Discovering Temporal Associations among Significant Changes in Gene Expression

Hashmat Rohian, Aijun An, Jiashu Zhao, Jimmy Huang
Department of Computer Science and Engineering, York University
Toronto, Canada
{rohian, aan, jessie, jhuang}@cse.yorku.ca

***Abstract*— One of the most demanding problems in mining temporal data is to identify how multivariate change associations might be discovered and used to better understand data interactions and dependencies. This paper introduces a framework to mine associations among significant changes in multivariate time-series data. Building on statistical methods, we detect significant changes in time-series data and use marginal change rates to qualify the direction of change at significant change points. Furthermore, a propositional confirmation-guided rule discovery method is used to discover associations among these significant changes. We apply our approach to gene expression data measured in yeast cell cycles and demonstrate that our method can learn novel and high-quality significant change associations among different genes. Such associations can be used to cluster genes and build gene interaction networks.**

Keywords- Biological Data Mining and Visualization, Microarray Data Analysis, Change Association Mining

I. INTRODUCTION

Gene expression data, produced by DNA microarray experiments, have enabled us to gain insight into the fundamental aspects underlining the growth and development of life. Unlike traditional methods in molecular biology, which are limited to investigating one gene at a time, microarray experiments allow us to monitor and analyze the expression levels of thousands of genes at once and in an efficient manner. By analyzing the results of such experiments, we can better understand the interactions among genes. As a matter of fact, a goal in analyzing gene expression data is to determine how the expression of any particular gene or a group of genes might affect the expression of other genes [4].

There are two types of gene expression data: static and time-series data [2]. Static gene expression data consist of snapshots of the expressions of genes in different and independent samples, while time-series expression data are obtained by measuring the expressions of genes in the same sample over a time course. In this paper, we are concerned with time-series data. Table 1 shows an example time-series expression data set, in which the value corresponding to time t_i and $gene_j$ is the log ratio of the amount of gene products (such as mRNA) produced by $gene_j$ at time t_i to the average amount of gene products over the entire period.

Various data mining tools have been applied to gene-expression data to discover relationships among genes. For example, association rule mining has been applied to extract associations among subsets of genes. Before mining association rules, most of the methods convert the continuous expression data into discrete data by computing the difference between the expression value of $gene_j$ at time t_i and its value at time t_{i-1}. If the difference is greater than a threshold v, $gene_j$ is considered to be *up-regulated* at t_i; if it is smaller than $-v$, the gene is *down-regulated* at t_i; otherwise, no regulation occurs at time t_i. Based on the discretized data, association rules of the form:

$$\{gene\ A\uparrow, gene\ B\downarrow\} \rightarrow \{gene\ C\uparrow\}$$

can be discovered, which means that when gene A is up-regulated and gene B is down-regulated, it is very likely to observe an up-regulation of gene C. Such associations have been shown to be useful. One might infer that the genes involved in an association rule participate in some type of gene network [4], in which they interact with each other indirectly through their RNA and protein expression products. However, there are limitations with the existing methods. First, there is no convention on the definition of up- and down-regulation. Different researchers may use different thresholds to determine up- and down-regulations.

TABLE I. AN EXAMPLE TIME-SERIES EXPRESSION DATA SET

Time	**$gene_1$**	**$gene_2$**	**$gene_3$**	**…**	**$gene_m$**
t_1	0.01	-0.01	0.05	…	-0.44
t_2	0.03	0.0	0.06	…	-0.90
t_3	0.07	-0.05	0.05	…	0.00
t_4	0.00	-0.08	0.05	…	0.05
t_5	-0.05	-0.07	0.09	…	0.10
…	…	…	…	…	…
t_n	0.05	0.04	0.02	…	0.09

Second, most applications of association rule mining to gene expression data either were on static data or did not consider the sequential information embedded in time-series data. That is, most of them treated the data at different time points independently. Thus, temporal relationships were not revealed by these methods. Third, too many rules may be discovered from a gene expression data set [3]. The generated rules can be overwhelming and hard to analyze.

