Detecting Coherent Local Patterns from Time Series Gene Expression Data by a Temporal Biclustering Method

Ji-Bin Qu
Xiang-Sun Zhang
Ling-Yun Wu
Yong Wang
Institute of Applied Mathematics Academy of Mathematics and Systems Science, CAS, Beijing 100190
Email:qujibin@amss.ac.cn
zxs@amt.ac.cn
wulingyun@gmail.com
ywang@amss.ac.cn

Luonan Chen
Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Center for Pre-diabetes, Shanghai Institutes for Biological Sciences, CAS.
Shanghai 200031
Email: lnchen@sibs.ac.cn

***Abstract*—Time-series gene expression data analysis plays an important role in bioinformatics. In this paper, we propose a biclustering method to detect local expression patterns in time-series gene expression data by performing clustering on both gene and time dimensions. Our method aims to find gene subsets which show coherent expression profiles in some time subsets which have a consecutive order in a bicluster. Specifically, our temporal biclustering method is composed of a discretization procedure and a follow-up sequence alignment, which can identify similar local expression profiles and further reveal coherent local relations such as complementary and time-lagged coherence. We apply our method to yeast cell cycle data, and find several biologically important biclusters.**

I. Introduction

Gene expression data records the concentration of mRNAs in given conditions and plays an important role in understanding large-scale biological systems. In particular, time-series gene expression data is composed of a series of experiments recording the mRNA concentration in different time points. This type of data can describe the dynamical changes of gene expression, and thus it is helpful to characterize the time-dependent biological processes, for example cell cycle, rhythmicity, development, disease progression, and so on [1][2]. Computational methods to analyze time-series gene expression data are in pressing need.

A great number of clustering methods, such as hierarchical clustering and k-means algorithm, have been designed to group genes or conditions into subsets based on their gene expression profiles [3]. The underlying assumption is that genes in the same subset perform the coherent functions or regulatory mechanisms, and experimental conditions in the same subset are coherent(for example the similar growing environment or the same disease). Traditional clustering methods group either genes or conditions, but it is more meaningful to cluster the two factors simultaneously, that is, some genes characterizing a special cellular processes share similar expression patterns at a specific period. In many situations, biologists believe that a cellular process is active only under a subset of conditions [4].

As a result, biclustering methods have been suggested to identify coherent local profiles in gene expression data. The resulting bicluster is defined as a subset of genes that exhibit compatible expression pattern over a subset of conditions, which may be a transcription module or an active pathway. In time-series gene expression data, the condition is time point, and thus it is naturally to require that the time points in a subset must be consecutive [5]. We note that some temporal biclustering algorithms have been proposed to analyze time series data and provides in biological meaningful results [6][7].

In this paper, we propose a new biclustering method by considering the time consecutiveness in time-series gene expression data. This method is based on discretization preprocessing [4][8] and sequence alignment [9]. In next section, we describe the framework of our method. Some computational results in yeast cell cycle data are shown in the third section. Finally we analyze the results and discuss the biological insights of our results.

II. Method

We denote a microarray dataset as a $N \times M$ matrix. Each row is the profile of a gene in all conditions and each column is an array for all the genes in a condition. Suppose there are N genes and M time points (generally $N \gg M$). Each bicluster is described as a submatrix $\{N_i, M_i\}$ for $i = 1, \cdots, K$, N_i is the subset of genes and M_i is the subset of conditions for the ith bicluster. In our model, the condition is time point, and therefore the time points in a bicluster must be consecutive, i.e., M_i should consist of consecutive time points.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

A. Discretization preprocessing

Firstly we preprocess the raw gene expression matrix by a discretization technique. There have been many discretization techniques specifically for time series gene expression data to detect the transitions in expression patterns between successive time points. Regarding to the impact of discretization on biclustering, we find that the techniques based on transformations between time-points obtain better results than those using absolute values [10].

