A Least Squares Fitting-Based Modeling of Gene Regulatory Sub-networks

Ranajit Das[1], Sushmita Mitra[1], C.A. Murthy[1], and Subhasis Mukhopadhyay[2]

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

Abstract. This paper presents a simple and novel least squares fitting-based modeling approach for the extraction simple gene regulatory sub-networks from biclusters in microarray time series gene expression data. Preprocessing helps in retaining the strongly interacting gene regulatory pairs. The methodology was applied to public-domain data sets of Yeast and the experimental results were biologically validated based on standard databases and information from literature.

Keywords: Biclustering, transcriptional regulatory network, least squares, gene interaction network.

1 Introduction

During the recent years, rapid development in DNA microarray technology have resulted in the parallel generation of expression data of thousand of genes, of various organisms, under several experimental conditions. Genome expression profiling of many organisms have been completed in the past few years. The latest Affymetrix gene chips accommodate 750,000 unique 25-mer oligonucleotide features constituting more than 28,000 mouse gene-level probe sets. It is known that mRNA profiles are prone to different kinds of noise and ambiguity, and may be unequally sampled over time. Time series gene expression data is also essentially under-determined, involving high-dimensional genes with very few time-points. Clustering is one way of estimating such noisy expression data, by grouping co-expressed genes with the assumption that they are co-regulated. However, it is observed that a subset of genes is co-regulated and co-expressed over a subset of experimental conditions only. Biclustering (or co-clustering) aims at bringing out such local structure inherent in the gene expression data matrix. It refers to some sort of feature selection and clustering in the space of reduced dimension, at the same time [1].

To quantify the similarity among the co-expressed genes in a bicluster several distance measures have been employed. However, it is to be noted that any apparent similarity of expression profiles between a pair of genes need not always signify direct regulation. It may denote an indirect coregulation by other genes,

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 165–170, 2009.

or it may also be due to a mere coincidence involving no causal relationship. The indirect interaction between two genes may result from the regulation mediated by proteins, metabolites and non-coding RNAs (ncRNAs). Transcription factor (TF) is a protein that interacts directly with its target gene(s) (T) by up regulating (down regulating) its gene expression – resulting in activation (inhibition) of the target. There may also exist regulatory cascades (of first-order interactions between gene pairs), whereby the product of one gene influences the transcription rate of the second one, and so on [1].

In this paper we propose the method of least squares fitting using polynomials in the framework of continuous-column multiobjective evolutionary biclustering [2] to extract the interaction between gene pairs. Preprocessing, involving the discretization of the error matrix (using quantile partitioning) and the subsequent elimination of links with higher errors, is employed to retain strongly regulated gene pairs. An adjacency matrix is formed from the resulting error matrix, based on which the regulatory network is generated and biologically validated. The usefulness of the model is demonstrated, using time-series gene expression data from Yeast.

2 Reverse Engineering of Gene Regulatory Sub-networks

The various properties of the genome, along with the expression of a gene (which is the amount of mRNA it produces) are addressed in an important group of biological networks known as the genetic regulatory network (GRN). A GRN comprises of a complicated structure involving different gene products that activate or inhibit other gene products [1]. The Multi-objective evolutionary algorithm (MOEA), in association with the local search, were used for the generation of the set of biclusters. The algorithm followed is discussed in details in [2].

2.1 Algorithm for Extraction of Gene Gene Regulatory Sub-networks

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

1. Extraction of biclusters by MOEA.
2. Computation of pairwise error for least squares fitting between gene pairs.
3. Discretization of the error matrix for eliminating the higher errors.
4. Network generation from the connectivity matrix.
5. Biological validation.

2.2 Least Squares Fitting of Gene Pairs

The huge size of the gene expression data and the associated combinatorial problems in high-dimensional space, have opened up a challenge to traditional techniques. It emphasizes the need for dimensionality reduction in this context. The major problem with DNA microarray data analysis is that the data is essentially under-determined, with very few samples (or time-points) compared to the

high-dimensional attributes/features (or genes) to be estimated; thus creating an additional constraint. Although standard statistical techniques for extracting relationships have proposed multiple models to fit the data [3], they often require additional data to resolve the ambiguities. These make the regression problem an ill-posed one; therefore a regularization technique is desired to be developed.

