An Efficient Algorithm for Finding Continuous Coherent Evolution Bicluster in Time-series Data

Yun Xue[1], Jie Luo[1], Haolan Zhang[2,*], Zhengling Liao[1], Meihang Li[1], Qiuhua Kuang[1]
[1]School of Physics and Telecommunication Engineering
South China Normal University
Guangzhou, China
xueyun@scnu.edu.cn

[2]Center for SCDM, NIT
Zhejiang University
Hangzhou, China
haolan.zhang@nit.zju.edu.cn

***Abstract*—Most traditional biclustering algorithms focus on biclustering model on non-continuous column, which are not suitable for the analysis of the time series gene expression data. We proposes an effective and exact algorithm, which can be used to mine biclusters with coherent evolution on the contiguous columns as well as the complementary biclusters and time-lagged biclusters for the analysis of time series gene expression data. The experimental results show that the algorithm can find biclusters with statistical significance and strong biological relevance. We extend it to the currency data analysis in the financial field and obtain meaningful results.**

Keywords—time-series data; bicluster; coherent evolution; complementary; time-lagged

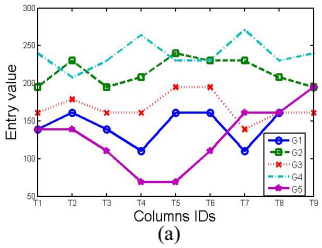
I. INTRODUCTION

It is an upsurge of gene analysis since the development of DNA microarray technology allows simultaneous measurement of a large number of gene expression levels under the given experimental conditions, which results in a large number of significant gene expression data. To extract relevant biological information from gene expression data is a very important and challenging task [1]. The gene expression data are usually stored in matrices whose rows and columns represent genes and conditions respectively, and each entry represents the expression level of a gene under a given experiment condition.

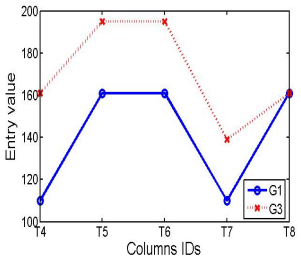
The traditional clustering techniques such as hierarchical clustering [2], self-organizing maps [3] and K-means clustering [4] have been widely used in the analysis of gene expression data. They are based on the similarity of gene expression patterns in all conditions. The genes are divided into mutually disjoint subsets. Each subset corresponds to a cluster, and those genes in the same cluster have the same regulation mechanism or biological function. However, the traditional methods have some disadvantages. First of all, according to the understanding of the cellular processes, some genes only have similar expression patterns under a particular subset of conditions with independent expression level under the other conditions. For example, TABLE I is a 5×9 matrix A. If considering the rows in all columns, as shown in Fig. 1(a), we cannot find the common patterns. However, if choosing columns from the fourth to the eighth, we can see the values of the first row and the third row show the same tendency, as shown in Fig. 1(b). In addition, a gene may participate in more than one biological process, so a gene may belong to more than one clusters, i.e. different subsets may overlap, but not disjoint.

TABLE I. THE ORIGINAL TIME SERIES GENE EXPRESSION DATA MATRIX A

Col / Row	T1	T2	T3	T4	T5	T6	T7	T8	T9
G1	139	161	139	110	161	161	110	161	195
G2	195	230	195	208	240	230	230	208	195
G3	161	179	161	161	195	195	139	161	161
G4	240	208	230	264	230	230	271	230	240
G5	139	139	110	69	69	110	161	161	195

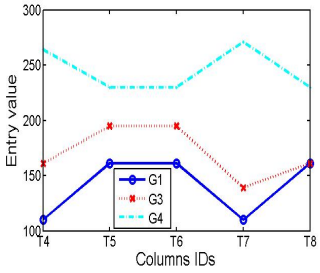


(a)



(b)

978-1-4799-4274-9/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDMW.2014.123



(c)

Fig. 1. (a) Visualization for the raw data matrix. (b)Row 1 and row 3 exhibit a coherent evolution pattern under 5 contiguous columns. (c)A complementary bicluster consisted of row 1, row 3 and row 4.

