Constructing Time-Varying Gene Regulatory Networks

Xiaohong Ni1 Yingfei Sun2

School of Electtonic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, 100049, P. R. of China.
1e-mail: nixh10b@mails.ucas.ac.cn 2e-mail: yfsun@ucas.ac.cn

Abstract

Discovered gene regulation networks are very helpful to understand biological mechanisms and predict unknown gene functions. Little literature has been done toward estimating time-varying networks from time series data. In this paper, we develop a new method to discover the dynamic, and delay interactions between genes based on gene expression data. The method builds on time lagged mutual information and a temporally smoothed $l_1$-regularized logistic regression. To evaluate our method, we use both simulated and biological data from the yeast cell cycle in regulatory networks modeling. Results demonstrate the superior performance of our method.

Keywords: gene regulatory networks; mutual information; $l_1$-regularization.

1. Introduction

Most methods of graph inference work very well on stationary time-series data, in that the generating structure for the time series does not exhibit switching[1]. Time-Varying networks have advantage over conventional methods in applications where the intrinsic regulatory networks are subject to changes for adapting to internal or external stimuli[2-3]. Recent work on Time-Varying networks could be found in [4-5]. Ahmed[4] address the case of discrete time-varying graphs modeled as evolving Markov random fields. One related approach is the Keller algorithm by Song[5]. However, Keller aims at recovering bursty rather than smoothly varying networks.

There are several limitations on the existing methods that are discussed above. First, they don’t consider an essential factor - time lag. The concept of time lags was first introduced by Zou [6]. Applying time-delayed property is important to model relations between genes [7]. Second, they construct time varying undirected graphs. It is difficult to infer functional relationships between interacting genes.

In this paper, we propose a new Time-Varying directed networks approach, which extends the work presented in [4]. Our method modifies $l_1$-regularized logistic regression by using $l_1$ variable selection to identify each variable’s Markov blanket. In addition, we design a lag determine mechanism by considering time lagged Mutual Information between genes. In this approach we efficiently reduce the model optimizing space, capture the dynamic, causal and delay interactions between genes based on their time series of expression values.

2. Methodology

2.1. Time Lagged Mutual Information(TLMI)

If gene $y$ regulates gene $x$ at time slot
with a latency $\tau$, $x_{t+\tau}$ has to depend on $y_t$. Given $m$-sample time series $x$ and $y$, we compute TLMI between gene $y$ and gene $x$ as following:

$$M_t^\tau(x, y) = \frac{1}{m-\tau} \sum_{t=\tau}^{m-1} I_{(j)}(x_t)$$

where $P(x_{t+\tau}, y_t)$ and $P(x_{t+\tau}, y_t)$ are cross time joint and marginal probabilities. We quantize the gene expression levels into two levels, 0 for downregulated and 1 for upregulated. The marginal and joint probabilities are given by the following equation:

$$P(x_t = j) = \frac{1}{m-\tau} \sum_{i=1}^{m-\tau} I_{(j)}(x_t)$$

where $I_{(j)}(\cdot)$ stands for the indicator function and is defined as follows:

$$I_{(j)}(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases}$$

Large TLMI supports the proposition that $y$ regulates $x$ in $\tau$ time point with a high probability[8]. Therefore, time dependent TLMI is calculated for each expression profile obtained by shifting genes by one time step till the defined maximum time delay is reached. After this we introduce the max lag over all possible delays $\tau$:

$$\tau_{\max} = \max\{M_t^\tau(x, y)\}, \quad \tau = 1, \ldots, \frac{m}{3}$$

2.2. Modeling time series of gene expression

Denote by $G' = (V, E')$ the graph structure at time $t$ with gene set $V$ of size $|V| = p$ and edge set $E'$. At a particular time point $t$, we denote the microarray measurements for $p$ genes as a vector of random variables $X^{(t)} \in \{0,1\}^p$. Given time lag $\tau$, we consider the random variables generated under the Ising model:

$$P(x^{(t)}|\theta^{(t)}, \tau) = \exp\sum_{i=t}^{t+\tau} \theta_i x^{(t)} \cdot A(\theta^{(t)})$$

where the parameters $\{\theta_0^{(t)}\}_{i,j \in E}$ capture the regulation strength between genes $x_i^{(t)}$ and $x_j^{(t)} \cdot A(\theta^{(t)})$ is the normalization constant. Here, we want to apply the pseudo-likelihood approach of Besag[9] to our problem. We use $\nabla$ to denote the set of genes except gene $i$, i.e. $\nabla = \nabla - i$. The log-pseudo-likelihood is then given by:

$$\ell(x^{(t)}_i, \theta^{(t)}_i) = \log P(x^{(t)}_i | x^{(t-1)}_1, \ldots, x^{(t-\tau)}_i, \theta^{(t)}_i)$$

Where

$$P(x^{(t)}_i | x^{(t-\tau)}_1, \ldots, x^{(t-\tau)}_i, \theta^{(t)}_i) = \log[1 + \exp(x^{(t-\tau)}_1 \theta^{(t)}_1 + \ldots + x^{(t-\tau)}_i \theta^{(t)}_i)]$$