In this paper we propose a least squares fitting-based approach for the reconstruction of interactions in gene regulatory networks. Commonly used similarity measures like the Euclidean distance, Pearson correlation or the Spearman's rank correlation do not take into consideration the non-linear effects among genes. The above distance measures can serve as a satisfactory measure of relationship between two genes only when the two are linearly related. A low value of the correlation coefficient also does not rule out the possibility that the genes may be related in some other manner. Again, the fact that the coefficient of correlation between two genes is higher does not necessarily mean that they are causally related. Filkov *et al.* have designed an edge detection function for identifying regulation of genes, which demonstrates that less than 20% of the known regulatory gene pairs exhibit strong correlations [4].

The least-squares method, a very simple form of the regularization technique, can be helpful for model selection. Using the method of least squares we minimize the sum of squares of the error of estimation (S^2) of fitting one gene to another. If the computed error for fitting a particular gene $G1$ with another gene $G2$ (say), within the reduced localized domain of biclusters, be less than that for fitting $G2$ with $G1$ then we infer that $G1$ affects $G2$ and vice versa. The relationship is represented in terms of rules, linking the gene which regulates, to the regulated gene.

A gene expression profile e is represented over a series of n time points. Let S_1^2 denote the residual or the sum of squares of the error of estimation for fitting gene e_1 to e_2 (sampled at e_{1i} and e_{2i} over n time intervals). This is given by

$$S_1^2 \equiv \frac{1}{n}[\sum_i \{e_{2i} - (a_0 + a_1 e_{1i} + a_2 e_{1i}^2 + \ldots + a_k e_{1i}^k)\}^2]. \tag{1}$$

where a_k's denote the k coefficients of the kth order polynomial fitted. Analogously, the residual for fitting e_2 to e_1 can be represented as

$$S_2^2 \equiv \frac{1}{n}[\sum_i \{e_{1i} - (b_0 + b_1 e_{2i} + b_2 e_{2i}^2 + \ldots + b_k e_{2i}^k)\}^2]. \tag{2}$$

The coefficients, a_k's and b_k's, are obtained by solving the corresponding k normal equations

$$\frac{\partial S_1^2}{\partial a_k} = 0; \tag{3}$$

and

$$\frac{\partial S_2^2}{\partial b_k} = 0. \tag{4}$$

The corresponding errors $\xi(e_1, e_2)$ are represented as $\xi_1(e_1, e_2) = \sqrt{S_1^2}$ and $\xi_2(e_2, e_1) = \sqrt{S_2^2}$, which again are evaluated using eqns. (1), (2), (3) and (4), respectively.

The minimum error between the two fittings of the genes is selected, *i.e.*, $\xi(m,n)$ is chosen, if $\xi(m,n) < \xi(n,m)$ for the mn-th gene pair, identifying the gene that affects the other one more (for the pair). This results in the formation of the $ErrorMatrix$, $Err(i,j)$ (with $1 \leq i,j \leq N$, N being the total number of genes obtained in the bicluster) which is eventually discretized using quantile partitioning [5] for automatically eliminating the higher errors, which contribute less towards regulation. Only those error values are retained for which the absolute value is below a certain threshold, implying a larger interaction between the gene pairs. The entire range $[Err_{\max}(i,j), Err_{\min}(i,j)]$ is divided into three partitions each, using *quantiles* or *partition values*[1] so that the influence of noisy gene patterns are lessened. Error values smaller than Q_1^+ are indicative of less error and high interaction, while those with values in $(Q_1^+, Q_2^+]$ indicate moderate interaction.

An adjacency matrix is calculated as follows:

$$A(i,j) = \begin{cases} +1 \text{ if} & Err(i,j) \leq Q_1^+ \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

Here we have avoided fitting a gene with itself, assuming self interaction to be absent. Thereafter, a network connecting the various genes is generated. Note that the threshold Q_1^+ is automatically determined from the data distribution, thereby eliminating the need for user-defined parameters.

3 Results

Yeast cell-cycle CDC28 data [6], a collection of 6220 genes for 17 time points, taken at intervals of 10-minutes, were chosen for applying our methodology. Genes that were not annotated and those with more than 30% missing expression values were first removed from the data set. Eventually, a total of 6029 genes were taken for imputation of missing values according to the methodology provided in [7]. The minimum of the errors for fitting one gene with another and *vice-versa* is chosen based on least squares fitting using eqns. (1) – (4). The orders of the polynomial, k are chosen as 1 (linear), 2 (quadratic) and 3 (cubic). It is noticed that better fitting (lower error) is provided by choosing a cubic polynomial to fit a pair of genes. The stronger interactions, using thresholding by quantile partitioning, were retained for forming a network connecting the various genes.