In order to overcome the defects of the traditional clustering, biclustering methods have been introduced in gene expression to find the local patterns. Many biclustering algorithms have been proposed [5-7].With simultaneous clustering of rows and columns, it could find out the subsets of genes of similar patterns under the specific conditions.

In this paper, we use biclustering techniques to analyze the time series gene expression data, which records cycles of cell activity or another biological process and the element values are the gene expression levels at different time points. These data describes the dynamic changes in the gene expression levels [8,9], so it could help us to study the time-dependent biological process, such as cell cycle, rhythmicity, development, disease progression, etc. Many biclustering algorithms have been proposed, but most of them are not suitable for the analysis of the time series gene expression data. Because they ignore some important characteristics of this kind of data, such as time point dependency and biological process inherent time contiguity. One of the most important is that, a cellular process usually lasts for a continuous time interval, so the genes involved in the cellular processes have similar expression patterns in a continuous time interval, i.e. local patterns found in the data are usually required across a continuous time interval. Therefore algorithms to find biclusters with non-continuous time intervals are meaningless and algorithms to analyze this kind of data are in pressing need.

Since the typical microarray data has a high level of noise and its values usually span a large range, this article focus on the coherent evolutionary trend among the rows. In other words, we are looking for biclusters whose rows show the same tendency (up or down) in the specific column sets with regardless of their actual value, as shown in Fig. 1(b). In addition, biclusters with complementary relationship are found, i.e. the cluster includes the same and opposite local gene expression patterns and the latter shows a reverse-regulatory process, as shown in Fig.1(c).

In order to solve the above problems, we propose a new biclustering algorithm on time series gene expression data with consideration of time-continuity, which can find all the biclusters whose number of rows and continuous columns meet the threshold. The main work in this paper is shown as following: (1) Take advantage of a difference matrix and a matching matrix to search the biclusters with coherent evolution on the contiguous k columns as well as the complementary biclusters and time-lagged biclusters. (2) Extend biclusters with coherent evolution on the contiguous k columns to gain biclusters with coherent evolution on the contiguous k+1 columns by jointing patterns. (3) Validate the algorithm's performance on the real yeast and extend it to the currency data analysis in the financial field.

The rest of the paper is as follows: In Section 2 some related works are briefly reviewed; the algorithm is detailed in Section 3; experimental results are given in Section 4; Section 5 concludes this paper.

II. RELATED WORKS

Hartigan [10] first proposed the idea of simultaneous clustering rows and columns. Later Cheng and Church [11] applied it to the analysis of gene expression data for the first time. They defined a mean square residue to measure the consistent correlation of the bicluster, and proposed a greedy algorithm to alternatingly delete genes and conditions. Therefore, the found submatrices correspond to highly correlated biclusters. The algorithm has been validated to have good performance to find a biclusters with coherent values in gene expression data.

After this algorithm, a large number of excellent biclustering algorithms have been proposed to analyze gene expression data. The algorithms define a variety of bicluster models and have good effects. In recent years, the biclustering algorithms for the time series gene expression data clustering have attracted widespread attention.

Zhang [12] proposed CC-TSB biclustering algorithm to analyze the time series gene expression based on the CC algorithm, which requires the increased/deleted columns should be continuous with the columns of the existing biclusters. So the results are time continuous biclusters. However, local optimal biclusters are obtained by its greedy search. Besides, once a bicluster is found and random numbers are introduced to substitute the corresponding original data, there is loss of information.

Ji and Tan [13] paid the same attention to mine biclusters with continuous columns. The original matrix is discretized and the sliding window is applied to reduce the complexity of the algorithm. Finally the time-lagged coherent evolution biclusters are searched and a reliable time-lagged coherent evolutionary information among genes is provided [13]. However, the algorithm's time complexity and space complexity will increase exponentially with the number of columns, so it is not suitable for realistic problems with many columns.

