System Identification and Nonlinear Factor Analysis for Discovery and Visualization of Dynamic Gene Regulatory Pathways

A.Darvish* , K.Najarian*, D. H. Jeong* and W. Ribarsky*
*College of Information Technology, University of North Carolina at Charlotte,
Charlotte, NC 28213
{adarvish,knajaria,dhjeong,ribarski}@uncc.edu

Abstract-**DNA microarray time-series provide the information vital to estimate the dynamic regulatory pathways and therefore predict the dynamic interaction among genes in time. While dynamic system identification theory has been applied to many fields of study, due to some practical limitations, this theory has been widely used to analyze DNA microarray time series. In this paper, we describe some of these limitations and propose a hierarchical model utilizing nonlinear factor analysis methods to analyze time-series DNA microarray data and identify the dynamic regulatory pathways. The proposed model is applied to model the eukaryotic cell cycle process using a popular dataset of cell cycle time-series. The results indicate that the proposed method can successfully predict the dynamic pathway involved in the process.**

I. INTRODUCTION

Inferring the gene regulatory network is one of the major steps in addressing many problems in molecular biology. Discovery of gene regulatory network involves the estimation of the interactions among the genes involved in a given biological pathway using molecular biology data sets often provided by new high-throughout assays [1]. The main objective of such studies is to identify the genes and regulatory networks affecting the expression level of a particular gene and/or the concentration of a given molecule. In the majority of the existing methods, the regulatory pathways are identified using the mRNA expression data produced by DNA microarray machines. Modeling of the regulatory networks allows quantitative estimation of the gene expressions, which in turn facilitates and expedites studies such as drug discovery. This problem is often too complicated to have a definite solution. In the recent years several methods have been introduced to address this problem. These methods include Boolean Networks [2,3], Bayesian Networks [4,5], Dynamic Bayesian Networks [6], linear models [7], compartment modeling using differential equations [8], techniques based on control theory [9], full biochemical interaction models [10] and methods using metabolic regulation concepts [11,12,13,14]. The above-mentioned methods have span wide range of computational complexity. While some of these techniques such as Boolean networks are too abstract and simplified, others are computationally complex. The complex models include techniques modeling all biochemical interactions among genes based on a large number of differential equations.

In many drug discovery applications, while knowing the steady-state effects of drug is vital, the drug's short-term activities and potential transient effects on the molecular level must also be thoroughly studied and analyzed. As a result, inferring dynamic gene regulatory networks that can be extracted from time series DNA microarray data has recently attracted a tremendous amount of attention [15,16].

In a typical Auto Regressive (AR) model the main variables, e.g. gene expressions, are considered as variables of the AR model, and the values of these variables for the future time samples are estimated using the model. There are some practical issues in directly using AR for modeling of microarray data. In direct modeling of the interactions among a large number of genes with AR models, one needs an extremely large number of training points (in time) to reliably estimate all model coefficients. Due to several factors including 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 long to estimate a large number of model parameters. In addition, a blind application of AR models to molecular biology problems would not provide insightful clustering of the genes involved in a biological process, i.e. the model would fail to display the massive grouping and parallelism in genetic network. 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.

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. In [17], we used k-means algorithm to cluster all genes to five cluster and then applied the AR algorithm to prototypes of clusters. This process significantly reduces the parameters of the AR model. A disadvantage of this method is that it finds the model between prototypes of different classes of genes, and therefore, may not predict the expression value of individual genes with high accuracy.

In this paper we use non-linear component analysis instead of k-means algorithm. This method first applies nonlinear component analysis to extract the main components that represent the trends of gene expressions, and then employs an AR method to model the interactions among these components. The advantage of this method is that once the values of components in future time step

0-7803-9387-2/05/$20.00 ©2005 IEEE

are estimated using the AR model, the non-linear component analysis model can predict the expression of individual genes in future time steps. We also present a highly interactive visualization technique to allow users effectively perceive the structure and predictions of the resulting models.

The rest of the paper is organized as follows. Section 2 provides a brief explanation of the nonlinear component analysis method and AR model. In Section 3, the eukaryotic cell cycle process and the database used for this study are briefly described and this section presents the result of the model and describes the proposed visualization method for dynamic gene regulatory network. Finally, Section 4 concludes the results.

