The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

Finite-Horizon Control of Genetic Regulatory Networks with Multiple Hard-Constraints

Wai-Ki Ching* Yang Cong†

Advanced Modeling and Applied Computing Laboratory,
Department of Mathematics, The University of Hong Kong, Hong Kong.

Abstract Probabilistic Boolean Networks (PBNs) provide a convenient tool for studying the interactions among different genes while allowing uncertainty. This paper deals with the issue of finite-horizon control with multiple hard-constraints in a PBN. More precisely, under the constraint of the number of times that each control method can be applied, we develop a control strategy by which the state of a given genetic network falls into a desired state set with a prescribed minimum probability. We propose an efficient algorithm to find the feasible solutions. An upper bound for the computational cost is also given. An numerical experiment is then conducted to demonstrate the efficiency of our proposed method.

Keywords Probability Boolean Networks; Finite-Horizon; Multiple Hard-Constraints; Intervention; Markov Chain; Optimal Control

1 Introduction

In computational system biology, building mathematical models and efficient numerical algorithms to study regulatory interactions among DNA, RNA, proteins and small molecules are important issues [2, 18]. There have been many mathematical models proposed to study genetic regulatory networks such as Bayesian networks [17], Boolean networks (BNs) [13, 14], multivariate Markov chain model [4], regression model [7], Probabilistic Boolean Networks(PBNs) [22, 23, 24, 25] and reviews on other mathematical models can be found in [19, 27]. Among these models, BN and its extension PBN have received much attention as they can capture the switching behavior of the biological process [18].

Boolean networks (BNs) are introduced by Kauffman [13, 14, 15, 16]. A Boolean Network $G(V,F)$, consists of a set of vertices $V = \{v_1, v_2, \ldots, v_n\}$, where $v_i(t)$ is the expression state for gene i at time t. We quantize $v_i(t)$ to only two levels: on and off (represented by 1 and 0). We define $F = \{f_1, f_2, \ldots, f_n\}$ as a set of Boolean functions $(f_i : \{0,1\}^n \to \{0,1\})$ to represent the rules of the regulatory interactions among the genes: $v_i(t+1) = f_i(\mathbf{v}(t))$. Here $\mathbf{v}(t) = (v_1(t), v_2(t), \ldots, v_n(t))^t$ is called the Gene Activity Profile (GAP). The GAP can take any possible form from the set

$$S = \{(v_1, v_2, \ldots, v_n)^T : v_i \in \{0,1\}\} \tag{1}$$

*Email: wching@hkusua.hku.hk
†Email: congyang0305@yahoo.com.cn

and totally there are 2^n possible states.

A BN is indeed a deterministic model and the only randomness comes from its initial state. Given an initial state, the BN will eventually enter into a set of state(s) called attractor cycle and stay there forever [1, 13, 14]. The attractors have biological significance such as states of cell proliferation or cell apoptosis [12]. However, genetic regulation process exhibits uncertainty and microarray data sets used to infer the model have errors due to experimental noise in the complex measurement process. Thus it is more realistic to consider stochastic models. To extend BNs to PBNs, the main idea is as follows. To determine v_i the state of gene i, $i = 1,2,\ldots,n$, there can be more than one Boolean function $f_j^{(i)}(j = 1,2,\ldots,l(i))$ to be chosen. Here $1 \le l(i) \le 2^{2^n}$ is the total number of possible Boolean functions for gene i. The probability of choosing $f_j^{(i)}$ as the predictor function is $c_j^{(i)}$, where $0 \le c_j^{(i)} \le 1$ and

$$\sum_{j=1}^{l(i)} c_j^{(i)} = 1. \tag{2}$$

One can estimate the probability $c_j^{(i)}$ by the statistical method Coefficient of Determination (COD) with real gene expression data sets [11]. There are at most $N = \prod_{i=1}^{n} l(i)$ different possible realizations of BNs. Let f_j be the jth possible realization,

$$f_j = (f_{j_1}^{(1)}, f_{j_2}^{(2)}, \ldots, f_{j_n}^{(n)}), \quad 1 \le j_i \le l(i). \tag{3}$$