In this paper, we propose a framework for finding strong associations among significant changes of different genes in time-series gene expression data. A significant change is a change point in a data sequence where an abrupt variation occurs. In the context of time-series gene expression data, a significant change in the expression sequence of a gene is an abrupt increase or decrease in the

gene's expression level over the time course. Detecting significant changes in gene expression levels and finding associations among them can reveal new knowledge about the data that has not been discovered. Our framework can also overcome some limitations of the existing methods for mining association rules from gene expression data. First, our method does not depend on the thresholds for up- and down-regulations. Instead, it uses statistical tests to detect the significant change points in the gene expression sequences and finds associations among them. Second, our method can not only discover the associations among the significant changes occurring at the same time among different genes, but can also discover the associations of these changes at different time points. Thus, temporal relationships or dependencies can be discovered. Third, since we focus only on significant changes, the number of discovered associations is much smaller than the one from the previous methods. Although rules that associate non-significant changes are missed, we believe that the rules relating significant changes are among the most interesting ones. It is often preferable to present the user with a small set of more interesting rules so that they can be easily analyzed.

The paper is organized as follows. In Section 2, we introduce the related work. Our proposed framework is presented in Section 3. In Sections 4, 5 and 6, we describe our methods for change point detection, change point measurement and association rule mining, respectively. We provide experimental evaluations in Section 7. The paper is concluded in Section 8.

II. Related Work

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Agrawal et al. [1] defined the problem of association rule mining as follows. Let $I=\{i_1,i_2,\ldots,i_n\}$ be a set of items. Let $D=\{t_1,t_2,...,t_m\}$ be a transaction database containing a set of transactions. Each transaction in D contains a set of the items in I. An association rule is defined as an implication of the form X=>Y where $X, Y \subseteq I$ and $X \cap Y = \varnothing$ representing that Y is likely to occur in a transaction whenever X occurs. When applying to gene expression data, the items in an association rule are genes being up-regulated or down-regulated.

There have been a number of applications of association rule mining to gene expression data [3, 4, 10]. In [4], association rules are learned from static gene expression data. P. Carmona-Saez et al. [3] integrate multiple gene annotations and expression data in the same analytic framework and extract meaningful associations among heterogeneous sources of data. The method does not consider the temporal relationship among expression values of different times. Nam et al [10] proposed a temporal association rule mining method that extracts temporal dependencies among related genes. A temporal association rule in [10] has the form {gene A↑, gene B↓} → (7 min) {gene C↑}, which means that high expression level of gene A and significant repression of gene B is followed by significant expression of gene C after 7 minutes. The significance of expression or repression is determined by a threshold, which is determined by trial-and-error using a measure of correctness of the extracted candidate rules that tests whether the extracted rules is matched with known domain knowledge. A problem with such a strategy is that if the discovered patterns are true but unknown, the corresponding parameter setting may not be considered the best and the patterns can be missed. Also, it is time-consuming to test various parameter setting and time intervals.

There is also a considerable amount of prior work on finding surprising patterns and change-points in time series. For example, Keogh et al. [9] described a technique that represents a real-valued time-series by quantizing into it a finite set of symbols and then uses a Markov model to detect surprising patterns in the symbol sequence. Guralnik and Srivastava [7] proposed an iterative likelihood-based method for segmenting a time-series into piecewise homogeneous regions.

Wan and An [14] introduce the concept of significant milestones for a transitional pattern, which are time points at which the frequency of the pattern changes most significantly. The above approaches share a common goal with that of this paper, namely detection of novel and unusual data points or segments in time-series. However, none of these earlier works focus on the specific problem we address here, namely mining associations among significant changes.

III. Proposed Framework

In this section, we present our framework for finding associations among significant changes in a multivariate time series data set that contains multiple sequences of data measured at the same series of successive time points. The main idea of our framework is to observe the trend of each data sequence, detect the significant changes in the data sequence, and find associations of the significant changes among different sequences.

The framework consists of three main steps as illustrated in Figure 1. In the first step, a change point detection algorithm (to be described in Section 4) is used to detect significant changes in each data sequence. The time point where a significant change takes place in a data sequence is called a *change point* in the data sequence.

