RECONSTRUCTION OF GENETIC REGULATORY NETWORKS BASED ON THE POSTERIOR PROBABILITIES OF GENE REGULATIONS

Wentao Zhao[1]*, Kwadwo Agyepong*[1]*,Erchin Serpedin*[1]*, and Edward R. Dougherty*[1,2]

[1]Texas A&M University
Dept. of Electrical and Computer Engineering
College Station, Texas, 77843-3128

[2]Translational Genomics Research Institute
400 North Fifth Street, Suite 1600
Phoenix, Arizona 85004

ABSTRACT

Recent advances in high throughput microarray data have enabled the learning of the structure and operation of gene regulatory networks. This paper proposes a novel approach for reconstruction of gene regulatory networks based on the posterior probabilities of gene regulations. Built within the framework of Bayesian statistics and exploiting efficient computational Monte Carlo techniques, the proposed approach prevents the dichotomy of classifying gene interactions as either being connected or disconnected, and thereby it reduces significantly the inference errors. Simulation results corroborate the superior performance of the proposed approach relative to the existing state-of-the-art algorithms.

Index Terms— Monte Carlo Methods, Genetics, Biological Systems

1. INTRODUCTION

Currently, one of the most important research problems in molecular biology and bioinformatics consists of finding out the mechanisms that lie at the basis of gene regulatory networks. The importance of gene regulatory networks is due to their fundamental role in the control and operation of all the processes taking place in the living cell. Therefore, learning the structure and operation of gene regulatory networks opens up the possibility for understanding and controlling the functioning of organisms at the molecular level, and for designing intelligent therapies and drugs.

Gene regulatory networks have been employed to model the multivariate gene interactions in systems biology. Recent years have witnessed a number of different frameworks for gene regulatory network modeling, ranging from fine-scale modeling of biological interactions at the molecular level (using partial differential equations and stochastic equations) to large scale modeling at the gene and protein-level (Boolean and probabilistic Boolean networks, and (dynamic) Bayesian networks) [1] - [6]. The small scale modeling techniques are used to obtain a detailed biochemical description of molecular interactions and are in general very computational demanding. On the other side, the large scale models provide a global view of the interactions among the constituent elements of gene regulatory networks and are generally represented in terms of graphs. For example, gene regulatory networks can account for the rate at which genes, i.e., DNA segments, are transcribed into mRNA with the involvement of other genes, which can be either activators or repressors.

To obtain a correct description of the interactions between genes, it is necessary to design metrics for assessing not only the direct connectivity but also the regulating orientations. There are two types of DNA microarray data sets: time series (or time dependent) and time independent (also called steady state or single point time series) data sets. In general, the time independent gene expression profiles are capable of recovering steady state attractors, but fail to recover the direct and oriented (temporal regulating) relationships. On the other side, time series data sets can improve the inference greatly in contrast to time independent data sets [7]. However, the formidable cost is one major factor in collecting time series data. Recent statistics show that about 70% of published data are time independent data [8]. Therefore, the steady state analysis is highly valuable despite the difficulty of making accurate inference of temporal relationships.

This paper proposes a Bayesian approach to analyze the steady state data and establishes a confidence measure of gene interactions. The proposed scheme possesses five key features which make it different from the existing algorithms. First, most of the current schemes infer a unique genetic network represented by a graph which best fits the observed data in a certain metric, while the proposed approach determines the posterior probabilities for all gene-pair interactions and avoids to make a dichotomy decision, i.e., to classify each gene interaction as either being connected or disconnected. The proposed gene reconstruction approach can be easily transformed into a dichotomy scheme by only preserving the highly probable gene interactions. Second, the underlying structural model is assumed to be a directed cyclic graph, which allows cycles (feedback loops) and directed acyclic graphs are

This work was supported by the National Cancer Institute (CA-90301) and the National Science Foundation (ECS-0355227 and CCF-0514644).

treated as special cases. Third, the proposed approach assumes continuous-valued variables, and this prevents the information loss incurred by data quantization. This represents an extension of the discrete-valued Bayesian networks [9, 10]. Fourth, the proposed connectivity score is oriented and has a very clear meaning, in the sense of posterior probabilities, while the existing scores based on the mutual information [11] are vague and lack orientation information. Fifth, in the present approach the system kinetics can be assumed to be nonlinear, while linear models are commonly utilized for the purpose of simplification [12]. The proposed scheme establishes a general framework whose components can be customized to fit the nature of the underlying biological system.