Then in an independent PBN (the selection of the Boolean function for each gene is assumed to be independent), the probability of choosing the corresponding BN is given by $q_j = \prod_{i=1}^{n} c_{j_i}^{(i)}, j = 1,2,\ldots,N$. We note that the transition process among the states in the set S is a Markov chain process. Let $\mathbf{a}$ and $\mathbf{b}$ be any two column vectors (can be the same) in the set S. Then the transition probability from state $\mathbf{b}$ to state $\mathbf{a}$ is

$$\begin{aligned} & \text{Prob}\,\{\mathbf{v}(t+1) = \mathbf{a} \mid \mathbf{v}(t) = \mathbf{b}\} \\ = & \sum_{j=1}^{N} \text{Prob}\,\{\mathbf{v}(t+1) = \mathbf{a} \mid \mathbf{v}(t) = \mathbf{b}, \text{the } j\text{th network is selected}\} \cdot q_j. \end{aligned} \tag{4}$$

By computing the transition probability for all the possible states in S, we can obtain the transition probability matrix A of the PBN. In fact, the transition probability matrix A can be written as the sum of the Boolean network matrices A_i, $A = \sum_{i=1}^{N} q_i A_i$, see for instance [5].

In a PBN, the steady-state probability distribution provides its first-order statistical information, through which one can understand a genetic network and identify the influence of different genes in such a network. Power method has been used to compute the steady-state probability distribution with an efficient construction of the transition probability matrix [28]. A matrix approximation method has been proposed in [8] to get an approximation of the steady-state probability distribution.

Furthermore, it is possible to control some genes in a network to drive the whole network into a desirable steady-state probability distribution. In [26], the potential effect of individual gene on the global dynamical network behavior is studied, by means of random gene perturbation and intervention. The effect of altering the the rule-based structure is discussed in [23]. To achieve relatively more permanent effect of intervention, optimal control theory finds its application. In [9], an optimal finite-horizon control problem for gene intervention is formulated as a minimization problem with penalty costs. The penalty costs include both control cost and cost of the terminal states. The control cost is defined as the cost of applying control inputs in some particular states. Relatively higher terminal costs are assigned to those undesirable states. Thus the optimal control policy is the one which minimizes the overall expected costs. One can obtain the optimal control strategy by dynamic programming method. Other control problems such as imperfect information, context-sensitive PBN and infinite-horizon control are discussed in [10, 20, 21] separately. In [3], an algorithm is proposed to study the problem of controlling a gene network (without state feedback) such that it reaches a target state set with a prescribed maximum or minimum probability. The algorithm in [3] stops whenever a optimal solution is obtained regardless of the length of the control horizon. If there is no optimal solution, the algorithm will run infinite times.

All the above optimal control formulations did not consider the case of hard-constraint, i.e. to include an upper bound for the number of controls. In case of disease such as cancer, control inputs can be medication, radiation etc. They are typically applied during a period. And some treatments such as radiation can not be applied too many times. [6] fills that blank by studying an optimal finite-horizon problem with hard-constraint. It discusses the problem with one control variable. Observing that usually there are more than one treatment methods applied together, we study finite-horizon external control problem with multiple hard-constraints. In [3], the authors focus on leading the network to fall into a desirable state set with a prescribed minimum or maximum probability. Here we set minimizing the cost of the control strategy as the objective, meanwhile adopt the idea in [3] as a criteria. Apart from the finite-horizon control problem with multiple hard-constraints, we provide a more effective algorithm in generating all the feasible control strategies than that proposed in [6]. We remark that our proposed formulation can be applied to both perturbed and context-sensitive PBNs, though we only discuss examples of instantaneously random PBNs. Here we point that the number of possible states in the network increases exponentially with respect to the number of genes n, thus the computational cost for solving the optimal control problem can be enormous even for moderate n. It has been shown that finding a control strategy for a BN to a global state is actually NP-hard [1].

The remainder of the paper is structured as follows. In Section 2, we introduce our optimal finite-horizon control problem with multiple hard-constraints. In section 3, numerical examples are given to demonstrate the efficiency of our proposed method. Finally, concluding remarks are given to address further research issues in Section 4.

2 Problem Formulation

In this section, we give a mathematical formulation for our optimal control problem with multiple hard-constraints. Our goal is to find an optimal strategy for manipulating

external control variables to desirably affect the dynamic evolution of a random PBN over a finite time horizon with the minimum corresponding cost.

