Information Fusion 10 (2009) 242–249

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Gene interaction – An evolutionary biclustering approach

Sushmita Mitra [a], Ranajit Das [a,*], Haider Banka [b], Subhasis Mukhopadhyay [c]

[a] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India
[b] Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700 108, India
[c] Bioinformatics Center, Department of Bio-Physics, Molecular Biology and Genetics, Calcutta University, Kolkata 700 009, India

ARTICLE INFO

Article history:
Received 17 October 2006
Received in revised form 9 August 2007
Accepted 12 November 2008
Available online 6 December 2008

Keywords:
Biclustering
Transcriptional regulatory network extraction
Bioinformatics
Microarray data
Gene expression profile
Gene interaction network

ABSTRACT

DNA Microarray experiments form a powerful tool for studying gene expression patterns, in large scale. Sharing of the regulatory mechanism among genes, in an organism, is predominantly responsible for their co-expression. Biclustering aims at finding a subset of similarly expressed genes under a subset of experimental conditions. A small number of genes participate in a cellular process of interest. Again, a gene may be simultaneously involved in a number of cellular processes. In cellular environment, genes interact among themselves to produce enzymes, metabolites, proteins, etc. responsible for a particular function(s).

In this study, a simple and novel correlation-based approach is proposed to extract gene interaction networks from biclusters in microarray data. Local search strategy is employed to add (remove) relevant (irrelevant) genes for finer tuning, in multi-objective biclustering framework. Preprocessing is done to preserve strongly correlated gene interaction pairs. Experimental results on time-series gene expression data from *Yeast* are biologically validated using benchmark databases and literature.

1. Introduction

Advent of DNA microarray technology have led to complete-genome expression profiling. This, coupled with various analytical methods, provides a lot of insight into the functioning of a cell and forms an indispensable tool for the system-level exploration of transcriptional regulatory networks. This represents a major advancement in the analysis of genomes, functionally, and is useful for exploring various complex cellular interactions [26]. Biological networks relate genes, gene products or their groups (like protein complexes) to each other. They harbour information about a gene, i.e., which pathway it represents and which other genes it affects. Gene clusters can be interpreted as a network of co-regulated genes encoding interacting proteins that are involved in the same biological processes. Clustering of gene expression patterns is used to identify groups of co-expressed genes [12] and generate gene interaction/gene regulatory networks [38].

Sharing of the regulatory mechanism among genes at the sequence level, in an organism, is predominantly responsible for them being co-expressed. Genes having similar gene expression profiles are more likely to regulate one another or be regulated by some other common parent gene [14]. Clustering of co-expressed genes, into biologically meaningful groups, helps in inferring the biological role of an unknown gene that is co-expressed with a known gene(s) [15,39]. Usually, it is found that a small set of genes are co-regulated and co-expressed over a subset of conditions, behaving almost independently for rest of the conditions. More often than not, genes are grouped into biclusters using continuous columns biclustering since biological processes initiate and terminate over a continuous interval of time [17,23,40]. The role of biclustering is to unravel such local structure inherent in the gene expression data matrix. It refers to the clustering of both rows (genes) and columns (conditions) of a data matrix (gene expression matrix), simultaneously, during knowledge discovery about local patterns from microarray data [8].

In this paper we propose an approach involving continuous column multi-objective evolutionary biclustering followed by simple extraction of correlated gene pairs, for an automated generation of gene interaction subnetworks based on regulatory information among genes. Preprocessing is done to preserve strongly correlated (positive or negative) gene interaction pairs. The rest of the sections is organized as follows. Section 2 introduces the basics of biological networks, gene interaction networks, identification of regulatory elements and biclustering. The proposed gene interaction network extraction is described in Section 3. The effectiveness of the model is demonstrated in Section 4, using time-series gene expression data from *Yeast*. The article is concluded in Section 5.

2. Preliminaries

In this section we provide a brief outline on the basic concepts of biological networks (followed by a short survey on

* Corresponding author. Tel.: +91 33 2575 3100; fax: +91 33 2578 3357.
E-mail addresses: sushmita@isical.ac.in (S. Mitra), ranajit_r@isical.ac.in (R. Das), hbanka2002@yahoo.com (H. Banka), smbmbg@caluniv.ac.in (S. Mukhopadhyay).

1566-2535/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2008.11.006

gene interaction networks, and the identification of regulatory elements) and biclustering.

2.1. Biological networks

Biological pathways can be conveniently represented as networks. They may be broadly categorized into *metabolic pathways*, *signal transduction pathways* and *gene regulatory networks or gene interaction networks*, though the prevalent databases might resort to a classification scheme involving a slightly different nomenclature. The repository of information about various biological pathway data are available in some databases like *The Kyoto Encyclopedia of Genes and Genomes* (KEGG) database,[1] BioCyc,[2] *Yeast Biochemical Pathways*,[3] *What Is There* (WIT) system,[4] etc.

The metabolic pathways facilitate mass generation, energy production, information transfer and cell-fate specification, in a cell or micro-organism; they are seamlessly integrated through a complex network of cellular constituents and reactions. Such a metabolic network consists of nodes, i.e., substrates (genes or proteins), that are interconnected through links, i.e., metabolic reactions in which enzymes provide the catalytic scaffolds [16].

Signal transduction is the process by which a cell converts one kind of signal or stimulus into another by a series of steps, causing functional changes inside the cell. The signal may pass from one cell to another (Hormone-Receptor concept), from extracellular environment to inside the cell (through plasma membrane) or from one compartment inside the cell to another compartment (i.e., from cytoplasm to nucleus). A signal transduction pathway can be considered as a biological network of biomolecules connected by various kind of interactions (protein–protein interactions, protein–ion interactions, etc.) among them.

A gene codes for a protein by prompting its production through *transcription*[5] followed by *translation*.[6] An enzyme, making such a reaction possible, is responsible for the production of the protein. In turn, a protein is responsible for a particular biological function. Proteins are bound to the promoter region of a gene to make it expressed or repressed, thereby acting as both activators or repressors. Every gene has one or more activators, i.e., biochemical signals which are necessary to start transcription of the gene. Without activators, a gene can attain only a low level expression value. A gene also has repressors, or biochemical signals which prevent its expression even in the presence of activators. As only a small number of genes produce proteins that function as activators or repressors, their identification is an important and difficult problem.