TABLE II. Data set with marginal change rates

Change point	seq_1	seq_2	seq_3	...	seq_m
c_1	r(1,1)		r(1,3)	…	
c_2		r(2,2)		…	
c_3	r(3,1)	r(3,2)		…	r(3,m)
…	…	…	…	…	…
c_k		r(n,2)	r(n,3)	…	

In the second step, we measure the degree of change at each change point of a sequence using the marginal change rate (to be defined in Section 5), and convert each data sequence into a sequence of change rates at its change points. We then combine the sequences of change rates for all the original data sequences into a data set as shown in Table 2, where $\{c_1, c_2, \ldots, c_k\}$ is the union of the sets of change points of all the data sequences and $r(i, j)$ is the marginal change rate for sequence seq_j at change point c_i. An empty cell at (i, j) means c_i is not a change point for seq_j.

TABLE III. DATA SET WITH UPS AND DOWNS

Change point	seq_1	seq_2	seq_3	…	seq_m
c_1	↑		↓	…	
c_2		↓		…	
c_3	↑	↑		…	↓
…	…	…	…	…	…
c_k		↓	↑	…	

TABLE IV. DATA SET FOR TEMPORAL ASSOCIATION RULES

Change point	seq_1	seq_2	seq_3	…	seq_m
c_1	↑		↓	…	
c_2	↑	↓	↓	…	
c_3	↑	↑	↓	…	↓
…	…	…	…	…	…
c_k	↑	↓	↑	…	↓

In the third step, two types of association rules are generated by using an association rule learning algorithm. The first type of association rules describe the associations among the significant changes occurring at the same time. To learn such type of association rules, the change rates in the data set in Table 2 are further mapped into UPs (↑,) and DOWNs (↓), denoting that the value at the corresponding change point is increasing or decreasing, respectively. This results in the data set in Table 3, which is for mining associations among the significant changes of different sequences occurring at the same time.

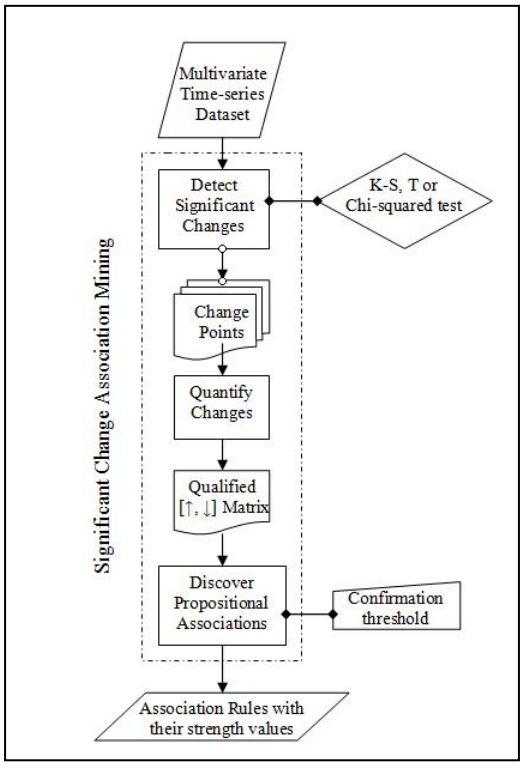

Figure 1. Framework for Mining Significant Change Associations

The second type of association rules that we discover is temporal associations among significant changes in different sequences which may occur at different times. To learn such type of association rules, we create another data set containing UP and DOWN values for all the change points, as shown in Table 4. This data set is the same as the data set in Table 3, except that an empty cell in Table 3 is filled with the value in the nearest non-empty cell above it. That is, if c_i is not a change point for seq_j, the value for cell (i, j) is the value of seq_j at seq_j's closest change point before c_i. This allows us to apply an association rule learning algorithm to find associations among significant changes at different change points. For example, an association rule, $\{seq\ 1\uparrow, seq\ 2\downarrow\} \rightarrow \{seq\ 3\uparrow\}$, discovered from such type of data sets means that if the values in sequence 1 have been significantly increasing but the values in sequence 2 have been significantly decreasing, it is very likely to observe a significant increase of values in sequence 3. Such association rules describe temporal relationships between a significant change in one or more sequences and the most *recent* significant changes in some other sequences.