Let $X = \{x_{i,j}\}_{N\times M}$ be the raw data. We aim to group genes having similar expression profiles, i.e., the abundant of these genes changed synchronously. As the first step, we transform X to matrix $Y = \{y_{i,j}\}_{N\times(M-1)}$, $y_{i,j} = x_{i,j+1} - x_{i,j}$. Matrix Y describes that how genes varies in different time intervals. Next step, the matrix Y is transformed to $Z = \{z_{i,j}\}_{N\times(M-1)}$. In matrix Z, K symbols are used to represent the varieties of the expression profiles. In other words, we divide the values of matrix Y into K bins. To avoid the impact of extreme values, each bin has the similar number of figures and is represented by a symbol (see Figure 1). To realize this we divide the normal distribution of Y into intervals and identify each interval that contains $\frac{1}{K}$ of the values from the normal distribution. Finally values in each interval of Y are represented by a single symbol in Z.

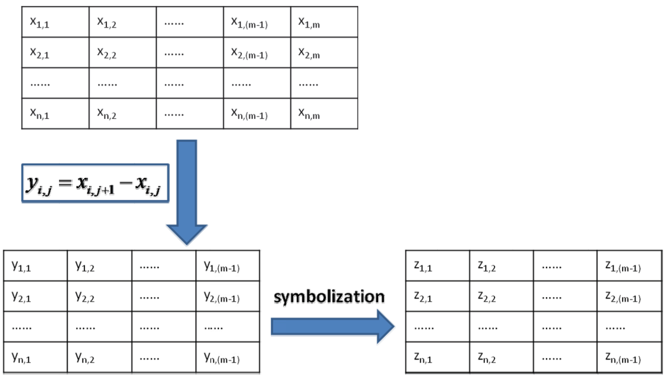

Fig. 1. The discretization preprocessing procedure for time series gene expressing data.

B. Revealing the local expression patterns by sequence alignment

From the symbolized matrix Z, we try to identify local expression patterns in biclusters. Here we name a kind of expression profiles with specific shape as an expression pattern. In order to get all the potential patterns, we utilize the sequence alignment method. The expression profile of each gene can be seen as a sequence of symbols, then we can find the common sub-sequences between every pair of genes by pairwise alignment, and every sub-sequence corresponds to a local expression pattern.

In our model, the alignment ignores insertion, deletion, and replacement. We only consider whether the same positions of both sequences are repeated. Therefore, there are only diagonal scores in the dot matrix $s(i), i = 1, 2, \cdots, M-1$. Given two sequences $A = (a_1, a_2, \cdots, a_{M-1})$ and $B = (b_1, b_2, \cdots, b_{M-1})$, we define the score function of the ith position in diagonal of dot matrix as follows:

$$s(i) = \begin{cases} s(i-1) + m(a_i, b_i), & |s(i-1) + m(a_i, b_i)| \\ & > |s(i-1)| \\ 0, & otherwise \end{cases} \tag{1}$$

where

$$m(a_i, b_i) = \begin{cases} 1, & a_i = b_i \\ -1, & a_i = -b_i \\ 0, & otherwise \end{cases} \tag{2}$$

When we calculate all the scores in the diagonal, the maximum is the length of longest common sub-sequence. After pairwise alignment through all pairs of genes, we get a pattern list P which stores all longest common sub-sequences between gene pairs and their positions. Our score function can find not only same sub-sequences, but also complementary sub-sequences which mean that some local profiles of both genes have inverse shape.(see Figure 2)

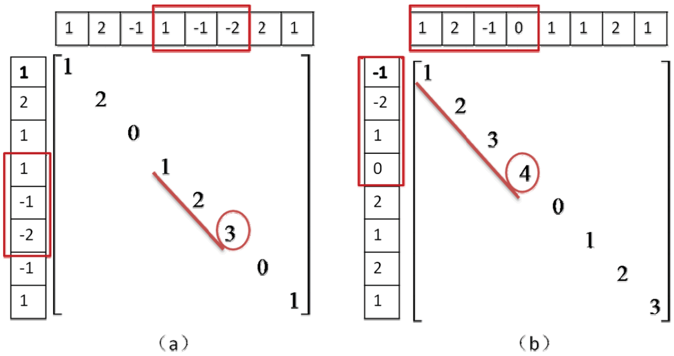

Fig. 2. Some examples to illustrate the sequence alignment procedure: (a) The longest common sub-sequence with length 3 and (b) A reverse longest common sub-sequence with length 4.

