J. Biomedical Science and Engineering, 2008, 1, 15-21

Scientific Research Publishing

Construction and control of genetic regulatory networks:a multivariate Markov chain approach

Shu-Qin Zhang[1], Wai-Ki Ching[2], Yue Jiao[2], Ling-Yun Wu[3] &Raymond H. Chan[4]

[1]School of Mathematical Sciences, Fudan University, Shanghai, 200433, China. [2]Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong. [3]Institute of Applied Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences. [4]Department of Mathematics, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong. *Correspondence should be addressed to Shu-Qin Zhang (zhangs@fudan.edu.cn).

ABSTRACT

In the post-genomic era, the construction and control of genetic regulatory networks using gene expression data is a hot research topic. Boolean networks (BNs) and its extension Probabilistic Boolean Networks (PBNs) have been served as an effective tool for this purpose. However, PBNs are difficult to be used in practice when the number of genes is large because of the huge computational cost.In this paper, we propose a simplified multivariate Markov model for approximating a PBN. The new model can preserve the strength of PBNs, the ability to capture the inter-dependence of the genes in the network, and at the same time reduce the complexity of the network and therefore the computational cost. We then present an optimal control model with hard constraints for the purpose of control/intervention of a genetic regulatory network. Numerical experimental examples based on the yeast data are given to demonstrate the effectiveness of our proposed model and control policy.

Keywords: Gene expression sequences; Multivariate Markov chain; Optimal control policy; Probabilistic Boolean networks

1. INTRODUCTION

An important issue in systems biology is to understand the mechanism in which cells execute and control a huge number of operations for normal functions, and also the way in which the cellular systems fail in disease, eventually to design some control strategy to avoid the undesirable state/situation. Many mathematical models such as neural networks, linear model, Bayesian networks, non-linear ordinary differential equations, Petri nets, Boolean Networks (BNs) and its generalization Probabilistic Boolean Networks (PBNs), multivariate Markov chain model etc. [1,2,4,11,15,16,17,21] have been proposed. Among all the models, BNs and PBNs have received much attention. The approach is to model the genetic regulatory system by a Boolean network and infer the network structure from real gene expression data. Then by using the inferred network model, the underlying gene regulatory mechanisms can be uncovered. This is particularly useful as it helps to make useful predictions by computer simulations. We refer readers to the survey paper by Shmulevich *et al.* [18, 19] and the book by Shmulevich and Dougherty [20].

The BN model was first introduced by Kauffman [12, 13, 14]. The advantages of this model can be found in Akutsu *et al.* [1], Kauffman [14] and Shmulevich *et al.* [17]. Since genes exhibit switching behavior [10], BN models have received much attention. In a BN, each gene is regarded as a vertex of the network and is quantized into two levels only (expressed (1) or unexpressed (0)). We remark that the idea and the model can be extended easily to the case of more than two states. The target gene is predicted by several genes called its input genes through a Boolean function. If the input genes and the Boolean functions are given, a BN is defined. The only randomness involved here is the initial system state. However, the biological system has its stochastic nature and the microarray data sets used to infer the network structure are usually not accurate because of the experimental noise in the complex measurement process. Thus stochastic models are more reasonable choices. To overcome the deterministic nature of a BN, Akutsu *et al.* [1] proposed the noisy Boolean networks together with an identification algorithm. In their model, they relax the requirement of consistency imposed by the Boolean functions. Regarding the effectiveness of a Boolean formalism, Shmulevich *et al.* [17] proposed a PBN that can share the appealing rule-based properties of Boolean networks and it is robust in the presence of uncertainty. The model parameters can be estimated by using Coefficient of Determination (COD) [8].

The dynamics of the PBN can be studied in the context of standard Markov chain [17, 18, 19]. This makes the analysis of the network easy. However, the number of parameters (state of the system) grows exponentially with respect to the number of genes n. Therefore it is natural to develop heuristic methods for model training or to consider other approximate model. Here we propose a simplified multivariate Markov model, which can capture both the intra- and inter-associations (transition probabilities) among the gene expression sequences. The number of parameters in the model is only $O(n^2)$ where n is the number of genes in a captured network. We remark that this order is already minimal. We then develop efficient model parameters estimation methods based on linear programming. We further propose an optimal control formulation for regulating the network so as to avoid some undesirable states which may correspond to some disease like cancer.