2.1.1. Gene interaction networks

A gene interaction network defines the complicated structure involving the gene products which activate or inhibit other gene products. The synthesized proteins may bind to regulatory sites of other genes functioning as transcription factors (TFs), as biological catalysts (enzymes) mediating chemical reactions of the metabolic pathway, or as activators/repressors of signaling pathways. A gene coding for the protein that regulates is termed the transcription factor, while the gene that is regulated is called its target. A TF, if present, can alternatively switch "ON" some genes while making others "OFF", thereby simultaneously orchestrating many genes in this manner. Understanding of gene interaction networks is crucial to the understanding of fundamental cellular processes involving growth, development, hormone secretion and cellular communication. Determination of TFs that control gene expression can offer further insight into the misregulated expressions common in many human diseases. In the next section we discuss, in brief, about some experimental techniques for the identification of TFs and their binding sites in the DNA.

In this section we provide a brief survey on existing literature along this direction. Regulatory network is predicted by SVMs [29] for the budding yeast genome by mining the gene expression data from different physiological conditions. Relationship between the expression timecourse of a TF and its target factor not being a simple correlation, SVMs are found to fare better than conventional hierarchical clustering. SVMs are trained using both positive and negative examples from the data set. A negative example is a gene pair that definitely has no regulatory relationship. The training set consists of a pair of genes, with the first being the known TF (R) and the second being the target gene (T) that is potentially regulated by R. After training, the system determines the probabilities of each $R \Rightarrow T$ pairing in order to construct parts of a regulatory network. The data set consists of 209 TFs $\times$ 6128 genes, resulting in mining among 1,280,752 combinations to determine which of these pairs represent a true regulatory relationship. The accuracy of the prediction is reported to be 93%, involving both positive and negative examples.

Regulatory relationships were deduced from the correlation of co-expression between a DNA-binding transcription regulator and its target gene, by using a probabilistic expression model [33]. Gene perturbation is employed to discover the direction of the regulatory effects in gene interaction networks [18,32]. However, there exist limitations in this sort of reverse-engineering approach to deriving transcriptional regulatory networks from gene expression data [2]. Correlation matching alone does not distinguish regulators from target genes, and it is difficult to discern whether the correlated target is directly or indirectly regulated. Additional information like protein–DNA binding has been integrated into transcriptional regulatory networks [2] for validating direct regulator–target interaction.

Often other biological conditions have a role in activating a regulator–target gene link in a known regulatory network. Gene expressions and TF binding data were used with the united signature algorithm [19], in order to reveal such new condition-specific links in known gene transcriptional regulatory networks. Consideration of both negatively- and positively-correlated genes, under specific experimental condition, allows identification of inhibitory and stimulatory regulatory interactions. The approach utilizes knowledge about the expression and binding patterns of the neighboring nodes of each link in existing experimentally-based literature-derived gene transcriptional regulatory networks, while extending them *in silico* using TF-gene binding motifs and *Yeast* gene expression data. It enables breaking down a composite transcriptional regulatory network into condition-specific networks.

2.1.2. Identification of transcription factor binding sites

Within a genome, the two most important components are the TFs (which are still mostly partially known) and their binding sites within the DNA, the prediction of which remains quite a challenging task [6]. The transcription factor binding sites (TFBSs) are approximately 5–15 base-pairs in length, and form degenerate sequence motifs. This degeneracy is responsible for different promoters having different activity levels, resulting in the transcription of certain genes at higher levels than others [36]. There are a number of methodologies, both *in vitro* and *in vivo*, developed for the identification of TFBSs. Traditional experimental methods involve footprinting methods which locate that DNA region which is defended by the binding protein, nitrocellulose binding assays, etc. [6]. Other high-throughput methodologies for determining the relationship

[1] http://www.genome.ad.jp/kegg/.
[2] http://www.biocyc.org/.
[3] http://pathway.yeastgenome.org/biocyc/.
[4] http://wit.integratedgenomics.com/.
[5] where many copies of the information encoded in a gene are produced in the form of many mRNA molecules.
[6] where many protein molecules of a single type are produced from the mRNAs.

between TFs and their binding sites in the DNA consist of the most widely used chromatin immunoprecipitation assays (ChIP-chip) for, *in vivo*, genome-wide location analysis [5,31] and systemic evolution of ligands by exponential evolution (SELEX) and its variants for finding, *in vitro*, those binding sequences having high protein affinity [27]. Homology prediction of sequence-specific DNA-binding TFs have been undertaken with the help of profile hidden Markov models (HMMs), which is more successful than methods based on pairwise sequence comparisons [20]. In recent years, DNA-binding proteins and their sequence specificities were extracted using the binding information of proteins, directly with DNA microarrays [7]. ChIP-chip methodology play a potentially important role in identifying the TFs in *Yeast* and the genomic targets in the mammalian cell [30,31].

Once the TFs have been identified, the next aim is the identification of target genes regulated by the TF. *In silico* techniques involving Position Weight Matrix (PWM) is used for representing degenerate sequence preferences of the binding protein, the elements of which correspond to the likelihood scores of obtaining the nucleotide of interest at the position of the TFBS [36]. TRANSFAC database [24] comprises of a repertory of TFs with their corresponding binding sites. Reference may be made to [6] for more details on the above methodologies.

2.2. Biclustering

One of the important problems in extracting and analyzing information from large databases is the associated high complexity. Feature selection is helpful as a preprocessing step for reducing dimensionality, removing irrelevant data, improving learning accuracy and enhancing output comprehensibility. Microarray data is a typical example presenting an overwhelmingly large number of features (genes), the majority of which are not relevant to the description of the problem and could potentially degrade the classification performance by masking the contribution of the relevant features. The key informative features represent a base of reduced cardinality, for subsequent analysis aimed at determining their possible role in the analyzed phenotype. This highlights the importance of feature selection, with particular emphasis on microarray data.