II. PROPOSED METHOD

The block diagram of the proposed model is shown in Figure 1. As shown in the block diagram of Figure 1, we first use nonlinear factor analysis (NLFA) to extract the major time components of all genes and then apply the AR model on the components.

Applying AR on a few components as opposed to a large number of genes dramatically reduces the number of model parameters and results to a much more reliable model. Once these two steps are performed, the AR model can predict the future values of the components. Then using these predicted values for the components and the NLFA formulation in the inverse form, one can predict the future values of individual genes.

Next, each step of the algorithm is described in more details.

A. Nonlinear Component Analysis

One of the most challenging problems in signal processing is blind separation of sources from their nonlinear mixtures. In this problem it is assumed that we have access only to the observation $x(t)$ that is consisted to nonlinear mixtures of sources signal $y(t)$, i.e.

$$x(t) = f[y(t)] + n(t) \tag{1}$$

where $n(t)$ is an additive noise and $f(.)$ describes the nonlinear mapping (mixture). In our gene expression method:

$$x(t) = [x_1(t), x_2(t), \cdots, x_K(t)] \tag{2a}$$

$$y(t) = [y_1(t), y_2(t), \cdots, y_p(t)] \tag{2b}$$

where $x_i(t)$ represents the expression level of gene i at time t and $y_j(t)$ denotes the value of the jth component at time t. The main challenge of blind separation of sources is that both the mapping function and sources must be estimated from the observed data. A method to address this problem that will be used in our method has been introduced by Lappalainen *et. al.* [18]. This method is a nonlinear counterpart of principal component analysis (PCA). Since this model includes a noise term it is often called nonlinear factor analysis. In this method, a multilayer perceptron (MLP) network is used to model the nonlinearity of the system. The main reason for using MLP is the fact that MLP network can model fairly accurately many naturally occurring multidimensional nonlinear processes. In particular, in many cases MLP gives a simpler parameterization for the involved nonlinearity than Taylor or Fourier series explanations. The sources are on the output layer of MLP and observations are on the input layer and the middle layer contains hidden neurons to compute a nonlinear function of inputs. The equation of mapping is:

$$\begin{aligned} x(t) &= f(y(t)) + n(t) \\ &= C\tanh(Dy(t) + g) + h + n(t) \end{aligned} \tag{3}$$

The matrices C and D are the weights of first and second layers, and vectors g and h are the corresponding bias terms. The main goal of learning is to approximate the posterior pdf of all the unknown values in the model by minimizing a cost function defined based on posterior pdf's. After obtaining the optimal set of parameters A, B, a, and b, the sources can be estimated. The algorithm is a reversible process because after applying any change on the sources the new observation can be reconstructed by the network.

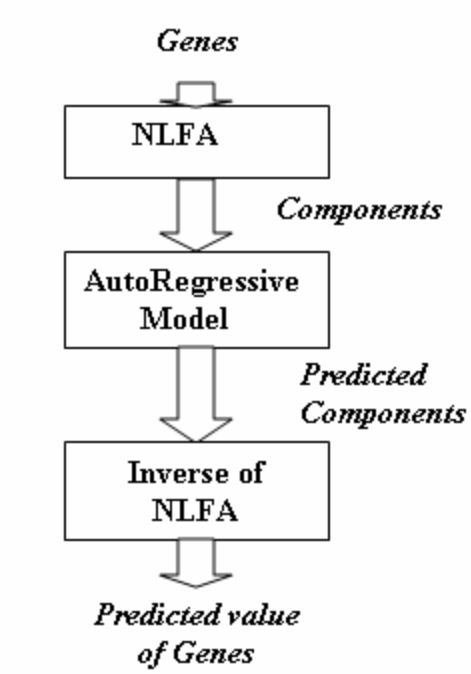

Figure1: Block diagram of the proposed algorithm.

We apply this technique to obtain major trends which then represent "state variables" facilitating the estimation of the expression values for all genes. So the expression value of all genes in a specific number of time steps plays the role of observations in non-linear components analysis technique and the major trends of the genes are the sources.

