Genetic Regulatory Network Modeling Using Network Component Analysis and Fuzzy Clustering

Fatemeh Bakouie, Mohammad H.Moradi
Department of Biomedical engineering Amirkabir University of technology Tehran., Iran
emails: fbakouie@cic.aut.ac.ir , mhmoradi@aut.ac.ir

Abstract— **Gene regulatory network model is the most widely used mechanism to model and predict the behavior of living organisms. Network Component Analysis (NCA) as an emerging issue for uncovering hidden regulatory signals, has attracted significant trends in the research community. The common scheme in NCA is to model the controlling behavior of some proteins on the expression value of genes. However, this modeling requires performing certain experiments which are expensive in terms of time and feasibility. In this paper, we employ simple and effective data mining algorithm to obtain a purely gene- to gene model which predicts the effect of certain genes on the whole system. In order to accomplish this goal we employ Fuzzy clustering and Mutual Information (MI) for determining regulator genes resulting in two methods named as: Mutual Information based NCA (MINCA) and Fuzzy based NCA (FNCA). Simulation results validated using Coefficient of Determination (CoD), show that our methods model the system simpler and more accurate than conventional schemes.**

I. Introduction

Traditional molecular biology typically focuses on a single gene, protein, reaction or pathway, and follows a reductionism approach to study biological systems. Over the years, this practice has led to remarkable achievements. However, biological processes are inherently integrated and interactive. Therefore traditional studies cannot resolve the complex relationships among biological entities. In order to understand the nature of cellular function, it is necessary to study the behavior of genes in a holistic rather than an individual manner. Since the expressions and activities of genes are not isolated or independent of each other[1] . networks, neural networks, differential equations, and models including stochastic components on the molecular level [1] .

High-throughput techniques in biology, such as DNA microarray have generated a large amount of data that can potentially provide systems-level information regarding the underlying dynamics and mechanisms. These high- dimensional output data are typically the end products of low-dimensional regulatory signals driven through an interacting network. As illustrated in figure 1, the relationship between the lower dimensional regulatory signals (or states) and output data can be modeled by a bipartite networked system, where the output signals (e.g., gene expression levels) are generated by weighted functions of the intracellular states (e.g., the activity of the transcription factors or expression of regulatory genes). A major challenge in systems biology is to develop methodologies for simultaneous reconstructions of the hidden dynamics of the regulatory signals [9].

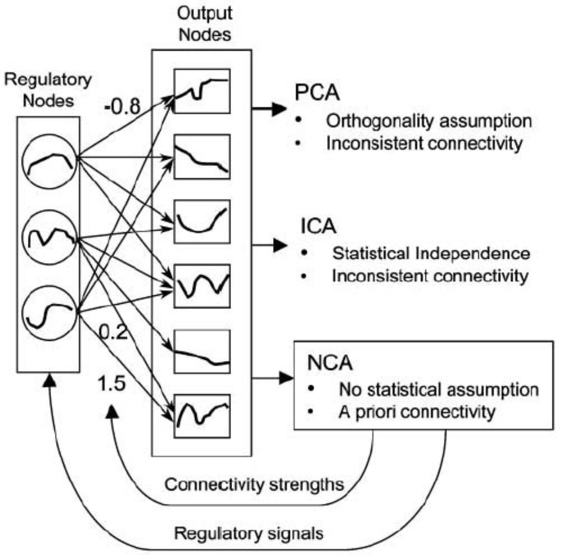
Output Nodes
Regulatory Nodes
-0.8
0.2
1.5
PCA
• Orthogonality assumption
• Inconsistent connectivity
ICA
• Statistical Independence
• Inconsistent connectivity
NCA
• No statistical assumption
• A priori connectivity
Connectivity strengths
Regulatory signals

Fig. 1. A regulatory system in which the output data are driven by regulatory signals through a bipartite network.

In this paper we use Network Component Analysis (NCA) for constructing Genetic regulatory Network. In common works, expression values of different genes are linked to the concentration values of activating form of regulatory proteins and a prior knowledge about connectivity pattern in the network obtain from some biological experiment. Besides in this paper our network models interaction between genes and without any biological knowledge about connectivity pattern among genes. So we use two data mining methods to obtain this knowledge about connectivity pattern of gene, which are: Mutual Information (MI) and a Fuzzy clustering algorithm.