The rest of the paper is structured as follows. In section 2, we present the simplified multivariate Markov model. In section 3, the estimation method for model parameters is given. In section 4, an optimal control formulation is proposed. In section 5, we apply the proposed model and method to some synthetic examples and also the gene expression dataset of yeast. Concluding remarks are then given to address further research issues in section 6.

2. THE MULTIVARIATE MARKOV CHAIN MODEL

In this section, we first review a multivariate Markov chain model proposed in Ching, *et al.* [3] for modeling categorical time series data. We remark that the model has been first applied to predicting demand of inventory of correlated products. Later the model was applied to the building of genetic regulatory networks [4] from gene expression data. However, the number of parameters is still large and further reduction of the model parameters is necessary and a simplified model was proposed in [5]. In the remainder of this section, we present the simplified multivariate Markov chain model.

Given n categorical time sequences, we assume they share the same state space M. We denote the state probability distribution of Sequence j at time t by $\boldsymbol{V}_t^{(j)}$, j=1,2,$\cdots$,n . In Ching, *et al.* [3], the following first-order model was proposed to model the relationships among the sequences:

$$\mathbf{V}_{t+1}^{(i)} = \sum_{j=1}^{n} \lambda_{ij} P^{(ij)} \mathbf{V}_t^{(j)}, \quad i = 1,2,...,n \qquad (1)$$

Where

$$\lambda_{ij} \geq 0 \quad \text{for} \quad 1 \leq i, j \leq n \quad \text{and} \quad \sum_{j=1}^{n} \lambda_{ij} = 1. \qquad (2)$$

Here λ_{ij} is the non-negative weighting of Gene j to Gene i. The matrix $P^{(ij)}$ is a transition probability matrix for the transitions of states in Sequence j to states in Sequence i in one step, see for instance [3]. In matrix form we have

$$\mathbf{V}_{t+1} \equiv \begin{pmatrix} \mathbf{V}_{t+1}^{(1)} \\ \mathbf{V}_{t+1}^{(2)} \\ \vdots \\ \mathbf{V}_{t+1}^{(n)} \end{pmatrix} = Q \begin{pmatrix} \mathbf{V}_t^{(1)} \\ \mathbf{V}_t^{(2)} \\ \vdots \\ \mathbf{V}_t^{(n)} \end{pmatrix} \equiv Q\mathbf{V}_t$$

where

$$Q = \begin{pmatrix} \lambda_{11}P^{(11)} & \lambda_{12}P^{(12)} & \cdots & \lambda_{1n}P^{(1n)} \\ \lambda_{21}P^{(21)} & \lambda_{22}P^{(22)} & \cdots & \lambda_{2n}P^{(2n)} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{n1}P^{(n1)} & \lambda_{n2}P^{(n2)} & \cdots & \lambda_{nn}P^{(nn)} \end{pmatrix}.$$

We note that the column sum of Q is not equal to one (the column sum of each $P^{(ij)}$ is equal to one). The followings are two propositions [3] related to some properties of the model.

Proposition 2.1 If $\lambda_{ij} > 0$ for $1 \leqslant i, j \leqslant n$, then the matrix Q has an eigenvalue equal to 1 and the eigenvalues of Q have modulus less than or equal to 1.

Proposition 2.2 Suppose that $P^{(ij)}$ ($1 \leqslant i, j \leqslant n$) are irreducible and $\lambda_{ij} > 0$ for $1 \leqslant i, j \leqslant n$. Then there is a vector

$$\bar{\mathbf{V}} = [\bar{\mathbf{V}}^{(1)}, \bar{\mathbf{V}}^{(2)}, \cdots, \bar{\mathbf{V}}^{(n)}]^T$$

such that

$$\bar{\mathbf{V}} = Q\bar{\mathbf{V}}$$

and

$$\sum_{i=1}^{m} [\bar{\mathbf{V}}^{(j)}]_i = 1, 1 \leq j \leq n$$

where m is the number of states.