C. Detecting the biclusters

A naturally assumption is that longer patterns tend to have real biological meanings. So we determine biclusters by a greedy strategy to pick out long local expression patterns. In beginning, we pick out the longest sub-sequence from the pattern list, and then the time interval is confirmed, i.e., the position of the sub-sequence. In next step we will find genes with the same or complementary sub-sequences in this time intervals, and derive the gene subset. The gene subset and time interval together constitute a bicluster. We then delete this pattern from the pattern list and repeat the above procedures until the list is empty. Finally, we get a series of biclusters.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

The resulting biclusters are in large number and we adopt some strategies to narrow down the biclusters. Since it is more likely that short patterns appear by chance, we remove the biclusters whose time points are less than N_1. We also remove the biclusters with the number of genes less than N_2. And then, there are overlaps between biclusters. We combine those similar biclusters into a representative one. The following score is defined to measure the similarity of two biclusters.

Given two biclusters $B_1 = (I_1, J_1)$ and $B_2 = (I_2, J_2)$, I is the row index set and J is the column index set. $|B_1| = |I_1| \times |J_1|$ and $|B_2| = |I_2| \times |J_2|$ mean the number of elements in the two biclusters. The similarity score is defined as follows:

$$J(B_1, B_2) = J((I_1, J_1), (I_2, J_2)) = \frac{B_1 \cap B_2}{B_1 \cup B_2} \tag{3}$$

$|B_1 \cap B_2| = |I_1 \cap I_2| \times |J_1 \cap J_2|$ is the number of elements in the intersection set of B_1 and B_2, and $|B_1 \cup B_2| = |B_1| + |B_2| - |B_1 \cap B_2|$ is the number of elements in the union set of B_1 and B_2. We set a threshold and merge the biclusters with scores larger than the threshold.

III. Computational results

To validate our method, we test it in the yeast gene expression datasets by Spellman [11]. This data was produced to study the temporal expression profiles of genes involved in cell cycle. We select the α factor dataset, which contain 6178 genes in 18 time points. After removing the missing data, 4489 genes are left. In our experiments, we set $K = 5$, $N_1 = 8$, and $N_2 = 10$, similarity threshold 0.5. Finally our method reveals 128 biclusters in this dataset.

These biclusters have several types. Some include genes with similar local patterns (see Figure 3) and some include genes with both similar local patterns and complementary local patterns (see Figure 4). We note that genes in the same bicluster have high coherence in the time interval of the bicluster (more than 0.9), but low coherence in all time points (less than 0.4). Genes with similar local patterns mean that they are co-expressed in the specific time intervals and may involve in a similar regulatory process, and genes with complementary local patterns are in reversed regulatory process. For example some genes are activated while others are inhibited.

One advantage of our method is that we can reveal the time-lagged patterns. It is apparent that many genes do not regulate each other simultaneously but after a certain time lag, so the expression profiles should have time-delay[12]. By our method it is easy to detect this case: every bicluster corresponds to a sub-sequence, we just compare the sub-sequences among these biclusters and pick out those with coherent sub-sequences but delayed for a time interval. Commonly the delayed time is less than two hours. In our result, we find many time-lagged biclusters. (see Figure 5)

CCC-biclustering method [6] is an efficient temporal biclustering method developed in recent years. The aim of our method and CCC is similar, that is, detecting the coherent sub-sequences among genes in specific time intervals from discretized data. Therefore, our method performs similarly in