Given $n$ samples $\{X^{(t)}\}_{i=1}^n$ drawn from an unknown distribution $P(x^{(t)} | \theta^{(t)}, \tau)$ of the form Eq(5). The goal is to estimate the structure of the graph, i.e. to estimate $\{E^{(t)}\}_{i=1}^n$. We decouple this problem into a set of $p$ separate smoothed $l_1$-regularized logistic regression problems, one for each variable. For each node $i$ in the graph, we solve the following problem:

$$\hat{\theta}^{(t)}_i, \ldots, \hat{\theta}^{(t-n+\tau)}_i = \arg \min_{\theta^{(t-\tau)}_i, \ldots, \theta^{(t-n+\tau)}_i} \sum_{t=\tau}^{t+n-\tau} \ell(x^{(t)}_i, \theta^{(t)}_i) + \lambda \sum_{t=\tau}^{t+n-\tau} \|\theta^{(t)}_i - \theta^{(t-1)}_i\|$$

where $\theta^{(t)}_i = \{\theta^{(t)}_j | j \in \nabla_i\}$ is the sub-vector of parameters associated with gene $i$, and the regularization parameter $\lambda$ controls the number of nonzero el-
ments. The second term $\lambda_2$ penalizes the difference between parameters that are adjacent in time and, as a result, the estimated parameters have infrequent changes across time. A good choice of the tuning parameters is essential in obtaining a good estimator that does not overfit the data, and balances between the pseudo-likelihood and the degrees of freedom. We adopt a criteria for selecting the hyperparameters using the BIC score (See for Ahmed[4] details). In order to accelerate the optimization, we use l1 variable selection. This process finds a set that is as small as possible, but that contains all of gene $i$’s parents, children and co-parents (its Markov blanket). This pre-processing yields a small set of potential parents that are then used to restrict the edges when optimizing regression model. The directed edges $\hat{E}'$ can be recovered via

$$\hat{E}' = \{(i, j) \in V \times V | i \neq j, |\hat{\theta}_{ji}'| \neq 0\}. \quad (8)$$

3. Experiments

3.1. Recovering synthetic networks

Based on a simple model as Eq(9), we simulated 20 time points of rewiring networks and their corresponding node states.

$$\begin{align*}
time & = 1 \rightarrow 2 \rightarrow \cdots \rightarrow 10 \rightarrow 11 \cdots \rightarrow 20 \\
x_i(t) &= f(t); \\
x_i(t) &= a_1 x_i(t-1) + \xi_1; a_2 x_i(t-1) + \xi_2; \\
x_i(t) &= b_1 x_i(t-1) + \eta_1; b_2 x_i(t-1) + \eta_2; \\
\end{align*} \quad (9)$$

In the first 10 discrete time points, the network have two edges: $x_3 \leftarrow x_1 \rightarrow x_2$. After $t=11$, 1 existing edge is deleted, 1 new edge is added and the final network with two edges: $x_1 \rightarrow x_2 \rightarrow x_3$. When the true network is known, this allows us to construct the Receiver Operating Characteristic (ROC) curve, and to assess the network reconstruction accuracy in terms of the areas under these graphs (AUROC). The result is given in Table1 (we average AUROC from $G^1$ to $G^9$, and from $G^{10}$ to $G^{19}$, lag=1).

Table 1 AUROC values on synthetic data:

<table>
<thead>
<tr>
<th></th>
<th>time 2:10</th>
<th>time 11:20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average AUROC</td>
<td>1</td>
<td>0.929</td>
</tr>
</tbody>
</table>

3.2. Recovering time-evolving interactions between genes in yeast cell

We next apply our method to uncover networks describing the regulation of transcription during the cell cycle in yeast[10]. we extract a set of 25 genes previously used in [11], and use alpha time series subset (with 18 time points) in this analysis. The gold standard comes from a compiled list of evidence in the literature and from the Saccharomyces Genome Database[12]. In Table 2, we compare the experimental results of three approaches: dynamic Bayesian networks (DBN) model, DBN model by combining prior knowledge [11], and our method on yeast cell data.

Table 2 Comparison of the experimental results of three approaches with the gold standard network. TP, true positive; FP, false positive; TN, true negative; FN, false negative. Sensitivity = TP/(TP+FN); Specificity = TN/(TN+FP); Accuracy= (TP+TN)/(TP+FP+TN+FN).

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>Sp</th>
<th>Se</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBN</td>
<td>7</td>
<td>20</td>
<td>0.9005</td>
<td>0.1795</td>
<td>0.7833</td>
</tr>
<tr>
<td>DBN + Prior</td>
<td>14</td>
<td>64</td>
<td>0.8762</td>
<td>0.1687</td>
<td>0.7783</td>
</tr>
<tr>
<td>Our method</td>
<td>17</td>
<td>27</td>
<td>0.9499</td>
<td>0.1977</td>
<td>0.8464</td>
</tr>
</tbody>
</table>

4. Conclusion

In this paper, we propose time-varying directed networks for modeling the varying network structures underlying non-
stationary biological time series, we design a lag determine mechanism by considering TLMI between genes. By using l1 variable selection to identify each variable’s Markov blanket, we efficiently reduce the model optimizing space, capture the dynamic interactions between genes based on their time series of expression values. Given the rapid advances in data collection technologies for biological systems, we expect that complex, high-dimensional, and feature rich data from complex dynamic biological processes, such as cancer progression, immune responses, and developmental processes, will continue to grow. Thus, we believe our new method is a timely contribution that can narrow the gap between imminent methodological needs and the available data and offer deeper understanding of the mechanisms and processes underlying biological networks.

5. References


