A Novel Method for Mining Temporally Dependent Association Rules in Three-Dimensional Microarray Datasets

Yu-Cheng Liu[1], Chao-Hui Lee[1], Wei-Chung Chen[1], J. W. Shin[2], Hui-Huang Hsu[3] and Vincent S. Tseng[1, 4]
[1]Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan, R.O.C
[2]Department of Microbiology & Immunology, National Cheng Kung University, Taiwan, R.O.C
[3]Department of Computer Science and Information Engineering, Tamkang University, Taiwan, R.O.C
[4]Institutes of Medical Informatics, National Cheng-Kung University, Taiwan, R.O.C.
{uchenliu, lobby, weichung}@idb.csie.ncku.edu.tw, hippo@mail.ncku.edu.tw, hhsu@cs.tku.edu.tw,
tsengsm@mail.ncku.edu.tw

***Abstract*—Microarray data analysis is a very popular topic of current studies in bioinformatics. Most of the existing methods are focused on clustering-related approaches. However, the relations of genes cannot be generated by clustering mining. Some studies explored association rule mining on microarray, but there is no concrete framework proposed on three-dimensional gene-sample-time microarray datasets yet. In this paper, we proposed a temporal dependency association rule mining method named 3D-TDAR-Mine for three-dimensional analyzing microarray datasets. The mined rules can represent the regulated-relations between genes. Through experimental evaluation, our proposed method can discover the meaningful temporal dependent association rules that are really useful for biologists.**

Keywords- Data Mining; Microarray; Gene Expression Analysis; Association Rule Mining

I. INTRODUCTION

Recently, microarray technology has become an important tool to elucidate many genes simultaneously, and data mining is a popular technology used for microarray analysis to elucidate gene expression information efficiently. Moreover, a new microarray technology can observe a set of genes under a set of samples during a series of time points. This kind of data is called gene-sample-time, three dimensional or 3D microarray datasets.

Most of the 3D microarray data analyses are focused on clustering. However, the regulation mechanism between genes cannot be generated by clustering mining. Nevertheless, to discover the relations between genes can depend on association rule mining. Although, some studies apply association rule mining [3], but there has no a concrete framework proposed on 3D microarray datasets yet. Therefore, we proposed a method named 3D-TDAR-Mine to discover the Temporal Dependency Association Rules on 3D microarray datasets. Represent the regulated-relations between genes.

The remainder of this paper is organized as follows. In Section 2, we give a problem definition about our research. The proposed method is detailed in Section 3. Section 4 gives the experimental evaluation, and the concluding remarks are made in Section 5.

II. PROBLEM DEFINITION

A. *Frequently Coherent Pattern*

The purpose of this research is use 3D-TDAR-Mine to discover the Temporal Dependency between gene expressions reaction. We define the Frequently Coherent Pattern as reaction in this paper. Indicate as the Figure 1, gene expressions of Pho5 in the time points T6 to T9 are rise coherently in the samples S4 to S7.

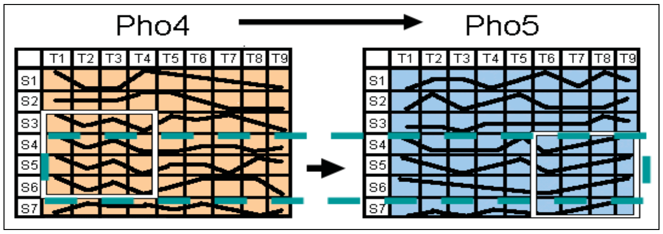


Figure 1. Temporal dependency between Frequently Coherent Patterns.

Hence, Coherent Pattern is focus on one gene in one continuous time segment to compute the similarity of gene expression value between any two samples. If similarity value between two samples series are large than Coherence Threshold that user defined. Then, we can define these two samples series in this time segment as Coherent Pattern for this gene. For example, we defined the gene expression series S3 and S3 from T1 to T4 of gene Pho4 is Coherent Pattern if this two series satisfy the Coherence Threshold.

Besides, we define two sample series as Support Sample if the two sample series are a Coherent Pattern in a sample set in this time segment. Therefore, all sample series in the Support Sample are similar. Moreover, in the cause of discover meaningful Coherent Pattern, the ratio of Support Sample in all samples must large than Minimum Support Threshold that user defined. And the ratio of Support Sample in all samples we named Support Sample Ratio.

Furthermore, the time length between the times points in each microarray data probably different. User can define the

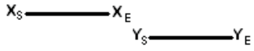
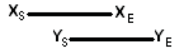
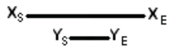
978-1-4244-7640-4/10/$26.00 ©2010 IEEE

Pattern Minimum Length Threshold to restrict the meaningful minimum continual time point number of a Coherent Pattern. If a Coherent Pattern satisfies all of above thresholds, then we can define it as a Frequently Coherent Pattern and a reaction.