2. SYSTEM FORMULATION

Genetic regulatory networks can be represented by a parameterized graph $(\mathbf{G}, \mathbf{\Theta})$, where $\mathbf{G}$ and $\mathbf{\Theta}$ stand for the graph structure and parameter set, respectively. The graph structure qualitatively explains the direct gene interactions, while the parameter set quantitatively describes the system kinetics.

2.1. Structural Model

The graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ is employed to map gene interactions at transcriptional level, where $\mathbf{V}$ denotes the set of vertices (genes) and $\mathbf{E}$ stands for the set of edges (regulation relationships). If gene X regulates gene Y, graphically such a relation is represented in terms of an oriented edge $X \rightarrow Y$, where X is a parent of Y and Y is considered a child of X. All genes that present incidence edges with gene X represent the set of parent genes of X, and are compactly denoted in terms of the notation Π_X. If two genes X and Y interact with each other but the regulation orientation can not be determined, a disoriented edge is laid between the two genes (as $X - Y$). A sequence of consecutive oriented edges represents a directed path. If there is no directed path which starts and ends at the same vertex, in other words the graph contains no loops, the graph is called a directed acyclic graph (DAG). DAGs lie at the basis of Bayesian networks, which are commonly employed to model causal relationships.

General directed graphs (with possibly cycles) will serve as our structural model since they are consistent with the features exhibited by biological systems, in which loops account for system redundancy and stability. Given the graph structure $\mathbf{G}$, the parent set Π_X is specified for any gene X. Next we discuss the system kinetics parameters defined in $\mathbf{\Theta}$.

2.2. System Kinetics

The system kinetics represents the dynamics that governs the gene's mRNA concentrations in terms of gene-gene interactions. It can be modeled by a set of differential equations (DE). A simplified form is a set of linear DEs. However, we accept the more complicated form which were employed by [13, 14] since it is much more realistic and accounts for the expression saturation. Given a gene X, its parent set Π_X can be further partitioned into two disjoint subsets: the activator set A_X and the repressor set R_X, i.e., $\Pi_X = A_X \cup R_X$ and $A_X \cap R_X = \phi$. The kinetics of gene X can be explained by the following differential equation:

$$\frac{dx}{dt} = -\lambda x + \frac{\delta + \sum_{i=1}^{|A|} a_i^{\alpha_i}}{1 + \sum_{i=1}^{|A|} a_i^{\alpha_i} + \sum_{j=1}^{|R|} r_j^{\gamma_j}} \qquad (1)$$

where x is the concentration of gene X's product, namely, mRNA. The change rate of gene X is controlled by its activating and repressing parents, denoted individually by $a_i \in A$ and $r_j \in R$. α and γ serve as the regulating factors corresponding to each activator and repressor. α and γ assume positive values, and hence can be modeled by Gamma distribution with shape and scale parameters (κ, β). λ stands for the gene degradation rate and the time scale can be properly chosen in order to normalize $\lambda = 1$. δ represents the expression baseline rate, i.e., the expression rate when there is neither activator nor repressor regulating the target gene X. Suppose y represents the observation of x, then y has the form $y = x + \varepsilon$, where ε incorporates all sources of noise and is modeled by a Gaussian random variable with mean and variance $(0, \sigma^2)$.

In general, biological systems always converge to some steady states. In a steady state, all genes stay in equilibrium and do not change their expressions. By setting $dx/dt = 0$ and $\lambda = 1$, the observation y of the steady-state gene expression for gene X can be expressed as:

$$y = \frac{\delta + \sum_{i=1}^{|A|} a_i^{\alpha_i}}{1 + \sum_{i=1}^{|A|} a_i^{\alpha_i} + \sum_{j=1}^{|R|} r_j^{\gamma_j}} + \varepsilon \qquad (2)$$