This paper is organized as follows. Section II describes the Network Component Analysis, our approach in modeling by using NCA and two data mining methods to obtain prior knowledge about connectivity pattern of our model and Coefficient of Determination (CoD) that is used for evaluating our models, Section III lists some related work about modeling problem, Section IV provides experimental results. In section V we discuss about the result and future work.

1-4244-0710-9/07/$20.00 ©2007 IEEE

II. RELATED WORK

In order to construct formal models of genetic interactions, in related studies variety of mathematical and computational methods have been developed. There have been a number of attempts to model gene regulatory networks, including linear models, Bayesian networks, neural networks, differential equations, and models including stochastic components on the molecular level [1], [2].

During recent years, statistical techniques for determining low dimensional representations of high-dimensional data sets, e.g., principal component analysis (PCA) or singular value decomposition and independent component analysis (ICA), have been applied successfully to deduce biologically significant information from high-throughput data sets. PCA and ICA both would generate linear networks for interpreting the observed data set, where the regulatory signals are constrained to be mutually orthogonal and statistically independent, respectively. However, both the reconstructed signals and the networks do not match the real system and provide only a phenomenological modeling of the observed data. A semi-blind deconvolution method for analyzing gene expression data have been proposed recently in [9], [11]. This method has some advantage: Traditional statistical methods for computing low-dimensional or hidden representations of these data sets, such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), ignore the underlying network structures and provide decompositions based purely on a priori statistical constraint on the computed component signals. The resulting decomposition thus provides a phenomenological model for the observed data and does not necessarily contain physically or biologically meaningful signals. In this method named as Network Component Analysis (NCA) the expression values of different genes are linked to the concentration values of activating form of regulatory proteins according to a simple model described in following section. Besides in Genetic Regulatory Networks these links are purely between genes. We assume that some of genes have regulatory affection on other genes in the network. So we can model this affection by means of linking which we define in NCA. The only problem is to recognize the genes that have regulatory affection on others. We solve this problem by means of two data mining methods: Mutual Information and Fuzzy clustering. Therefore we model interactions between genes without any prior biological knowledge that usually obtained from some certain experiments.

III. METHODOLOGY

A. Mathematical framework of NCA

The multidimensional data are organized in a format where M samples (or time points) of N output variables (such as the expression level of genes) is collected in the rows of a matrix $[E]$ (size: N rows and M columns). We seek to reconstruct a model of the type:

$$[E] = [A] \times [P] \tag{1}$$

Here the matrix $[P]$ (size: LM) consists of samples of L regulatory signals, where L is in general smaller than N, thus resulting in the reduction in dimensionality. The matrix $[A]$ $(size : NL)$ encodes the connectivity strength between the regulatory layer and the output signals(figure 1). Eq. 1 represents the linear approximation of any detailed mechanistic model and is commonly used as the first approximation when the latter is unavailable [9], [11].

In this paper, expression values of different genes in the network are linked to regulator genes according to a simple model described in the following.

Once the identifiably of a given system has been established, the regulatory signals, [P], and connectivity strength, [A], can be reconstructed trough the following procedure. An initial guess for the connectivity matrix A is formed by setting to zero all of the elements corresponding to missing edges between the regulatory layer and the output layer. In fact genes that have controlling and regulatory roles on expression of other genes in the network should be determined. These genes have the ability of predicting expression of other genes. In this paper we use two data mining methods for obtaining these genes and connectivity pattern of network. They are based on Mutual information (MI) and Fuzzy clustering that will be described briefly following.

Because the experimental measurements are noisy, an exact solution to the decomposition problem does not exist in general. However, when the NCA criteria are satisfied, the estimation problem becomes well posed, and a solution that provides the best fit in the least squares sense can be computed. We proceed by minimizing the following objective function:

$$min|[E] - [A][P]|^2 \tag{2}$$

For detail about solving this problem, refer to [9].