CCC-Biclustering algorithm [1] is raised by Sara C. Madeira et al. to look for all the maximum continuous consistently biclusters. The original matrix is discretized and the suffix tree is applied to linearize the algorithm's time complexity with the size of matrix. E-ccc bicluster algorithm [14] is the improvement of ccc-bicluster algorithm, which allows those genes supporting the pattern to have a given limited error. The experimental results

proved that the biological significance of the biclusters obtained by E-ccc bicluster algorithm is more significant. However, these two algorithms are unable to find the complementary continuous coherent evolutionary biclusters.

III. THE ALGORITHM

The paper presents a new biclustering algorithm on the time series gene expression data, which could find all the biclusters with coherent evolution on the contiguous columns as well as the complementary biclusters and time-lagged biclusters.

The original matrix is converted into the difference matrix at first, then the k-continuous common subsequences(k-CCS) between two sequences are found by the matching matrix, where k is the column threshold. The next step is to get the supporting row set of the CCS, which are combined with the columns belonging to the CCS, thus the biclusters with coherent evolution on the contiguous columns and complementary biclusters are formed. According to the merging rules, biclusters with continuous k+1 columns could be generated by biclusters with continuous k columns. The algorithm is simple and easy to implement without the introduction of complex data structure.

In this paper, the gene expression data is represented as a $n \times m$ data matrix A, where the row set represents the genes $G = \{G_1, G_2, G_3, \ldots, G_n\}$ and the column set represents the time points $T = \{T_1, T_2, T_3, \ldots, T_n\}$, $a_{i,j}$ is the element value of the ith row and jth column in A, which corresponds the gene G_i's expression level under the time point T_j.

Definition 1.Provided a submatrix $A'_{IJ} = (I, J)$, where $I = \{i_1, \ldots, i_k\}$ is the subset of the row set G, $J = \{j_1, \ldots, j_s\}$ is the subset of the column set T. If all the rows in A'_{IJ} show the same tendency(up or down) in any two continuous columns, then A'_{IJ} is a continuous coherent evolutionary bicluster.

To avoid meaningless biclusters with trivial rows or columns, the row threshold and the column threshold are set. Therefore, to find the continuous coherent evolution biclusters, one need to find the biclusters satisfying the row and column threshold in the gene expression data matrix A. As shown in Fig. 1(b), the elements in the row 3 and the row 4 show the same tendency from the column 4 to the column 8. If the row and the column threshold are set as 2, then $A'_{IJ} = (\{3,4\}, \{4,5,6,7,8\})$ is a bicluster with coherent evolution on the contiguous 5 columns.

The detail of the algorithm is described as the following:

A. Matrix Discretization

Coherent evolutionary biclusters focus on the trend of genes at different time points, therefore, we first transform the matrix A into a $n \times (m-1)$ difference matrix B, which reflects the trend of genes at different time points. The element $b_{i,j}$ in the matrix B records the trend of $a_{i,j+1}$ from the time point T_j to the time point T_{j+1}. We use the digits 1, 0, and -1 to represent the trend up, unchanged and trend down respectively. The difference matrix B is a $n \times (m-1)$ matrix which has three values -1, 0 and 1.

As there exists certain measure errors in the experiment, in order to avoid the influence of extreme values, we set a rate threshold of change θ to limit the range. If the values of two columns changes in a certain range , it is considered as unchanged, otherwise it is considered as increasing or decreasing. θ can be adjusted according to the actual value of the data matrix. Then the entries of difference matrix B could be calculated as (1), where $C = \frac{A_{i,j+1} - A_{i,j}}{|A_{i,j}|}$, θ is the threshold which can be modified.

$$B_{ij} = \begin{cases} 1, & if \quad C > \theta; \\ 0, & if \quad -\theta < C < \theta; \\ -1, & if \quad C < -\theta; \end{cases} \qquad (1)$$

B. Build the Matching Matrix

Each row in the matrix B could be regarded as a sequence composed of {-1,0,1}. The common subsequences between the two rows correspond to the potential local expression patterns. Therefore, in order to find all the continuous common subsequences between any two rows, the matching matrix C is constructed by any two rows in the matrix B.