In other words, biclustering refers to some sort of feature selection and clustering in the space of reduced dimension, at the same time. Biclustering has been applied to gene expressions from cancerous tissues [22], mainly for identifying co-regulated genes, and for classification of samples. A bicluster can be defined as a pair (g, c), where $g \subseteq \{1, \ldots, m\}$ represents a subset of genes and $c \subseteq \{1, \ldots, n\}$ represents a subset of conditions (or time points). The optimization task [8] involves finding the maximum-sized bicluster not exceeding a certain homogeneity constraint mentioned below. The size (or volume) $f(g, c)$ of a bicluster is defined as the number of cells in the gene expression matrix E (with values e_{ij}) that the bicluster covers. The homogeneity $\mathscr{G}(g, c)$, of the bicluster, is expressed as a mean squared residue score. The task is to maximize

$$f(g, c) = |g| \times |c|, \tag{1}$$

subject to a small $\mathscr{G}(g, c) \leqslant \delta$ for $(g, c) \in X$, with $X = 2^{\{1,\ldots,m\}} \times 2^{\{1,\ldots,n\}}$ being the set of all biclusters, where

$$\mathscr{G}(g, c) = \frac{1}{|g| \times |c|} \sum_{i \in g, j \in c} (e_{ij} - e_{ic} - e_{gj} + e_{gc})^2. \tag{2}$$

In the above expression $e_{ic} = \frac{1}{|c|} \sum_{j \in c} e_{ij}$ and $e_{gj} = \frac{1}{|g|} \sum_{i \in g} e_{ij}$ denote the mean row and column expression values, respectively, for (g, c) while $e_{gc} = \frac{1}{|g| \times |c|} \sum_{i \in g, j \in c} e_{ij}$ denotes the mean expression value over all cells within the bicluster (g, c). The threshold δ, a user-defined quantity, represents the maximum allowable dissimilarity within the bicluster [25]. A good bicluster is one for which $\mathscr{G}(g, c) < \delta$ for some $\delta \geqslant 0$.

The optimization task of obtaining one or more biclusters, preserving the two competing constraints, *viz.*, homogeneity and size, is reported to be NP-complete [28]. The high complexity of this problem has motivated researchers to apply various approximation techniques to generate near optimal solutions. A number of good surveys on biclustering is available in literature [1,22,37].

Multi-objective evolutionary algorithm (MOEA) [11], a global search heuristic, has been used for biclustering [25]. The representation for each bicluster is by a fixed sized binary string called chromosome or individual. The chromosome has one bit string for genes appended by another for conditions, and represents a solution for the optimal bicluster generation problem. A bit is set to one when the corresponding gene and/or condition is present in the bicluster, and reset to zero otherwise. Fig. 1 pictorially depicts the encoding of genes and conditions in such a chromosome. Though the initial population of genes is generated randomly, yet in order to maintain the continuity among the time points one needs to impose certain constraints. Hence consecutive ones in the figure correspond to the continuous columns of the bicluster. The maximal set of genes and conditions were generated keeping the "homogeneity" criteria of the biclusters intact. Since these two characteristics of biclusters are conflicting to each other, multi-objective optimization provides an alternative, more efficient, approach to model them. To optimize this conflicting pair, the fitness function f_1 is always maximized while function f_2 is maximized as long as the residue is below the threshold δ. The formulation is as follows:

$$f_1 = \frac{g \times c}{|G| \times |C|}, \tag{3}$$

$$f_2 = \begin{cases} \frac{\mathscr{G}(g,c)}{\delta} & \text{if } \mathscr{G}(g, c) \leqslant \delta \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where g and c represent, respectively, the number of ones in the genes and conditions within the bicluster, $\mathscr{G}(g, c)$, δ are as defined earlier, and G and C are the total number of genes and conditions of the initial gene expression array. The algorithm followed is discussed in details in [25].

There exists a number of investigations dealing with time-series data [21,40]. Here the added constraint is to ensure the time locality existent in the data. Hence the starting and ending points in a time interval are the candidates for elimination from the bicluster, while its number of genes as well as the length of time interval are simultaneously maximized.

Initially, since the generation of biclusters is random, some extraneous genes and/or conditions may get included even though their expression values lie far apart in the feature space. Similar case may also arise during crossover and mutation in each generation. These extraneous genes and conditions need to be excluded deterministically. Moreover, good biclustering requires that some genes and/or conditions with similar expression values be added as well. Local search strategies [3,8] can be employed in such situations, to add or eliminate multiple genes and/or conditions (or time points). The Multi-objective GA (NSGA II), in association with the local search procedure, discussed in details in [25], were used for the generation of the set of biclusters.

3. Gene interaction network extraction

The definition of biological networks, in terms of gene pairs demonstrating regulatory relationship, is an important research problem. In spite of the availability of large volumes of experimental data, gene regulation as a complex dynamical process is not yet

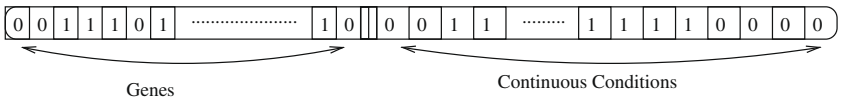

Fig. 1. An encoded chromosome representing a bicluster.

well understood. The large number of theoretical studies conducted in predicting various properties have been successful to a very limited extent, in the sense that they agree only qualitatively with experiments.

While considering transcription of gene expression one has to note that specific groups of genes may be made active by certain signals which once activated, may regulate common processes. The genes may also regulate the transcription of one another. Signal transduction pathways often activate genes via stimuli (e.g., hormones or neurotransmitters) leading to the production of second-messengers and to activation of TFs, often via phosphorylation (i.e., the addition of a phosphate group to a protein or a small molecule). Genes targeted by the same TFs often demonstrate similar expression patterns along time. Analyzing such similar expression profiles reveals several complex relationships between co-regulated gene pairs, including co-expression, simultaneous, time-shifted, and inverted relationships [29]. A suitable metric is also required to capture these similarities, in order to assist in clustering the genes with coupled regulatory mechanisms. There can also exist regulatory cascades (of first-order interactions between gene pairs), whereby the product of one gene influences the transcription rate for a second, the product of the second gene influences the transcription rate for a third, and so on. All these play a major role in the process of determining global regulatory networks.