B. Auto-Regressive Model

Once the major components of the set of all genes have been identified, an Auto-Regressive (AR) model is applied to relate the expression levels of each of the components to each other. The model relates the future expression level of each component to the values of other components 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.:

$$\begin{aligned} & y_i(t) = -a_{i11}y_1(t-1) - \ldots - a_{i1n_1}y_1(t-n_1) \\ & -a_{i21}y_2(t-1) - \ldots - a_{i2n_2}y_2(t-n_2) \\ & \ldots \\ & -a_{ip1}y_p(t-1) - \ldots - a_{ipn_p}y_p(t-n_p) + e \end{aligned} \tag{4a}$$

where: y_i is the value of the component *i*, coefficients a_{ij} are the parameters of the model, n_j is the degree with respect to component *i* and component *j* and *e* is the noise factor. The above AR model can also be shown as:

$$\begin{aligned} y(t) \quad & = A_1 y(t-1) + A_2 y(t-2) + \ldots \\ & + A_p y(t-n_p) + e(t) \end{aligned} \tag{4b}$$

where A_i's are the coefficients for "*t-i*". Now, define θ as the vector of all parameters, i.e.

$$\begin{aligned} \theta = \quad & (a_{i11}, a_{i12}, \ldots, a_{i1n_1}, \\ & a_{i21}, a_{i22}, \ldots, a_{i2n_2}, \ldots, \\ & a_{ip1}, a_{ip2}, \ldots, a_{ipn_p}) \end{aligned} \tag{5}$$

Next, in order to predict the coefficients of equation (4), for each time step, the function $\varepsilon(t,\theta)$ is defined as the prediction error:

$$\varepsilon(t,\theta) = y(t) - \hat{y}(t \mid \theta) \tag{6}$$

where $\hat{y}(t \mid \theta)$ is the predicted output given the set of parameters θ. The prediction-error sequence can be processed through a stable linear filter $L(q)$ to further specialize the error function:

$$\varepsilon_F(t,\theta) = L(q)\varepsilon(t,\theta) \tag{7}$$

where "*q*" stands for an element of delay. Next, we define $V_N(\theta, Z^N)$ as the measure the total error (averaged over all *N* time points):

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} l(\varepsilon_F(t,\theta)) \tag{8}$$

The function $l(.)$ is any scalar-valued (positive) measure function (often defined as the square function). The parameter estimation is then defined as finding a set of parameters that minimizes the total error function, i.e.:

$$\hat{\theta}_N = \hat{\theta}_N(Z^N) = \arg\min_{\theta \in D_M} V_N(\theta, Z^N) \tag{9}$$

In this paper, the square function is used as $l(.)$ and the least-squares method is applied to obtain the best set of parameters. Assume that:

$$\hat{y}(t \mid \theta) = \varphi^T(t)\theta \tag{10}$$

where φ is the regression vector defined as:

$$\varphi(t) = [-y(t-1) \quad -y(t-2) \quad \ldots \quad -y(t-n)]^T \tag{11}$$

From (10), the prediction error becomes:

$$\varepsilon(t,\theta) = y(t) - \varphi^T(t)\theta \tag{12}$$

Assuming $L(q) = 1$ (i.e. identify filter), and $l(\varepsilon) = \frac{1}{2}\varepsilon^2$ (i.e. the square function), the total averaged error criterion function becomes:

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{2}[y(t) - \varphi^T(t)\theta]^2 \tag{13}$$

This is the least squares criterion for the linear regression and it can be minimized analytically, which gives the following solution:

$$\begin{aligned} \hat{\theta}_N^{LS} \quad & = \arg\min V_N(\hat{\theta}_N, Z^N) \\ & = [\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)]^{-1}\frac{1}{N}\sum_{t=1}^{N}\varphi(t)y(t) \end{aligned} \tag{14}$$

III. MODELING AND PREDICTION OF CELL CYCLE DYNAMIC PATHWAY

The dataset we use to validation our method is budding yeast S. Cerevisiae cell cycle dataset introduced by Cho et. al. [19]. In [19], the genes in the cell cycle data are clustered according to their known biological functions, e.g. the stage at which the genes are active. It is known that there are five major phases in cell cycle development: Early G1 phase, Late G1 phase, S phase, G2 phase, and M phase. The functional gene clusters are formed based on the activation of genes in one of the five phases, i.e. the genes in each clusters are the ones active in only one of the five stages of cell cycle. The dataset used in [19] to create the clusters comprises the mRNA transcript levels of all studied genes during the cell cycle of the budding yeast *S. Cerevisiae*. To obtain synchronous yeast culture, cdc28-13 cells were arrested in late G1, raising the temperature to C^o 37, and the cell cycle was reinitiated by shifting cells to C^o 25 . Cells were collected at 17 time points taken at ten-minute intervals, covering nearly two cell cycles.