The construction rule of the matching matrix is as following: Given any two rows u, v in the matrix B, to construct the matching matrix $C_{<u,v>}$, whose row index is from the row u and the column index from the row v, then the element $C_{<u,v>}(i, j)$ could be derived from (2):

$$C_{<u,v>}(i,j) = \begin{cases} 0, & \text{if } B_{v,i} = B_{u,j} = 0 \\ 1, & \text{if } B_{v,i} = B_{u,j} = -1 \text{ or } B_{v,i} = B_{u,j} = 1 \\ -1, & \text{if } B_{v,i} = 1, B_{u,j} = -1 \text{ or } B_{v,i} = -1, B_{u,j} = 1 \\ null, & otherwise \end{cases} \qquad (2)$$

To construct the matching matrix C by all the pair rows in matrix B, there could be C_n^2 combinations. S_i is the ith sequence(row) in the B matrix. For example, the sequence $S_3 = \{01100(-1)0(-1)001\}$ and $S_7 = \{001100(-1)(-1)00(-1)\}$ could be used to form the matching matrix $C_{3,7}$ according to the above rules, as shown in TABEL II.

TABLE II. THE MATCHING MATRIX $C_{3,7}$ FORMED BY THE 3RD AND THE 7TH ROW IN MATRIX B

	0	**1**	**1**	**0**	**0**	**-1**	**0**	**-1**	**0**	**0**	**1**
0	0			0	0		0		0	0	
0	0			0	0		0		0	0	
1		1	1			-1		-1			1
1		1	1			-1		-1			1
0	0			0	0				0	0	
0	0			0	0				0	0	
-1		-1	-1			1		1			-1
-1		-1	-1			1		1			-1
0	0			0	0				0	0	
0	0			0	0				0	0	
-1		-1	-1			1		1			-1

C. *Recognize the Continuous K-columns Bicluster*

The column threshold K is set as 3. We search column patterns composed of continuous k elements from the principal diagonal and other secondary diagonals in matching matrix. The information in principal diagonal is used to search biclusters with coherent evolution on the contiguous columns and the complementary biclusters. The information in secondary diagonal is used to search time-lagged biclusters. In particular, 1 and -1 cannot occur at the same time in any contiguous column patterns because of the construction rule for matching matrix. As the value of corresponding position in the matching matrix is 1 when the trend of two subsequences is the same. Otherwise if it is opposite, the value is -1. If 1 and -1 appear in the diagonal simultaneously, which means that the two sequences show neither same nor opposite pattern on the continuous k columns, so it is out of interest and not necessary to store the information. As shown in Fig. 3, the line G2 and G4 have different trends. After differential discretization the trends of the two lines is (1, -1) and (1, 1) respectively. In the matching matrix, the diagonal elements in the matrix are (1, -1), which shows the two lines don't contain an identical pattern or an opposite pattern in the three continuous columns.

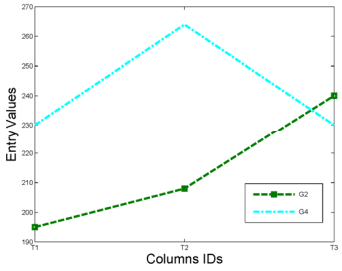


Fig. 2. Patterns without interest.

The information of useful patterns in the diagonals are stored. The subsequences corresponding to each diagonal including its starting column's index and its supporting rows' index are saved. Each subsequence corresponds to a potential pattern composed of continuous k columns. For example, As showed in TABLE III, the corresponding patterns of the first continous sequence (0, 0, 1) in the second diagonal in TABLE II are {0,0, -1} and {0,0,1}. The beginning columns' index of the two subsequence are 4 and 1, while the supporting rows' index are 3 and 7 respectively.

The above operations is performed C_n^2 in the matrix B and the information of the patterns composed of the continuous k elements is stored in TABLE III. As the subsequence of a row could match the subsequences of multi-rows or the other subsequences starting from the different time point of another row, there are redundant rows which should be removed in TABLE III. In order to get the target biclusters, the following two steps are implemented:

In the first step, to get all the biclusters formed by continuous k columns, we compute the intersection of the corresponding supporting rows of all the same pattern pairs whose starting columns' index are identical in TABLE III.

