SPARSE BIOLOGICALLY-CONSTRAINED OPTIMAL PERTURBATION OF GENE REGULATORY NETWORKS

Haoyu Wang[1], Nidal Bouaynaya[2], Member, IEEE, Roman Shterenberg[3], and Dan Schonfeld[4], Fellow, IEEE

[1,4]Department of Electrical and Computer Engineering, University of Illinois at Chicago, USA
[2]Department of Systems Engineering, University of Arkansas at Little Rock, USA
[3]Department of Mathematics, University of Alabama at Birmingham, USA

ABSTRACT

This paper derives a sparse optimal perturbation of gene regulatory networks by determining the optimal perturbation of the minimal number of individual genes that force the network to settle into desired equilibrium states. Previous efforts have led to intervention in gene regulatory networks by deriving the optimal perturbation of the state probability transition matrix. Current technology in molecular biology, however, is limited to perturbation of the state of individual genes, not the state probability transition matrix. Our computer simulation experiments on the Human melanoma gene regulatory network demonstrate the superiority of the proposed approach to gene regulation in comparison to the previous methods based on the marginal of the optimal perturbation of the probability transition matrix of the network.

1. INTRODUCTION

1.1. Motivation

The biological mechanisms that govern our development are amazingly integrated and complex. They regulate the expression of thousands of genes and proteins in any given cellular function. The spatial and temporal interactions of these genes and proteins encode the developmental processes of the cell. Understanding these genomic regulatory networks can substantially enrich our knowledge of health and disease. Subsequently, designing intervention strategies to control the behavior of these networks and reach desirable cellular states lies at the heart of modern therapeutic methods. Moreover, such an optimal intervention strategy must be biologically-viable given the current technology in molecular biology, where only dowregulation or upregulation of gene expression levels can be experimentally implemented.

1.2. Related Work

Previous control models, discussed in the literature, relied mainly on the classical theory of optimal stochastic control in engineered systems, where external inputs are injected into the system through some known targets in order to optimize a specific objective or cost function [1]. The resulting control policy is iterative and does not guarantee (at least for the finite-horizon control) that the steady-state dynamics of the network have actually changed [2–9]. Once the external input is withdrawn, the network is prone to go back to its original (undesirable) steady-state. Various heuristic interventions, which alleviate the computational burden of the optimal stochastic control and guide the time evolution of the network in a heuristic manner, have been proposed [10], [11], [12], [13], [14], [15].

In biological control, it is essential to be able to control the underlying rules of the network in order to alter its long-run or steady-state behavior. There is increasing evidence that steady-states of biological systems, particularly, genomic regulatory networks, correspond to phenotypic characteristics, such as cell proliferation and apoptosis [16]. In particular, disease and health can be mathematically modeled as steady-states of corresponding genomic regulatory networks. In contrast to engineered systems, where cost and minimal error are the main control variables, the effective biological objective function should be related to the steady-state behavior of the genomic network. This suggests a different control framework for biological intervention from traditional control in engineering systems [1].

Bouaynaya *et al.* formulated optimal perturbation in gene regulation as a solution to an inverse perturbation problem, which finds the required perturbation in order to reach a desired stationary state [17], [18]. The solution to the inverse perturbation problem, casted as a strictly convex optimization problem, is demonstrated to be unique, globally optimum, and non-iterative. In particular, it can be solved efficiently using standard convex optimization methods, [17], [18]. However, the optimal perturbation control framework in [17], [18] is formulated in terms of probability transition matrices of network states and not in terms of perturbations of the gene expression levels. Given current biotechnology techniques, it is only possible to experimentally control gene expression levels. Thus, biological design rules are still needed to trans-

This project is supported by Award Number $R01GM096191$ from the National Institute Of General Medical Sciences (NIH/NIGMS).

late perturbations at the network state levels into actual perturbations of the gene expression levels.