In Proposition 2.2, we require all $P^{(ij)}$ are irreducible. But actually, if Q is irreducible, we can get the same conclusion. If the model is applied to gene expression data sequences, one may take M={0,1} and $\boldsymbol{V}_t^{(i)}$ to be the expression level of the i-th gene at the time t. From (1), the expression probability distribution of the i-th gene at time (t+1) depends on the weighted average of $P^{(ij)}\boldsymbol{V}_t^{(j)}$. We remark that this is a first-order model and λ_{ij} actually give the weighting of how much Gene i depends on Gene j. In Ching, *et al.* [4], this model has been used to find cell cycles. The most proper parent genes for the i-th gene (i.e., $\boldsymbol{V}_{t+1}^{(i)}$) can be retrieved from the corresponding

λ_{ij}. The higher the value of λ_{ij}, the stronger the parent and child relationship between i-th and j-th gene will be. When this process is repeated for each j, the whole genetic network can be constructed. Given a set of genes

$$\{V^{(j_h)} : h = 1,2,...,w \quad \text{and} \quad j_h \in (1,2,...,n)\}$$

If for any gene in this set, the rest genes are the only candidates being a corresponding parent gene, then this set of genes forms a cycle.

A simplified model was proposed in Ching *et al.* [5] by assuming

$$P^{(ij)} = I \quad \text{if} \quad i \neq j. \tag{3}$$

The simplified model has smaller number of parameters and it has been shown to be statistically better in terms of BIC, see for instance [5]. Moreover, Propositions 1 and 2 still hold for the simplified model.

3. ESTIMATION OF MODEL PARAMETERS

In this section, we present methods to estimate $P^{(ij)}$ and λ_{ij}. We estimate the transition probability matrix $P^{(ij)}$ by the following method. First we count the transition frequency of the states in the i-th sequence. After making a normalization, we obtain an estimate of the transition probability matrix. We have to estimate n such m-by-m transition probability matrices to get the estimate for $P^{(ij)}$ as follows:

$$F^{(ii)} = \begin{pmatrix} f_{11}^{(ii)} & \cdots & f_{1m}^{(ii)} \\ \vdots & \ddots & \vdots \\ f_{m1}^{(ii)} & \cdots & f_{mm}^{(ii)} \end{pmatrix},$$

From $F^{(ij)}$, one can obtain the estimate for $P^{(ij)}$ as follows:

$$\hat{P}^{(ii)} = \begin{pmatrix} \hat{p}_{11}^{(ii)} & \cdots & \hat{p}_{1m}^{(ii)} \\ \vdots & \ddots & \vdots \\ \hat{p}_{m1}^{(ii)} & \cdots & \hat{p}_{mm}^{(ii)} \end{pmatrix},$$

Where

$$\hat{p}_{ab}^{(ii)} = \begin{cases} \dfrac{f_{ab}^{(ii)}}{\sum_{a=1}^{m} f_{ab}^{(ii)}}, \text{if } \sum_{a=1}^{m} f_{ab}^{(ii)} \neq 0 \\ \dfrac{1}{m}, \text{otherwise.} \end{cases}$$

Besides $\hat{p}^{(ii)}$, we need to estimate the parameters λ_{ij}. It can be shown that the multivariate Markov model has a "stationary vector" $\overline{\mathbf{V}}$ in Proposition 2. The vector $\overline{\mathbf{V}}$ can be estimated from the gene expression sequences by computing the proportion of the occurrence of each gene and we denote it by

$$\hat{\mathbf{V}} = (\hat{\mathbf{V}}^{(1)}, \hat{\mathbf{V}}^{(2)}, ..., \hat{\mathbf{V}}^{(n)})^T.$$

We therefore expect that

$$Q\hat{\mathbf{V}} \approx \hat{\mathbf{V}}.$$

$$\begin{pmatrix} \lambda_{11}\hat{P}^{(11)} & \lambda_{12}I & \cdots & \lambda_{1n}I \\ \lambda_{21}I & \lambda_{22}\hat{P}^{(22)} & \cdots & \lambda_{2n}I \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{n1}I & \lambda_{n2}I & \cdots & \lambda_{nn}\hat{P}^{(nn)} \end{pmatrix} \hat{\mathbf{V}} \approx \hat{\mathbf{V}}.$$