B. Temporal Dependency

In our research, we use the Temporal Dependency between Frequently Coherent Pattern from Tatavarty [6]. Table 1 show the definition about Temporal Dependency between Frequently Coherent Patterns. X and Y are the leading and following Frequently Coherent Pattern. Xs and Xe are the beginning and end time point of X. Besides, Ys and Ye are the beginning and end time point of Y. And use a user defines Time-delay Threshold to constraint the Frequently Coherent Patterns possible regulation time.

TABLE I. TEMPORAL DEPENDENCY BETWEEN FREQUENTLY COHERENT PATTERNS.

Dependency	Pictorial Example	Constraint
Followby		$X_E < Y_S$, $Y_S - X_E \leqq$ Time-delay Threshold
Overlaps		$X_S < Y_S \leqq X_E < Y_E$
Contains		$X_S \leqq Y_S < X_E \leqq Y_E$

C. Proposed TS³ Similarity Measurement

In this paper, we define the Frequently Coherent Pattern as gene expressions reaction. Furthermore, Coherent Pattern is focus on one gene in one continuous time segment to compute the gene expression value similarity between any two samples. Hence, user can depend on their required feature of Coherent Pattern to choice the similarity measure method. If user wants to discover the Coherent Pattern between two samples that have identical shape in gene expression value series. They can use the PCC (Pearson correlation coefficient). But, in the real life reaction, it not always has identical shape. The expression value series between samples also have Shifting, Scale and Trend relation. Therefore, we proposed the TS^3 similarity measurement to estimate the Coherent Pattern that considers the Shifting, Scale and Trend factors.

To take a example, we have two gene expression series $S_i=\{s_{i1}, s_{i2}, ..., s_{in}\}$ and $S_j=\{s_{j1}, s_{j2}, ..., s_{jn}\}$. Besides, S_{i_max} and S_{i_min} is the maximum and minimum value in S_i respectively. Furthermore, S_{j_max} and S_{j_min} is the maximum and minimum value in S_j respectively. Moreover, S'_i and S'_j is the slope series after Min-max normalization from S_i and S_j. Therefore, $S'_i=\{ (s_{i2} - s_{i1}) / (S_{i\_max} - S_{i\_min}), (s_{i3} - s_{i2}) / (S_{i\_max} - S_{i\_min}), ..., (s_{in} - s_{i(n-1)}) / (S_{i\_max} - S_{i\_min}) \}=\{ s'_{i1}, s'_{i2}, ..., s'_{i(n-1)} \}$ and $S'_j=\{ (s_{j2} - s_{j1}) / (S_{j\_max} - S_{j\_min}), (s_{j3} - s_{j2}) / (S_{j\_max} - S_{j\_min}), ..., (s_{jn} - s_{j(n-1)}) / (S_{j\_max}-S_{j\_min}) \}=\{s'_{j1}, s'_{j2}, ..., s'_{j(n-1)}\}$. Each element in S'_i and S'_j was normalizing the ranges of value between 0 and 1. Besides, S'_D is the difference series that from S'_i and S'_j. Then, The difference series $S'_D=\{s'_{i1} - s'_{j1}, s'_{i2} - s'_{j2}, ..., s'_{i(n-1)} - s'_{j(n-1)}\}$. Further, R_i is the rank series from S'_i. And R_j is the rank series from S'_j. From minimum value to maximum value in one series is assign as 1 to n-1 in the rank series. The rank series such as $R_i=\{r_{i1}, r_{i2}, ..., r_{i(n-1)}\}$ and $R_j=\{r_{j1}, r_{j2}, ..., r_{j(n-1)}\}$, respectively. R_D is a rank difference series that from R_i and R_j, to compare the rank difference between these two series. Such as $R_D=\{ |r_{i1} - r_{j1}| + 1, |r_{i2} - r_{j2}| + 1, ..., |r_{i(n-1)} - r_{j(n-1)}| + 1 \}=\{ r_{d1}, r_{d2}, ..., r_{d(n-1)} \}$. Each element in R_D is adding 1 as the base value. And $|R_D|=| r_{d1} + r_{d2}, …, + r_{d(n-1)} |$.