Given a parent structure Π_X for gene X, the parameters in Θ_X can be summarized as follows:
1) For each parent $\pi \in \Pi_X$, a binary variable is demanded to specify whether the parent is an activator or repressor. That is $\mathbf{1}_{A_X}(\pi)$, where $\mathbf{1}$ is the indicator function and it assumes the value 1 when $\pi \in A_X$, and 0 otherwise. It can be modeled by a Bernoulli variable with known success probability ρ.
2) For each activator $a \in A_X$ and repressor $r \in R_X$, it is assumed that the regulating factors $\alpha, \gamma \sim Gamma(\kappa, \beta)$, where κ, β are known.
3) The baseline parameter δ is usually known.
4) The noise $\varepsilon \sim N(0, \sigma^2)$, where σ^2 can be set to a specific value or estimated.

It is worth to note that the choice of the nonlinear equation and parameter priors does not influence the flow of analysis. Our scheme stands for a particular framework and the detailed parameters can be easily customized to other scenarios.

3. INFERENCE METHOD

Consider a system composed of n genes $\{X_1, X_2, \cdots, X_n\}$, and assume that m observations of expression vector are ob-

tained and stored in matrix $D^{n\times m}$. Next, we develop a computational approach to establish the posterior probability of the regulation $X_i \rightarrow X_j$, i.e., the probability of the existence of the edge e_{ij}, which is represented by $p(e_{ij}|D)$. This posterior can be obtained through integrating over the whole parent gene set and parameter space for gene X_j:

$$\begin{aligned} p(e_{ij}|D) &= \sum_{\Pi_j} \int_{\Theta_j} p(e_{ij}, \Pi_j, \Theta_j|D) d\Theta_j \\ &= \sum_{\Pi_j} \int_{\Theta_j} \mathbf{1}_{\Pi_j}(i) p(\Pi_j, \Theta_j|D) d\Theta_j \end{aligned} \quad (3)$$

Applying Bayes theorem, $p(\Pi_j, \Theta_j|D)$ can be expressed as

$$\begin{aligned} p(\Pi_j, \Theta_j|D) &= \frac{p(D|\Pi_j, \Theta_j) p(\Pi_j, \Theta_j)}{p(D)} \\ &= \frac{p(D|\Pi_j, \Theta_j) p(\Pi_j, \Theta_j)}{\sum_{\Pi_j} \int_{\Theta_j} p(D|\Pi_j, \Theta_j) p(\Pi_j, \Theta_j) d\Theta_j} \\ &= \frac{p(D_j|D_{\bar{j}}, \Pi_j, \Theta_j) p(\Pi_j, \Theta_j)}{\sum_{\Pi_j} \int_{\Theta_j} p(D_j|D_{\bar{j}}, \Pi_j, \Theta_j) p(\Pi_j, \Theta_j) d\Theta_j} \end{aligned} \quad (4)$$

where D_j denotes the observations of gene X_j, and $D_{\bar{j}}$ represents the collection of all the observations pertaining to all genes excluding those of gene X_j. $p(\Pi_j, \Theta_j)$ denotes the probability density of the high-dimensional parental model which is a subgraph of the whole network, and $p(D_j|D_{\bar{j}}, \Pi_j, \Theta_j)$ stands for the data likelihood given the parental values and the graphical model. It is a Gaussian distribution with mean determined by the first part of equation (2) and known variance. By plugging (4) into (3), it can be inferred that

$$p(e_{ij}|D) = \frac{\sum_{\Pi_j} \int_{\Theta_j} \mathbf{1}_{\Pi_j}(i) p(D_j|D_{\bar{j}}, \Pi_j, \Theta_j) p(\Pi_j, \Theta_j) d\Theta_j}{\sum_{\Pi_j} \int_{\Theta_j} p(D_j|D_{\bar{j}}, \Pi_j, \Theta_j) p(\Pi_j, \Theta_j) d\Theta_j} \quad (5)$$

The integrations at the numerator and denominator of (5) can not be generally performed in a closed-form expression. However, the Monte Carlo methods enable to numerically evaluate the posterior probabilities. We can generate Monte Carlo samples based on the model probability density $p(\Pi, \Theta)$ and the integration can be obtained by averaging over these samples. Then the posterior probabilities can be estimated by

$$p(e_{ij}|D) \approx \frac{\sum_{\Pi_j, \Theta_j} \mathbf{1}_{\Pi_j}(i) p(D_j|D_{\bar{j}}, \Pi_j, \Theta_j)}{\sum_{\Pi_j, \Theta_j} p(D_j|D_{\bar{j}}, \Pi_j, \Theta_j)} \quad (6)$$

