A Simplified Multivariate Markov Chain Model for the Construction and Control of Genetic Regulatory Networks

Shu-Qin Zhang
School of Mathematical Sciences,
Fudan University,
Shanghai,200433, China
E-mail: zhangs@fudan.edu.cn

Wai-Ki Ching
and Yue Jiao
Advanced Modeling and Applied Computing Laboratory,
Department of Mathematics,
The University of Hong Kong,
Pokfulam Road, Hong Kong
E-mail: wching@hkusua.hku.hk
E-mail: jiaoyue@hkusua.hku.hk

Ling-Yun Wu
Institute of Applied Mathematics,
Academy of Mathematics and System Sciences,
Chinese Academy of Sciences
E-mail:lywu@amt.ac.cn

Raymond, H. Chan
Department of Mathematics,
The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong
E-mail:rchan@math.cuhk.edu.hk

***Abstract*—The construction and control of genetic regulatory networks using gene expression data is an important research topic in bioinformatics. 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 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 then given to demonstrate the effectiveness of our proposed model and control policy.**

I. 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. 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], [7], [11], [12], [13], [14] 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, one can uncover the underlying gene regulatory mechanisms. This is particularly useful as it helps to make useful predictions by computer simulations.

The BN model was first introduced by Kauffman [9], [10]. 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)). 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. To overcome the deterministic nature of a BN, Shmulevich *et al.* [13] proposed a PBN that can share the appealing rule-based properties of Boolean networks and it is robust in the presence of uncertainty.

The dynamics of the PBN can be studied in the context of standard Markov chain. 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 develop efficient model parameters estimation methods based on linear programming. We then 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 organized 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 the gene expression data of yeast. Concluding remarks are then given to address further research issues in Section 6.

II. Multivariate Markov Chain Models

In this section, we first review a multivariate Markov chain model proposed in Ching, *et al.* [3] for modeling categorical time series data. The model has been used for building genetic regulatory networks [4]. We then present our 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 $\mathbf{V}_t^{(j)},\ j = 1, 2, \ldots, n$. In Ching, *et al.* [3], the following first-order model was proposed to model the relationships among the sequences:

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

where

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

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 $1 \leq i, j \leq n$, 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 $P^{(ij)}$ $(1 \leq i, j \leq n)$ are irreducible and $\lambda_{ij} > 0$ for $1 \leq i, j \leq n$. 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 sequences, we may take $M = \{0, 1\}$ and $\mathbf{V}_t^{(i)}$ is 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 $P^{(ij)}\mathbf{V}_t^{(j)}$. In Ching, *et al.* [4], this model has been used to find cell cycles.

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. Moreover, Propositions 1 and 2 still hold for the simplified model.

III. Estimation of Model Parameters

In this section, we present methods to estimate $P^{(ij)}$ and λ_{ij}. We estimate the transition probability matrix $P^{(ii)}$ 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 $P^{(ii)}$ as follows:

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

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

where

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

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" $\bar{\mathbf{V}}$ in Proposition 2. The vector $\bar{\mathbf{V}}$ can be estimated from the gene expression sequences by computing the proportion of the occurrence of each gene and we denote it by

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

We therefore expect that

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

¿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 $B_i = [\hat{\mathbf{V}}^{(1)} \mid \hat{\mathbf{V}}^{(2)} \mid \cdots \mid \hat{P^{ii}}\hat{\mathbf{V}}^{(i)} \mid \cdots \mid \hat{\mathbf{V}}^{(n)}]$, $\mathbf{e} = (1, 1, ..., 1)^T$, and $\lambda_{i,\cdot}$ is the i-th row of Λ. 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.