The main ideal of TS^3 is use Min-max normalization to normalize each series in the same range to handle the Shifting and Scale factor. And use rank difference series R_D to be the weight of Trend to estimate the similarity between two series. As the function (1), TS^3 computes the S'_{Dk} in the absolute value symbol, to estimate the difference value of expression. The range of $|S'_{Dk}|$ is between 0 and 2. Then, use r_{d1} as the weight in Trend difference. If the elements S'_{ik} and S'_{jk} have rank difference r_{dk}, then the difference S'_{Dk} will be enlarge as the weight r_{dk}. After sum up the weighted value, it use $2\times|R_D|$ to normalize the value between 0 and 1. After all of above stage, we get the dissimilarity between S_i and S_j. In the last stage, Subtract dissimilarity from 1 and get the TS^3 similarity.

$$\begin{aligned} TS^3\ Simility &= 1-(\frac{1}{2*|R_D|}\sum_{k=0}^{n-1} r_{dk}\left|\frac{s_{i(k+1)}-s_{ik}}{s_{i\_\max}-s_{i\_\min}}-\frac{s_{j(k+1)}-s_{jk}}{s_{j\_\max}-s_{j\_\min}}\right|) \\ &= 1-(\frac{1}{2\times|R_D|}\sum_{k=0}^{n-1} r_{dk}|s'_{ik}-s'_{jk}|) = 1-(\frac{1}{2\times|R_D|}\sum_{k=0}^{n-1} r_{dk}|s'_{Dk}|) \end{aligned} \tag{1}$$

III. PROPOSED METHOD

In this research, we proposed the 3D-TDAR-Mine to discover the Temporal Dependency Association Rules in 3D microarray datasets. Figure 2 is the flow chart of our method; it can divide into two phases. First, the Coherent Pattern Phase discovers the Frequently Coherent Pattern for each gene from the 3D gene expression data. We define the Frequently Coherent Pattern as reaction in this paper. The second phase discovers the Temporal Dependency Association Rules rely on the Frequently Coherent Pattern. We define the Temporal Dependency Association Rules as the reaction relationship between genes. Detailed illustrations of the method are given below.

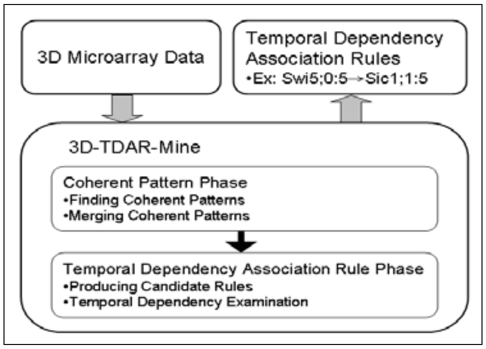

3D Microarray Data
Temporal Dependency Association Rules
•Ex: Swi5;0:5→Sic1;1:5
3D-TDAR-Mine
Coherent Pattern Phase
•Finding Coherent Patterns
•Merging Coherent Patterns
Temporal Dependency Association Rule Phase
•Producing Candidate Rules
•Temporal Dependency Examination

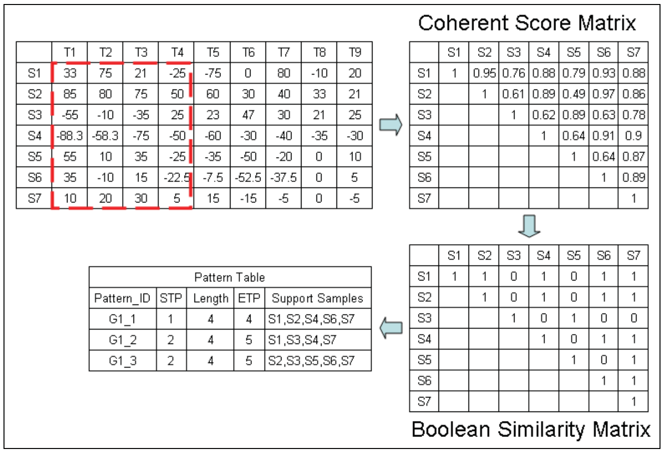
Figure 2. 3D-TDAR-Mine algorithm flow chart

A. Finding Coherent Patterns

The purpose of finding Coherent Patterns stage is aim to discover the Frequently Coherent Pattern for each gene. And

Frequently Coherent Patterns is focus on one gene in one continuous time segment have similar expression value in enough samples.

Suppose the method discover the 3D microarray dataset that has 3 genes, 7 samples and 9 time points. We named these 3 genes as G1, G2 and G3. And we named these 7 samples as S1 to S7, sequentially. Then, we named these 9 time points as T1 to T9. And the user defined Pattern Minimum Length Threshold is assuming as 4 time points length. Figure 3 is the flow chart of discover Frequently Coherent Patterns. In the first step, the method discovers the Frequently Coherent Patterns in the Sample-Time microarray of G1. In the beginning, time window cover time points T1 to T4. Then, the method estimates the Coherent Patterns similarity between each two samples under window. In the second step, the method annotates the similarity between each two samples in the Coherent Score Matrix. In the third step, estimates the similarity in the Coherent Score Matrix. The sample similarity of Coherent Pattern must large than user define threshold. If satisfy the threshold then annotate 1 in the Boolean Similarity Matrix. Otherwise, annotate 0. In the fourth step, the method discovers the Frequently Coherent Patterns from the Boolean Similarity Matrix. The Support Sample Ratio for a Frequently Coherent Pattern must large than Minimum Support Threshold that user defined.



