

A New Hierarchical Method for Identification of Dynamic Regulatory Pathways from Time-Series DNA Microarray Data

Alireza Darvish¹, E. Bak², K. Gopalakrishnan¹, R.H. Zadeh³, and Kayvan Najarian¹

¹ College of Information Technology

²Electrical engineering Department

³Physics Department

University of North Carolina at Charlotte

Email: {adarvish, ebak, kgopala, rhhkimza knajaria}@uncc.edu

Abstract

A new hierarchical method is proposed to analyze time-series DNA microarray data to identify dynamic genetic pathways. Initially the hierarchical method applies a specialized clustering technique to incorporate the available heuristic information about biological system. Then, the prototypes of the resulting clusters are used as time-variables to develop an Auto Regressive model relate the expression of the prototypes to each other. The resulting model also allows the prediction of gene expressions for the next time steps. The developed AR model can then be used to relate the expression value of each single gene to the genes of other clusters. The proposed method was applied to the cell-cycle dataset containing the DNA microarray time-series of a large number of genes involved in the eukaryotic cell-cycle. The technique resulted to a network of interactions among five clusters of genes in which the genes of each cluster have a biologically-meaningful trend in time.

1. Introduction

In recent years, many methods have been proposed and used to extract gene regulatory network. They include methods such as clustering, Boolean network, differential equation, reverse engineering and etc. But almost all above-mentioned techniques suffer from a significant shortcoming; while discovering dynamic models that describe the time-based interactions among genes is the true objective of the pathway identification paradigm, the existing methods are mainly designed to estimate “static” models from the steady-state data. In many drug discovery applications, while knowing the steady-state effects of drug is vital, the drug’s short-term activities and potential short transitional side effects on the molecular level must also be thoroughly studied and analyzed. In this paper we introduce a hierarchical dynamic modeling method based on a specialized clustering

technique and Auto Regressive (AR) models to address the above issue. In the proposed model, gene

expressions are considered as variables of the AR model, and the genes expression for the future time samples are estimated using the model. A major difference between the proposed method and other traditional applications of AR models is the way the large number of variable (i.e. genes) in the system is handled. For direct modeling of the interactions among a large number of genes via AR models, one would need an extremely large number of training points (in time) to reliably estimate all coefficients of the model. Due to the cost of conducting molecular biology experiments, many such time-series have only a few time steps in them, and therefore may not be sufficiently large to estimate a large number of parameters.

To address these issues, we exploit the fact that many genes behave very similarly in a biological study and therefore the role and effects of these genes can be somehow combined by a suitable clustering technique before dynamic modeling. More specifically, we group the genes in a small number of clusters and then use the prototype of each cluster as a single input in the AR model. This means that the AR model is developed among the prototypes of the clusters resulting from the clustering step as opposed to all individual genes. The model allows prediction of expression levels of each gene in the next steps based on the collective effects of other genes in the past samples. The distance measure used in the K-means algorithm was chosen to be “1-Correlation Coefficient” and we adopted the k-means clustering based on the biological knowledge we had from the trend of genes.

Cho *et al.* [1] have introduced and clustered near to 200 genes involved in the cell cycle of the budding yeast *S. Cerevisiae*. The clusters are created according to the known biological functions of the genes and the stage at which the genes are active. It is known that there are five major phases in cell cycle

development: EarlyG1 phase, LateG1 phase, S phase, G2 phase or M phase. The functional gene clusters are often formed based on the activation of genes in one of the five phases, i.e. the genes in each cluster are the ones active in only one of the five stages of cell cycle.