From the above equation, it suggests one possible way to estimate the parameters $\Lambda = \{\lambda_{ij}\}$ as follows:

$$\min_{\lambda} \max_{k} \left| \left[\lambda_{ii}\hat{P}^{(ii)}\hat{\mathbf{V}}^{(i)} + \sum_{j=1, i\neq j}^{n} \lambda_{ij}\hat{\mathbf{V}}^{(j)} - \hat{\mathbf{V}}^{(i)} \right]_k \right| \tag{4}$$

subject to

$$\sum_{j=1}^{n} \lambda_{ij} = 1, \quad \text{and} \quad \lambda_{ij} \geq 0, \quad \forall j.$$

We note that the following formulation of n linear programming problems can give the necessary solutions of Problem (4). For each i:

$$\min_{\lambda} w_i$$

Subject to

$$\begin{cases} w_i\mathbf{e} \geq \hat{\mathbf{V}}^{(i)} - B_i\lambda_{i,\cdot} \\ w_i\mathbf{e} \geq -\hat{\mathbf{V}}^{(i)} + B_i\lambda_{i,\cdot} \end{cases} \tag{5}$$

Where

$$B_i = [\hat{\mathbf{V}}^{(1)} | \hat{\mathbf{V}}^{(2)} | \cdots | P^{ii}\hat{\mathbf{V}}^{(i)} | \cdots | \hat{\mathbf{V}}^{(n)}],$$

and

$$\mathbf{e} = (1,1,...,1)^T.$$

Here λ_{ij} is the i-th row of Λ.

We remark that the estimation method can be applied to the simplified model (3). We remark that other vector norms such as $\|.\|_2$ and $\|.\|_1$ can also be used but they have different characteristics. The former will result in a quadratic programming problem while $\|.\|_1$ will still result in a linear programming problem. The main computation cost comes from solving the linear programming problem. In the estimation of $\hat{P}_{ii}$, it involves only counting frequencies of transitions and therefore the cost is minimal. Once the model parameters are available, one can then construct the underlying genetic network easily. We will demonstrate this in the section of numerical examples. The model can also be further modified to include extra conditions such as some λ_{ij} are known

to be zero. Such information can be included by adding the constraints $\lambda_{ij}=0$. Furthermore, for large network, it is known that the in-degree follows the Poisson distribution while the out-degree follows the power-law, i.e., the number of out-degree to some negative power. These important properties can also be easily included in our proposed model [24].

4. THE OPTIMAL CONTROL FORMULATION

In this section, we present the optimal control problem based on the simplified multivariate Markov model (3) and formulate it based on the principle of dynamic programming. In the simplified model (3) we proposed above, the matrix Q can be regarded as a "transition probability matrix" for the multivariate Markov chain in certain sense, and $\boldsymbol{V}_t$ can be regarded as a joint state distribution vector. We then present a control model based on the paper by Ching, *et al.*[6]. Beginning with an initial joint probability distribution $\boldsymbol{V}_0$ the gene regulatory network (or the multivariate Markov chain) evolves according to two possible transition probability matrices Q_0 and Q_1. Without any external control, we assume that the multivariate Markov chain evolves according to a fixed transition probability matrix Q_0 ($\equiv Q$). When a control is applied to the network at one time step, the Markov chain will evolve according to another transition probability Q_1 (with more favorable steady states or a more favorable state distribution). It will then return back to Q_0 again if there is no control. We note that one can have more than one type of controls, i.e., more than one transition probability matrix Q_1 to choose in each time step. For instance, in order to suppress the expression of a particular gene, one can directly toggle off this gene. One may achieve the goal indirectly by means of controlling its parent genes which have a primary impact on its expression too. But for the simplicity of discussion, we assume that there is only one direct possible control here. We then suppose that the maximum number of controls that can be applied to the network during a finite investigation period T (finite-horizon) is K where $K \leqslant T$. The objective here is to find an optimal control policy such that the state of the network is close to a target state vector $\mathbf{Z}$. Without loss of generality, here we focus on the first gene among all the genes. Accordingly, we remark that the sub-vector $\mathbf{Z}^{(1)}$ denotes the vector containing the first two entries in $\mathbf{Z}$. It can be a unit vector (a desirable state) or a probability distribution (a weighted average of desirable states). The control system is modeled as:

$$\mathbf{v}(i_t i_{t-1}...i_1) = Q_{i_t} \cdots Q_{i_1} \mathbf{v}_0,$$
$$i_1,\ldots,i_t \in \{0,1\} \quad \text{and} \quad \sum_{j=1}^{t} i_j \leq K,$$