	T1	T2	T3	T4	T5	T6	T7	T8	T9
S1	33	75	21	-25	-75	0	80	-10	20
S2	85	80	75	50	60	30	40	33	21
S3	-55	-10	-35	25	23	47	30	21	25
S4	-88.3	-58.3	-75	-50	-60	-30	-40	-35	-30
S5	55	10	35	-25	-35	-50	-20	0	10
S6	35	-10	15	-22.5	-7.5	-52.5	-37.5	0	5
S7	10	20	30	5	15	-15	-5	0	-5

Coherent Score Matrix

	S1	S2	S3	S4	S5	S6	S7
S1	1	0.95	0.76	0.88	0.79	0.93	0.88
S2		1	0.61	0.89	0.49	0.97	0.86
S3			1	0.62	0.89	0.63	0.78
S4				1	0.64	0.91	0.9
S5					1	0.64	0.87
S6						1	0.89
S7							1

	S1	S2	S3	S4	S5	S6	S7
S1	1	1	0	1	0	1	1
S2		1	0	1	0	1	1
S3			1	0	1	0	0
S4				1	0	1	1
S5					1	0	1
S6						1	1
S7							1

Boolean Similarity Matrix

Pattern Table

Pattern_ID	STP	Length	ETP	Support Samples
G1_1	1	4	4	S1,S2,S4,S6,S7
G1_2	2	4	5	S1,S3,S4,S7
G1_3	2	4	5	S2,S3,S5,S6,S7

Figure 3. Flow chart of discover Frequently Coherent Patterns

After these four steps, the method sliding the window begins from T2. Then repeat these four steps to discover the Frequently Coherent Patterns, until the window contains T9. After all of these steps, the Frequently Coherent Pattern of G1 has been discovered. Then the method repeat the steps like G1 to discovers the Frequently Coherent Pattern for other genes, respectively.

B. *Merging Coherent Patterns*

Different gene reaction maybe has different time length in the real life. Use a fixed Pattern Minimum Length Threshold for each gene will cute a complete reaction into many sub reactions. Therefore, in this stage, the method will try to merge the Frequently Coherent Patterns in the adjacent window for the same gene. Then, the method will estimate the common Support Sample Ratio. If the ratio still satisfies the Minimum Support Threshold then merge these two patterns as a longer one. The method will repeat the merge procedure, until check all of the Frequently Coherent Patterns in the adjacent window for the same gene. After merge pattern stage, each sub patterns will be deleting when merge be the longer Frequently Coherent Pattern.

C. *Producing Candidate Rules*

In the real life, each gene reaction will regulate other gene reaction in the finite time length. Therefore, any two Frequently Coherent Patterns have regulation mechanism will not separate in a long time range. For this purpose, the method discovers candidate rules that have Temporal Dependency.

On the purpose of discover Temporal Dependency between Frequently Coherent Patterns. In the first step, the method sort the patterns depend on their beginning time point. Figure 4 show the diagram for the Frequently Coherent Patterns after sorting. In the second step, the method producing the candidate rules that perhaps have Temporal Dependency between two Frequently Coherent Patterns. On the purpose of avoid producing the candidate rules that impossible have Temporal Dependency. The method restricts the adjacent beginning times point lengths between the patterns in the candidate rule. The adjacent beginning times point lengths cannot longer than leading pattern length adds the user defined Time-delay Threshold. Figure 4 show the candidate pattern in the restrict time point length with dotted line window for the leading pattern G1_1. In the third step, the method estimate the Temporal Dependency between the patterns in the restrict time point length. The process discovers the candidate rules from the first time point to the last time point; Table 2 shows the producing result of candidate rules.

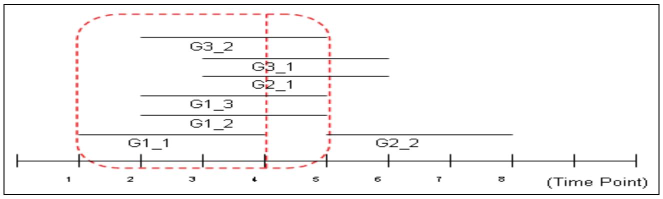


Figure 4. Diagram for the Frequently Coherent Patterns after sorting.

TABLE II. DIAGRAM FOR THE CANDIDATE RULES PRODUCING FROM THE EXAMPLE.