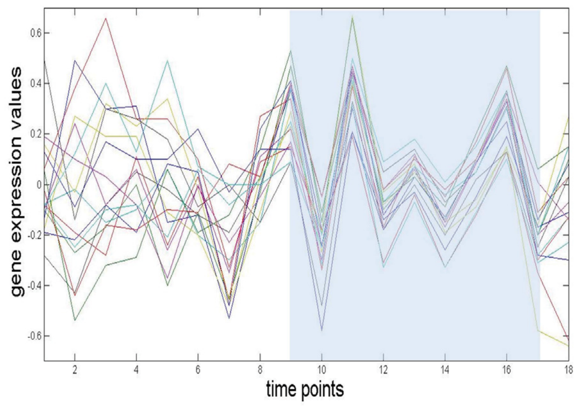

Fig. 3. A bicluster with similar gene expression pattern. Bicluster 12 contains 17 genes: YBR280C, YDL075W, YDL081C, YDR047W, YDR073W, YDR169C, YDR187C, YDR455C, YER044C, YER090W, YER162C, YGL067W, YHR142W, YIL076W, YIL091C, YJL014W, and YOL053W. The time interval is from time point 9 to 17.

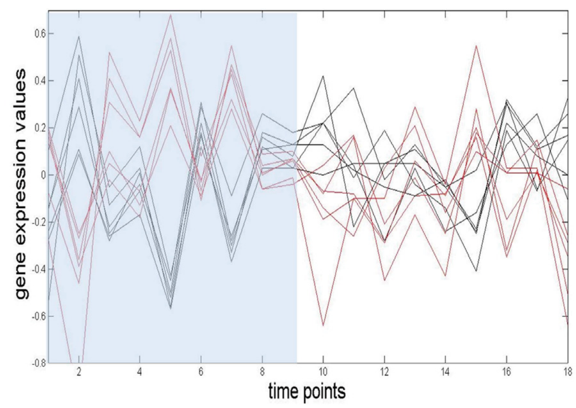

Fig. 4. A bicluster with complementary gene expression pattern. Bicluster 9 contains 12 genes: YBL030C, YCRX07W, YDL012C, YDL080C, YDR355C, YDR410C, YER161C, YGR129W, YJL197W, YJR141W, YKR088C, and YMR303C. The time interval is from time point 1 to 9.

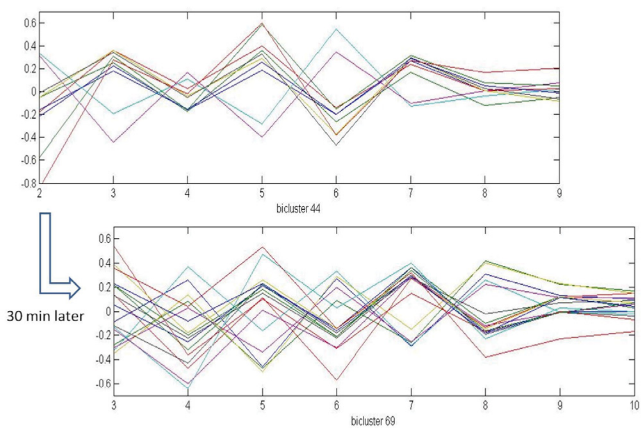

Fig. 5. Two biclusters with time-lag relationship. Bicluster 44 and 69 have the same expression pattern and a time-lag of 30 minutes.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

terms of accuracy comparing with CCC. However, our method has several advantages. Firstly our discretization method divides the raw data homogeneously by the normal distribution, rather than CCC which sets cutoffs to divide raw values. Our method can avoid the impact of extreme values and the discretized matrix Z is more homogeneous. On the other hand, CCC uses the theory of suffix trees. However our method is easily interpreted and accessible to biologists.

IV. Discussions and conclusions

In this section, we analyze the biological significance inside the biclusters identified by our method. As we know, genes sharing coherent expression patterns are thought to have co-expression relationships, and this phenomenon may be caused by co-regulated behavior. Co-expression genes have tight relations such as sharing similar regulatory mechanisms or executing similar functions.