In this paper we model the relationship between the expression level variation over time of a transcription factor and that of its target, in the framework of the generated biclusters. Since a bicluster represents a subset of highly correlated genes and conditions, the extraction of the relationship between the gene pair is more meaningful and computationally less expensive in this decomposed domain. The relationship is represented in terms of rules, linking a TF to the target gene that it regulates. Subsequently these rules are mapped to generate parts of the entire regulatory network. It may be noted that intra-pathway gene interactions, performing a certain biological function and possibly within a bicluster, are generally stronger than any inter-pathway interactions.

3.1. Algorithm for extraction of gene interaction network

The main steps of the proposed algorithm are outlined as follows.

(1) Extraction of biclusters to reduce complexity.
(2) Computation of pairwise correlation among gene pairs.
(3) Discretization of the correlation matrix, using quantile partitioning.
(4) Elimination of weak correlation links.
(5) Computation of adjacency matrix, and subsequent network generation.
(6) Biological validation.

3.2. Correlation between gene pairs

A gene expression profile e is represented over a series of n time points. Genes within a bicluster being co-expressed, we use the concept of correlation to quantify their similarity.

The *Pearson correlation coefficient* $C(e_1, e_2)$ between gene expression profile pair e_1 and e_2 provides a similarity measure between the two time-series curves, sampled at e_{1i} and e_{2i} over n time intervals. This is expressed as

$$C(e_1, e_2) = \frac{\sum e_{1i}e_{2i} - \sum e_{1i} \sum \frac{e_{2i}}{n}}{\sqrt{\left(\sum e_{1i}^2 - \frac{\left(\sum e_{1i}\right)^2}{n}\right)\left(\sum e_{2i}^2 - \frac{\left(\sum e_{2i}\right)^2}{n}\right)}}. \quad (5)$$

The first preprocessing step is to filter correlation coefficients which contribute minimally towards regulation. This is because often an exhaustive search of the possible interactions between genes is intractable. For example, a graph-based approach [35] resorts to thinning by removing edges according to conditional independence relations in various orders. Since we restrict our domain to a bicluster, therefore those genes that show little or no variation over time are automatically omitted. This physically signifies a gene's probable inactiveness or non-involvement in regulation.

Next we select those coefficients having absolute values above a detection threshold, implying larger correlation between the gene pairs. This enables us to focus on a few highly connected genes, that possibly link the remaining less connected genes. For example, let us consider the top and bottom view of the pairwise correlation coefficient plot for a sample bicluster consisting of 14 genes as shown in Fig. 2a and b respectively. In the plot the *X*- and *Y*- axes denote the 14 genes with the numbers 1–14 corresponding to the genes *YBL*014*C*, *YGL*110*C*, *YGL*164*C*, *YGR*077*C*, *YIR*012*W*, *YKL*002*W*, *YMR*047*C*, *YMR*060*C*, *YNL*250*W*, *YOL*005*C*, *YOL*135*C*, *YOR*244*W*, *YPL*055*C* and *YPL*088*W*, respectively, while the *Z*-axis denotes their pairwise correlation coefficient. We observe that only a few genes are strongly correlated, either positively or negatively, while the pairwise correlation (or dependence) among the rest is pretty poor.

We divide a correlation range $[C_{\max}, C_{\min}]$ into three partitions each, using *quantiles* or *partition values*[7] [10] in order to minimize the influence of extreme values or noisy patterns. Let $C_{\max}$ and $C_{\min}$ denote the maximum and minimum values for the correlation coefficients, $C_{\min} < 0 < C_{\max}$, considering all gene pairs from the adjacency matrix as determined from the bicluster. We split these values into two sets, *viz.*, those having positive $C^+ \in [0, C_{\max}]$ and negative $C^- \in [C_{\min}, 0]$ values respectively. Thereby, both positive and negative regulation is considered between the interacting genes. Let these values be sorted in the ascending order. The first positive quantile (Q_1^+) is the value of C^+ that exceeds one-third of the positive measurements and is less than the remaining two-thirds. The second positive quantile (Q_2^+) is the value of C^+ that exceeds two-thirds of the positive measurements and is less than the remaining one-third. An analogous computation generates the negative quantile pair Q_1^- and Q_2^-.

Note that negative correlation between two gene profiles is essentially not zero correlation between them. Moreover, any correlation coefficient (i) having value above Q_2^+ (below Q_2^-) indicates high positive (negative) correlation, and (ii) having value in $[Q_1^+, Q_2^+)$ ($[Q_2^-, Q_1^-)$) indicates moderate positive (negative) correlation.

In order to determine the two quantiles, we divide the measurements into a number of small class intervals of equal width d' and

[7] Quantiles or partition values are the values of a variate which divide the total frequency into a number of equal parts.

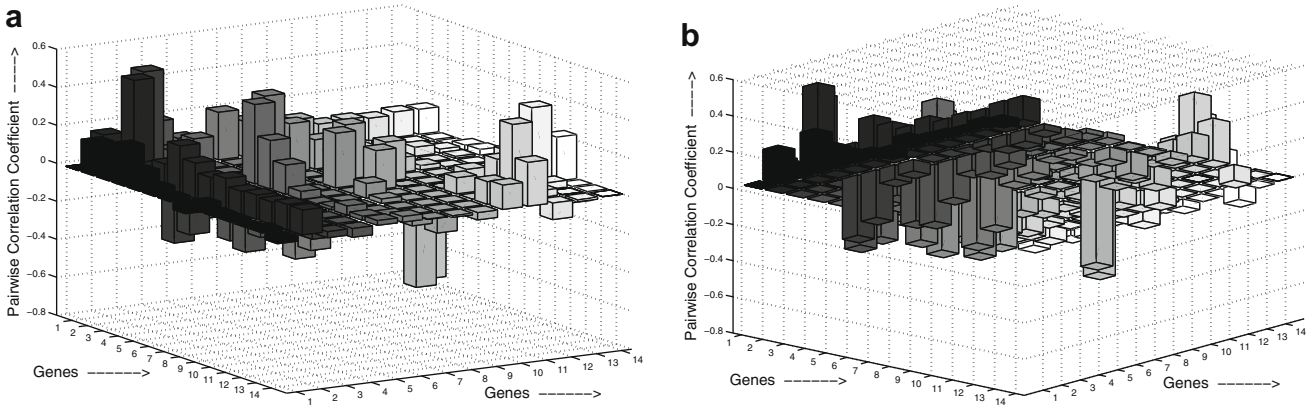

Fig. 2. Pairwise correlation coefficient plots for a 14-gene bicluster depicting (a) top view and (b) bottom view.