G1_1, G1_2 (Overlaps) (Supp: 0.43, Conf: 0.6)	G3_2, G1_2 (Contains) (Supp: 0.43, Conf: 0.75)	G1_3, G2_2 (Overlaps) (Supp: 0.43, Conf: 0.6)
G1_1, G1_3 (Overlaps) (Supp: 0.43, Conf: 0.6)	G1_2, G2_1 (Overlaps) (Supp: 0.43, Conf: 0.75)	G3_2, G2_1 (Contains) (Supp: 0.43, Conf: 0.75)
G1_1, G3_2 (Overlaps) (Supp: 0.43, Conf: 0.6)	G1_2, G3_1 (Overlaps) (Supp: 0.57, Conf: 1.0) ☆	G3_2, G3_1 (Contains) (Supp: 0.57, Conf: 1.0) ☆
G1_1, G2_1 (Overlaps) (Supp: 0.43, Conf: 0.6)	G1_2, G2_2 (Overlaps) (Supp: 0.57, Conf: 1.0) ☆	G3_2, G2_2 (Overlaps) (Supp: 0.43, Conf: 0.75)
G1_1, G3_1 (Overlaps) (Supp: 0.57, Conf: 0.8) ☆	G3_2, G1_3 (Contains) (Supp: 0.43, Conf: 0.75)	G2_1, G3_1 (Contains) (Supp: 0.57, Conf: 0.8) ☆
G1_1, G2_2 (Followedby) (Supp: 0.57, Conf: 0.8) ☆	G1_3, G2_1 (Overlaps) (Supp: 0.43, Conf: 0.6)	G2_1, G2_2 (Overlaps) (Supp: 0.43, Conf: 0.6)
G1_2, G1_3 (Contains) (Supp: 0.29, Conf: 0.5)	G1_3, G3_1 (Overlaps) (Supp: 0.43, Conf: 0.6)	G3_1, G2_2 (Overlaps) (Supp: 0.57, Conf: 0.8) ☆

D. *Temporal Dependency Examination*

The first step in this stage, the method estimates the common Support Sample Ratio from the candidate rules that producing in the producing candidate rules stage. The candidate rule must large than Minimum Support Threshold that to represent the patterns in this rule is in the same regulation mechanism. In the second step, the method will estimate Confidence in the candidate rule. The ratio of following pattern Support Samples in the leading pattern Support Samples named Confidence for this candidate rule. The Confidence large than the user defined Minimum Confidence Threshold represent that have enough dependent relation between these two patterns. The candidate rules satisfy above two thresholds will enter the third step to estimate the Temporal Dependency. The candidate rule satisfies one of the Temporal Dependencies that will be named as Temporal Dependency Association Rule. The rule also named as Temporal Dependency Association Rule with 2-Length.

Table 2 shows the Support and Confidence of candidate rules. Supposed user defines Minimum Support Threshold as 0.5 and Minimum Confidence Threshold as 0.8 in this example. If satisfy these two thresholds, then will to estimate the Temporal Dependency. Figure 5 shows the estimative result of Temporal Dependency Association Rule with 2-Length.

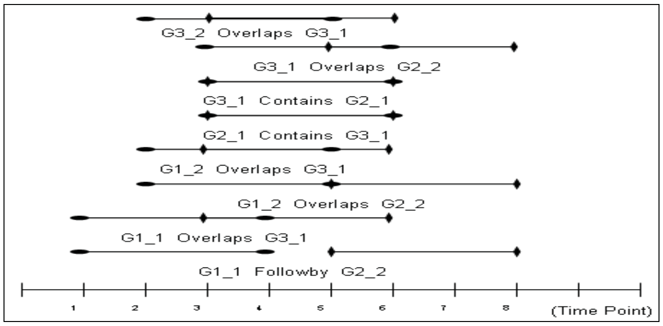


Figure 5. Diagram for the rule with 2-length

After these steps, the method will merge the Temporal Dependency Association Rule with longer length when the longer rule still satisfies these thresholds. The stage will terminate when the method can not to merge Temporal Dependency Association Rule longer. Figure 6 shows the estimative result that longer than 2-length in this example.

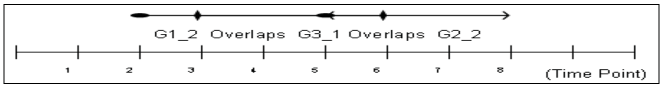


Figure 6. Diagram for the estimative result that longer than 2-length.

After producing one rule, the method will delete the rule that contain in other longer rules. The process will avoid to producing redundant rule with the same regulation mechanism. Therefore, the rule G1_2 Overlap G3_1 is the sub rules in the rule G1_2 Overlaps G3_1 Overlaps G2_2. And the rule G3_1 Overlaps G2_2 is the sub rules in the rule G1_2 Overlaps G3_1 Overlaps G2_2. The redundant rules will be deleting. Table 3 shows the final result of this example.

TABLE III. FINAL RESULT OF TEMPORAL DEPENDENCY ASSOCIATION RULES.