1.3. Main Contributions

Recent efforts have led to intervention in gene regulatory networks by deriving the optimal perturbation of the state probability transition matrix. Current technology in molecular biology, however, is limited to perturbation of the state of individual genes, not the state probability transition matrix. The previous efforts therefore relied on the optimal perturbation of the state probability transition matrix to compute a perturbation of the state of individual genes by determining the marginal of the perturbed state probability transition matrix. However, there is no guarantee that the resulting perturbation of the state of individual genes is optimal. Indeed, it is possible that other perturbations of the state of individual genes exist which may lead to a superior performance in gene regulation. In this paper, we provide a sparse biologically-viable optimal perturbation of gene regulatory networks by determining the optimal perturbation of the minimal number of individual genes that are consistent with the desired state distribution. We conduct computer simulation experiments that demonstrate the superiority of the proposed approach to gene regulation in comparison to the previous methods based on the marginal of the optimal perturbation of the state probability transition matrix.

2. THE OPTIMAL GENE PERTURBATION PROBLEM

2.1. Markov Chain Dynamics

We consider a network with p nodes (here genes), where the expression level of each gene is quantized to l values. Kim et *al.* showed that the dynamics of the network can be modeled by a homogeneous Markov chain with probability transition matrix (p.t.m.) $P \in \mathbb{R}^{n\times n}$, where $n = l^p$ [19]. A probability vector $\boldsymbol{\pi} = (\pi_1, ..., \pi_n)^t$ is called a *steady-state* distribution or a *stationary* distribution of P_0 if $\boldsymbol{\pi}^t P = \boldsymbol{\pi}^t$. Since P is stochastic, stationary distributions always exist.

If the probability transition matrix P is irreducible (i.e., all states communicate with each other) and aperiodic, it is called *ergodic*. A network governed by an ergodic p.t.m. converges to a unique, strictly positive stationary distribution π, in the following sense:

$$\lim_{n\to\infty} P^n = \mathbf{1}\pi^t. \tag{1}$$

It is important to notice that uniqueness of the stationary distribution does not imply convergence in the sense defined by (1). In fact, a network may admit a unique stationary distribution but fails to converge to it [18]. Ensuring convergence towards the steady-state distribution is essential in the framework of optimal control of genomic networks. Merely reshaping the steady-state landscape of the network is not enough, as the network may not settle in the long-run in the stationary distribution. It can be shown that a necessary and sufficient condition for convergence of the p.t.m. P towards its unique steady-state distribution is that its Second Largest Eigenvalue Modulus (SLEM) is strictly less than unity [18].

2.2. The Feasible Control

We consider the scenario where the original network, governed by the p.t.m. P_0, admits at least one undesirable steady-state distribution. We would like to linearly perturb the p.t.m. P_0 so that the perturbed matrix converges to the unique desired steady-state distribution. Let us write the perturbed probability transition matrix P as $P = P_0 + C$, where C is a zero-row sum perturbation matrix. The zero row-sum condition ensures that the perturbed matrix P is stochastic. We denote by $\boldsymbol{\pi}_d$ the desired stationary distribution. The goal is, therefore, to design a perturbation C, which ensures convergence of the perturbed network towards the unique desired distribution $\boldsymbol{\pi}_d$. A feasible perturbation matrix, C, must satisfy the followings four constraints:

$$(i) \quad \boldsymbol{\pi}_d^t(P_0 + C) = \boldsymbol{\pi}_d^t; \;\; (ii) \quad C\mathbf{1} = \mathbf{0};$$
$$(iii) \quad P_0 + C \geq \mathbf{0}; \;\; (iv) \quad \text{SLEM}(P_0 + C) < 1.$$