count the corresponding class frequencies f_i. The position of the kth positive partition value (here quantile, as $k = 1, 2$ for three partitions) is calculated as

$$Q_k^+ = l_i + \frac{R_k - cf_{i-1}}{f_i} \cdot d', \tag{6}$$

where l_i is the lower limit of the ith class interval (for C^+), $R_k = \frac{N.k}{3}$ is the rank of the kth partition value, and cf_{i-1} is the cumulative frequency of the immediately preceding class interval, such that $cf_{i-1} < R_k < cf_i$. Analogous computation is performed for the negative correlation factors, considering their absolute values.

An adjacency matrix is computed as

$$A(i,j) = \begin{cases} -1 & \text{if } C \leqslant Q_2^- \\ +1 & \text{if } C \geqslant Q_2^+ \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where we assume absence of self correlations among the genes. Thereafter, a network connecting the various genes is generated.

4. Experimental results

Yeast cell cycle CDC28 data [9] is a collection of 6178 genes (attributes) of the budding *yeast Saccharomyces cerevisiae* under 17 conditions (time points), taken at 10-minute time intervals covering nearly two complete cycles of cell cycle. The synchronization of the *yeast* cell cultures were done using the so-called CDC28 arrest. The experiments were performed using Affymetrix oligonucleotide array. This data set is attractive because a large number of genes contained in it are biologically characterized and have been assigned to different phases of the mitotic cell cycle, with specific functional activities [9,34]. Each phase of the cell cycle includes genes that show higher expression levels at that cycle time than the others. The missing values are imputed according to the methodology provided in [4].[8]

Current microarray technology is limited in sensitivity when comparing the relative concentrations of different genes in a single experiment. The associated costs make it difficult for these experiments to be repeated. Therefore, the available gene expression data is intrinsically noisy and difficult to analyze [13]. Filkov et al. have shown that less than 20% of the known regulatory gene pairs exhibit strong correlations in the Cho/Spellman data sets.

The multi-objective biclustering algorithm [25] was executed on the *Yeast* microarray data set, discussed above. The parameter values chosen in the experiment were as follows: population size = 50, mutation probability = 0.05, crossover probability = 0.75, $\delta = 20$, $\alpha = 1.2$. The values for the crossover and mutation probabilities were chosen after several trials and errors with random seeds. No significant changes in the results were obtained with their variations. The program was implemented in C and the environment used was SunOS Release 5.9 Generic_117171-07. The system configurations are: Sun System Name: Sun Blade 2000; Platform Name: SUNW, Sun-Blade-1000; sparc; Platform Group: sun4u, Physical Memory (RAM) = 2048 MB.

4.1. Missing values

Missing values often appear in gene expression data, due to various experimental limitations. Since many methods for the analysis of gene expression data require a complete data matrix as input, either these missing values have to be estimated or those genes with missing expression values need to be eliminated. However, it is not desirable that genes with only a few missing values get eliminated in the process. Hence, it is imperative to estimate those missing values. The accuracy of such an estimation process needs to be good enough for the subsequent analysis to be informative [4]. We begin by eliminating from the data set those genes with more than half missing expression values. Eventually, a total of 6132 genes were taken for further analysis. Thereafter, the remaining missing values are estimated using the LSimpute software.

4.2. Network extraction

To generate the network architecture from the extracted biclusters, we first compute the pairwise correlation coefficients by Eq. (5). Next the quantile partitioning is employed to select the strong positive as well as negative correlation links, using Eq. (6). Thereby, the top 1/3 of the positive and negative links are chosen to be connected in a network. This implies strong correlation (or dependence), either positive or negative, between gene pairs. A sample network consisting of three biclusters of sizes 7, 10, and 14, respectively, are shown in Fig. 3. A transcription factor is connected to its target gene by an arrow if a TF-target pair exists within any of the biclusters. Gene pairs having positive correlation are connected by solid lines, while those that are negatively correlated are connected by dashed lines. TFs external to the network, but having targets within the network, are connected to the corresponding vertices by dotted arrows. For example, the gene *YHR084W* (encircled with solid lines) is a TF belonging to the network of 10 genes and has targets in all three networks. An external TF *YJL056C* (encircled with dotted lines) has targets in networks of 7 and 10 genes. A gene

[8] LSimpute: accurate estimation of missing values in microarray data with least squares methods.