A. biclusters in biological networks

If genes are regulated by the same mechanism, a intuitive hypothesis is that either some genes are regulated by the same transcription factors (TFs), or some belong to a protein complex. In the former case, the distances among genes in a protein-protein interaction networks (PPI) are 2, i.e., they have common upstreams; and in the latter case, the genes in a complex are adjacent, the distances among them are 1. We compute the average shortest paths (ASPs) for genes in the same biclusters. Almost all ASPs are between 2 and 3 (see Figure 6), and most distances of genes in the same bicluster are 1 or 2.

To check the significance of this conclusion, randomized biclusters are sampled from the whole PPI network of BioGRID [13]. 98 biclusters in all 128 have significantly shorter ASPs than randomized cases.

Let's take bicluster 5 as an example. There are 10 genes in this bicluster: $YAR009C$, $YLR109W$, $YLR179C$, $YLR197W$, $YLR406C$, $YLR441C$, $YNL096C$, $YNL244C$, $YOL139C$, and $YPR044C$. 9 of them appear in the whole PPI network. The distances among these genes are completely 1 or 2. $YLR406C$, $YLR441C$, and $YNL096C$ constitute the ribosomal protein complexes and they are adjacent in the network. We also find the common TFs regulating these genes in YEASTRACT database [14], Yap1p regulates all 9 genes in this bicluster and many co-regulatory relationships can be observed. (see Figure 7)

B. Gene functions enriched in biclusters

Next we check the coherence of gene functions in biclusters. There exist many methods and computational tools to annotate gene functions and compute the significance of functions in gene sets [15][16]. We choose the g:profiler software [15] to find the enriched functions in our identified biclusters.

As we discussed above, long patterns are more likely to have real coherent relationships. In all the 128 biclusters, the longest one contains 11 time points and the second longest bicluster has 10 time points. Both biclusters have many significant

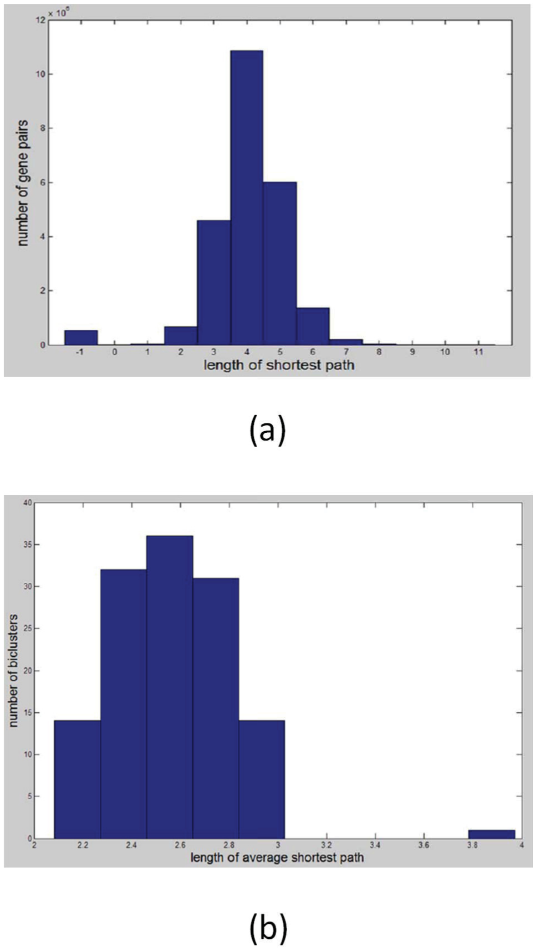

Fig. 6. (a) The shortest path distribution for the whole PPI network, the average value isabout 4; (b) The average shortest path of our 128 biclusters. Most APSs are near 2.5 and shorter than the random case.

