Cross-Correlation and Evolutionary Biclustering: Extracting Gene Interaction Sub-networks

Ranajit Das[1], Sushmita Mitra[1], and Subhasis Mukhopadhyay[2]

[1] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India
{ranajit_r,sushmita}@isical.ac.in
[2] Department of Bio-Physics, Molecular Biology and Bioinformatics, Calcutta University, Kolkata 700 009, India
sm.bmbg@gmail.com

Abstract. In this paper we present a simple and novel time-dependent cross-correlation-based approach for the extraction of simple gene interaction sub-networks from biclusters in temporal gene expression microarray data. Preprocessing has been employed to retain those gene interaction pairs that are strongly correlated. The methodology was applied to public-domain data sets of Yeast and the experimental results were biologically validated based on standard databases and information available in the literature.

Keywords: Biclustering, transcriptional regulatory network, time-delay, time-lagged correlation, gene interaction network.

1 Introduction

With the current development in microarray technology (gene chips), today researchers in Bioinformatics have, at their disposal, expression data of thousand of genes of different organisms under various experimental conditions. DNA microarray technology, thus, forms an indispensable tool for exploring transcriptional regulatory networks from the system level and is very helpful when one dwells into the cellular environment to investigate various complex interactions. Biological pathways can be conveniently characterized as networks and broadly classified as *metabolic pathways*, *gene regulatory networks or gene interaction networks* and *signal transduction pathways*. Gene regulatory networks connect genes, gene products (in the form of protein complexes) or their groups to one another. A network of coregulated genes may form gene clusters that can encode proteins, which interact amongst themselves and take part in common biological processes. Clustering has been applied to locate co-expressed groups of genes and extract gene interaction/gene regulatory networks from gene expression data [1].

Genes with similar expression profiles may regulate one another or be regulated by some other common parent gene. However, one need to observe that a subset of genes is co-regulated and co-expressed only over few conditions. The genes also share local rather than global similar patterns in their expression profiles [2]. Such sets of genes may be identified in the form of biclusters [3] using continuous columns, to represent a continuous interval of time [4].

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 199–204, 2009.

Evolutionary biclustering has been used for extracting gene interaction networks from time series gene expression data [5]. The networks were validated incorporating domain knowledge from transcription factor (TF)-target (T) databases like $TRANSFAC$[1] and literature [6]. Co-expressed gene pairs were seldom found to exhibit simple simultaneous relationships. On closer inspection of their profiles it could rather be noted that there exists a time shifted response of the target gene to its TF [6]. Time-lagged correlation or cross-correlation helps in analyzing the positive or negative correlation among time-lagged profiles of gene pairs. This motivated us to explore the use of cross-correlation with time-delay for the appropriate modeling of temporal interactions.

In this paper continuous-column multiobjective evolutionary biclustering [4] has been used for extracting time-correlated gene pairs. A gene pair having correlation magnitude above a detection threshold was considered to be interacting or regulating each other. Preprocessing was done to eliminate the weakly correlated (positive or negative) gene interaction pairs. An adjacency matrix was constituted from the resulting cross-correlation matrix, which was eventually used for reverse engineering transcriptional gene interaction sub-networks using regulatory information among genes. The usefulness of the model is demonstrated, using time-series gene expression data from Yeast and biologically validated.

2 Gene Interaction Sub-network Extraction: A Multi-objective Evolutionary Approach

A gene produces a protein by *transcription* (formation of many copies of mRNA molecules) followed by *translation* (resulting into production of protein), the reaction taking place in the presence of an enzyme. In turn, a protein is responsible for a particular biological function. A TF is a gene product that binds to the promoter region of a target gene, up-regulating or down-regulating its expression. Every gene has one or more such activators/repressors. Their identification, and the subsequent elucidation of the biological networks demonstrating TF-T relationship is quite a challenging task. Analysis of their profiles brings out several complex relationships between the co-regulated gene pairs, including co-expression, time shifted, and inverted relationships [6].