Without loss of generality, here we consider the case of two control methods. Then at each time point t, $t = 1,2,\ldots,T$, one of the followings three control options will be conducted: Control 0 (i.e. no control), Control 1 and Control 2, represented by u_0, u_1 and u_2 respectively. Their corresponding transition probability matrices P_0, P_1, P_2 are given. The optimal control problem can be stated as follows. Given an initial probability distribution $\mathbf{x}_0$ and a set of target states $S' \subseteq S$, our goal is to find a sequence of actions σ that lead the system reaching a target state with a minimum probability $\bar{p}$ over a finite time horizon T (i.e. $\sum_{i\in S'}[\mathbf{x}_T]_i \geq \bar{p}$) while minimizing the sum of the costs of the control actions applied at each time point $\sum_{i=1}^{T} C(\sigma_i)$. Thus we obtain the following optimal control problem:

$$\min_{\sigma} \sum_{i=1}^{T} C(\sigma_i)$$

subject to

$$\begin{cases} \sum_{i\in S'}[\mathbf{x}_T]_i \geq \bar{p}, \\ 0 \leq s_1 \leq K_1, \\ 0 \leq s_2 \leq K_2. \end{cases} \tag{5}$$

Here s_i is the number of times that Control i is conducted, and K_i is the maximum number of times that Control i can be applied, $i = 1,2$. We use $i_j \in \{0,1,2\}$ to represent that Control i_j is applied to the network at time j. Then control string $i_1 i_2 \ldots i_k$ represents the control actions conducted from time 1 to time k. We define set

$$U = \{\sigma = i_1 i_2 \ldots i_T : i_j \in \{0,1,2\}, \quad \text{and} \quad 0 \leq s_i \leq K_i\}$$

as the set containing all the possible control strategies satisfying the multiple hard-constraints. Given the initial probability distribution vector $\mathbf{x}_0$, state probability distribution vector $\mathbf{x}_T = P_{i_T} \ldots P_{i_1} \mathbf{x}_0$ represents the state distribution vector at time T obtained by control strategy $\sigma = i_1 i_2 \ldots i_T$. The feasible solution set V is a subset of set U, where

$$V = \{\sigma \in U : \sum_{i\in S'}[\mathbf{x}_T]_i \geq \bar{p}\}.$$

Finally, optimal solution exists if the set V is not empty, and there can be more than one optimal solutions. Any control strategy in the set V with minimum corresponding cost is an optimal solution.

The main computational cost comes from the matrix-vector multiplication. For each control strategy, the number of matrix-vector multiplication is T. If we search an optimal solution in the set $W = \{\sigma = i_1 i_2 \ldots i_T : i_j \in \{0,1,2\}\}$ the cost is $O(T3^T 2^{2n})$, where n is the number of genes in the network. However, we only need to consider those strategies in set V, this reduces the computational and storage costs. It's hard to estimate the number of elements in the set V. But we know there are totally

$$M = \sum_{j=0}^{K_2} \left(C_T^j \sum_{i=0}^{K_1} C_{T-j}^i \right) = \sum_{j=0}^{K_2} \sum_{i=0}^{K_1} \frac{T!}{i!j!(T-i-j)!} \tag{6}$$

control strategies in set $U \supseteq V$. Thus $MT2^{2n}$ is an upper bound of the computational cost.

2.1 Algorithm for Finding Feasible Solutions

In order to find the feasible solution set for the optimal control problem with hard-constraint, [6] applied a recursive method as follows. They first start with set $\{0,1\}$, which contains all the possible control strategies at time $t = 1$, then one can obtain set $\{00,10,01,11\}$ for time $t = 2$. Recursively one can get the feasible solution set. However, our problem involves more than one control methods under hard-constraints. Thus here we introduce a more efficient method. Note that

$$V = \{\sigma \in U : \sum_{i \in S'} [\mathbf{x}_T]_i \geq \bar{p}\},$$

one can get the feasible solution set V by checking whether the state probability distribution obtained by any control string in the set U satisfies the constraint $\sum_{i \in S'} [\mathbf{x}_T]_i \geq \bar{p}$. Thus the key point is to generate the set U, the set of all possible control strategies satisfying the hard-constraints.

We first assume that the number of times that Control 2 is applied is fixed as $k, 0 \leq k \leq K_2$. We reserve k places in the control string of length T for Control 2, there are totally C_T^k cases. Then we only need to find all the control strings of length $T-k$ where Control 0 (i.e. no control) and Control 1 can be applied and the maximum number of times that Control 1 can be applied is K_1. Now the method in [6] can be applied. However, in order to save memory and promote efficiency, we apply the following method. We note that among all the possible control strings, binary string $\underbrace{11\ldots1}_{K_1}\underbrace{00\ldots0}_{T-K_1-k}$ is the biggest one. Thus by translating decimal digits from 0 to $2^{T-k}-1$ to binary digits and checking the number of times that Control 1 is applied, we can generate all the control strings of length $T-k$ satisfying the hard-constraint for Control 1. Finally we can obtain the set U by increasing k from 0 to K_2.

