Integer Programming and Dynamic Programming-based Methods of Optimizing Control Policy in Probabilistic Boolean Networks with Hard Constraints

Xi Chen ,[†1] Tatsuya Akutsu ,[†2] Takeyuki Tamura [†2]
and Wai-Ki Ching [†1]

Control problems of Boolean Networks (BNs) and Probabilistic Boolean Networks (PBNs) are studied in this paper. BN CONTROL is formalized to derive the network to the desired state within a few time steps by external control. PBN CONTROL is formalized to find a control sequence such that the network will terminate in the desired state with a maximum probability. Furthermore, we propose to minimize the maximum cost of the terminal state to which the network will enter. For solving the above problems, integer linear programming and dynamic programming-based methods in conjunction with hard constraints are developed. A hardness result suggesting that PBN CONTROL is harder than BN CONTROL is also presented.

1. Introduction

In bioinformatics, it is important to develop efficient algorithms for controlling genetic regulatory networks. Many formalisms have been developed for modeling genetic regulation processes, such as Bayesian networks, multivariate Markov chain model[6)], Boolean networks and probabilistic Boolean networks[10)]. Boolean networks (BNs) and their extension probabilistic Boolean networks (PBNs) have received much attention among all the models since they are able to capture the switching behavior of the genetic process.

In 1969, Kauffman firstly introduced Boolean network (BN)[11)]. BN is a very simple model: each gene is quantized to only two levels – on and off (represented as 1 and 0). Target genes are regulated by several genes called its input genes via its Boolean function (predictor function). If all the input genes and Boolean functions are given, a BN is said to be well defined. But a BN is a deterministic model. Randomness comes only from the initial state. From this reason, it is more realistic to extend a BN to a stochastic one, namely, Probabilistic Boolean Network (PBN). Instead of having only one Boolean function, each gene in a PBN can have multiple Boolean functions with selecting probabilities assigned to them. The dynamics of a PBN can be studied and analyzed by the theory of Markov chain. Moreover, it is possible to control one or more genes in a network such that the whole network is derived into a desired state or a steady-state distribution. Then we can develop therapeutic gene intervention or gene control policy[1),2),7)].

We propose in this paper to solve the control problem of PBNs by using integer linear programming and dynamic programming in conjunction with hard constraints. Kobayashi et al. applied an integer programming approach to solve the control problem of PBN[12)]. We consider adding hard constraints (i.e. adding an upper bound for the number of controls that can be applied to the network[7)]) into the problem and propose an integer linear programming based method with hard constraints to solve the control problem of BN and PBN. Introducing hard constraints is important for medical applications because the number of treatments such as radiation and chemo-therapy is usually limited[7)]. Furthermore, given the terminal cost for each state, we want to derive the network into the state with the minimized maximum cost by applying external control.

Besides development of algorithms, we study the time complexity of control problems for PBN. We prove both minimizing the maximum cost and minimizing the average cost are Σ_2^p-hard, where the latter problem corresponds to the original control problem for PBN[8)]. Note that control of BN is NP-complete[⋆1] and control of PBN is NP-hard[1)]. Because it is believed that Σ_2^p-hard problems are much harder than NP-complete problems[9)], this result suggests that control of PBN is much harder than control of BN. Moreover, this result suggests that such methods as integer linear programming cannot be effectively applied to solve the

[†1] AMAC Laboratory, Department of Mathematics, The University of Hong Kong
[†2] Bioinformatics Center, Institute for Chemical Research, Kyoto Univerty

[⋆1] Control of BN is NP-complete if the number of time steps is polynomially bounded. Otherwise, it is PSPACE-complete[3)]. However, it is not usual to consider an exponential number of time steps.

control problem of PBN because (a decision problem version of) integer linear programming is known to be NP-complete[9]. Therefore, integer programming-based approach can only be applied to control of BN[2] and special restricted variants of control of PBN[12].

2. Problems

2.1 Boolean Networks and Probabilistic Boolean Networks

A Boolean network (BN) is represented by a set of nodes (genes) $V = \{v_1, v_2, \ldots, v_n\}$ and a list of Boolean functions $F = \{f_1, f_2, \ldots, f_n\}$ where a Boolean function $f_i(v_{i_1}, \ldots, v_{i_k})$ with inputs from specified nodes $v_{i_1}, \ldots, v_{i_k}$ is assigned to v_i. We use $IN(v_i)$ to represent the set of input nodes $v_{i_1}, \ldots, v_{i_k}$ to v_i. The number of inputs to v_i is called the *indegree* of v_i. We define K as the *maximum indegree* of a BN.