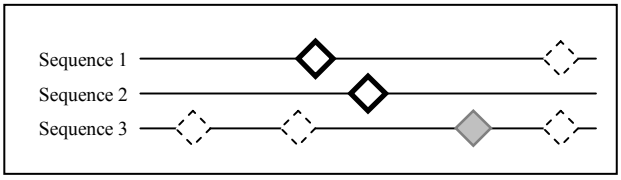

Figure 2. Discovering Delayed Change Associations

Figure 2 illustrates such a relationship, where diamonds represent change points, and dark bold diamonds on Sequences 1 and 2, which happen earlier than the grey diamond on Sequence 3, are associated to the change at the grey diamond on Sequence 3. This type of associations may reveal temporal causal relationship among different sequences. By finding such type of temporal relationships, we assume that a significant change in some sequences is mostly influenced by the most recent significant changes in some other sequences.

IV. CHANGE POINT DETECTION

Many efforts have been devoted to change point detection [e.g., 6, 7, 9]. They can be classified into online and offline methods or parametric and non-parametric methods. We adopt an online non-parametric method [6], which is faster than offline methods and more importantly it does not assume that the data follows a particular distribution. The method is based on the cumulative sum of consecutive measurements, which offers the most visually descriptive way to detect changes in the process or distribution generating the measurements. Given this metric, any change can be considered as a change in the first derivative of the cumulative record. The algorithm can be summarized as follows, in which a cumulative record is a data sequence containing the running sums of the input data sequence:

For each point p_i in the cumulative record,

1) Find an earlier candidate change point c_i. To identify this candidate point, a straight line is drawn between the starting point of the cumulative record and p_i. The point in the cumulative record that has the maximum deviation from the straight line is a candidate change point as depicted in Figure 2.
2) Test whether c_i is a true change point by running a non-parametric statistical test to determine if the data before c_i significantly differ from the data after c_i (but before p_i), making no assumption about the distribution of data.

3) If the test failed, the procedure moves to the next point p_{i+1}, that is, it goes back to Step 1 to find and test the candidate change point with respect to p_{i+1}.
4) Otherwise, the candidate point ci is considered as a true change point, and the data before ci is removed. The algorithm recursively restarts with the remaining data sequence, treating the newly-identified change point ci as the new starting point of the data sequence, to find other change points.

Skinner's work reveals that the use of cumulative sums is the best way to visualize changes in measurements, since changes in behavior appear as readily apparent changes in the slope of the cumulative record [13]. When behavior does not change, the running sum of the measurements can be estimated as a straight line. However, if there has been a change, earlier points will diverge steadily from that straight line. Thus, the point of greatest departure from the straight line is the best estimation of a change point. In Step 2, we can use the Kolmogorov-Smirnov test (K-S test) to determine if the data before and after the candidate change point differ significantly. Alternatively, we can use t-test to assess whether the means of the aforementioned data groups are statistically different from each other. Moreover, chi-square can be used that tests a null hypothesis that the frequency distribution of certain events observed in a sample is consistent with a theoretical distribution using goodness of fit and independence.

In our experiment, K-S test is used since it gives the best performance in our preliminary study on synthetic data. The only parameter that the user needs to supply is the significance level α of the statistical test, which determines the sensitivity of change detection. We use the commonly-used significance level, 0.05. The above procedure is applied to each of the data sequences in the multivariate time-series data. Thus, a list of change points is detected for each data sequence.

V. Change Point Measurement

Our objective is to find the associations among the change points in the data sequences of a multivariate time-series data set. After the change points are detected, we need to quantify each change point to determine the degree and direction of the change at a change point.

One way to do so is to compare the value at a change point to the value right after it to determine whether the change is an increase or decrease and how much the change is. We call such a method a local change measurement method. The problem with such a method is that it is sensitive to the perturbation of the data and has the risks of incorporating noise. Since the data we deal with (i.e., the gene expression data) are usually noisy, a method that looks over the trend in a broader range is better.

To this end, we measure the change at a change point c_i by computing the marginal change rate between the value at c_i and the value at the next change point c_{i+1}:

$$MCR(c_i, c_{i+1}) = \frac{y_{c_{i+1}} - y_{c_i}}{c_{i+1} - c_i}$$

VI. Association rule mining

After a change point is represented by its marginal change rate, we convert the original data into two sets of transaction data to learn two types of association rules as discussed in Section 3. The first type of rules describes the associations among the significant changes occurring at the same time. The second type of rules reveals temporal associations among significant changes that may occur at different times.