2.1 Evolutionary Biclustering

Biclustering refers to the simultaneous clustering and redundant feature reduction involving both attributes and samples. This results in the extraction of biologically more meaningful, less sparse partitions from high-dimensional data, that exhibit similar characteristics. A bicluster may be defined as a pair (g, c), where $g \subseteq \{1, \ldots, m\}$ denotes a subset of genes and $c \subseteq \{1, \ldots, n\}$ denotes a subset of conditions (or time points). The optimization task [3] involves finding the maximum-sized bicluster subject to a certain homogeneity constraint. The size (or volume) $f(g, c)$ of a bicluster is defined as the number of cells in the gene

[1] http://www.gene-regulation.com/pub/databases.html

expression matrix E (with values e_{ij}) that are covered by it. The homogeneity $\mathcal{G}(g, c)$ is expressed as a mean squared residue score (or error). Since these two characteristics of biclusters are conflicting to each other, multi-objective evolutionary algorithms, in association with local search, was applied to provide an alternative, more efficient approach [4].

2.2 Time-Lagged Cross-Correlation between Gene Pairs

Often genes are found to share similar sub-profiles (over a few time points) instead of the complete gene expression profiles. Considering the global correlation among gene pairs, *i.e.*, computation of correlation amongst genes employing the complete gene expression data matrix, may not reveal the proper relationship between them. Since the transcriptional response of a gene can occur from tens of minutes to several hours, time delay correlation may help determine the underlying causal relationship. The concept of cross-correlation has been introduced to take into account the time-shifted behaviour between TF-T pairs. This allows a more realistic modeling of gene interactions within the reduced localized domain of biclusters. In this work we extend our earlier network extraction algorithm [5] to include such temporal influence.

The expression profile e of a gene may be represented over a series of n time points. The cross-correlation $CC_d(e_1, e_2)$ between gene pair e_1 and e_2, with delay d,is expressed as

$$CC_d(e_1, e_2) = \frac{\sum e_{1i} e_{2i-d} - \sum e_{1i} \sum \frac{e_{2i-d}}{n}}{\sqrt{(\sum e_{1i}^2 - \frac{(\sum e_{1i})^2}{n})(\sum e_{2i-d}^2 - \frac{(\sum e_{2i-d})^2}{n})}}. \tag{1}$$

Here we select that delayed time $d = \Delta$t which maximizes the correlation in absolute value by eqn. (2) as

$$CC(e_1, e_2) = \max |CC_d(e_1, e_2)| \qquad where - 2 \leq d \leq 2 \tag{2}$$

The maximum delay of two time point are allowed as longer shifts are hard to explain from a biological point of view [2].

Filtering out the weaker correlation coefficients, which presumably contribute less towards regulation, serves as the first preprocessing step. This allows us to avoid an exhaustive search of all possible interactions among genes. The remaining coefficients, having absolute values above a detection threshold, imply a larger correlation among the gene pairs. The correlation range $[CC_{\max}, CC_{\min}]$ is divided into three partitions each, using *quantiles* or *partition values* [5] to reduce the influence of extreme values or noisy patterns. Negative correlation for a gene pair is not zero correlation. Time lagged-correlation coefficients with values greater than Q_2^+ (less than Q_2^-) indicate high positive (negative) correlation, while those with values in $[Q_1^+, Q_2^+)$ ($[Q_2^-, Q_1^-)$) indicate moderate positive (negative) correlation.

An adjacency matrix is calculated as follows:

$$A(i,j) = \begin{cases} -1 \text{ if } & CC \leq Q_2^- \\ +1 \text{ if } & CC \geq Q_2^+ \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where self correlations among the genes are assumed to be absent. Next the extraction of gene interaction sub-networks is attempted. Biological validation is made in terms of ontology study.