3 Numerical Examples

In this section, we present an numerical example using a hypothetical gene network to illustrate the application of the proposed algorithm. The network is consist of two genes denoted by A and B, induced by a certain biological signal. The states of genes A and B are given in Table 1. There are three external control methods: (i) Control 0 (no control): chemical signal absent, (ii) Control 1: chemical signal present in low concentration, and (iii) Control 2: chemical signal present in high concentration. Their corresponding transition probability matrices are given as follows.

$$P_0 = \begin{pmatrix} 0.7 & 0.4 & 0.4 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.5 \\ 0.3 & 0.0 & 0.5 & 0.3 \\ 0.0 & 0.3 & 0.1 & 0.2 \end{pmatrix}, P_1 = \begin{pmatrix} 0.3 & 0.1 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.4 & 0.1 \\ 0.4 & 0.5 & 0.4 & 0.1 \\ 0.2 & 0.4 & 0.2 & 0.8 \end{pmatrix},$$
$$P_2 = \begin{pmatrix} 0.1 & 0.0 & 0.0 & 0.0 \\ 0.3 & 0.1 & 0.2 & 0.3 \\ 0.3 & 0.4 & 0.5 & 0.4 \\ 0.3 & 0.5 & 0.3 & 0.3 \end{pmatrix}. \tag{7}$$

State	A	P
1	Off	Off
2	Off	On
3	On	Off
4	On	On

Table 1: State of Gene A and Product P.

k	Control Strategy $\sigma = i_1 i_2 \dots i_T$										Cost
$k=0$	0	0	0	1	0	0	1	1	1	1	12.5
	0	0	0	0	1	0	1	1	1	1	
	0	0	0	0	0	1	1	1	1	1	
$k=1$	0	0	0	0	0	0	2	1	1	1	11.5
$k=2$	0	0	0	0	0	2	2	1	1	1	15.5
	0	0	0	0	2	0	2	1	1	1	
	0	0	0	2	0	0	2	1	1	1	
	0	0	2	0	0	0	2	1	1	1	
	0	2	0	0	0	0	2	1	1	1	
	2	0	0	0	0	0	2	1	1	1	

Table 2: Sub-optimal Control Strategies Under Different k.

Our objective is to find a control strategy that ensures the total probability of gene A being expressed is at least 0.8 (i.e., $\mathbf{x}_3 + \mathbf{x}_4 \geq 0.8$) with the minimum cost, given an initial state distribution of $\mathbf{x}_0 = (0.1, 0.4, 0.3, 0.2)^t$. The maximum numbers of times that Control 1 and Control 2 can be conducted are $K_1 = 5$ and $K_2 = 2$ respectively. The cost for conducting Control 1 is 2.5, the cost for Control 2 is 4, and no cost for Control 0. Table 2 lists the strategies obtained with minimum cost for each fixed k from 0 to $K_2 = 2$, where k is the number of times that Control 2 is conducted.

From Table 2, there is only one optimal control strategy: conduct Control 2 at time point $t = 7$ and Control 1 at time point $t = 8, 9, 10$, the total cost is 11.5 and the corresponding state probability distribution vector is $\mathbf{x}_T = (0.0275, 0.1682, 0.2679, 0.5364)^t$.

4 Concluding Remarks

In this paper, we introduce a new optimal finite-horizon control problem with multiple hard-constraints. Beyond this originality, we proposed an efficient algorithm to generate all the feasible solutions. An upper bound for the computational cost is also given. Our formulation can be applied to both perturbed and context-sensitive PBNs though we only test it with the instantaneously random PBNs. Since the control problem is NP-hard, in further research we will consider the control problem with multiple hard-constraints for large scale genetic networks. Extending our formulation to infinite-horizon is another future research topic.

Acknowledgement

Research supported by HKRGC Grant No. 7017/07P, HKU CRGC Grants and HKU Strategy Research Theme fund on Computational Sciences and HKU Hung Hing Ying Physical Science Research Grant.

References

[1] T. Akutsu, M. Hayasida, W, Ching and M. Ng. *Control of Boolean Networks: Hardness Results and Algorithms for Tree Structured Networks*, Journal of Theoretical Biology, 244: 670-679, 2007.

[2] J. E. Celis, M. Kruhøfferm I. Gromova, C. Frederiksen, M.Østergaard, T.. F. Ørntoft. *Gene Expression Profiling: Monitoring Transcription and Translation Products Using DNA Microarrays and Proteomics*, FEBS Lett. 480 (1) 2-16, 2000.