Temporal Dependency Association Rules	Regulation Temporal Dependency Association Rules
G3_2 Contains G3_1	G3;2;5→G3;3;6
G3_1 Contains G2_1	G3;3;6→G2;3;6
G2_1 Contains G3_1	G2;3;6→G3;3;6
G1_2 Overlaps G2_2	G1;3;5→G2;5;8
G1_1 Overlaps G3_1	G1;1;4→ G3;3;6
G1_1 Followedby G2_2	G1;1;4→G2;5;8
G1_2 Overlaps G3_1 Overlaps G2_2	G1;2;5→G3;3;6→G2;5;8

IV. EXPERIMENTAL EVALUATION

Our experimental evaluation on real life datasets will demonstrate that the proposed 3D-TDAR-Mine to discover the Temporal Dependency Association Rules. And to prove the rules really have Temporal Dependency and biological meaning in the real life.

A. *Experimental Dataset*

In the experiment, we use the Yeast *Saccharomyces cerevisiae* microarray datasets from the Stanford University yeast cell cycle analysis project website on http://genome-www.stanford.edu/cellcycle/. We used 6810 yeast genes, 13 samples and 14 time points in this datasets. And the time length between adjacent time points is 30 minutes. Several previous researches also use the same yeast datasets [5] [10] [11].

The KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway Database [4] is a collection of manually drawn pathway maps depend on experts. It annotates the validly metabolism, genetic information processing, environmental information processing such as signal transduction, various other cellular processes and human diseases regulation mechanism. We use this database to estimate the rules that we discover is really have Temporal Dependency and biological meaning in pathway. The yeast KEGG Pathway database that use to verify is downloading from the official website on http://www.genome.jp/kegg/pathway.html.

However, it still have many regulation mechanisms may not be confirmed. Besides, the regulation mechanisms may not exist between any genes. Therefore, we will use the Gene Ontology (GO) Semantic Similarity method that Wang proposed [9] to verify the gene relations with GO database [1]. The yeast annotate GO database that use to estimate is downloading from the official website on http://www.geneontology.org/.

B. *Describe of Parameters*

Table 4 shows the parameters that user define in this method. The strictest value of Minimum Support Threshold, Coherence Threshold and Minimum Confidence Threshold is 1. Moreover, user can define the Pattern Minimum Length Threshold to restrict the meaningful minimum continual time point number of a reaction. Furthermore, Time-delay Threshold to restrict the continual time point number of a regulation.

TABLE IV. PARAMETERS OF 3D-TDAR-MINE.

Parameter	Range
Minimum Support Threshold	0~1
Coherence Threshold	0~1
Pattern Minimum Length Threshold	2~ Total time point numbers
Minimum Confidence Threshold	0~1
Time-delay Threshold	0~ Total time point numbers

C. *Coherence Threshold Analysis*

In this section, we discuss the PCC and proposed TS^3 influence of pattern and rule numbers satisfied different Coherence Threshold. Figure 7 and Figure 8 show the result that Minimum Support Threshold defines as 0.7 and 0.8 respectively. Pattern Minimum Length Threshold defines as 5. Minimum Confidence Threshold defines as 0.9. Time-delay Threshold defines as 3.

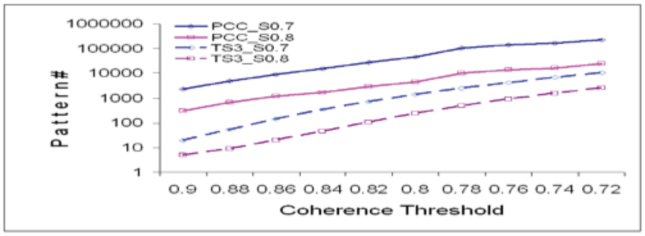


Figure 7. Pattern number under different Coherence Threshold.

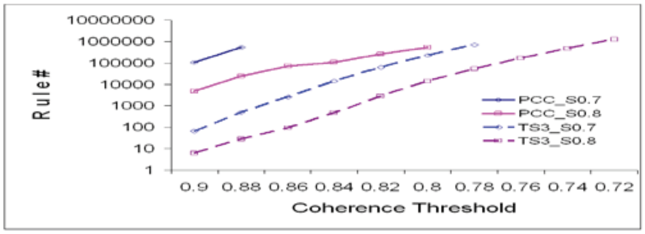


Figure 8. Rule number under different Coherence Threshold.

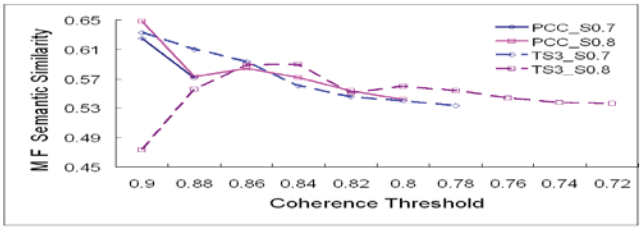


Figure 9. MF Semantic Similarity under different Coherence Threshold.