Coefficient of Determination (CoD) is employed to compute merit of our prediction. The determination coefficient is defined in accordance with the degree to which a filter estimates a target variable (by using some predictor variables) beyond the degree to which the target variable is estimated by its mean (without any predictor variable)[7],[8].

$$CoD = \frac{\varepsilon_\mu - \varepsilon}{\varepsilon_\mu} \tag{3}$$

Where $\varepsilon_\mu$is error of prediction in the absence of predictor variables and $\varepsilon$is error of prediction when we use predictor variables. So the bigger CoD shows better filtering and prediction.

B. Selecting Predictor Genes Using Mutual Information

The motivation for considering mutual information is its capacity to measure a general dependence among random variables. The main idea of this technique is to identify a relatively small number of candidate parents for each gene based on statistics such as correlation. Shannon's information theory provides a suitable formalism for quantifying the above concepts. The mutual information (MI) between X and Y is a measure of information about X (or Y) contained in Y (or X) and given by:

$$I(X;Y) = H(X) - H(X|Y) \tag{4}$$

Where H(X) is entropy of X [1], [4], [5]. We assume that regulator genes contain the most information content of other genes in the network. For each gene we compute sum of pair wise MI between this gene and other genes. Genes that have the most MI with other genes can be considered as regulator genes that can predict expression levels of other genes. For obtaining initial connectivity pattern, we use a threshold. We consider a link between each gene in the network and each of predictor gene if they have a pair wise MI more than this threshold. We name this method of modeling as Mutual Information based NCA (MINCA).

C. Selecting Predictor Genes using Fuzzy Variable Neighborhood Search Clustering

The VNS algorithm is a proposed metaheuristic for solving combinatorial and global optimization problems. The basic goal of the method is to proceed to a systematic change of neighborhood within a local search algorithm. This algorithm remains in the same locally optimal solution exploring increasingly far neighborhoods by random generation, until another solution better than the incumbent is found. When so, it jumps to the new solution and proceeds from there. The neighborhood centroid structures are obtained by replacing at random some predetermined number k of existing centroids of clusters with k randomly chosen patterns, i.e., genes. For a more detailed analysis of VNS metaheuristic, see references [6, 10]. In this paper we use Fuzzy General VNS (FGVNS) method for clustering. Indeed for local search we use fuzzy clustering because Fuzzy logic can effectively model gene regulation and interaction to accurately reflect the underlying biology. Fuzzy clustering method allows genes to interact between regulatory pathways and across different conditions at different levels of detail. Fuzzy cluster centers can be used to quickly discover causal relationships between groups of co regulated genes [2, 6]. Fuzzy clustering algorithms assign a value named membership function, to each gene that show how much this gene belong to specific cluster. We assume gene that has the most membership function belong to that cluster the most and so can predict other genes that exist in the cluster well. We consider these genes as predictor genes. In fuzzy clustering each gene belong to all clusters but with different weights. So if we use a threshold on membership function, genes will just belong to some clusters not to all of them. By using this, we can obtain prior knowledge about connectivity pattern of network. This is done by considering linking between genes that are co-cluster with predictor genes. We name this method of modeling as Fuzzy based NCA (FNCA).

IV. Results

Our experiments are based on the observations in transcription level in the context of responsiveness to genotoxic stresses. The ternary data of the survey (14 genes and 30 samples) are given in [1, 7]. We compare our results to a linear modeling named as perceptron which is mentioned in ref [1]. we consider precision of regulatory network just for 4 genes: x12, x11, x10 and x2. For modeling the network, at first we should find parent genes that can predict other genes in the network. It is done via Fuzzy clustering and MI. We have done clustering for different number of clusters and selecting different number of genes that have the most pair wise MI with other genes. In this work the number of parent genes (regulator genes) varies from 3 to 8, so we have different set of parent genes for each target gene and we can model network in different level of dependencies. Table I, II and III, show results of networking for each method of modeling.