The second step, to get all the complementary biclusters formed by continuous k columns, we find out all the opposite pattern pairs whose starting columns' index are identical and only record one pattern of them. In order to identify which part of row set is in support of the recorded or opposite pattern, the symbol "|" is introduced to set the two row sets apart. The ahead of it is the row set supporting the recorded pattern and the hinder is row set supporting the opposite pattern.

After the above two steps, the results are shown in TABLE IV. If the supporting rows of a pattern satisfy the supporting threshold, it could form a continuous-k-columns coherent evolution bicluster. If the union set of supporting rows of a pattern and its opposite pattern satisfy the supporting threshold, then the two patterns and the union set could make up a complementary bicluster.

TABLE III. THE INFORMATION OF THE PATTERNS IN DIAGONALS

Pattern	**The beginning column**	**Supporting row**
00-1	4	3
001	1	7
011	1	3
011	2	7
110	2	3
110	3	7
-100	8	3
-100	8	7
001	9	3
00-1	9	7
…	…	…

TABLE IV. THE INFORMATION OF CONTINUOUS K-COLUMNS PATTERNS

Pattern	**The beginning column**	**Supporting row**
00-1	2	(2,4|3,5,8)
100	1	(2,3,4|5,8,9)
0-11	3	(2,4|3,8,9)
011	4	(1,5,7|3,8,9)
-1-11	5	(3,9|1,5,6)
011	6	(3,7|null)
…	…	…

D. Joint

The continuous k+1 columns coherent evolution biclusters could be obtained on the basis of the continuous k column coherent evolution biclusters, which should fulfill the following conditions:

Given two continuous k-column patterns,

a) The difference of the initial columns' index of the two continuous k-patterns is 1;

b) The tail k-1 elements of the pattern are the same or opposite as the former k-1 elements of the other pattern;

c) The row support of the continuous k+1 column pattern should be greater than the support threshold.

According to the above conditions, for each row in TABEL III, we are looking for the rows whose starting column number's difference is 1 and the tail k-1 elements of the pattern are the same or opposite as its former k-1 elements. Although only one of the complementary patterns is recorded, the row sets of both complementary patterns are stored, so each row corresponding to a pair of opposite patterns could be jointed with other rows. For example, the fourth row's pattern in TABEL III is {0,1,1} and the fifth row's is {-1,-1,1}. The posterior 2 elements of the fourth row {1,1} are opposite as the previous 2 elements of the fifth row {-1,-1}. There is a need to joint the pattern of the fourth row {0,1,1} with the opposite pattern of the fifth row {1,1,-1} and the opposite pattern of the fourth row {0,-1,-1} with the pattern of the fifth row {-1,-1,1}. The results are {0,1,1,-1} and {0,-1,-1,1} respectively. It can be seen that the two patterns are complementary.

In order to verify the conditions c, the intersection of the corresponding row sets is executed. For example, to judge whether the support rows of {0,1,1, -1} and {0,-1,-1,1} meet the threshold, the row set (1,5,7) is intersected with the row set (1,5,6) and the same operation is done between the row sets (3,8,9) and (3,9) . The results are shown in TABLE V.

TABLE V. EXAMPLE OF (K+1) PATTERN

Pattern	The beginning column	Supporting row
011-1	4	(1,5|3,9)

After each row in TABLE IV has been jointed together, TABLE VI could be obtained.

TABLE VI. THE INFORMATION OF CONTINUOUS (K+1)-COLUMNS PATTERNS

Pattern	The beginning column	Supporting row
100-1	1	(2,4|5,8)
1001	1	(3|null)
0101	2	(3,6,9|5)
…	…	…
011-1	4	(1,5|3,9)
-1-110	5	(3,6,7|5,8)
…	…	…

If the row support of the merged pattern meet the threshold, then the supporting rows could form a continuous k+1 columns coherent evolutionary bicluster. In addition, if the intersection of the row sets of the merged pattern and its opposite pattern satisfy the threshold, the couple of patterns and their row sets' union make up a complementary continuous k+1 columns coherent evolutionary bicluster.