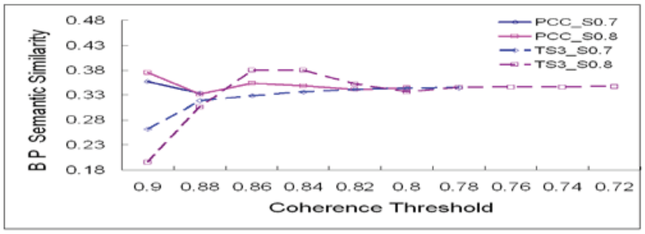


Figure 10. BP Semantic Similarity under different Coherence Threshold.

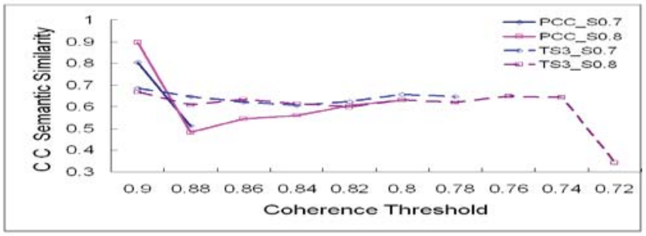


Figure 11. CC Semantic Similarity under different Coherence Threshold.

It was observed that PCC produced more patterns and rules than TS^3 in the same parameters. Further, when Minimum Support Threshold raises, satisfied pattern and rule numbers sink.

Furthermore, Figure 9, Figure 10 and Figure 11 show the result that we verify the rule with GO Semantic Similarity under MF, BP and CC of GO respectively.

It was observed that when define Coherence Threshold strictly, the gene expression are more coherently. Therefore, the GO Semantic Similarity rises. But, due to satisfied pattern and rule numbers sink. It will be more easily to be influence of meaningless rules. Therefore, we recommend the Coherence Threshold should be defined between 0.88 and 0.8 in this datasets.

D. *Rule Analysis*

In this section, we verify the rules that discovering depend on proposed TS^3 similarity measurement. First, we used the parameters on Table 5 to discover rules. After that, we observe these rules mapping to this Pathway. The Overlapping type rule "Fus3;0;4→Far1;3;7" can mapping to the Cell Cycle Pathway. Gene Fus3 (YBL016W) will regulate gene Far1 (YJL157C) to active a Phosphorylation [8]. Moreover, the Overlapping type rule "Swi5;0;4 → Sic1;1;5" also can mapping to this Pathway. Gene Swi5 (YDR146C) will regulate gene Sic1 (YLR079W) in the cell division stage [7]. This is an important regulation to the next Cell Cycle.

Secondly, we used the parameters on Table 5 to discover rules. After that, we observe these rules mapping to the Glycolysis Pathway. The rule "PGI1;0;4 → PFK2;0;4 → PFK1;3;7" can mapping to this Pathway. Gene PGI1 will regulate gene PFK2 and gene PFK2 will regulate gene PFK1 [2].

TABLE V. PARAMETERS FOR RULE DISCOVERING ON DIFFERENT PATHWAY VERIFY.

Parameter	Cell Cycle	Glycolysis
Minimum Support Threshold	0.6	0.6
Coherence Threshold	0.8	0.8
Pattern Minimum Length Threshold	5	5
Minimum Confidence Threshold	0.9	0.9
Time-delay Threshold	3	3

E. *Similarity Factors Analysis*

In this section, we will discuss the similarity factors in the rules that we discovered. Figure 12 show the Coherent Patterns in the rule "Fus3;0;4→Far1;3;7" that measure with

TS^3 when the Coherence Threshold define as 0.9. We observed that it has Scale relations between the samples in this Coherent Patterns for gene Fus3. And in the Coherent Patterns for gene Far1, it has Shift relation between samples. However, if we want to discover this rule that measure the Coherent Patterns with PCC. The Coherence Threshold must define less than 0.75. It will discover a lot of rules that maybe without biological meaning.

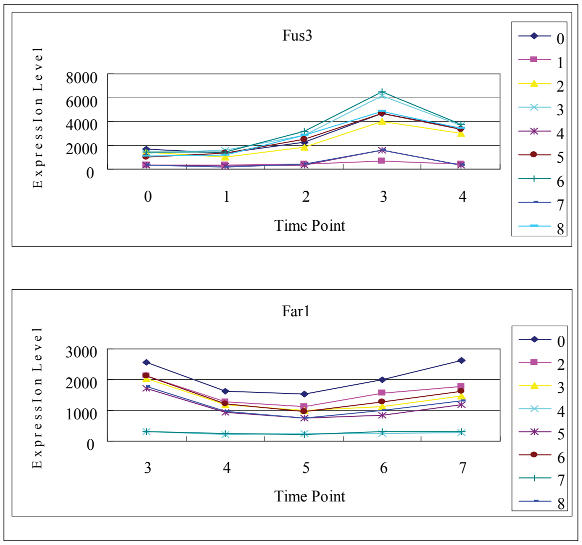


Figure 12. Similarity factors in the Coherent Patterns that TS^3 discovered.