where $\mathbf{v}(i_t\, i_{t-1} \cdots i_1)$ represents all the possible network state probability distribution vectors up to time t. We define

$$U(t) = \{\mathbf{v}(i_t i_{t-1}...i_1) : i_1,...,i_t \in \{0,1\}$$
$$\text{and} \quad \sum_{j=1}^{t} i_j \leq K\}$$

to be the set which contains all the possible state probability vectors up to time t. We note that one can conduct a forward calculation to compute all the possible state vectors in the sets $U(1), U(2),\ldots U(\mathrm{T})$ recursively. Here the main computational cost is the matrix-vector multiplication and the cost is $O((2n)^2)$ where n is the number of genes in the network. We note that some state probability distribution actually does not exist because the maximum number of controls is K, the total number of vectors involved is only

$$\sum_{j=0}^{K} \frac{T!}{j!(T-j)!}.$$

For example if K=1, the complexity of the above algorithm is $O(T(2n)^2)$.

Returning to our original problem, our purpose is to make the system go to the desirable states. The objective here is to minimize the overall average of the distances of the state vectors $\mathbf{v}(i_t...i_1)$ (t=1,2,...,T) to the target vector $\mathbf{z}$, i.e.,

$$\min_{\mathbf{v}(i_T i_{T-1}...i_1)\in U(T)} \frac{1}{T}\sum_{t=1}^{T} \| \mathbf{v}(i_t...i_1) - \mathbf{z} \|_2 . \qquad (6)$$

To solve (6), we have to define the following cost function

$$D(\mathbf{v}(\mathbf{w}_t), t, k), \quad 1 \leq t \leq T, \quad 0 \leq k \leq K$$

as the minimum total distance to the terminal state at time T when beginning with state distribution vector $\mathbf{v}(\mathbf{w}_t)$ at time t and that the number of controls used is k. Here W_t is a Boolean string of length t. Given the initial state of the system, the optimization problem can be formulated as:

$$\min_{0\leq k\leq K} \{D(\mathbf{v}_0, 0, k)\} \qquad (7)$$

subject to:

$$D(\mathbf{v}(\mathbf{w}_t), t, K+1) = \infty, \quad \text{for all } \mathbf{w}_t \text{ and } t,$$
$$D(\mathbf{v}(\mathbf{w}_T), T, k) = \| \mathbf{v}(\mathbf{w}_T) - \mathbf{z} \|_2,$$
$$\text{for} \quad \mathbf{w}_T = i_T...i_1, \sum_{j=1}^{T} i_j \leq K, k = 0,1,...,K.$$

To solve the optimization problem, one may consider the following dynamic programming formulation:

$$D(\mathbf{v}(\mathbf{w}_{t-1}),t-1,k)=$$
$$\min\{\| \mathbf{v}(0\mathbf{w}_{t-1})-\mathbf{z}\|_2 + D(\mathbf{v}(0\mathbf{w}_{t-1}),t,k),$$
$$\| \mathbf{v}(1\mathbf{w}_{t-1})-\mathbf{z}\|_2 + D(\mathbf{v}(1\mathbf{w}_{t-1}),t,k+1)\}. \quad (8)$$

Here $0\mathbf{w}_{t-1}$ and $1\mathbf{w}_{t-1}$ are Boolean strings of size t. The first term in the right-hand-side of (8) is the cost (distance) when no control is applied at time t while the second term is the cost when a control is applied. The optimal control policy can be obtained during the process of solving (8). We remark that instead of considering the objective (6), one can consider

$$\min_{V(i_T i_{T-1},\ldots,i_1)\in U^T} \sum_{t=1}^{T} \alpha_t \| v(i_t \ldots i_1) - z \|_l$$

With $\{\alpha_i\}$ a new weighting and a different vector norm $\|.\|_l$. Furthermore, it is interesting to study the case of infinite horizon. In this case α_t is chosen to be $(1-\alpha)\alpha^{t-1}$ for some discount factor $\alpha \in (0,1)$.