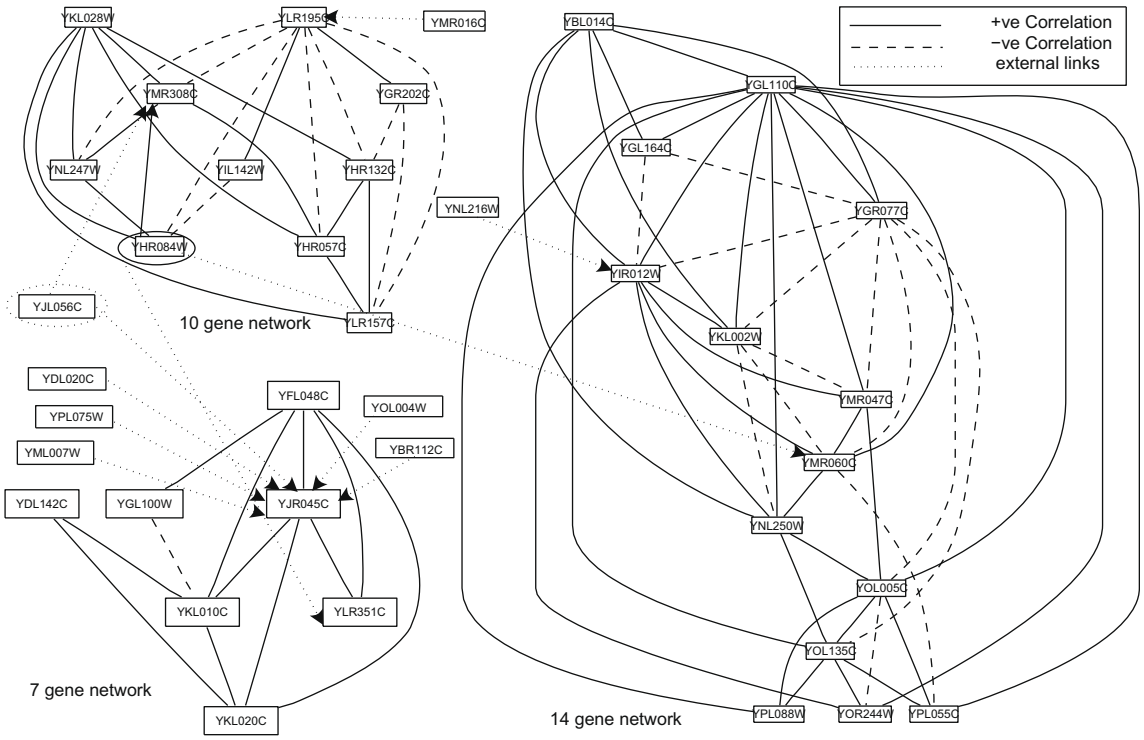

Fig. 3. Network (bicluster) of 10 genes connected by transcription factor *YHR*084*W* to networks (biclusters) of 7 and 14 genes.

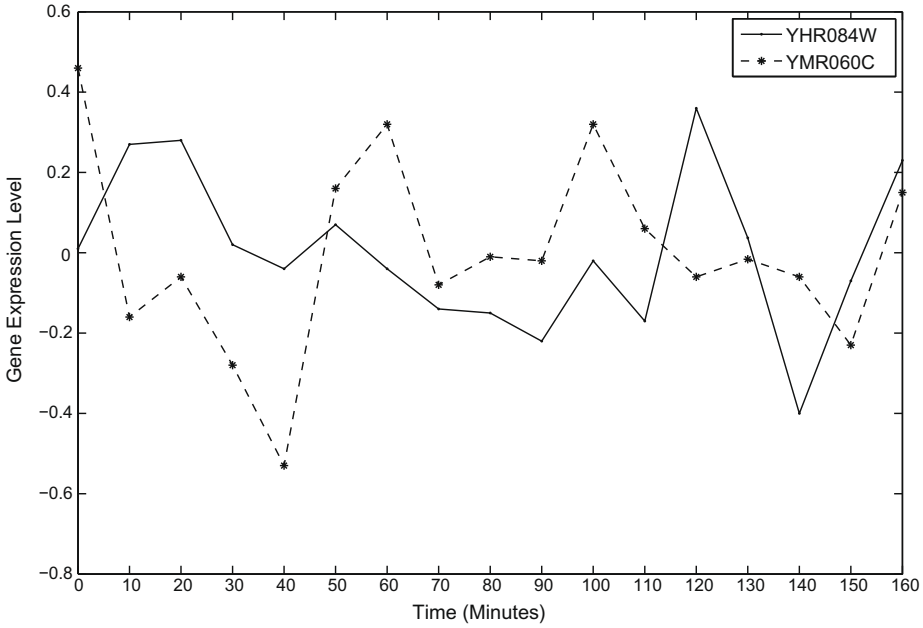

Fig. 4. Expression profile of transcription factor *YHR*084*W* (10-node network) and its target *YMR*060*C* (14-node network).

ontology study was made to biologically validate the biclusters, in terms of the statistically significant GO annotation database.[9]

Fig. 4 depicts the pairwise profile relationship between the TF *STE*12/*YHR*084*W* of the 10-node bicluster network (Fig. 3) with the target in the 14-node network. The solid line represents the expression profile for the TF while the dashed line represents that of the corresponding regulated gene. We observe that the two genes pose an inverted relationship over time span 0–10 min, while there exists simultaneous relationship over the time spans 10–50 min and 80–110 min.

Fig. 5 illustrates the expression profiles of the six external TFs *GCR*1/*YPL*075*W*, *RPN*4/*YDL*020*C*, *YAP*1/*YML*007*W*, *SIN*3/*YOL*004*W*, *ZAP*1/*YJL*056*C* and *CYC*8/*YBR*112*C* along with the target gene *SSC*1/*YJR*045*C*, corresponding to the 7-node bicluster network (generated in terms of pairwise correlation values) in Fig. 3.

[9] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.