Constraint (i) implies that $\boldsymbol{\pi}_d$ is a stationary distribution of the perturbed matrix $(P_0 + C)$ (not necessarily unique). Constraint (ii) is equivalent to the stochasticity of the p.t.m. P. Constraint (iii) is an elementwise inequality and simply ensures that all entries of P are non-negative. Constraint (iv) implies that the stationary distribution $\boldsymbol{\pi}_d$ is unique and the perturbed matrix, P, will converge towards it. Let us denote by $\mathscr{F}$ the set of matrices satisfying constraints (i) though (iv), i.e., $\mathscr{F} = \{C \in \mathbb{R}^{n*n} : \pi_d^t(P_0 + C) = \pi_d^t, C\mathbf{1} = \mathbf{0}, P_0 + C \geq \mathbf{0}, \text{SLEM}(P_0 + C) < \mathbf{1}\}$. It is easy to check that $(1\pi_d^t - P_0) \in \mathscr{F}$ and thus, the feasible set $\mathscr{F} \notin \varnothing$. Therefore, there exists at least one perturbation, which forces the network to converge towards the desired steady-state.

2.3. The Gene Optimal Perturbation Control

The Markov probability transition matrix, describing the dynamics of the network at the state level, is related to the actual gene network by observing that the probability law describing the genes' dynamics can be obtained as the marginal distribution of the state transition probabilities:

$$\Pr(g_i = x_i | g_1 \cdots, g_m) = \tag{2}$$
$$\sum_{\tilde{x}_i} \Pr(g_1 = x_1, \cdots, g_m = x_m | g_1 \cdots, g_m),$$

where $\tilde{x}_i$ denotes the set of all x_j's except x_i. We define the gene network matrix, G, as the matrix whose entries are the conditional probabilities of the individual genes expression levels given the current network state, i.e., given the expression levels of all other genes. We order the columns of

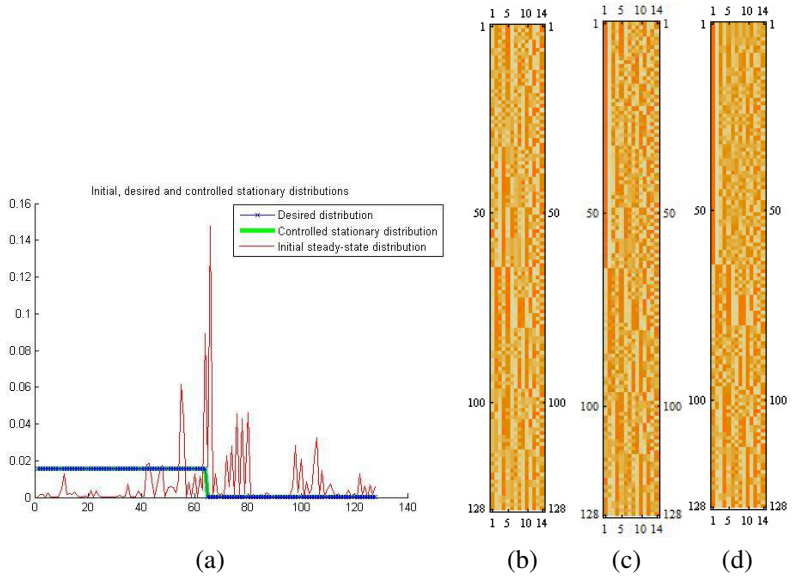

Fig. 1: Optimal gene perturbed matrix of the Human melanoma gene regulatory network. The matrix plots are obtained using the function *MatrixPlot* in MATHEMATICA. They provide a visual representation of the values of elements in the matrix. The color of entries varies from white to red corresponding to the values of the entries in the range of 0 to 1. (a) The initial steady-state distribution (red) of the melanoma network, the desired steady-state distribution (blue) and the controlled steady-state distribution (green). (b) The original Human melanoma gene network matrix, $\boldsymbol{G}_0$; (c) The optimal melanoma gene network matrix, $\boldsymbol{G}^*$, corresponding to the steady-state distribution $\boldsymbol{\pi}_d$; (d) The melanoma gene network matrix obtained as a marginalization of the perturbation in [17] corresponding to the same steady-state distribution.