5. NUMERICAL EXPERIMENTS

5.1. A Simple Example

In this subsection, we consider a small five-gene network whose gene expression series can be found in the Appendix. **Figure 1** shows the five-gene network. We note that Gene 1 and Gene 4 depends on all the other genes, Gene 2 depends on Gene 1 and Gene 3 only, Gene 3 depends on Gene 1 and Gene 2 only, while Gene 5 depends on itself only.

To solve the linear programming problem in equation (5), infinity norm is chosen for all numerical experiments. The matrices Λ, P, and Q_0 (without control) are obtained from the proposed model as follow:

$$P=\begin{pmatrix} P_1 & I_2 & I_2 & I_2 & I_2 \\ I_2 & P_2 & I_2 & I_2 & I_2 \\ I_2 & I_2 & P_3 & I_2 & I_2 \\ I_2 & I_2 & I_2 & P_4 & I_2 \\ I_2 & I_2 & I_2 & I_2 & P_5 \end{pmatrix}$$

Where

$$P_1=\begin{pmatrix} 0.6000 & 0.4286 \\ 0.4000 & 0.5714 \end{pmatrix} \quad P_2=\begin{pmatrix} 0.2857 & 0.6667 \\ 0.7143 & 0.3333 \end{pmatrix}$$

$$P_3=\begin{pmatrix} 0.5000 & 0.4467 \\ 0.5000 & 0.5333 \end{pmatrix} \quad P_4=\begin{pmatrix} 0.3571 & 0.6000 \\ 0.6429 & 0.4000 \end{pmatrix}$$

$$P_5=\begin{pmatrix} 0.4000 & 0.3158 \\ 0.6000 & 0.6842 \end{pmatrix} \quad I_2=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Lambda=\begin{pmatrix} 0.3369 & 0.2604 & 0.2604 & 0.1417 & 0.0005 \\ 0.5000 & 0.0000 & 0.5000 & 0.0000 & 0.0000 \\ 0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 \\ 0.2045 & 0.2045 & 0.2045 & 0.2028 & 0.1838 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

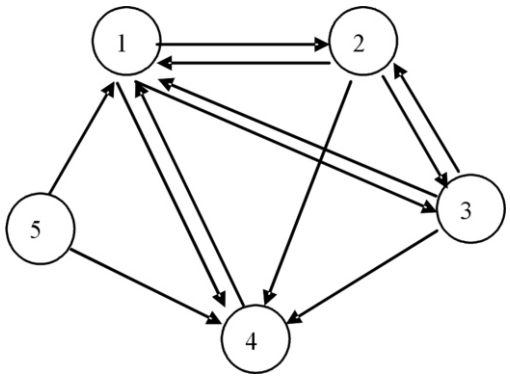

Figure 1. The Five-gene Network.

and

$$Q_0=\begin{pmatrix} 0.3369P_1 & 0.2604I_2 & 0.2604I_2 & 0.1417I_2 & 0.0005I_2 \\ 0.5000I_2 & 0.0000P_2 & 0.5000I_2 & 0.0000I_2 & 0.0000I_2 \\ 0.5000I_2 & 0.5000I_2 & 0.0000P_3 & 0.0000I_2 & 0.0000I_2 \\ 0.2045I_2 & 0.2045I_2 & 0.2045I_2 & 0.2028P_4 & 0.1838I_2 \\ 0.0000I_2 & 0.0000I_2 & 0.0000I_2 & 0.0000I_2 & 1.0000P_5 \end{pmatrix}.$$

The target here is to suppress the first gene but no preference on other genes. The control we used is to suppress the first gene directly. Thus the control matrix is as follows:

$$Q_1=\mathbf{Diag}\left(\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, I_2, I_2, I_2, I_2\right).$$

Without loss of generality, we assume that the initial state vector is the uniform distribution vector (for each gene), that is

$$\mathbf{v}_0=\frac{1}{2}(1,1,1,1,1,1,1,1,1,1)^T.$$

Moreover, we assume that the total time T is 12 and we try several different numbers of controls K=1,2,3,4,5. **Table 1** shows the numerical results. All the computations were done in a PC with Pentium D and Memory 1GB with MATLAB 7.0. In **Table 1**, "Policy" represents the optimal time step at the end of which a control should be applied. For instance, means that the optimal control policy is to apply the control at the end of the t=1,2,3-th time step. From **Table 1**, observable improvements of the optimal value is obtained when K increases from 1 to 5.