The number of total genes considered in this study is 40. In order to train the model, the first ten time points of each component are used as training data to find the NLFA and AR models. Then, the expression values of each component and as a result all individual genes are predicted for all time steps.

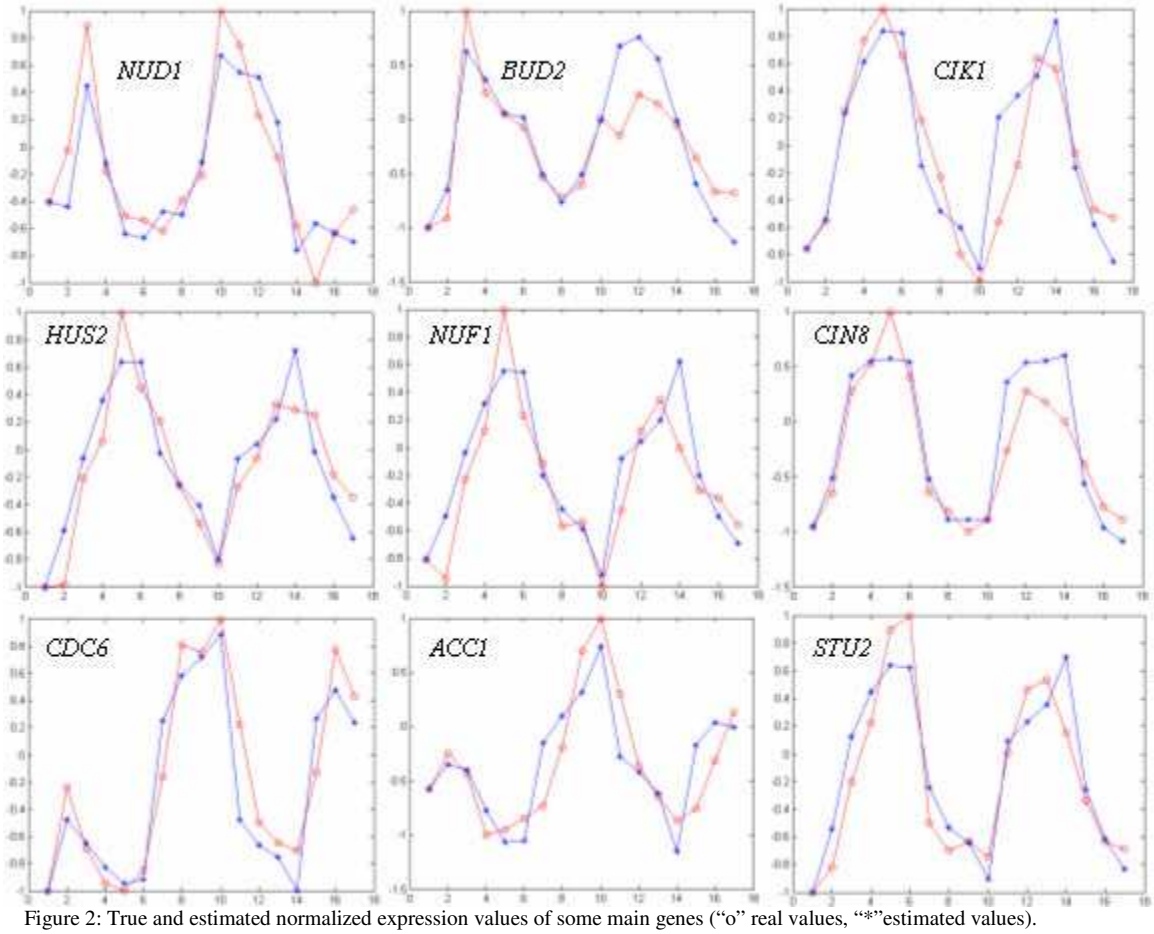

Figure 2: True and estimated normalized expression values of some main genes ("o" real values, "*"estimated values).