TABLE I
CoD FOR MINCA

No. parent genes	CoD. G2	CoD. G10	CoD. G11	CoD. G12
3	0.440	0.370	0.360	0.306
4	0.491	0.411	0.467	0.440
6	0.604	0.591	0.626	0.623
8	0.958	0.949	0.956	0.950

TABLE II
CoD FOR FNCA

No. parent genes	CoD. G2	CoD. G10	CoD. G11	CoD. G12
3	0.387	0.555	0.398	0.583
4	0.325	0.567	0.489	0.594
6	0.473	0.692	0.580	0.668
8	0.516	0.715	0.667	1

V. Discussion

Result of our simulation is reported in table I to **??**. In first table precision of modeling using Mutual Information based NCA (MINCA) is reported. Table II shows result of modeling using Fuzzy NCA (FNCA). In table **??** we compare our result with a nonlinear model introduced in [1]. In that work, predictor design has been done by using: reversible-jump Markov-chain-Monte-Carlo (MCMC). Results show that although our model is simpler than MCMC but it requires more accurate modeling. In the case of modeling finding predictor genes is important. Sometimes we can obtain this knowledge by accomplishing some biological experiments that are usually expensive. Therefore using data mining methods can help us for obtaining this information.

Acknowledgment

This work is supported by Iran Telecommunication Research Center (ITRC).

TABLE III
CoD FOR FNCA

No. parent genes	CoD. G2	CoD. G10	CoD. G11	CoD. G12
MCMC	0.607	0.385	0.804	0.608
MINCA	0.958	0.949	0.956	0.950
FNCA	0.516	0.715	0.668	1

REFERENCES

[1] X. Zhoua, X Wangb, E. Doughertya; ”Construction of genomic networks using mutual-Information clustering and reversible-jump Markov-chain-Monte-Carlo predictor design,” Signal Processing, vol. 83, pp. 745 - 761, 2003

[2] P. Du, J. Gong, E.S. Wurtele, and J.A. Dickerson, ”Modeling Gene Expression Networks Using Fuzzy Logic” IEEE transactions on system,man,and cybernetics-partB:cybernetics,vol. 35, no.6, pp 1351-1359, december 2005.

[3] I. Shwulevich, E.R. Dougherty, S. Kim and W. Zhang, ”Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks,” Bioinformatics, vol.18. no.2, pp. 261-274, 2002

[4] A. Kraskov, H. Stogbauer, R. G. Andrzejak, and P. Grassberger, ” Hierarchical Clustering Using Mutual Information ”,2003.

[5] X, Zhou, X. Wang, E.R. Dougherty, D. Russ, and E. Suh,” Gene Clustering Based on Clusterwide Mutual Information,” Journal of computational biology, Volume 11, Mary Ann Liebert, Inc. Pp. 147-161, Number 1, 2004.

[6] N. Belacel, M. Cuperlovic-Culf, R. Ouellette, and M. Boulassel, ”The Variable Neighborhood Search Metaheuristic for Fuzzy Clustering cDNA Microarray Gene Expression Data,” Proceedings of IASTED-AIA-04 Conference. Innsbruck, Austria, February 16-18, 2004.

[7] S. Kim, E. R. Dougherty, M. L. Bittner, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, ”General nonlinear framework for the analysis of gene interaction via multivariate expression arrays,” Journal of Biomedical Optics, vol.5, pp. 411-424, October 2000

[8] E. R. Dougherty, S. Kim, Y. Chen, ”Coefficient of determination in nonlinear signal processing,” Signal Processing, vol. 80, pp. 2219-2235, 2000.

[9] J. C. Liao, R. Boscolo, Y.L. Yang, L. M. Tran, C. Sabatti and V. P. Roychowdhury, ”Network Component Analysis: Reconstruction of Regulatory Signals in Biological Systems ”, PNAS, vol. 100, no.26, pp. 15522-15527, December 2003.

[10] F. Bakouie, MH. Moradi, ” application of Variable Neighborhood Search (VNS) methods in microarray data clustering,” ICEE2006, 14th Conference on Electrical Engineering, Iran, Tehran, Amirkabir University of technology.

[11] R.Boscolo, C. Sabatti, J. C. Liao and V. P. Roychowdhury, ” A Generalized Framework for Network Component Analysis ” IEEE Transaction. ON Computational Biology and Bioinformatics,vol. 2, no.4, October-December 2005.