5.2. The Yeast Example

Table 1. Numerical results for the 5-gene network.

K	1	2	3	4	5
Control Policy	[1]	[2]	[1,2,3]	[1,2,3,7]	[1,2,3,7,8]
Objective Value	0.5628	0.4277	0.3379	0.2717	0.2090
Time in Seconds	0.02	0.02	0.06	0.15	0.23

In this subsection, we apply our proposed simplified multivariate Markov models to the yeast data sequences [23]. Genome transcriptional analysis is an important analysis in medicine, etiology and bioinformatics. One of the applications of genome transcriptional analysis is used for eukaryotic cell cycle in yeast. The fundamental periodicity in eukaryotic cell cycle includes the events of DNA replication, chromosome segregation and mitosis. It is suggested that improper cell cycle regulation leads to genomic instability, especially in the etiology of both hereditary and spontaneous cancers [9, 22]. Eventually, it is believed to play one of the important roles in the etiology of both hereditary and spontaneous cancers. The dataset used in our study is the selected set from Yeung and Ruzzo (2001) [23]. In the discretization, if an expression level is above (below) a certain standard deviation from the average expression of the gene, it is over-expressed (under-expressed) and the corresponding state is 1 (0) [4].

To solve the linear programming problem in (5), infinity norm is chosen for all numerical experiments. The matrices Λ, P, and Q_0 (without control) are obtained from the proposed model. The initial state vector is assumed to be the uniform distribution (for each gene) vector

$$\mathbf{v}_0 = \frac{1}{2}(1,1,\cdots,1)^T.$$

In addition, we assume that the total time T is 12 and several different maximum numbers of controls K=1,2,3,4,5 are tried in our numerical experiments. The target is to suppress the first gene but no preference on other genes. That is the target state vector $\mathbf{Z}^{(1)}$ is $(1,0)^T$. The control we used is to suppress the first gene directly. Thus the control matrix Q_1 takes the same form as the following:

$$Q_1 = \mathbf{Diag}\left(\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, I_2, I_2, ..., I_2, I_2\right).$$

It means that we want to control the first gene such that it will be unexpressed with more probabilities. The transitions of all the other genes will not be changed. **Table 2** reports the numerical results and the computational time for different numbers of controls K. From **Table 2**, observable improvements of the optimal value is obtained when K increases from 1 to 5. For example, if we will conduct 4 controls totally in the 12 time steps, we need to suppress the first gene in the first 4 steps, and will not control it in other steps. These experiments show that even the number of genes (384 genes in this data set) is comparatively large, the method still can find the control policies fast.

Table 2. Numerical results for the yeast data set.

K	1	2	3	4	5
Control Policy	[1]	[2]	[1,2,3]	[1,2,3,4]	[1,2,3,4,5]
Objective Value	0.6430	0.5751	0.5165	0.4582	0.4000
Time in Seconds	4.00	20.60	67.90	152.88	245.95

6. CONCLUDING REMARKS

In this paper, we proposed a simplified multivariate Markov model for approximating PBNs. Efficient estimation methods based on linear programming method are presented to obtain the model parameters. Methods for recovering the structure and rules of a PBN are also illustrated in details. We then give an optimal control formulation for control the network. Numerical experiments on synthetic data and gene expression data of yeast are given to demonstrate the effectiveness of our proposed model and formulation.

For future research, we will extend the control problem to the case of having multiple control policy. We will develop efficient heuristic methods for solving the control problem and genetic algorithm is a possible approach [7]. Extension of the study to the case of infinite horizon is also interesting. Finally, we will also apply our model to more real world datasets.

APPENDIX

The five gene expression sequences.

*Gene*1 : 11000011111000001010001111010 1

*Gene*2 : 01010111010011001001101101010 0

*Gene*3 : 01101100110001100001111101010 0

*Gene*4 : 11010101010111100100100011100 1

*Gene*5 : 11111101101110111000101000111 1

ACKNOWLEDGEMENTS

Research supported in part by RGC Grants 7017/07P and HKU CRCG Grants. The preliminary version of the paper has been published in the proceedings of the 2nd International Conference on Bioinformatics and Biomedical Engineering, Shanghai, China, 2008.