We define $v_i(t)$ to be the state (0 or 1) of the gene i at time t. The rules of the regulatory interactions among the genes can then be represented by Boolean functions: $v_i(t+1) = f_i(v_{i_1}(t), \ldots, v_{i_k}(t)), i = 1, 2, \ldots, n$. Here we let $\mathbf{v}(t) = (v_1(t), v_2(t), \ldots, v_n(t))^T$ which is called the *Gene Activity Profile* (GAP). The GAP can take any possible states from the set $S = \{(v_1, v_2, \ldots, v_n)^T : v_i \in \{0, 1\}\}$ and thus totally there are 2^n possible states in the network. We then define $z(t) = 1 + \sum_{i=1}^{n} 2^{n-i} v_i(t)$. As $v_1(t)v_2(t)\ldots v_n(t)$ ranges from $00\ldots0$ to $11\ldots1$, $z(t)$ will take on all values from 1 to 2^n. Clearly, there is a one-to-one map from $x(t)$ to $z(t)$. Hence instead of the binary representation for the global state, one can use equivalent decimal representation $z(t)$.

To extend the concepts of a BN to a stochastic model, for each vertex v_i in a PBN, instead of having only one Boolean function as in BN, there are a multiple of Boolean functions (predictor functions) $f_j^{(i)}(j = 1, 2, \ldots, l(i))$ to be chosen for determining the state of gene v_i and usually $l(i)$ is not very large. The probability of choosing $f_j^{(i)}$ as the predictor function is $c_j^{(i)}$, $0 \le c_j^{(i)} \le 1$ and $\sum_{j=1}^{l(i)} c_j^{(i)} = 1$ for $i = 1, 2, \ldots, n$.

We let f_j be the jth possible realization, where $f_j = (f_{j_1}^{(1)}, f_{j_2}^{(2)}, \ldots, f_{j_n}^{(n)})$, $1 \le j_i \le l(i)$, $i = 1, 2, \ldots, n$. Suppose that the selection of the Boolean function f_{j_i} for each gene i is an independent process, then the probability of choosing the corresponding BN with Boolean functions $f_j = (f_{j_1}^{(1)}, f_{j_2}^{(2)}, \ldots, f_{j_n}^{(n)})$ is given by $q_{j_1 j_2 \cdots j_n} = \prod_{i=1}^{n} c_{j_i}^{(i)}$. There are at most $N = \prod_{i=1}^{n} l(i)$ different possible realizations of BNs. Let $\mathbf{a}$ and $\mathbf{b}$ be any two column vectors in the set S. Then the transition probability $\mathrm{P}\ \{\mathbf{v}(t+1) = \mathbf{a} \mid \mathbf{v}(t) = \mathbf{b}\} = \sum_{j=1}^{N} \mathrm{P}\ \{\mathbf{v}(t+1) = \mathbf{a} \mid \mathbf{v}(t) = \mathbf{b}, \text{the } j\text{th BN is selected}\} \cdot q_j = \sum_{j \in \mathcal{I}} q_j$ where $\mathcal{I}$ is the set of BNs of which the transition probability from state $\mathbf{b}$ to state $\mathbf{a}$ is 1. Here we let $q_j = q_{j_1 j_2 \cdots j_n}$ and $j = j_1 + \sum_{i=2}^{n} \left((j_i - 1)(\prod_{k=1}^{i-1} l(k)) \right)$. We can then use both of them when there is no confusion.