Since the number of time points in the training data is small (i.e. eight steps in each cycle), in order to minimize the number of parameters, the degree of the model is set to one.

A. Results

Following the formulation of Section 2, training of an AR model is equivalent to the estimation of a_{ij} coefficients, where i =1,…,5, and j =1,…,5. Since the degree of the model is set to 1 for all genes, the third index of the parameters is dropped. After training with the first ten time points the following values for a_{ij} coefficients re obtained as:

$$A = \begin{bmatrix} -1.045 & -0.970 & 0.634 & -36.803 & -19.920 \\ 0.848 & 0.033 & -0.606 & 28.340 & 40.800 \\ 0.0860 & 0.919 & -0.780 & 26.539 & 20.936 \\ -0.0022 & 0.0004 & 0.0135 & -0.480 & -0.285 \\ 0.0065 & 0.0124 & -0.0233 & 0.433 & 0.474 \end{bmatrix} \quad (15)$$

Using each row of this matrix one can predict the value of one component in the next time step based on the values of the all components in the previous times. For example, for the first component, the model can be written as:

$y_1(t) = -1.045\ y_1(t\text{-}1) - 0.970\ y_2(t\text{-}1)$
$+0.634\ y_3(t\text{-}1) - 36.803\ y_4(t\text{-}1)$
$-19.920\ y_5(t\text{-}1)$ (16)

Similar equations can be obtained from the matrix given above for other components. As mentioned above, the first 10 points of the data was used to train the model, and the capabilities of the model in predicting the correct values of expression in future samples are tested against all the time steps, which includes 17 time samples. To test and validate the performance of the model in estimating the future values of individual gene, we apply the nonlinear factor analysis method to all genes and after producing the major components of genes, the network of NLFA is used to predict the expression value of each single gene in the future time steps.

The prediction results for some of the genes are shown in Figure 2. As can be seen in Figure 2, the predicted values match very well with the true expression values of genes indicating that the model can successfully predict the trend of the genes

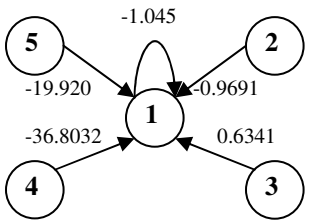

Figure 3: Effect of all components on component 1.