G such that the first l columns indicate the probabilities of gene $g_1 = 0, g_1 = 1, \cdots, g_1 = l-1$, respectively, given the network states; the next l columns provide the probabilities of gene $g_2 = 0, 1, \cdots, l$ given the network states, and so on. For instance, for a binary quantization ($l = 2$), we have $G(1,1) = \Pr(g_1 = 0 | g_1 = 0, g_2 = 0) = \Pr(g_1 = 0|00)$ and $G(1,2) = \Pr(g_1 = 1|00)$. Formally, the gene network matrix, for an l-quantization level is defined as

$$G(i, l(j-1)+k+1) = \Pr(g_j = k | \text{state } i), \tag{3}$$

$i \in \{1, \cdots, l^m\}, j \in \{1, \cdots, l \times m\}, k \in \{0, 1, \cdots, l-1\}$

Given an original gene network matrix G_0, the optimal perturbed gene expressions correspond to the "closest", in the Hamming sense, matrix G which satisfies the constrained dynamics of conditions $(i) - (iv)$ given in Section 2.1. The Hamming "norm" [1] is defined as:

$$\|G_o, G_p\|_H = \sum_{j=1}^{m} N(g_j), \tag{4}$$

where $N(g_j)$ is given by:

$$N(g_j) = \sum_{i=0}^{n} \sum_{k=0}^{l-1} p(i)\; h(G_0(i, l*(j-1)+k+1), \\ G(i, l*(j-1)+k+1)). \tag{5}$$

[1] The Hamming function is not a mathematical norm. However, by abuse of notation, we say *Hamming norm* and *Hamming distance*.

$p(i)$ is the probability of state i given by the desired stationary distribution $\boldsymbol{\pi}_d$ and the function $h()$ is the hamming "distance" defined as

$$h(x,y) = \begin{cases} 1, & \text{if } |x-y| \geq \varepsilon \\ 0, & \text{if } |x > y| < \varepsilon \end{cases} \tag{6}$$

with ε being a specified error tolerance threshold. For binary quantization ($l = 2$), $N(g_j)$ is the expected number of flips of gene j after perturbation. The optimization problem in (4) finds the perturbation matrix, which causes the minimum number of flips (for binary quantization) before and after control. $\|G_0, G_p\|_H$ is the expected number of flips of all genes in the network before and after control.

We still need to satisfy the steady-state constraints on the network to ensure that it converges to the desired steady-state distribution. In order to do so, we must formulate the constraints $(i) - (iv)$ in terms of the gene network matrices G. Let $b(j-1)$ be the binary representation of the number (j-1) using n bits. Then, $b(j-1)[1]$ denotes the most significant bit and $b(j-1)[n]$ denotes the least significant bit. Under the assumption of independence of gene expression levels, we have

$$P_{ij} = \prod_{k=1}^{n} G(i, l(k-1) + b(j-1)[k] + 1). \tag{7}$$

The optimization problem resulting from the Hamming distance objective can be shown to be an NP hard combinatorial problem. In fact, the Hamming "norm" is equivalent to

Table 1: l_1 distances between the gene matrices in Fig. 1

l_1 distance	G_0	G^*	G in [17]
G_0	0	192.054	289.601
G^*	192.054	0	196.580
G in [17]	289.601	196.580	0

the l_0 norm. We, therefore, propose to approximate the l_0 or "Hamming norm" by the convex l_1 norm. The optimization problem becomes then one of minimizing $\|G_0 - G\|_1$ subject to the same constraints. The optimal gene network perturbation problem can then be formulated as

$$\begin{aligned}
&\text{Minimize } \|G_0 - G\|_1 \quad \text{subject to} \qquad (8)\\
&\sum_{i=1}^{2^n} \boldsymbol{\pi}_d(i) \prod_{k=1}^{n} G(i, l(k-1) + b(j-1)[k] + 1) = \boldsymbol{\pi}_d(j),\\
&j = 1, \cdots, 2^n.\\
&\sum_{j=1}^{2^n} \prod_{k=1}^{n} G(i, l(k-1) + b(j-1)[k] + 1) = 1,\ i = 1, \cdots, 2^n.\\
&G \geq 0.
\end{aligned}$$