F. *Rule Redundant Rate Analysis*

One gene maybe can discover different Coherent Patterns with different sample set, in the same times points. But, these Coherent Patterns maybe have the same biological meaning. In this section, we will discuss the Rule Redundant Rate between PCC and TS^3. And observe the efficiency of rule discovery. We define the Rule Redundant Rate as follow function 2.

$$Rule\ Redundant\ Rate = \frac{Redundant\ Rule\ Number}{Total\ Rule\ Number} \qquad (2)$$

Figure 13 shows the result with different Minimum Support Threshold defines as 0.9 and 0.8 respectively. In any parameters, TS^3 is more efficiency then PCC manifestly.

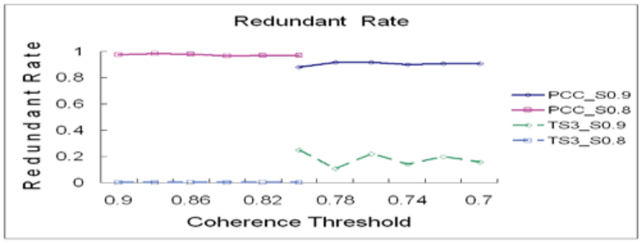


Figure 13. Redundant Rate of different similarity measurement method

V. Conclusion and Future Work

This research proposed the 3D-TDAR-Mine algorithm to discover the Temporal Dependency Association Rules in 3D microarray datasets. The rules discovered by 3D-TDAR-Mine can reveal gene expressions reactions and the regulation mechanism between reactions. Meanwhile, we proposed the TS^3 similarity measurement to estimate the Coherent Pattern that considers the Shifting, Scale and Trend factors in the real life. After rule discovering, we verify the Temporal Dependency Association Rules with KEGG Pathway and GO Semantic Similarity. It's also confirmed that the discovered rules are really with Shifting, Scale, Trend relation and biological meaning.

In the future, we will enhance the algorithm to discover the rules more efficiently in terms of time and memory cost. Furthermore, we will also try to use other data mining methods and biological information to filter out the rules without biologic meaning so that. we may help biologists use the rules more easily and efficiently.

References

[1] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.

[2] J. J. Foy and J. K. Bhattacharjee, "Gluconeogenesis in Saccharomyces cerevisiae: Determination of Fructose-1,6-Bisphosphatase Activity in Cells Grown in the Presence of Glycolytic Carbon Sources," American Society for Microbiology, February 1977; 129: 978 – 982.

[3] L. Ji, K. L. Tan and A. K. H. Tung, "Mining Frequent Closed Cubes in 3D Datasets," Proc. of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, pp 811 - 822. 2006.

[4] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, "The KEGG database at GenomeNet," Nucleic Acids Research, 30, 42-46.

[5] P. T. Spellman, G. Sherlock, M. Q. Zhang et al., "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization," Molecular Biology of the Cell, December 1998; 9: 3273 – 3297.

[6] G. Tatavarty, R. Bhatnagar and B. Young, "Discovery of Temporal Dependencies between Frequent Patterns in Multivariate Time Series," Proc. of the IEEE Symposium on Computational Intelligence and Data Mining, Honolulu, Hawaii, USA, pp 688 - 696. 2007.

[7] J. H. Toyn, A. L. Johnson, J. D. Donovan et al., "The Swi5 Transcription Factor of Saccharomyces cerezkiae Has a Role in Exit From Mitosis Through Induction of the cdk-Inhibitor Sicl in Telophase," Genetics Society of America, January 1996; 145: 85 – 96.

[8] M. Tyers and B. Futcher, "Farl and Fus3 Link the Mating Pheromone Signal Transduction Pathway to Three G1-Phase Cdc28 Kinase Complexes," Molecular and Cellular Biology, September 1993; 13: 5659 – 5669.

[9] J. Z. Wang, Z. Du, R. Payattakool et al., "A new method to measure the semantic similarity of GO terms," Bioinformatics, May 2007; 23: 1274 – 1281.

[10] X. Xu, Y. Lu, K. L. Tan and A. K. H. Tung, "Finding Time-Lagged 3D Clusters," Proc. of the 24th International Conference on Data Engineering, Shanghai, China, pp 445-456. 2009.

[11] L. Zhao and M. J. Zaki, "TRICLUSTER: An Effective Algorithm for Mining Coherent Clusters in 3D Microarray Data," Proc. of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, pp 694 - 705. 2005.