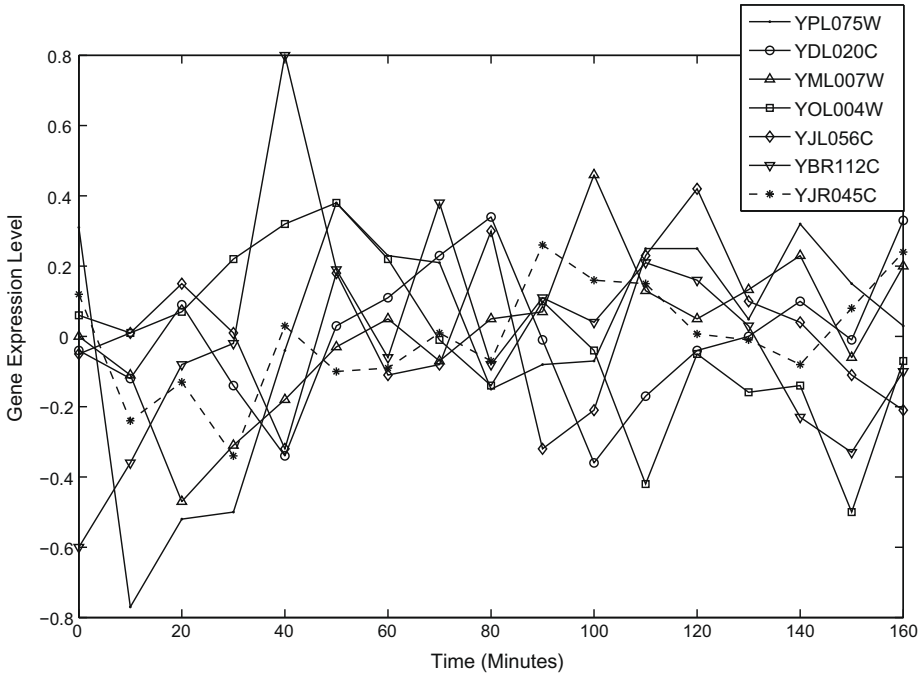

Fig. 5. Expression profile of six transcription factors and target YJR045C, in the 7-node network.

The solid lines connecting the various symbols, *viz.* $\bullet, \circ, \delta, \Box, \diamondsuit, \nabla$, are the expression profiles for the TFs, respectively, while the dashed line connecting the symbol ☆ is that corresponding to the regulated gene. Let us consider the behavior of *GCR*1/*YPL*075*W* with the target, as an example. Here we find almost simultaneous behavior over conditions 0–40 min, while conditions 40–80 min demonstrate a marked inverted relationship. Again the time span 110–130 min displays a shifted response of the target as compared to the TF. However, the expression profiles of *CYC*8/*YBR*112*C* and the target appear to share a strong simultaneous relationship almost along the total time span.

4.3. Biological validation

While predicting regulatory networks [29] it was reported that genes *YHR*084*W*–*YLR*351*C* form a TF-target pair. We also obtained the summary of the TF-target pair *YHR*084*W*–*YLR*351*C* (Fig. 3) in terms of *Molecular Function*, *Biological Process* and *Cellular Component* from the Saccharomyces Genome Database (SGD).[10] We have also confirmed from our calculations that there exists an interaction between the TF and its target. It is reported in the database that the biological process involving protein *YLR*351*C* is not fully understood as yet. It is also reported in the database that *YHR*084*W* has transcription factor activity. The problem is compounded whenever one tries to extract some biologically meaningful information involving these two entities. With this scanty information our method has been able to identify that there exists a link between a TF and its target.

Thus, considering the cellular component of *YHR*084*W* (residing in the nucleus) and *YLR*351*C* (residing in the cytoplasm and mitochondria) we model, as an efficacy of the biclustering, the transcription of *YLR*351*C* by *YHR*084*W* occurring in the nucleus, and the regular translation mechanism follows. Similarly for the TF-target pair of *YPL*075*W* and *YJR*045*C* (Fig. 3) reported in [29], we obtained their summary from SGD and performed subsequent analysis. From the database *YPL*075*W* is a transcriptional activator of genes involved in glycolysis while *YJR*045*C* has ATPase, enzyme regulator and protein transporter activity. Again we were able to predict that *YPL*075*W* is involved in the transcription of *YJR*045*C* and would go into the glycolysis process.

Similar kind of conclusions can be arrived at, for other TF-target pairs, with a certain definite degree of confidence. A large number negative results is expected as relevant literature in this area are really very few. One of the merits of our algorithm is that it has not detected any false positive or false negative TF-target pairs, which is consistent with the information available either in the literature or in the databases.

5. Conclusions and discussion

In this paper we have introduced an approach based on multiobjective evolutionary biclustering and subsequent extraction of correlated gene pairs for the generation of gene interaction networks. Biologically relevant small biclusters were obtained, using time-series gene expression data from *Yeast*. These were validated using the statistically significant GO annotation database. The pairwise correlation coefficients among gene pairs were computed by Eq. (5) followed by the quantile partitioning using Eq. (6) to select the strong positive as well as negative correlation links. The strongly correlated genes were then chosen to be connected in a network. TF-target gene pairs in the network, shown in Fig. 3, were found to exhibit strong correlations. We tried to model the interaction among them from information available in the literature/databases *viz.*, SGD. We have also analyzed the expression profiles of the regulator and the regulated genes which reveals several complex (simultaneous, time shifted, inverted, etc.) relationships between them.

During this analysis we noted that there may exist a delay in the regulation of a target by its TF. A TF firing at a particular instant may affect its target after a finite duration of time. Thus, the resultant protein would not be produced instantaneously. If this delay is comparable to the time interval between the gene expression time points, it would result in the correlation coefficient being low. In such cases we may employ time-lagged linear Pearson correlation

[10] A scientific database of the molecular biology and genetics of the yeast *S. cerevisiae* – http://db.yeastgenome.org/.

to extend this technique by determining the pairwise correlation among genes whose expression profiles are shifted over time. Furthermore, the TF-target relationship, among genes, for most of the organisms are not freely downloadable from databases or available in the literature. Due to extreme difficulty in accessing a variety of data sets, the construction of gene interaction networks and subsequently validating them in a realistic manner is an extremely difficult task. We are currently investigating the role of delays in the eventual correlation/adjacency matrix and other issues related to it.

Transcriptional regulation is fundamental in translating genetic information into biological function, and is crucial for understanding cell adaptation, differentiation, and pathological transformation. One challenge then is to decipher this intricate network of transcriptional interactions, for better appreciation of functional relationships and hence the discerning of various diseased states. The correlation-based biclustering method, as proposed by us, could be a step to address this challenge; and it is hoped that one may be able to apply this clustering technique to newer and newer data sets in order to analyze the efficacy of our method. In future, we would attempt to apply this technique to other categories of networks and also to other organisms, like human, where the transcriptional regulation is more complex.

Acknowledgements

The authors gratefully acknowledge Dr. R.K. De and Ms. L. Nayak for their helpful discussion during the progress of this work. Dr. S. Mukhopadhyay gratefully acknowledges the financial assistance received in the form of a grant, BT/BI/04/001/93 from the Department of Biotechnology, Government of India.