A sample extracted sub-network comprising 14 genes for $k = 3$ is depicted in Fig. 1. A transcription factor is connected to its target gene by an arrow if such a TF-T pair existed within the biclusters. The biclusters were biologically validated from gene ontology study based on the statistically significant GO annotation database[2].

From our computations we obtained a strong interaction between the TF-T pair $YHR084W$-$YNL192W$, indicated by the directed arrow in Fig. 1, which

[1] Quantiles or partition values denote the values of a variate which divide the entire frequency into a number of equal parts.

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

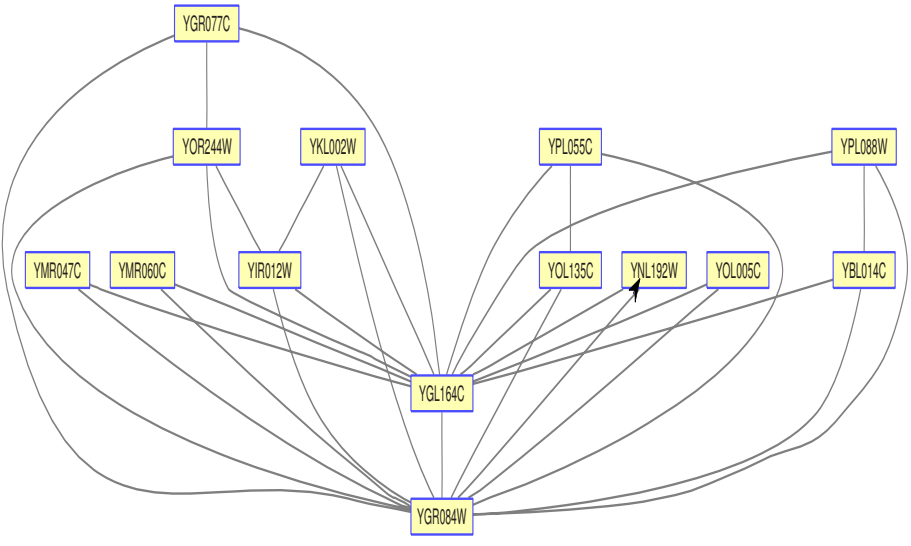

Fig. 1. A sample sub-network (bicluster) of 14 genes with transcription factor $YHR084W$ and target $YNL192W$

may be due to mating-specific binding behaviour or, they may belong to an additional mechanism for cell fusion. We also verified their summary from the *Saccharomyces Genome Database* (SGD)[3]. While identifying the hierarchical structure of regulatory networks [8] it was reported that $YHR084W$-$YNL192W$ forms a TF-T gene pair. One can also arrive at similar established conclusions for the other TF-T pairs (obtained in different biclusters). Our algorithm has not yet detected any false positive or false negative results.

4 Conclusions and Discussion

In this paper we have described an approach using the method of least squares fitting with polynomials, in the framework of continuous-column multiobjective evolutionary biclustering 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) – (4), followed by the quantile partitioning. The orders of the polynomial, $k = 3$ (cubic) leads to a better fitting (lower error) for a pair of genes. A sample TF-T gene interaction sub-network is depicted in Fig. 1. We tried to analyze the non-linear relationship between two genes which was validated using the statistically significant GO annotation database[4]. The least squares fitting-based method takes care of most

[3] http://www.yeastgenome.org/
[4] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder

higher-order dependencies between gene pairs, while automatically eliminating the need for user-defined parameters.

Acknowledgement

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. Mitra, S., Das, R., Hayashi, Y.: Genetic networks and soft computing. IEEE/ACM Transactions on Computational Biology and Bioinformatics (to appear)
2. Mitra, S., Banka, H.: Multi-objective evolutionary biclustering of gene expression data. Pattern Recognition 39, 2464–2477 (2006)
3. Kohane, I.S., Kho, A.T., Butte, A.J.: Microarrays for an Integrative Genomics. MIT Press, Cambridge (2003)
4. Filkov, V., Skiena, S., Zhi, J.: Analysis techniques for microarray time-series data. Journal of Computational Biology 9, 317–330 (2002)
5. Mitra, S., Das, R., Banka, H., Mukhopadhyay, S.: Gene interaction - An evolutionary biclustering approach. Information Fusion 10, 242–249 (2009)
6. 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)
7. Bo, T., Dysvik, B., Jonassen, I.: LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Research 32, 1–8 (2004)
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)