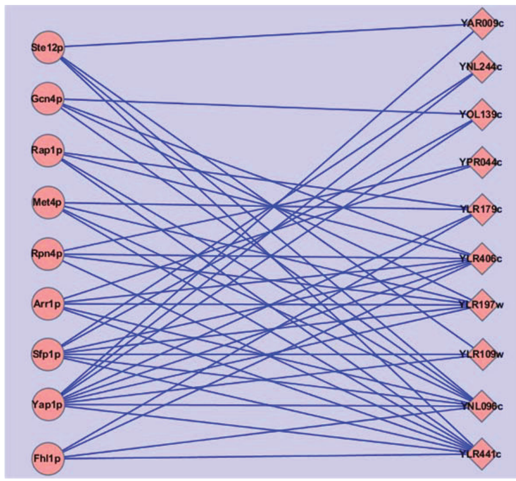

Fig. 7. The regulatory relationships in bicluster 5.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

TABLE I
LIST OF ENRICHED GO FUNCTION ANNOTATIONS FOR SOME EXAMPLE BICLUSTERS

Bicluster ID	Genes	Enriched functions		p-value
2	YAR009C YBL095W YBR112C YFL057C YJR139C YLR109W YLR179C YMR162C YNL096C YOL040C YOL139C YOR167C YOR285W YPL079W YPL081W YPR044C	GO:0022627	cytosolic small ribosomal subunit	1.40E-05
		BIOGRID:00000	BioGRID interaction data	1.98E-05
		GO:0022626	cytosolic ribosome	4.01E-05
		REAC:503952	Ribosomal scanning	6.75E-05
		REAC:502542	Formation of translation initiation complexes containing mRNA that does not circularize	6.75E-05
		REAC:504040	Start codon recognition	7.22E-05
		REAC:504769	Ribosomal scanning and start codon recognition	7.22E-05
		REAC:504671	Translation initiation complex formation	7.22E-05
		REAC:504643	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S	7.22E-05
		REAC:502544	Formation of translation initiation complexes yielding circularized Ceruloplasmin mRNA in a 'closed-loop' conformation	7.22E-05
		GO:0015935	small ribosomal subunit	7.59E-05
		REAC:501247	Association of phospho-L13a with GAIT element of Ceruloplasmin mRNA	8.25E-05
		REAC:504522	3' -UTR-mediated translational regulation	1.58E-04
		REAC:504521	L13a-mediated translational silencing of Ceruloplasmin expression	1.58E-04
		REAC:504611	Cap-dependent Translation Initiation	2.06E-04
		REAC:504612	Eukaryotic Translation Initiation	2.06E-04
		REAC:504507	Translation	2.47E-04
		KEGG:03010	Ribosome	4.88E-04
9	YBL030C YCRX07W YDL012C YDL080C YDR355C YDR410C YER161C YGR129W YJL197W YJR141W YKR088C YMR303C	KEGG:00010	Glycolysis / Gluconeogenesis	2.47E-03
44	YBR036C YDL165W YEL067C YIL041W YJL162C YLR155C YLR157C YLR393W YOL111C YPL251W	GO:0006530	asparagine catabolic process	2.07E-05
		GO:0004067	asparaginase activity	3.11E-05
		GO:0043562	cellular response to nitrogen levels	3.11E-05
		GO:0006995	cellular response to nitrogen starvation	3.11E-05
		KEGG:00460	Cyanoamino acid metabolism	6.81E-05
		GO:0006528	asparagine metabolic process	7.44E-05
		GO:0030287	cell wall-bounded periplasmic space	7.44E-05
		KEGG:00910	Nitrogen metabolism	2.26E-04
		KEGG:00250	Alanine, aspartate and glutamate metabolism	8.16E-04

biological functions such as ribosome, translation initiation, translational regulation, and so on. There are 19 biclusters containing 9 time points. In them, 14 biclusters are enriched by some gene functions. Taking bicluster 9 as an example, we find that the gene set is enriched by glycolysis and gluconeogenesis. Other biclusters in our result also have significant GO functions. In bicluster 44, the network positions of these genes are dispersive, but this set also has many common functions: cellular response to nitrogen, many metabolic process, cell wall-bounded periplasmic space, and so on. All these functions are important in cell cycle process (see table I).