References

[1] J.S. Aguilar-Ruiz, F. Divina, Evolutionary biclustering of microarray data, in: Applications on Evolutionary Computing/Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, 2005, pp. 1–10.
[2] Z. Bar-Joseph, G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo, F. Robert, D.B. Gordon, E. Fraenkel, T.S. Jaahhola, R.A. Young, D.K. Gifford, Computational discovery of gene modules and regulatory networks, Nature Biotechnology 21 (2003) 1337–1342.
[3] S. Bleuler, A. Prelić, E. Zitzler, An EA framework for biclustering of gene expression data, in: Proceedings of Congress on Evolutionary Computation, 2004, pp. 166–173.
[4] T.H. Bo, B. Dysvik, Inge Jonassen, Lsimpute: accurate estimation of missing values in microarray data with least squares methods, Nucleic Acids Research 32 (2004) 1–8.
[5] M.J. Buck, J.D. Lieb, Chip–chip: considerations for the design analysis and application of genome-wide chromatin immunoprecipitation experiments, Genomics 83 (2004) 349–360.
[6] M.L. Bulyk, Computational prediction of transcription-factor binding site locations, Genome Biology 5 (2003) 1–11.
[7] M.L. Bulyk, X. Huang, Y. Cho, G.M. Church, Exploring the DNA-binding specificities of zinc fingers with DNA microarrays, Proceedings of National Academy of Sciences USA 98 (2001) 7158–7163.
[8] Y. Cheng, G.M. Church, Biclustering of gene expression data, in: Proceedings of ISMB 2000, 2000, pp. 93–103.
[9] R.J. Cho, M.J. Campbell, L.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, R.W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, Molecular Cell 2 (1998) 65–73.
[10] G.R. Davies, D. Yoder, Business Statistics, John Wiley & Sons Inc., London, 1937.
[11] K. Deb, Multi-Objective Optimization using Evolutionary Algorithms, John Wiley, London, 2001.
[12] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proceedings of National Academy of Sciences USA 95 (1998) 14863–14868.
[13] V. Filkov, S. Skiena, J. Zhi, Analysis techniques for microarray time-series data, Journal of Computational Biology 9 (2002) 317–330.
[14] A.P. Gasch, M.B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy *k*-means clustering, Genome Biology 3 (2002). research0059.1–0059.22.
[15] E. Hartuv, R. Shamir, A clustering algorithm based on graph connectivity, Information Processing Letters 76 (2000) 175–181.
[16] H. Jeong, B. Tombor, R. Albert, Z.N. Oltval, A.-L. Barabasi, The large-scale organization of metabolic networks, Nature 407 (2000) 651–654.
[17] L. Ji, K.L. Tan, Identifying time-lagged gene clusters using gene expression data, Bioinformatics 21 (2005) 509–516.
[18] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, M. Tomita, Dynamic modeling of genetic networks using genetic algorithm and S-system, Bioinformatics 19 (2003) 643–650.
[19] H. Kim, W. Hu, Y. Kluger, Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces Cerevisiae, BMC Bioinformatics 7 (2006) 165.
[20] S.K. Kummerfeld, S.A. Teichmann, Dbd: a transcription factor prediction database, Nucleic Acids Research 34 (2006) 1–8.
[21] J. Liu, J. Yang, W. Wang, Biclustering in gene expression data by tendency, in: Proceedings of the 2004 Computational Systems Bioinformatics Conference (CSB 2004), 2004, pp. 1–12.
[22] S.C. Madeira, A.L. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE Transactions on Computational Biology and Bioinformatics 1 (2004) 24–45.
[23] S.C. Madeira, A.L. Oliveira, A linear time biclustering algorithm for time series gene expression data, in: R. Casadio, G. Myers (Eds.), WABI 2005, LNBI, vol. 3692, Springer Verlag, Berlin, 2005, pp. 39–52.
[24] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. Kel, et al., TRANSFAC: transcriptional regulation from patterns to profiles, Nucleic Acids Research 31 (2003) 374–378.
[25] S. Mitra, H. Banka, Multi-objective evolutionary biclustering of gene expression data, Pattern Recognition 39 (2006) 2464–2477.
[26] S. Mitra, W. Pedrycz, Special issue on bioinformatics, Pattern Recognition 39 (12) (2006).
[27] A. Oliphant, C. Brandl, K. Struhl, Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein, Molecular Cell Biology 9 (1989) 2944–2949.
[28] R. Peeters, The maximum edge biclique problem is NP-complete, Discrete Applied Mathematics 131 (2003) 651–654.
[29] J. Qian, J. Lin, N.M. Luscombe, H. Yu, M. Gerstein, Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data, Bioinformatics 19 (2003) 1917–1926.
[30] J.L. Reid, V.R. Iyer, P.O. Brown, K. Struhl, Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of esa1 histone acetylase, Molecular Cell 6 (2000) 1297–1307.
[31] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, et al., Genome-wide location and function of DNA binding proteins, Science 290 (2000) 2306–2309.
[32] J.J. Rice, Y. Tu, G. Stolovitzky, Reconstructing biological networks using conditional correlation analysis, Bioinformatics 21 (2005) 765–773.
[33] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, N. Friedman, Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, Nature Genetics 34 (2003) 166–176.
[34] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. 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 9 (1998) 3273–3297.
[35] P. Spirtes, C. Glymour, R. Scheines, From probability to causality, Philosophical Studies 64 (1991) 1–36.
[36] G. Stormo, DNA binding sites: representation and discovery, Bioinformatics 16 (2000) 16–23.
[37] A. Tanay, R. Sharan, R. Shamir, Biclustering algorithms: a survey, in: S. Aluru (Ed.), Handbook of Computational Molecular Biology, Computer and Information Science Series, Chapman & Hall/CRC, USA, 2005, p. 26-1.
[38] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, G.M. Church, Systematic determination of genetic network architecture, Nature Genetics 22 (1999) 281–285.
[39] Y. Xu, V. Olman, D. Xu, Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees, Bioinformatics 18 (2002) 536–545.
[40] Y. Zhang, H. Zha, C.H. Chu, A time-series biclustering algorithm for revealing co-regulated genes, in: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), 2005, pp. 32–37.