The above operation is continued until no rows could be jointed.

Another advantage of the method is that it could find the patterns with time delay. Because in the microarray experiments, the tests for different genes could start on different initial time and some genes begin to be co-regulated at different time points, but similarity will be exhibited after a certain period of time [8], so it is necessary to find time-lagged biclusters. By the method it is easy to find the time-lagged biclusters. As shown in TABLE VII, the pattern {0,1,1} has two beginning columns, which could be combined to form a time-lagged coherent evolutionary bicluster.

TABLE VII. TIME-LAGGED COHERENT EVOLUTIONARY BICLUSTER

Pattern	The beginning column	Supporting row
011	1	3
011	2	7

IV. THE EXPERIMENTAL RESULTS AND ANALYSIS

A. Results on Yeast Microarry Data

The time series gene expression data is yeast data in [7], which records 2884 genes' expression level at 17 sampling points in 10 minutes, almost covering 2 intact yeast cell cycle. The missing values are deleted and the final size is 2267×17 .

1) Statistical significance: Mean square residue(MSR) is emplyed in many papers to evaluate the biclusters. It is used to measure the degree of association of a group of genes in a set of time points. The smaller the MSR is, the higher degree of association the genes have. To consider a continuous coherent evolution bicluster $A'_{IJ}=(\mathrm{I},\mathrm{J})$, let a'_{ij} denote the element's value in the ith row and the jth column, a'_{Ij} is the mean of the Ith row, a'_{iJ} is the mean of the Jth column, a'_{IJ} is the mean of the biluster. The MSR of $A'_{IJ}=(\mathrm{I},\mathrm{J})$ is calculated as follows.

$$\mathrm{MSR}(\mathrm{I},\mathrm{J})=\frac{1}{|I||J|}\sum_{i\in I,j\in J}(a'_{ij}-a'_{iJ}-a'_{Ij}+a'_{IJ})^2 \quad (3)$$

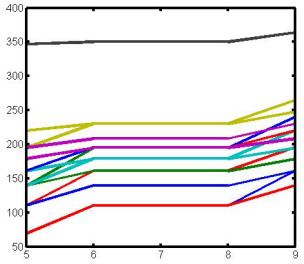
According to [7], a low mean square residue means the expression levels fluctuate consistently in the bicluster.

In experiment the row threshold is 20, the column threshold is 5 and θ is 0.001. MSR of each bicluster is computed and sorted in ascending order. TABLE VIII gives the information of the top 3 biclusters, which are plotted in

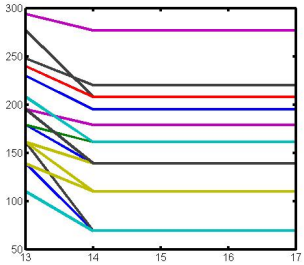
Fig.3. It can be seen that the biclusters have coherent evolutionary pattern in continuous columns.

TABLE VIII. THE TOP 3 BICLUSTERS WITH LOWEST MSR

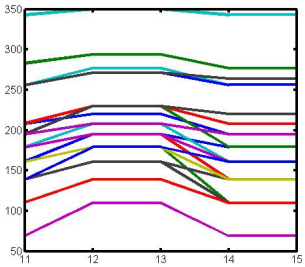
Bicluster ID	The number of rows	The number of column
1	26	5
2	42	5
3	36	5



a)Bicluster ID 1



b)Bicluster ID 2



c)Bicluster ID 3

Fig. 3. The top 3 biclusters with lowest MSR.

2) GO analysis: In order to analyze the biological significance of the biclusters, the GOToolbox [15] is applied to analysis of the biclusters. GOToolbox is a Web based application, allowing the users to query, browse and visualize the gene ontology annotation data. So far, this project contains most of the plant, animal and microbial genome database. We use this tool to analyze the biclusters including analysis of GO Term , GO level and other aspects.