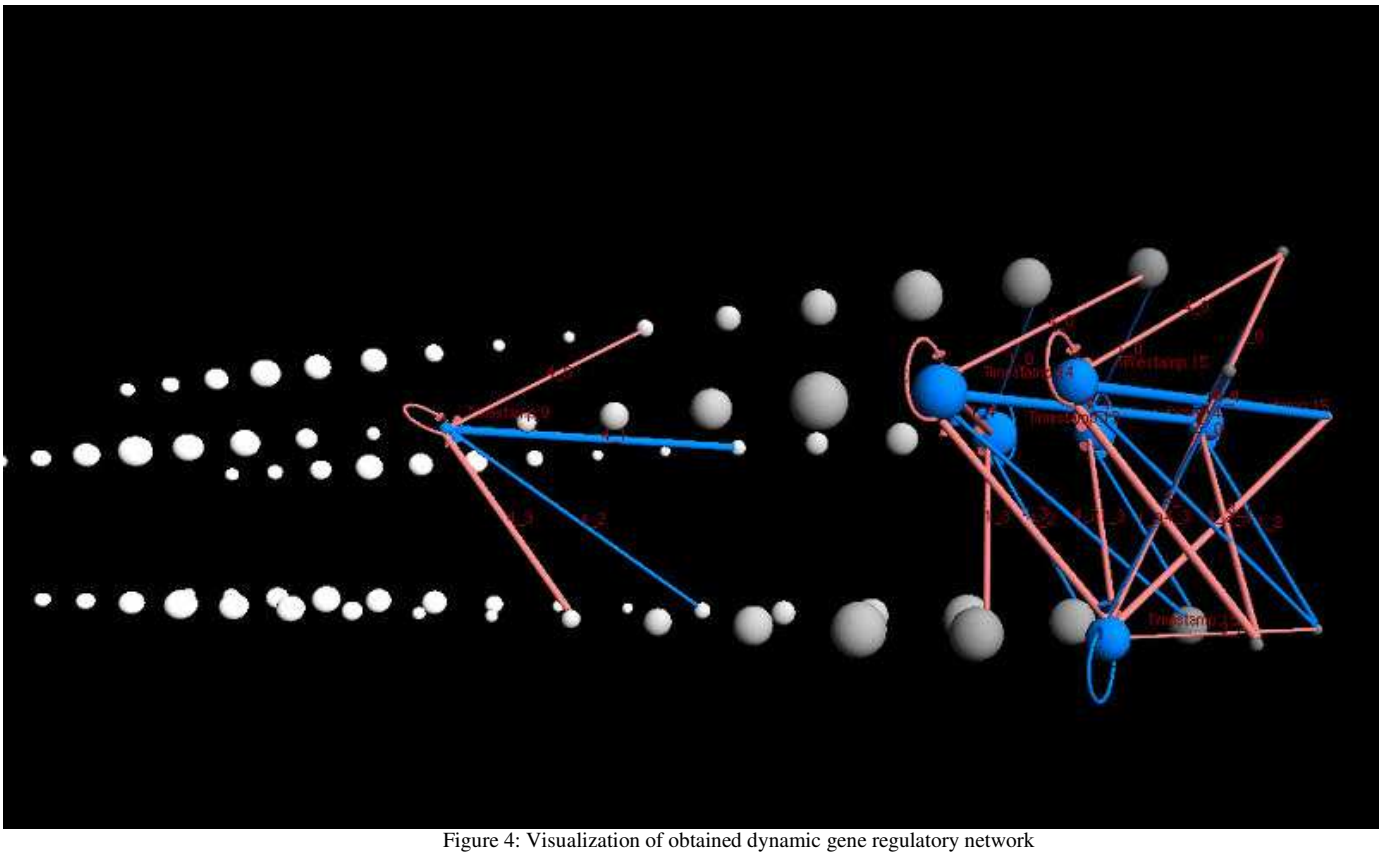

Figure 4: Visualization of obtained dynamic gene regulatory network

expressions based on the expression values of all genes at the previous times.

In addition, one can draw a network between five components to represent the dynamic pathway of cell cycle process. Such a network will show the effect of each component on itself and on other components in the next time step. Since showing all 25 links and arrows would complicate the graph, in Figure 3, we have only shown the effect of all components on the first component. An effective visualization method to show all links will be presented in the next section. It is important to note that in the resulting dynamic graph, there is a one-step delay in the effect of all components on component one (i.e. the components determine the expression value of the component one in the next time point). Such dynamic networks and pathways are extremely important in many bioinformatics applications including automated drug discovery.

B. Visualization

We also present a visualization technique to visualize the dynamic network between components. Figure 4 shows a snapshot of the method. As can be in Figure 4, the major components have been shown in time and the effect of all components on a specific component on the next time step has been shown by a dynamic network.

The visualization method shows a "conditional box" and links between 5 components. Time is laid out towards the user's viewpoint (farther objects are earlier in time). The interface is interactive. The user can rotate the whole scene, zoom in or out. The user can also select components at a given time step to show pathways and then move the conditional box along the time access to show dynamic behavior and links between components can be color and size-coded to show sign and strength

The conditional box can be moved back and forth in time. Selected links within the box will be displayed (but not outside). This gives a sense of the behavior over time while removing component. If the component time structure is positioned with a top-down view, one can get an animation over time when moving the box. There is a threshold selection so that only pathways with strength in a certain range are shown. In future, we are planning displays for structures within the clusters associated with individual genomes.

IV.CONCLUSIONS

We presented a hierarchical computational method to predict the expression value of genes in time-series microarray data. In this method, first we apply a nonlinear factor analysis algorithm that extracts the major components covering all considered genes. Then, an Auto Regressive (AR) model is used to quantitatively express the dynamic interactions of all components with each other. Then, using the network of

nonlinear factor analysis method the expression value of each gene can be obtained from predicted values of components. The method has been tested against the genetic study of the eukaryotic cell cycle system. The results witness to the successful modeling performance of the proposed method. Also a visualized method to present the obtained network was presented. The visualization technique shows the dynamic network among major components and the magnitude of effect the components have on each other in different time steps.