IV. 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 $\mathbf{V}_t$ 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 $\mathbf{v}_0$ 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 $K \leq T$. 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:

$$\mathbf{v}(i_t i_{t-1} \dots i_1) = Q_{i_t} \cdots Q_{i_1} \mathbf{v}_0,$$

$$i_1, \dots, i_t \in \{0, 1\} \quad \text{and} \quad \sum_{j=1}^{t} i_j \leq K,$$

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

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

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 $U(1), U(2), \dots, U(T)$ 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 $\mathbf{v}(i_t \dots i_1)$ $(t = 1, 2, \dots, T)$ to the target vector $\mathbf{z}$, i.e.,

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

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 $\mathbf{w}_t$ is a Boolean string of length t. Given the initial state of the system, the optimization problem can be formulated as:

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

subject to:

$$D(\mathbf{v}(\mathbf{w}_t), t, K+1) = \infty, \quad \text{for all } \mathbf{w}_t \text{ and } t,$$

$$D(\mathbf{v}(\mathbf{w}_T), T, k) = ||\mathbf{v}(\mathbf{w}_T) - \mathbf{z}||_2,$$

$$\text{for} \quad \mathbf{w}_T = i_T \dots i_1, \sum_{j=1}^{T} i_j \leq K, k = 0, 1, \dots, K.$$

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).

V. Numerical Experiments

In this subsection, we test our simplified multivariate Markov models for the yeast data sequences [16]. 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 [8], [15]. Eventually, it is believed to play one of the important roles in the etiology of both hereditary and spontaneous cancers. The data set used in our study is the selected set from Yeung and Ruzzo (2001) [16]. 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).

To solve the linear programming problem in (5), infinity norm is chosen for all numerical experiments. The matrices Λ, P, Q_0 (without control) and Q_1 (control matrix) are obtained from the proposed model. The control matrix Q_1 takes the same form as the following:

$$Q_1 = \mathbf{Diag}(\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, I_2, I_2, I_2, I_2). \quad (9)$$

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 number 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. The control we used is to suppress the first gene directly. The target state vector $\mathbf{z}^{(1)}$ is $(1, 0)^T$. Table 1 reports the numerical results and the computational time for different numbers of controls K. All the computations were done in a PC with Pentium D and Memory 1GB with MATLAB 7.0. In Table 1, "Control Policy" represents the optimal time step at the end of which a control should be applied. For instance, $[1, 2, 3]$ 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.

TABLE I
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

VI. Concluding Remarks

In this paper, we proposed a simplified multivariate Markov model for approximating PBNs. Efficient estimation methods 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 develop efficient heuristic methods for solving the control problem. We will also apply our model to more real world datasets.

References

[1] T. Akutsu, S. Miyano and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic arrays. *Bioinformatics*, 16: 727-734, 2000.

[2] J. Bower. Computational modeling of genetic and biochemical networks. MIT Press, Cambridge, M.A. 2001.

[3] W. Ching, E. Fung and M. Ng. A multivariate Markov chain model for categorical data sequences and its applications in demand predictions. *IMA Journal of Management Mathematics*, 13: 187-199, 2002.

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

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

[6] W. Ching, S. Zhang, Y. Jiao, T. Akutsu and A. Wong. Optimal Finite-Horizon Control for Probabilistic Boolean Networks with Hard Constraints. The International Symposium on Optimization and Systems Biology (OSB 2007), Lecture Notes in Operations Research, 2007.

[7] H. de Jong. Modeling and simulation of genetic regulatory systems: A Literature Review. *J. Comput. Biol.*, 9: 69-103, 2002.

[8] M. Hall, and G. Peters. Genetic alterations of cyclins, cyclin-dependent kinases, and Cdk inhibitors in human cancer. *Adv. Cancer Res.*, 68: 67-108, 1996.

[9] S. Kauffman. Metabolic stability and epigenesis in randomly constructed gene nets. *J. Theoret. Biol.*, 22: 437-467, 1969.

[10] S. Kauffman. The origin of orders. Oxford University Press, New York. 1993.

[11] S. Kim, S. Imoto and 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*, 2602: 104-113, 2003.

[12] F. Nir, L. Michal , N. Iftach and P. Dana. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4): 601-620, 2000.

[13] I. Shmulevich, E. Dougherty, S. Kim and W. Zhang, Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18: 261-274, 2002.

[14] P. Smolen, D. Baxter and J. Byrne. Mathematical modeling of gene network. *Neuron*, 26: 567-580, 2000.

[15] T. C. Wang, R.D. Cardiff, L. Zukerberg, E. Lees, A. Arnold and E.V. Schmidt. Mammary hyerplasia and carcinoma in MMTV-cyclin D1 transgenic mice. *Nature*, 369: 669-671, 1994.

[16] K. Yeung and W. Ruzzo. An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics*, 17: 763-774, 2001.