Time-lagged biclusters have additional advantage. For example, there are no significant GO functions for genes in bicluster 8 at first glance using g:profiler. However close check reveals that this bicluster is comprised of a time-lagged bicluster together with bicluster 19, i.e. bicluster 8 and 19 have the same expression pattern but the time interval of bicluster 8 is a little later than that of bicluster 19. If we do not consider the time-lagged case, bicluster 8 is not a significant bicluster and genes in it have no significant functions. But now genes in both biclusters are found to have many GO functions such as 3'-UTR-mediated translational regulation ($p-value < 6 \times 10^{-5}$), translation initiation complex formation ($p-value < 5 \times 10^{-5}$), start codon recognition ($p-value < 5 \times 10^{-5}$) and so on .

ACKNOWLEDGMENT

The authors would like to thank ZHANGroup members for insightful discussions. This work was supported by a fund from National Center for Mathematics and Interdisciplinary Sciences of CAS, the Knowledge Innovation Program of CAS with Grant No. KSCX2-EW-R-01, and by the Knowledge Innovation Program of SIBS of CAS with Grant No. 2011KIP203. Authors were also supported by NSFC under Grant No. 61072149 and No. 91029301. This research was partially supported by the Chief Scientist Program of SIBS of CAS with Grant No. 2009CSP002, and the FIRST program from JSPS, initiated by CSTP.

REFERENCES

[1] Bar-Joseph Z, *Analyzing time series gene expression data*, Bioinformatics 2004, 20(16):2493–2503.

[2] Androulakis IP, Yang E, Almon RR, *Analysis of time-series gene expression data: methods, challenges and opportunities*, Annual Review of Biomedical Engineering 2007, 9:205–228.

[3] Jiang DX, Tang C, Zhang AD, *Cluster analysis for gene expression data: a survey*, IEEE Trans. Knowl. Data Eng. 16(11):1370–86. 2004.

[4] Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E, *A systematic comparison and evaluation of biclustering methods for gene expression data*, Bioinformatics 2006, 22(10):1282–1283.

[5] Madeira SC, Oliveira AL, *Biclustering algorithms for biological data analysis: a survey*, IEEE/ACM Trans. Comput. Biol. Bioinform., 2004, 1:24–45.

[6] Madeira SC, Teixeira MC, S-Correia I, Oliveira AL, *Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm*, In IEEE/ACM Transactions on Computational Biology and Bioinformatics, 21 Mar. 2008 IEEE Computer Society Digital Library. IEEE Computer Society.

[7] Zhang Y, Zha H, Chu CH, *A time-series biclustering algorithm for revealing co-Regulated genes*, In Proc of the 5th IEEE International Conference on Information Technology: Coding and Computing 2005:32–37.

[8] Tanay A, Sharan R, Shamir R, *Discovering statistically significant biclusters in gene expression data*, Bioinformatics 2002, 18(Suppl 1):S136–S144.

[9] Needleman, Saul B, Wunsch, Christian D, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*, Journal of Molecular Biology 48 (3): 443–53. 1970.

[10] Madeira SC , Oliveira AL, *An evaluation of discretization methods for non-supervised analysis of time-series gene expression data*, Technical Report 42, INESC-ID, December 2005.

[11] PT Spellman, G Sherlock, MQ Zhang, VR Iyer, K Anders, MB Eisen, PO Brown, D Botstein, B Futcher, *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization*, Molecular Biology of the Cell, 1998, 9(12):3273–3297.

[12] Ji L, Tan K, *Identifying time-lagged gene clusters using gene expression data*. Bioinformatics 2005, 21(4):509–516.

[13] Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M, *The BioGRID interaction database: 2011 update*, Nucleic Acids Res. 2010 Nov 11.

[14] Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, Mira NP, Alenquer M, Freitas AT, Oliveira AL, Sa-Correia I, *The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae*, Nucleic Acids Research 2006, 34:D446–D451 [http://www.yeastract.com/].

[15] J Reimand, M Kull, H Peterson, J Hansen, J Vilo, *g:Profiler — a web-based toolset for functional profiling of gene lists from large-scale experiments*, NAR 35 W193-W200.2007

[16] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al, *Gene Ontology: tool for the unification of biology*, Nat. Genet. 2000;25:25–29

978-1-4577-1666-9/11/$26.00 ©2011 IEEE