2.2 Control of BN with Hard Constraints

In BN CONTROL, there are two types of nodes: *internal nodes* and *external nodes*, where internal nodes correspond to usual nodes in a BN and external nodes correspond to control nodes. Let a set V of $n+m$ nodes be $V = \{v_1, \ldots, v_n, v_{n+1}, \ldots, v_{n+m}\}$, where $v_1, \ldots, v_n$ are internal nodes and $v_{n+1}, \ldots, v_{n+m}$ are control nodes. Then the states of internal nodes at time $t+1$ are represented by $v_i(t+1) = f_i(v_{i_1}(t), \ldots, v_{i_k}(t))$, $i = 1, 2, \ldots, n$. where each v_{i_j} is either an internal node or a control node. Here we let $\mathbf{v}(t) = [v_1(t), v_2(t), \ldots, v_n(t)]$ and $\mathbf{u}(t) = [v_{n+1}(t), v_{n+2}(t), \ldots, v_{n+m}(t)]$. If $v_{n+i}(t) - v_{n+i}(t+1) \neq 0$, for some $i \in \{1, \ldots, m\}$, then we say that the external control is applied once to the network. Thus the number of controls applied to network is equal to $\sum_{t=0}^{M-1} \sum_{i=1}^{m} |v_{n+i}(t) - v_{n+i}(t+1)|$. Then the control problem of BN under hard constraints is as follows:

Definition 1: Suppose an initial state of the network is $\mathbf{v}^0$ and the desired state of the network is $\mathbf{v}^M$, find a control sequence $\langle \mathbf{u}(0), \mathbf{u}(1), \ldots, \mathbf{u}(M) \rangle$ such that $\mathbf{v}(0) = \mathbf{v}^0$ and $\mathbf{v}(M) = \mathbf{v}^M$, and the maximum number of controls applied to the network during the finite time period M is H.

2.3 Finding the Optimal Path with Hard Constraints

In a PBN, for each time step t, the network will choose one of the possible BNs (e.g., j_t-th possible BN) with the corresponding selecting probability q_{j_t} and enter into the next state $\mathbf{v}(t+1)$ from $\mathbf{v}(t)$. Given the initial state $\mathbf{v}^0$ and the desired state $\mathbf{v}^M$, we can define the probability of a path with $\mathbf{v}(0) = \mathbf{v}^0$ and $\mathbf{v}(M) = \mathbf{v}^M$ as $\prod_{t=0}^{M-1} q_{j_t}$. By applying external control to the network, we can derive the network into desired state $\mathbf{v}^M$ with different path probabilities. Then the problem of maximizing the highest probability of a path with the initial state $\mathbf{v}^0$ and the terminal (desired) state $\mathbf{v}^M$ can be described as follows:

Definition 2: Suppose an initial state of the network is $\mathbf{v}^0$ and the desired state of the network is $\mathbf{v}^M$, find a control sequence $\langle \mathbf{u}(0), \ldots, \mathbf{u}(M) \rangle$ such that the probability of the path with the initial state $\mathbf{v}^0$ and the terminal (desired) state $\mathbf{v}^M$ is maximized, and the maximum number of controls applied to the network during the finite time period M is H.

2.4 Minimizing the Maximum Cost

Suppose that a PBN with n internal nodes $\mathbf{v}(t) = [v_1(t), v_2(t), \ldots, v_n(t)]$ and m control nodes $\mathbf{u}(t) = [v_{n+1}(t), v_{n+2}(t), \ldots, v_{n+m}(t)]$. Let $z_t = 1 + \sum_{i=1}^{n} 2^{n-i} v_i$ which is the state of network at time step t, and $u_t = 1 + \sum_{i=1}^{m} 2^{m-i} v_{n+i}$ which is the control input of network at time step t. In a PBN, even if the network starts with the given initial state $z(0)$, the subsequent states will be random since the PBN is a stochastic model. That is, the terminal state z_M could take any possible values from 1 to 2^n. We assign a terminal cost $C_M(z_M)$ to each of states z_M at time step M. Note that, depending on the particular PBN and the control input used in each step, it is possible that the network can not enter some of the states at time step M. We define $C_t(z_t)$ as the maximum cost of which, beginning from z_t at time step t, the network can reach at the terminal time step. The problem of minimizing the maximum cost can be described as follows:

Definition 3: Given the terminal cost $C_M(z_M)$ for each of states $z_M \in \{1, 2, \ldots, 2^n\}$ at time step M, by applying external control, minimize the maximum cost $C_0(z_0)$ beginning from the given initial state z_0, and the maximum number of controls applied to the network is H.

