential pattern can describe how the expression of one gene may be associated with the expression of a set of genes over the time goes. However, the essential difference between makget data and gene expression data, understanding the meaning of market data's sequential patterns is straight forward but interpreting the biological meaning of discovered sequential patterns is not an easy job. In case of the stocks data, it is possible to accept the results of sequential patterns which are discovered under 1week, 1 month, 1 year cycle. On the other hand, in case of gene expression data, extra analysis step is necessary to make understand sequential patterns which are discovered under the various sampling time points cycle. With the discovered sequential patterns, one might easily infer that the genes involved take part in some type of gene network, but a sequential pattern may imply an temporal association between genes, it dose not imply a cause and effect relationship.

ACKNOWLEDMENT

This work was supported by the Korean Systems Biology Research Grant (2005-00343) from the Ministry of Science and Technology. We would also like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics and the IBM SUR program for providing research and computing facilities.

REFERENCES

[1] Chad Creighton, Samir Hanash. Mining gene expression database for association rules. Bioinformatics, **19**, 79-86, 2003

[2] Hughes, T.R., Marton, M.J., Jones, A.R., Roborts, C.J.,Stoughton, R., Armour; C.D., Bennett, H.A., Coffey, E., Dai, H., He,Y.D., Kidd, M.J., King, A.M., *et al.* Functional discovery via a compendium of expression profiles. Cell, **102**, 109-126, 2000

[3] Rakesh and Ramakrishnan. Mining sequential patterns, Eleventh International Conference on Data Engineering, 3-14, 1995

[4] Paul T. Spellman. Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiaes by Microarray Hybridization. Molecular Biology of the Cell, 3273-3297, 1998

[5] http://www.geneontology.org/

[6] Brown,M.P.S., Grundy, W.N., Cristianini, N., Sugent, C.W., Furey, T.S., Ares, M. and Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl Acad. Sci. USA, **97**, 262-267, 2000