2. Auto Regressive Method

Once the clusters involved in a biological study are identified, the genes of the relevant clusters are placed into a pool for further analysis. This pool of genes comprises a set of “potential members” of the pathway for the biological changes of interest. Next, an Auto-Regressive (AR) model is applied to relate the expression levels of each of the prototypes in all gene clusters to each other. The model relates the future expression level of the prototypes of gene clusters (y_i 's) to the values of other prototypes in the past time(s). The model also considers the uncertainty inherent to the model by considering a noise factor (e) in the equations. In its most general form, the model is a linear system of difference equations, i.e.:

$$y_i(t) = -a_{i11}y_1(t-1) - \dots - a_{i1n_1}y_1(t-n_{n_1}) - a_{i21}y_2(t-1) - \dots - a_{i2n_2}y_2(t-n_{n_2}) - a_{ip1}y_p(t-1) - \dots - a_{ipn_p}y_p(t-n_{n_p}) + e(t) \quad (1)$$

where: y_i is the expression of the prototype (gene) i , n_{ai} 's are the degrees with respect to gene cluster i and gene cluster j , coefficients a_{jj} 's are the parameters of the model, p is the dimension of the system and e is the noise factor.

In this paper, we consider only one degree delay so equation (1) changes as below:

$$y_i(t) = -a_{i11}y_1(t-1) - a_{i21}y_2(t-1) - \dots - a_{ip1}y_p(t-1) + e(t) \quad (2)$$

By this equation we have the relation between expression value of prototype i in time t with expression value of all prototypes in time $t-1$. Parameters of above equation can be estimated by least square(LS) estimation.[2]

3. Results

Here we show the result of our clustering method in figure1. Figure 1 shows the prototype of each cluster and as it can be seen in this figure each prototype trend has one peak in each cycle. By equation (2) we predicted the expression value of each single gene based on the expression value of all genes in previous time.

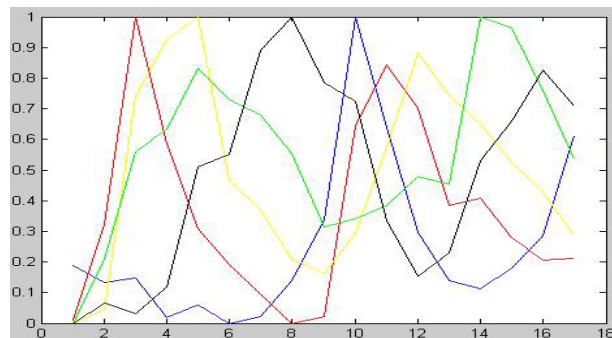


Fig1. Prototype of each cluster

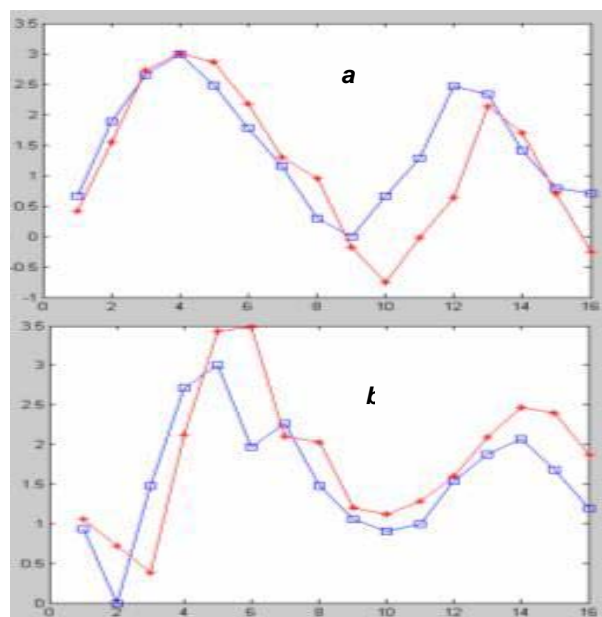


Fig2. . True and estimated expression values of some main genes in five clusters (“□”real values, “•”estimated values).(a. CIN8 b.CLG2)

4. Conclusion

The proposed method predicts the expression value of genes in time-series microarray data. In this method, we apply a K-means algorithm that considers the biological knowledge about the system and then an Auto Regressive (AR) model is used to quantitatively express the dynamic effect of all genes on each other. The method is tested the genetic study of the eukaryotic cell-cycle system. The results witness to the successful modeling performance of the proposed method.

Reference:

- [1] Cho,R.J. ,Campbell,M.J. , Winzeler,E.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” 1998, *Molecular Cell*, Vol. 2, 65–73.
- [2] Ljung,L.,”System Identification” Prentice-hall, Inc. 1987.