3. Algorithms

3.1 ILP with Hard Constraints for BN CONTROL

Let $x_{i,t}$ represent the Boolean value $v_i(t)$. Define

$$\sigma_b(x) = \begin{cases} x, \text{ if } b = 1. \\ \bar{x}, \text{otherwise.} \end{cases} \tag{1}$$

Then any Boolean function $f_i(x_{i_1,t}, \ldots, x_{i_k,t})$ is equivalent to $f_i(x_{i_1,t}, \ldots, x_{i_k,t}) = \bigvee_{b_{i_1} \ldots b_{i_k} \in \{0,1\}^k} \{f_i(b_{i_1}, \ldots, b_{i_k}) \wedge \sigma_{b_1}(x_{i_1,t}) \wedge \cdots \wedge \sigma_{b_k}(x_{i_k,t})\}$. Then we define binary variable $h_{i,t} \in \{0, 1\}$ $(i = n+1, \ldots, n+m)$ as the node control variable. If $h_{i,t} = 1$, we say the node i changes its value at time step t. Since the maximum number of controls applied to the network during the finite time period is H, we have $\sum_{t=0}^{M-1} \sum_{i=n+1}^{n+m} h_{i,t} \leq H$. Also, we define $\tau_b(x)$ as

$$\tau_b(x) = \begin{cases} x, \text{ if } b = 1. \\ 1 - x, \text{ otherwise.} \end{cases} \tag{2}$$

Then the ILP-Formulation for the BN CONTROL based on the method of[2] is to maximize $\sum_{i=1}^{N} x_{i,M}$. Methods of representing constraints are shown in Chen et al.[5].

3.2 ILP with Hard Constraints for PBN CONTROL

To extend the above ILP formulation for PBN CONTROL, we define $y_{r,t}$ as the selection variable. If $y_{r,t} = 1$, we say the rth BN is selected at time step t. Otherwise, we say it is not selected at time step t. Then we have $\sum_{r=1}^{R} y_{r,t} = 1$, for $t = 1, 2, \ldots, M-1$. Here R is the total number of possible realizations for the PBN. Define $f_{i,r}$ as the Boolean function for node v_i when the r-th BN is selected. Let $P = (p_1, p_2, \ldots, p_R)$ be the selecting probabilities for the R possible realizations. Then the objective function for the PBN control with hard constraints is to maximize $\sum_{t=0}^{M-1} \sum_{r=1}^{R} -\log(p_r) \cdot y_{r,t}$. Details are shown in Chen et al.[5].

3.3 Minimizing the Maximum Cost

Define $J(z_t, h_t)$ as the minimized maximum terminal cost $C_M(z_M)$ when the state is z_t, and the remaining number of external controls is h_t, at time step t. Define $u(z_t, h_t)$ as the control function when the state is z_t and the remaining number of external controls is h_t at time step t. Let $F(z_t, u_t)$ be the set of states at time step $t+1$ that can be reached from z_t with control u_t. Then dynamic programming for the PBN control with hard constraints is as follows:

Step 0: Set $t = M$; $J(z_M, h_M) = C_M(z_M)$ for all $h_M = \{0, \ldots, H\}$.
Step 1: $t := t - 1$.
Step 2: For any $z_t \in \{1, \ldots, 2^n\}$ and $h_t \in \{0, \ldots, H\}$, compute

$$J(z_t, h_t) = \min_{u_t \in \{1, \ldots, 2^m\}} \begin{cases} \max\limits_{z_{t+1} \in F(z_t, u_t)} J(z_{t+1}, h_t), \text{ if } u_t = u(z_{t+1}, h_t), \\ \max\limits_{z_{t+1} \in F(z_t, u_t)} J(z_{t+1}, h_t - 1), \text{ otherwise.} \end{cases}$$

and

$$u(z_t, h_t) = \mathrm{argmin}_{u_t \in \{1,\ldots,2^m\}} \begin{cases} \max\limits_{z_{t+1} \in F(z_t,u_t)} J(z_{t+1}, h_t), & \text{if } u_t = u(z_{t+1}, h_t), \\ \max\limits_{z_{t+1} \in F(z_t,u_t)} J(z_{t+1}, h_t - 1), & \text{otherwise.} \end{cases}$$

Step 3: If $t > 0$, go back to step 1; Otherwise, stop.