To learn association rules, we use the Tertius algorithm [5], which is a propositional confirmation-guided rule discovery algorithm based on novelty, satisfaction and confirmation. Given a rule B→H, *novelty* and *satisfaction* are defined as [P(H | B) – P(H)] P(B) and [P(H | B) – P(H)] / [1 – P(H)] respectively, and *confirmation* is the product of the novelty and satisfaction. Tertius is a heuristic algorithm using an optimal A* search that searches for rules whose confirmation value is above a threshold. We chose to use Tertius instead of other association rule algorithms (such as Apriori [1]) for the following reasons. First, in Tertius the strength of a rule is measured by two measures: the confirmation value and the frequency of counter-instances (i.e., the number of counter-instances divided by the total number of instances). Given these two measures, one can find better pruned, more descriptive, focused and shorter rules using Tertius. Second, Tertius is more flexible. For example, it can be set to find rules that predict a single condition or a predetermined attribute.

VII. Application to the Gene Analysis of Saccharomyces Cerevisiae

We applied our framework to the dye swap technical replicates dataset [11], which is an alpha-factor synchronized microarray time series spanning two cell cycles with a sampling interval of 5 minutes. The data set contains 4774 genes and 24 time points.

We begin by normalizing the values in the data set followed by detecting change points using the procedure described in Section 4. Given that we only have 24 time points in the data, we use the K-S test to determine whether the data before a candidate change point significantly differ from the one after the candidate point. This test does not depend on an adequate sample size for the approximations to be valid. The change point detection step is followed by calculating the marginal change rate for each change point. Without loss of generality, we focus on the first 20 genes with at least two significant changes in gene expression levels.

To evaluate how well our framework performs, we compare our methods for learning significant change associations to a method that learning associations based on every time point. The results are shown in Table 5. The first row in the table shows the result of our method for learning associations among significant changes that happen at the same time (i.e., the result learned from the data whose format is illustrated in Table 3). The second row shows the result of our method for learning temporal associations among significant changes that can happen at different times (i.e., the result learned from the data whose format is shown in Table 4). The third row shows the result of learning association rules from the data at all the time points (including non-change points). To show the quality and complexity of the learned rules from the three

methods, we measure the results using the number of discovered rules, the maximum number of literals in the rules, the average true positive rates, the average false positive rules and average confirmation values. The true positive rate (TPR) of a rule A→B is defined as the number of examples (i.e., transactions) containing A and B over the number of examples containing B; and the false positive rate is defined as the number of the examples containing A but no B over the number of the examples not containing B. The results show that our methods for mining significant change associations produce fewer rules with less complexity but better quality.

TABLE V. RESULTS FROM THE THREE METHODS

	Change Point Detection	Delayed Effect	# of rules	Max # literals in rules	Avg. TPR	Avg. FPR	Avg. C.V.
1	Yes	No	16	2	1.0	0.0	0.67
2	Yes	Yes	128	3	0.74	0.09	0.74
3	No	No	405	5	0.54	0.02	0.55

For learning associations, we use the Tertius algorithm implemented in Weka. We search for Horn clauses with up to 40 literals with possible negations and prune the resulting rules with a minimum confirmation value of 0.5.