3 Experimental Results

We applied our analysis to the Yeast cell-cycle CDC28 data gathered by Cho *et al.* [7]. It is a measure of the expression levels of 6220 gene transcripts (features/attributes) for 17 conditions (time points/samples), taken at 10-minute time intervals covering nearly two cycles. The synchronization of the yeast cell cultures was done using the so-called CDC28 arrest and the experiment was performed using Affymetrix oligonucleotide arrays. The missing values in the data set were imputed similar to that of our earlier network extraction technique [5] and the biclusters were extracted to detect sets of co-regulated genes. Pairwise time-lagged cross-correlation coefficients were calculated between gene pairs in the extracted biclusters using eqn. (1). The weaker interactions, as thresholded by quantile partitioning, were excluded.

A sample extracted sub-network comprising of four biclusters (modules or sub-networks) is depicted in Fig. 1. A transcription factor is connected to its target gene by an arrow if such a TF-Target pair existed within any of the biclusters. Gene pairs connected by solid lines depict positive correlation, while those connected by dashed lines are negatively correlated. As an example, the TF named $YHR084W$ (encircled with solid lines) belonging to the sub-network of 10 genes has targets in all the four sub-networks. These biclusters were biologically validated from gene ontology study, based on the statistically significant GO annotation database[2], and discussed below.

Fig. 2 demonstrates various complex relationships *viz.* and their interplay in the pairwise profile relationship of the TF and its target in Fig. 1. We observe that the target $YGR233C$ poses a mixed simultaneous (during 0-40 minutes) and time-shifted relationship (during 40-60 and 100-140 minutes) with the TF $YHR084W$ and that the relationship is not a simple direct one.

It has been reported during the prediction of regulatory network structure [8] that the gene pair $YHR084W$-$YGR233C$ (where both the TF and the target belongs to the 10-node network) form a TF-Target pair. We verified their GO summary in terms of *Molecular Function*, *Biological Process* and *Cellular Component* from the Saccharomyces Genome Database (SGD). Our computations also indicated an analogous interaction between the TF-Target pair and were supported by literature [9]. Based on the SGD it could be gathered that the $YHR084W$ is activated by a MAP kinase signaling cascade – that controls many important functions of living organisms like cell-cycle, apoptosis[3], differentiation, etc. while the protein $YGR233C$ has cyclin-dependent protein kinase inhibitor activity (TAS). One can think of several models considering the

[2] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder
[3] Programmed cell death.

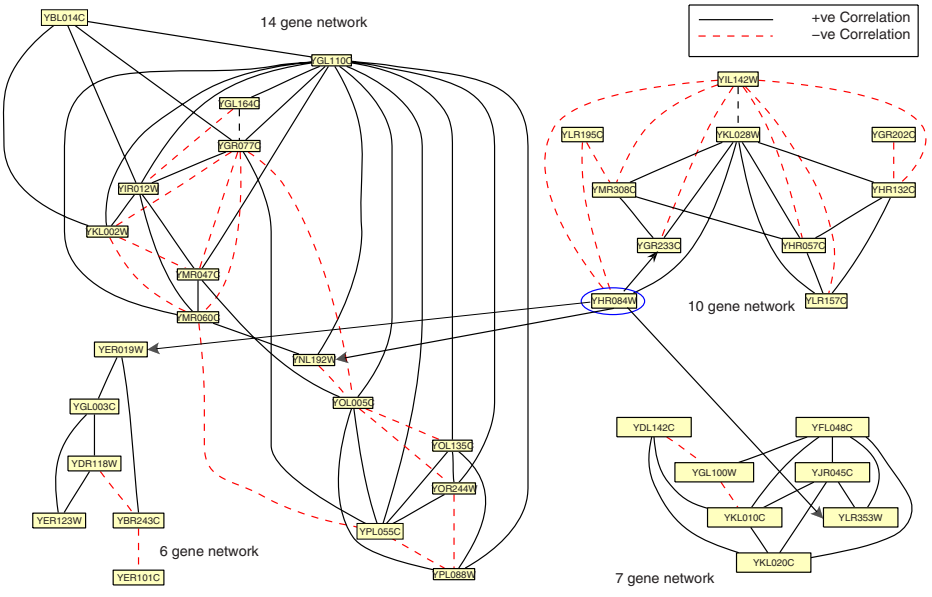

Fig. 1. Sub-network (bicluster) of 10 genes connected by transcription factor $YHR084W$ to sub-networks (biclusters) of 6, 7 and 14 genes