In the above, $u_t \neq u_{t+1}$ is counted as one control where we need to modify the algorithm for the case that the number of controls is defined as before. Finally, we take $\min_{h_0 \in \{0,\ldots,H\}} J(z_0, h_0)$ for computing the minimized maximum cost.[⋆1]

4. Complexity analysis

We give some analysis on the complexity of minimizing the maximum cost and minimizing the average cost in this section. We assume that a PBN is not given in the matrix form but in the form of $f_j^{(i)}$s and $c_j^{(i)}$s because A is of exponential size and thus it is almost meaningless to discuss the time complexity if we use A. Furthermore, we assume that it is only required to output $\mathbf{u}(0)$ for given z_0 and PBN (otherwise, we should output $\mathbf{u}(t)$s for an exponential number of GAPs). Then, we can keep both the sizes of input and output polynomial of n and thus can discuss the time complexity with respect to the network size. Moreover, we assume that the number of time steps (i.e., M) is polynomially bounded. Otherwise, both BN CONTROL and PBN CONTROL would be PSPACE-hard[3),4)]. Because it is not realistic to consider an exponential number of time steps, this is a reasonable assumption. Although details are omitted, we obtained following theoretical results[5)].

Theorem 1 Minimizing the maximum cost in control of PBN is $\sum_2^p$-hard.

Theorem 2 Minimizing the average cost in control of PBN is $\sum_2^p$-hard.

5. Conclusion

ILP-based methods for control of BN and for finding an optimal path for PBN are presented both with hard constraints. We have also presented a DP-based method for finding a control policy that minimizes the maximum cost for PBN under hard constraints, where it uses exponential size tables. The hardness results suggest that ILP cannot be effectively applied to minimization of the maximum or average cost for PBN.

[⋆1] We need to modify the algorithm if there exist multiple u_ts giving the minimum cost.

References

1) T.Akutsu, M.Hayashida, W.-K. Ching, and M.Ng, "Control of Boolean networks: Hardness results and algorithms for tree structured networks," *Journal of Theoretical Biology*, vol. 244, pp. 670–679, 2007.
2) T.Akutsu, M.Hayashida, and T.Tamura, "Integer programming-based methods for attractor detection and control of Boolean networks," in *Proc. the Combined 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference*, 2009, pp. 5610–5617.
3) C.L. Barrett, H.B. Hunt III, M.V. Marathe, S.S. Ravi, D.J. Rosenkrantz, and R.E. Stearns, "Complexity of reachability problems for finite discrete dynamical systems," *Journal of Computer and System Sciences*, vol.72, pp. 1317–1345, 2006.
4) C.L. Barrett, H.B. Hunt III, M.V. Marathe, S.S. Ravi, D.J. Rosenkrantz, R. E. Stearns, and M.Thakur, "Computational aspects of analyzing social network dynamics," in *Proc. 20th International Joint Conference on Artificial Intelligence*, 2007, pp. 2268–2273.
5) X. Chen, T. Akutsu, T. Tamura and W-K. Ching, "Finding optimal control policy in probabilistic Boolean networks with hard constraints by using integer programming and dynamic programming," IEEE International Conference on Bioinformatics and Biomedicine 2010 (BIBM 2010), pp. 240–246, 2010.
6) W.Ching, E.Fung, M.Ng, and T.Akutsu, "On construction of stochastic genetic networks based on gene expression sequences," *Journal of Neural Systems*, vol.15, pp. 297–310, 2005.
7) W.Ching, S.Zhang, Y.Jiao, T.Akutsu, and N.Tsing, "Optimal control policy for probabilistic Boolean networks with hard constraints," *IET Systems Biology*, vol.3, pp. 90–99, 2009.
8) A.Datta, A.Choudhary, M.L. Bittner, and E.R. Dougherty, "External control in Markovian genetic regulatory networks," *Machine Learning*, vol.52, pp. 169–191, 2003.
9) M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness.* NY: W. H. Freeman and Company, 1979.
10) H.deJong, "Modeling and simulation of genetic regulatory systems: A literature review," *Journal of Computational Biology*, vol.9, pp. 69–103, 2002.
11) S.A. Kauffman, *The Origins of Order: Self-organization and Selection in Evolution.* NY: Oxford Univ. Press, 1993.
12) K.Kobayashi and K.Hiraishi, "An integer programming approach to control problems in probabilistic boolean networks," in *Proc. 2010 American Control Conference*, pp. 6710–6715, 2010.