TABEL IX shows some genes with biological significance. It can be seen that the levels of the biclusters are high, the P-values are small and GO-Terms are apparent, which indicates that the found biclusters have high biological significance.

TABLE IX. GO ANALYSIS OF THE BICLUSTERS

Bicluster ID	Genes	Level	GO-Term	P-Value
1	GO:0008272	7,8	Sulfate transport	0.0015731
1	GO:0015698	6,7	Inorganic anion transport	0.0065902
2	GO:0006529	8,9	Asparaqine biosynthetic process	0.0081388
3	GO:0006003	8,9	Fructose 2,6-bisphosphate metabolic process	0.0025149
3	GO:0006110	8,7,9,10,11,6	Regulation of glycolysis	0.0037688

3) The complementary biclusters: This algorithm could also find the biclusters have both similar and opposite local patterns, as shown in Fig. 4.

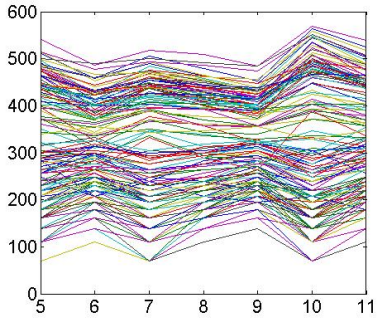


Fig. 4. The complementary bicluster.

A similar local expression pattern for genes may show the same regulation mechanisms or biological functions in the specific time period, while the complementary local pattern may show negative correlation. For example, some genes may inhibit others' expression levels or they go on the opposite paths.

B. Resluts on Exchange Rate Data

The exchange rate dataset (17×120) in [16] is used to validate the algorithm analysis in other fields of time series data. The relations among different exchange rates help to study economic relation among countries, which provides instruction for economists and governments to draft economic policies. At the same time, the analysis of the exchange rates provide guidance for investors to invest in the foreign exchange market. It includes 17 currencies, such as ARS (Argentine Peso), AUD (Australian Dollar), BRL (Brazilian Real), CAD (Canadian Dollar), CHF (Swiss Franc), EGP (Egyptian Pound), GBP (British Pound), IDR (Indonesian Rupiah), INR (Indian Rupee), JPY (Japanese Yen), MXN (Mexican Peso), PHP (Philippine Peso), RUB (Russian Rouble), SGD (Singapore Dollar), THB (Thai Baht), TWD (Taiwan Dollar) and ZAR (South African Rand). The time period spans 120 months from January 1, 1996 to December 31, 2005. Each element's value is a currency's average exchange rate valued against the USD (US Dollar).

1) The analysis: In the second experiment, the row threshold is 3, the column threshold is 10 and θ is 0.001. In Fig. 5, three kinds of exchange rates reveal coherent evolution in 11 consecutive months. The currencies are INR (Indian Rupee), THB (Thai Baht) and TWD (Taiwan Dollar) and the time period is consecutive 11 months from April, 2004 to February, 2005.

The bicluster reflects the potential close relation in the economic among India, Thailand and Taiwan. As we all know, the three countries locate in Southeast Asia. The close geographical position makes the climate, hydrology, topography and other features of the three countries quite similar, which leads to the close link in the economic development of the countries. In Fig. 5, the curves show same tendency. In addition, if the analysis goes further, the exchange rates of the three countries continued to decline from November 2004, the reason is speculated that the big tsunami happened in the India Ocean, which spread to Southeast Asia and South Asia countries. The disaster caused great damage to the Southeast Asia and South Asia.

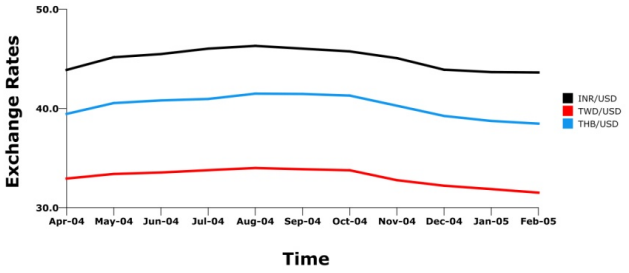


Fig. 5. Three exchange rates show the same tendency in 11 consecutive months.