[3] P. C. Y. Chen and J. W. Chen, *A Markovian Approach to the Control of Genetic Regulatory Networks*. Biosystems, 90(2): 535-545, 2006.

[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, M. Ng and T. Akutsu. *An Approximation Method for Solving the Steady-state Probability Distribution of Probabilistic Boolean Networks*, Bioinformatics, 23: 1511-1518, 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] S. Zhang, W. Ching, N. Tsing, H. Leung and D. Guo, *A Multiple Regression Approach for Building Genetic Networks*, Proceedings of the International Conference on BioMedical Engineering and Informatics (BMEI2008) Sanya, China (in CD-ROM).

[8] W. Ching, S. Zhang, M. Ng and T. Akutsu. *An Approximation Method for Solving the Steady-state Probability Distribution of Probabilistic Boolean Networks*, Bioinformatics, 23: 1511-1518, 2007.

[9] A. Datta, A. Choudhary, M. Bitter, and E. R. Dougherty. *External Control in Markovian Genetic Regulatory Networks*, Machine Learning, 52 : 169-191, 2003.

[10] A. Datta, A. Choudhary, M. Bitter, and E. R. Dougherty. *External Control in Markovian Genetic Regulatory Networks: The Imperfect Information Case*, Bioinformatics, 20: 924-930, 2004.

[11] E. Dougherty, S. Kim and Y. Chen. *Coefficient of Determination in Nonlinear Signal Processing*, Signal Processing, 80: 2219-2235, 2000.

[12] S. Huang and D.E. Ingber. *Shape-dependent Control of Cell Growth, Differentiation, and Apoptosis: Switching Between Attractors in Cell Regulatory Networks*, Exp. Cell Res., 261: 91-103, 2000.

[13] S. Kauffman. *Metabolic Stability and Epigenesis in Randomly Constructed Gene Nets*, J. Theoret. Biol., 22: 437-467, 1969.

[14] S. Kauffman. *Homeostasis and Differentiation in Random Genetic Control Networks*, Nature, 224: 177-178, 1969.

[15] S. Kauffman. *The Large Scale Structure and Dynamics of Genetic Control Circuits: An Ensemble Approach*, J. Theoret. Biol., 44: 167-190, 1974.

[16] S. Kauffman. *The Origins of Order: Self-organization and Selection in Evolution*, New York: Oxford Univ. Press, 1993.

[17] 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.

[18] S. Huang. *Gene Expression Profiling, Genetic Networks, and Cellular States: An Integrating Concept for Tumorigenesis and Drug Discovery*, J. Mol. Med. , 77, 469-480, 1999.

[19] H. D. Jong. *Modeling and Simulation of Genetic Regulatory Systems: A Literature Review*, J. Comp. Biol., vol. 9, pp. 67-103, 2002.

[20] R. Pal, A. Datta, M. L. Bittner and E. R. Dougherty. *Intervention in context-sensitive probabilistic Boolean networks*, Bioinformatics, 21(7): 1211-1218, 2005.

[21] R. Pal, A. Datta and E. R. Dougherty. *Optimal Infinite Horizon Control for Probabilistic Boolean Networks* IEEE Transactions on Signal Processing, 54(6): 2375-2387, 2006.

[22] I. Shmulevich, E. Dougherty, S. Kim and W. Zhang. *Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks*, Bioinformatics, 18(2): 261-274, 2002.

[23] I. Shmulevich, E. Dougherty, S. Kim and W. Zhang. *Control of Stationary Behavior in Probabilistic Boolean Networks by Means of Structural Intervention*, Journal of Biological Systems, 10: 431-445, 2002.

[24] I. Shmulevich, E. Dougherty, S. Kim and W. Zhang. *From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks*, Proceedings of the IEEE, 90: 1778-1792, 2002.

[25] I. Shmulevich and E. Dougherty, *Genomic Signal Processing*, Princeton University Press, U.S. 2007.

[26] I. Shmulevich, E. R. Dougherty and W. Zhang. *Gene Perturbation and Intervention in Probabilistic Boolean Networks*, Bioinformatics, 18(10): 1319-1331, 2002.

[27] P. Smolen, D. Baxter and J. Byrne. *Mathematical Modeling of Gene Network*, Neuron, 26: 567-580, 2000.

[28] S. Zhang, W. Ching, M. Ng and T. Akutsu. *Simulation Study in Probabilistic Boolean Network Models for Genetic Regulatory Networks*, Journal of Data Mining and Bioinformatics, 1 :217-240, 2007.