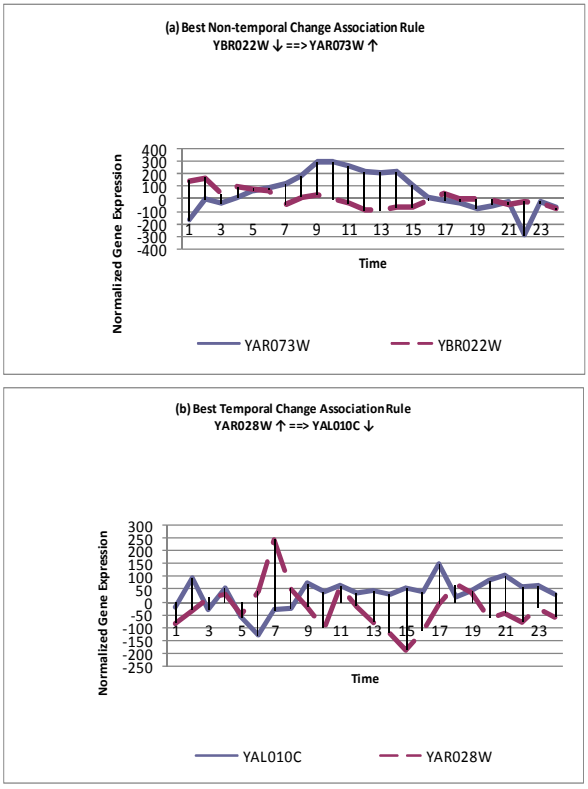

Figure 2. Visualization of the top rules

To see how the discovered rules match the underlying pattern in the data, we visualize in Figure 2 the expression levels of the genes appearing in the top-ranked rule (according to the confirmation measure) from each of our two methods for mining change associations. We can see that both rules exactly match the underlying pattern in the data. In addition, given the fact that our methods only focus on change points, the time taken to mine the associations among significant changes is much less than mining associations at all the time points. Since gene expression data usually contain thousands of genes, such a speed up is preferred.

VIII. CONCLUSION

Our goal in this paper is to design a practical and useful approach that can help us discover novel and interesting knowledge from large databases. Our framework finds the most interesting change associations among significant changes in gene expression, thus reducing the number and complexity of rules while increasing the quality of discovered associations. Past research has shown that less complex rules are surprisingly effective and easier to understand [8]. Our framework provides a way to learn such effective and useful knowledge. The associations discovered from the gene expression data show a new type of interactions among genes. They can be used to group genes into clusters and helps to build gene interaction networks. Future research directions includes clustering of genes based on discovered change associations, subgroup discovery, and building a interactive network of associations based on gene expression level changes. Besides, we also aim to employ regression and to validate our framework using real datasets in other domains.

REFERENCES

[1] R. Agrawal, T. Imielinski, A. Swami, Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference,1993, pp. 207-216

[2] Z. Bar-Joseph, Analyzing time series gene expression data, Bioinformatics, 20(16), 2004.

[3] P. Carmona-Saez, M. Chagoyen, et al., Integrated analysis of gene expression by association rules discovery, *BMC Bioinformatics*, Vol. 7, Feb.2006.

[4] C. Creighton, S. Hanash, Mining gene expression databases for association rules, Bioinformatics, 19(1), 2003, pp. 79-86.

[5] P. A. Flach, N. Lachiche, Confirmation-Guided Discovery of first-order rules with Tertius, Machine Learning, Vol. 42, 1999, pp. 61-95.

[6] Gallistel, C., Mark, T. A., King, A. P. & Latham, P, The rat approximates an ideal detector of changes in rates of reward: implication for law of effect, Journal of Experimental Psychology: Animal Behavior Processes, Vol.27, Washington, 2001, pp. 354-372.

[7] V. Guralnik and J. Srivastava, "Event detection from time series data", ACM SIGKDD, New York, NY, USA, 1999. pp. 33–42.

[8] [8] R.C. Holte, Very simple classification rules perform well on most commonly used datasets, Machine Learning 11, 1993, pp. 65–91.

[9] E. Keogh and S. Lonardi, Finding surprising patterns in a time series database in linear time and space, ACM SIGKDD, NY, USA, 2002, pp. 550–556.

[10] H. Nam, K. Lee and D. Lee, Identification of temporal association rules from time-series microarray data sets, *BMC Bioinformatics,* Vol. 10, Mar. 2009.

[11] T. Pramila, W. Wu, W. Noble, L. Breeden, Periodic genes of the yeast Saccharomyces cerevisiae: A combined analysis of five cell cycle data sets, Fred Hutchinson Cancer Research Center, WA, USA, 2009

[12] Piatetsky-Shapiro G., W. J. Frawley , Discovery, analysis, and presentation of strong rules, Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA., 1991

[13] Skinner, B. F., Farewell my LOVELY!, Journal of the Experimental Analysis of Behavior, Vol. 25, 1976, pp. 218-221.

[14] Q. Wan & A. An, Transitional Patterns and Their Significant Milestones, ICDM, 2007, pp. 691–696.