Assuming that the selection of a parent as an activator is performed in an independent manner, and that the selection of the regulation factor value is also performed independently, the model probability density $p(\Pi, \Theta)$ can be further expanded by using the chain rule:

$$\begin{aligned} p(\Pi, \Theta) &= p(\Theta|\Pi) p(\Pi) \\ &= \prod_{i=1}^{|A|} [\rho p(\alpha_i)] \prod_{j=1}^{|R|} [(1-\rho) p(\gamma_j)] p(\Pi) \end{aligned} \quad (7)$$

Algorithm 1 Inference of Connectivity Significance

1: Input gene expression data set $D^{n\times m}$ with n genes and m samples;
2: Initialize $n, \mathbf{L} = 0^{1\times n}, \mathbf{C} = 0^{n\times n}$;
3: **for** $k = 1$ to M **do**
4: Randomly create a directed graph and the adjacency matrix J;
5: **for** $i = 1$ to n **do**
6: For gene i's parents specified in $J(:, i)$, randomly assign them to be activators or repressors;
7: For each parent, randomly create their regulation factor α or γ;
8: $l \Leftarrow \text{likelihood}(D_i|D_{\bar{i}}, \Pi_i, \Theta_i)$;
9: **for** $j = 1$ to n **do**
10: **if** $j \in \Pi_i$ **then**
11: $\mathbf{C}_{ji} = \mathbf{C}_{ji} + l$;
12: **end if**
13: **end for**
14: $L_i = L_i + l$;
15: **end for**
16: **end for**
17: $\forall i, j, C_{ji} = C_{ji}/L_i$;
18: **return** $\mathbf{C}$.

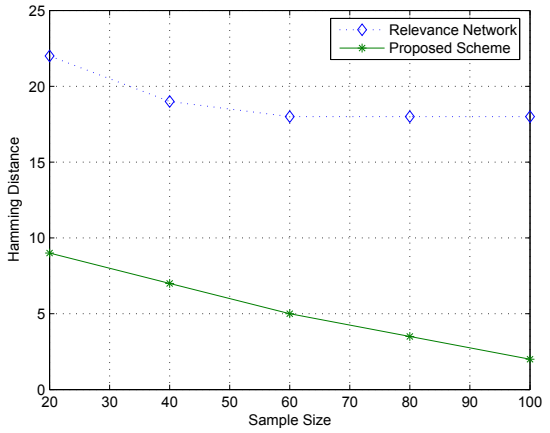
Equation (7) conveys the idea that the random samples of graphical models can be sequentially created and processed. First the network structure is created, then each parent is randomly assigned to represent an activator or repressor, and finally regulation factors are generated. Instead of separately creating parents for each node in a random way, random graphs are generated. The motivation is that generally we have some prior knowledge about the underlying graph, e.g., the sparsity of the graph, the statistic of the edges, whether the underlying graph is acyclic or not, etc. Such an approach enables to utilize informative priors for graphs rather than parental structures. In order to create multi-dimensional graphical models, we employ Markov chain Monte Carlo (MCMC) and sequential importance sampling (SIS) techniques. However, the details of this construction are omitted due to the space limitations. Therefore, our computational procedure can be briefly formulated in terms of the Algorithm 1, where the conventions from Matlab were used to write the pseudo-code, the output entry C_{ij} stands for $p(e_{ij}|D)$, and M denotes the number of Monte Carlo iterations.