V. CONCLUSION

In this paper, we propose an effective and exact algorithm, which can be used to mine biclusters with coherent evolution on the contiguous columns as well as the complementary biclusters and time-lagged biclusters for the analysis of time series gene expression data. We take advantage of a difference matrix and a matching matrix to find the patterns composed of k columns and then joint them together step by step into patterns with multiple columns. In order to verify the performance of the algorithm, we use the real gene expression data of yeast and exchange rate data as the experimental data, and carry out the corresponding analysis on the experimental results.

VI. ACKNOWLEDGMENT

This work is supported by Guangdong Science and Technology Department under Grant No.2012B091100349; Guangdong Economy & Trade Committee under Grant No. GDEID2010IS034; National Natural Science Foundation of China (Grant No.71102146); Research supported in part by the Guangdong Nature Science Fund (No. S2012010010661).

REFERENCES

[1] S. C. Madeira and A. L. Oliveira, "A linear time biclustering algorithm for time series gene expression data," 5th International Workshop on Algorithms in Bioinformatics, Spain,vol 3692, pp. 39-52, 2005.

[2] L. Kaufman and P.J. Rousseeuw, "Finding groups in data: an introduction to cluster analysis," Wiley-Interscience, 1990.

[3] P. Törönen, M. Kolehmainen, G. Wong and E. Castrén, "Analysis of gene expression data using self-organizing maps," vol.451, 1999, pp. 142 - 146.

[4] J. B. Macqueen, "Some methods for classification and analysis of multivariate observations", University of California Press, 1967, pp. 281 - 297.

[5] A. Ben-dor, B. Chor, R. Karp and Z. Yakhini, "Discovering local structure in gene expression data: the order-preserving submatrix problem". Journal of Computational Biology : a Journal of CO, vol. 10, 2003, pp. 373-384.

[6] G. Getz, E. Levine and E. Domany, "Coupled two-way clustering analysis of gene microarray data," Proceedings of the National Academy of Sciences, vol. 97, 2000, pp. 12079-12120.

[7] J. Yang, H. Wang, W. Wang and P. Yu, "Enhanced biclustering on expression data," In Proc. of 3rd IEEE Symposium on Bioinformatics and Bioengineering, 2003, pp. 321-327.

[8] J. Qu, X. Zhang, L. Wu, Y. Wang and L. Chen, "Detecting coherent local patterns from time series gene expression data by a temporal biclustering method," Systems Biology (ISB), 2011 IEEE International Con, 2011, pp. 388-393.

[9] J. Qi, R. Zhang, K. Ramamohanarao, H. Wang, Z. Wen, and D. Wu, "Indexable Online Time Series Segmentation with Error Bound Guarantee," World Wide Web Journal, 2013, pp.1-43.

[10] J. A. Hartigan, "Direct clustering of a data matrix," Journal of the American Statistical Association, vol. 67, 1972, pp. 123-129.

[11] Y. Cheng and G. M. Church, "Biclustering of expression data," Proceeding of Intelligent System for Molecule BIOL, vol. 8, 2000, pp. 93-103 .

[12] Y. Zhang, H. Zha and C.H. Chu, "A time-series biclustering algorithm for revealing co-regulated genes," In Proc of the 5th IEEE international Conference on Information Technology: Coding and Computing, 2005, pp. 32-37 .

[13] L. Ji and K. Tan, "Identifying time-lagged gene clusters using gene expression data," Bioinformatics, vol. 21, 2005, pp. 509-516.

[14] S. C. Madeira and A. L. Oliveira, "An efficient biclustering algorithm for finding genes with similar patterns in Time-series expression data," Imperial College Press, 2007, pp. 67 - 80.

[15] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry and B. Jacq, "GOToolBox: functional analysis of gene datasets based on Gene Ontology," Genome Biology, vol. 5, 2004.

[16] H. Li and H. Yan, "Bicluster Analysis of Currency Exchange Rates," Bhanu Prasad. Studies in Fuzziness and Soft Computing, 2008, pp.19-34.