REFERENCE

[1]T. Akutsu, S. Miyano & S. Kuhara. Inferring Qualitative Relations in Genetic Networks and Metabolic Arrays. *Bioinformatics* 2000, 16: 727-734.

[2]J. Bower. *Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge* 2001, M.A.

[3]Ching, W., E. Fung & M. Ng. A multivariate Markov Chain Model for Categorical Data Sequences and Its Applications in Demand Predictions. *IMA Journal of Management Mathematics* 2002, 13: 187-199.

[4]Ching, W., E. Fung, M. Ng & T. Akutsu. On Construction of Stochastic Genetic Networks Based on Gene Expression Sequences. *International Journal of Neural Systems* 2005, 15: 297-310.

[5]Ching, W., Zhang, S., & M. Ng. On Multi-dimensional Markov Chain Models. *Pacific Journal of Optimization* 2007, 3: 235-243.

[6]Ching, W., Zhang, S., Jiao, Y., T. Akutsu & Wong, A. *Optimal Finite-Horizon Control for Probabilistic Boolean Networks*

with Hard Constraints. *The International Symposium on Optimization and Systems Biology* 2007.
[7]Ching, W., H. Leung, Tsing, N. & Zhang, S. *Optimal Control for Probabilistic Boolean Networks : Genetic Algorithm Approach,* 2008.
[8]E. Dougherty, S. Kim & Chen, Y. Coefficient of Determination in Nonlinear Signal Processing. *Signal Processing* 2000, 80: 2219-2235.
[9]M. Hall, & G. Peters. Genetic Alterations of Cyclins, Cyclin-dependent Kinases, and Cdk Inhibitors in Human Cancer. *Adv. Cancer Res.* 1996, 68: 67-108.
[10]Huang, S. & D.E. Ingber. Shape-dependent Control of Cell Growth, Differentiation, and Apoptosis: Switching Between Attractors in Cell Regulatory Networks. *Exp. Cell Res.* 2000, 261: 91-103.
[11]H. de Jong. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *J. Comput. Biol.* 2002, 9: 69-103.
[12]S. Kauffman. Metabolic Stability and Epigenesis in Randomly Constructed Gene Nets. *J. Theoret. Biol.* 1969, 22: 437-467.
[13]S. Kauffman. Homeostasis and Differentiation in Random Genetic Control Networks. *Nature* 1969, 224: 177-178.
[14]S. Kauffman. *The Origin of Orders*, Oxford University Press, New York, 1993.
[15]S. Kim, S. Imoto & S. Miyano. Dynamic Bayesian Network and Nonparametric Regression for Nonlinear Modeling of Gene Networks from time Series Gene Expression Data. *Proc. 1st Computational Methods in Systems Biology, Lecture Note in Computer Science* 2003, 2602: 104-113.
[16]F. Nir, L. Michal, N. Iftach & P. Dana. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology* 2000, 7(3-4): 601-620.
[17]I. Shmulevich, E. Dougherty, S. Kim & W. Zhang. Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks. *Bioinformatics* 2002, 18: 261-274.
[18]I. Shmulevich, E. Dougherty, S. Kim & Zhang, W. Control of Stationary Behavior in Probabilistic Boolean Networks by Means of Structural Intervention. *Journal of Biological Systems* 2002, 10: 431-445.
[19]I. Shmulevich, E. Dougherty, S. Kim & W. Zhang. From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks. *Proceedings of the IEEE* 2002, 90: 1778-1792.
[20]I. Shmulevich & E. Dougherty. *Genomic Signal Processing*, Princeton University Press, USA, 2007.
[21]P. Smolen, D. Baxter & J. Byrne. Mathematical Modeling of Gene Network. *Neuron* 2002, 26: 567-580.
[22]Wang, T. C., R.D. Cardiff, L. Zukerberg, E. Lees, A. Arnold & E.V. Schmidt. Mammary Hyerplasia and Carcinoma in MMTV-cyclin D1 Transgenic Mice. *Nature* 1994, 369: 669-671.
[23]K. Yeung & W. Ruzzo. An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics* 2001, 17: 763-774.
[24]Zhang, S., Ching, W., Tsing, N., H. Leung & Guo, D. *A. Multiple Regression Approach for Building Genetic Networks. The Proceedings of the International Conference on BioMedical Engineering and Informatics* 2008, Sanya, China.