REFERENCES

[1] Hartemink AJ, Gifford DK, Jaakkola TS,Young RA , "Combining location and expression data for principled discovery of genetic regulatory network models", Pac Symp Biocomput. 437-49,2002

[2] Akutsu,T., Miyano,S. and Kuhara,S. "Identification of genetic networks from a small number of gene expression patterns under the boolean network model." In Proceedings of the 1999 Pacific Symposium in Biocomputing (PSB 99), pp. 17–28.

[3] Liang,S., Fuhrman,S. and Somogyi,R. "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures." In Proceedings of the 1998 PacificSymposioum in Biocomputing (PSB 98), pp. 18–29.

[4] Friedman,N. , Linial,M. , Nachman,I. and Pe'er,D. "Using Bayesian Network to Analyze Expression Data" Journal of Computational Biology, Vol. 7, pp. 601-620, 2000

[5] Setter,M., Deco,G. and Dejori,M. "Large-Scale Computational Modeling of Genetic Regulatory Networks" Artificial Intelligence Review ,Vol. 20, pp. 75-93 , 2003

[6] Thieffry,D. and Thomas,R. "Qualitative analysis of gene networks." In Proceedings of the 1998 Pacific Symposioum in Biocomputing (PSB 98), pp. 77–88.

[7] Dhaeseleer,P., Wen,X., Fuhrman,S. and Somogyi,R. "Linear modeling of mRNA expression levels during CNS development and injury". In Proceedings of the 1999 Pacific Symposioum in Biocomputing (PSB 99), pp. 41–52.

[8] Chen,T., He,H.L. and Church, G.M. "Modeling gene expression with differential equations". In Proceedings of the 1999 Pacific Symposioum in Biocomputing (PSB 99). pp. 29–40.

[9] Kholodenko,B.N.,Kiyatkin,A.,Bruggeman,F.J., Sontag,E. ,Westerhoff,H.V. and Hoek,J.B "Untangling the wires: A strategy to trace functional interactions in signaling and gene networks" PNAS, Vol. 99, No. 20, pp. 12841–12846,2002.

[10] Arkin,A., Ross,J. and McAdams,H.H. "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda infected Escherichia colli cells" Genetics,Vol. 149, pp. 1633-1648,1998.

[11] Klipp,E., Heunrich,R. and Holzhutter,H. "Prediction of temporal gene expression Metabolic optimization by re-distribution of enzyme activities" European journal of biochemistry, 2002, Vol. 269, pp.5406-5413.

[12] Heinrich, R. and Schuster, S. "TheRegulation of Cellular Systems", Chapman & Hall, 1996.

[13] de la Fuente,A., Brazhnik,P. and Mendes,P. "Linking the genes: inferring quantitative gene networks from microarray data" Trends In Genetics, Vol.18,No.18, pp.395-398,2002.

[14] Vo,T.D., Greenberg, H.J. and Palsson, B.Ø. " Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data", Journal of Biological Chemistry,Vol. 279, No. 38, pp.39532-40,2004.

[15] Li H, Luan Y, Hong F, Li Y. Statistical methods for the analysis of time course gene expression data. Front Biosci, Vol. 7, pp.90-98, 2002.

[16] Toh H, Horimoto K. . "Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling" Bioinformatics,2002, 18:287-297.

[17] Darvish,A., Hakimzadeh,R. , Najarian,K. "Discovering Dynamic Regulatory Pathway by Applying an Auto Regressive Model to Time Series DNA Microarray Data" Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, USA Sept. 1-5, 2004.

[18] Lappalainen,H. and Honkela,A. ," Bayesian Nonlinear Independent Component Analysis by Multi-Layer Perceptrons" In Advances in Independent Component Analysis,edited. by Mark Girolami, Springer , pp. 93-121, 2000.

[19] Cho,R.J. ,et. al. "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle" Molecular Cell, Vol. 2, pp. 65-73, 1998.