4. SIMULATION RESULTS

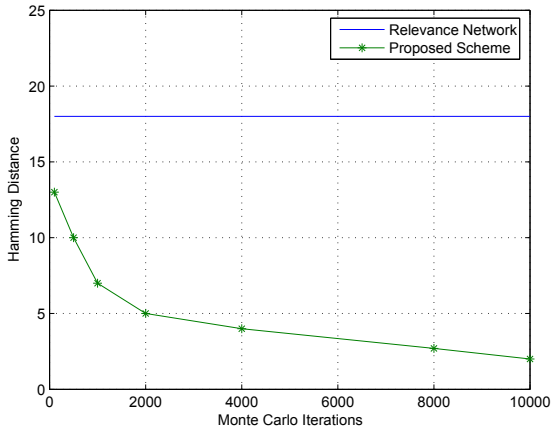
The performance of the proposed algorithm is next compared with one of the most representative algorithms available in the literature, namely the relevance network (RN) method [11]. In RN, the significance of gene interactions is measured in terms of the mutual information between the gene expressions. Hence, the RN is an undirected cyclic graph.

An artificial network which contains 10 vertices and 18 oriented edges was first created. Different numbers of steady-state samples were generated based on the adopted model. For both the RN method and the proposed scheme, the top 18 edges with highest scores (mutual information for RN and posterior probability for proposed scheme) were preserved. As a performance metric, the Hamming distance is used to compare the adjacency matrix of the inferred network with that of the original artificial network. Since RN is disoriented, we have to disregard the orientation information of the network identified by the proposed scheme.

Figure 1(A) compares the performance of RN method and the proposed scheme for different sample sizes. It is obvious that the proposed method provides much better inference accuracy. For small scale artificial networks, when the sample size is sufficiently large, the increase of sample size does not improve the performance of RN method. This is because the estimation of mutual information has been sufficiently accurate. The shortcoming of mutual information is apparent, and therefore it is not an appropriate metric for establishing the direct connectivity between genes. Figure 1(B) illustrates the impact of different number of Monte Carlo iterations on Hamming metric. More iterations surely improve the performance of the proposed scheme, while a mild number of iterations already guarantees the proposed scheme to outperform the RN method. When considering the orientation of the edges, we find that more than 90% of the inferred edges are correctly oriented by the proposed approach.



(A) Different Sample Sizes.



(B) Different Monte Carlo Iterations

Fig. 1. Simulation results. (a) The Monte Carlo iterations are fixed at 10,000. (b) The sampled data size is fixed at 100.

5. REFERENCES

[1] S.A. Kauffman, "Metabolic stability and epigenesist in randomly constructed genetic nets," *Theor. Biol.*, vol. 22, pp. 437–467, 1969.

[2] K. Murphy and S. Mia, "Modelling gene expression data using dynamic Bayesian networks," *Tech. Report* 1999, Berkeley Univ.

[3] P. Sebastiani, et al., "Bayesian Networks," *The Data Mining and Knowledge Discovery Handbook*, pp. 193–230, 2005.

[4] I. Shmulevich, et al., "Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks," *Bioinformatics*, vol. 18(2), pp. 261–274, 2002.

[5] I. Tabus and J. Astola, "Using MDL for Gene Expression Prediction from Microarray Measurements," *NSIP'01*, Baltimore, MD, June 2001.

[6] O. Yli-Harja, D. Nicorici, J. Astola et al., "A computational model for simulating continuous time Boolean networks," in em IEEE GENSIPS, 2004.

[7] W. Zhao, E. Serpedin, and E.R. Dougherty, "Inferring gene regulatory networks from time series data using the minimum description length principle," *Bioinformatics*, vol. 22, pp. 2129–2135, 2006.

[8] I. Simon, et al., "Combined static and dynamic analysis for determining the quality of time-series expression profiles," *Nature Biotechnology*, vol. 23, pp. 1503–1508, 2005.

[9] G.F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol.9, pp. 309–347,1992.

[10] D. Heckerman, D. Geiger, and D. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20(3), pp.197–243, 1995.

[11] A.J. Butte and I.S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pac. Symp. Biocomput.* 2000. pp. 418–429.

[12] S. Rogers, M. Girolami, "A Bayesian regression approach to the inference of regulatory networks from gene expression data," *Bioinformatics*, vol. 21(14), pp. 3131–3137, 2005

[13] J.J. Rice, et al., "Reconstructing biological networks using conditional correlation analysis," *Bioinformatics*, vol. 21(6), pp. 765–773, 2005.

[14] M.K.S. Yeung, et al., "Reverse engineering gene networks using singular value decomposition and robust," *Proc. Natl. Acad. Sci U.S.A.*, vol. 99(9), pp. 6163–6168, 2002.