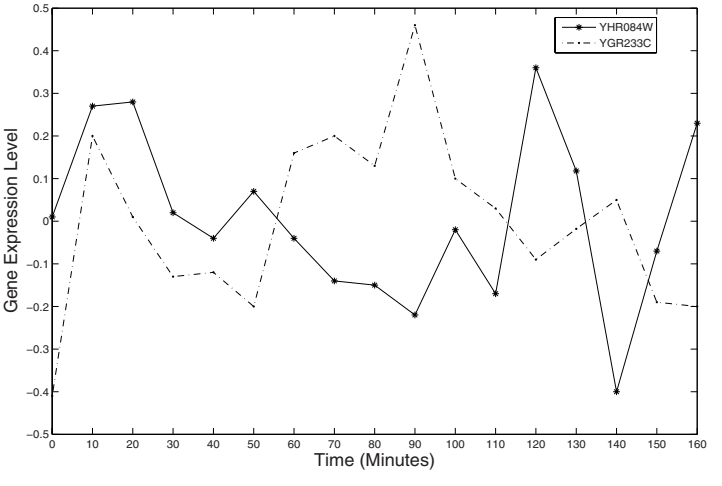

Fig. 2. Expression profile of transcription factor $YHR084W$ and its target $YGR233C$ (10-node network)

transcription of $YGR233C$ by $YHR084W$ to occur inside the nucleus, followed by the regular translation mechanism, based on their cellular component.

4 Conclusions and Discussion

In this paper we have described an approach for the extraction of cross correlated gene pairs for the generation of gene interaction networks. Biologically relevant biclusters were obtained using multiobjective biclustering, from time-series gene expression data from Yeast. The pairwise time-lagged correlation coefficients among gene pairs were computed by eqn. (1), followed by the quantile partitioning. Strongly correlated genes were considered for extracting sample TF-Target gene interaction sub-networks, as in Fig. 1. We tried to analyze the expression profiles of the regulator and the regulated genes to reveal several complex (time shifted, inverted, simultaneous, etc.) biological relationships from information available in the literature/databases. The sparsity and time-shifted behaviour between TF-T pairs in gene regulatory networks was reflected well on choosing cross-correlation as the similarity measure.

References

1. Mitra, S., Das, R., Hayashi, Y.: Genetic networks and soft computing. IEEE/ACM Transactions on Computational Biology and Bioinformatics (to appear)
2. Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., Kamper, J.: Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics 21, 1069–1077 (2005)
3. Cheng, Y., Church, G.M.: Biclustering of gene expression data. In: Proceedings of ISMB 2000, pp. 93–103 (2000)
4. Mitra, S., Banka, H.: Multi-objective evolutionary biclustering of gene expression data. Pattern Recognition 39, 2464–2477 (2006)
5. Mitra, S., Das, R., Banka, H., Mukhopadhyay, S.: Gene interaction - An evolutionary biclustering approach. Information Fusion 10, 242–249 (2009)
6. Qian, J., Lin, J., Luscombe, N.M., Yu, H., Gerstein, M.: Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19, 1917–1926 (2003)
7. Cho, R.J., Campbell, M.J., Winzeler, L.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2, 65–73 (1998)
8. Yu, H., Gerstein, M.: Genomic analysis of the hierarchical structure of regulatory networks. Proceedings of National Academy of Sciences USA 103, 14724–14731 (2006)
9. Zeitlinger, J., Simon, I., Harbison, C., Hannett, N., Volkert, T., Fink, G., Young, R.: Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell 113, 395–404 (2003)