Observe that (8) finds the optimal gene perturbation in the closure, $\bar{\mathscr{F}}$, of the feasible set $\mathscr{F}$, i.e., $\bar{\mathscr{F}}$ contains matrices satisfying constraints (i)-(iii) because SLEM$(P) \leq 1$ for all stochastic matrices P. If the optimal gene perturbation satisfies SLEM$(P) < 1$, then the network is forced to converge towards the desired steady-state distribution. Otherwise, the optimal solution is at the boundary $\partial\mathscr{F}$ of the feasible set; thus, the steady-state landscape of the network is modified to include the desired steady-state but the network does not converge to the desired equilibrium. One can, however, construct suboptimal solutions as proposed in [18].

Though the l_1 norm is convex, the constraints are nonlinear in the unknown and thus the problem is not convex and multiple local solutions exist. We use the interior-point method to solve this non-convex optimization problem, with a starting point given by the solution proposed in [17].

3. APPLICATION TO THE HUMAN MELANOMA GENE REGULATORY NETWORK

We consider the 7-gene Human melanoma gene regulatory network [2]. Upregulation of the gene WNT5A was found to be associated with the metastatic competence of cells. This implies that a system-level intervention that downregulates WNT5A while appropriately regulating the other genes could be a used as a molecular intervention against melanoma [20], [2], [21].

The human melanoma gene network was modeled as probabilistic Boolean network with seven gene: WNT5A, pirin, S100P, RET1, MART1, HADHB and STC2 [22]. Therefore, with the assumption that the expression level of each gene is either up (1) or down (0), the melanoma network has $2^7 = 128$ states ranging from 0000000 to 1111111. The gene network matrix is 128×14 and specifies the probability that each gene is up or down given the expression levels of the other genes. In the binary representation, the 7 genes are ordered as WNT5A, pirin, S100P, RET1, MART1, HADHB and STC2, i.e., the most significant bit corresponds to the expression level of WTN5A and the least significant bit is STC2.

In order to appropriately downregulate the WNT5A gene, we should assign a zero or near-zero probability of the steady-state states 64 to 127, which correspond to an upregulated level of WNT5A. In the computer simulations, we considered a desired steady-state distribution, π_d, where states 64-127 are assigned probability 10^{-4}, and states 0 to 63 have a uniform probability mass equal to 0.015525. The original and desired steady-state distributions are displayed in Fig. 1(a).

We compare our optimal gene perturbed matrix G^*, obtained as a solution of the optimization problem in (8), with the corresponding gene perturbation matrix G obtained by marginalizing the perturbed matrix in [17]. Observe that both matrices G^* and G satisfy constraints $(i)-(iv)$ and thus force the network dynamics to settle into the desired steady-state distribution. We found that SLEM $(P^*) = 0.57 < 1$, where P^* is the probability transition matrix corresponding to the optimal gene matrix G^*. Moreover, G^* is closer to the original melanoma matrix G_0 than G in [17]. MATHEMATICA plots of all gene network matrices are displayed in Fig. 1. The l_1 distances between the gene matrices are displayed in Table 1.

4. CONCLUSION

In this paper, we proposed a sparse biologically-viable optimal perturbation of gene regulatory networks in order to force the network to settle into a desirable equilibrium. The proposed perturbation affects the gene expression levels rather than the network states. The optimality criterion is defined in terms of minimizing the number of perturbations to achieve desired steady-state dynamics. This is equivalent to the sparsity of the perturbation. Deriving optimal interventions that minimally deviate from the original undesirable network is crucial in order to minimize adverse effects of the intervention. The proposed optimal perturbation is applied to the Human melanoma gene regulatory network and is shown to yield gene expression perturbations that are smaller than previous methods based on the marginal of the optimal perturbation of the state probability transition matrix.

References

[1] Dimitri P. Bertsekas, *Dynamic Programming and Optimal Control*, Athena Scientific, 3rd edition edition, 2007.

[2] A. Datta, R. Pal, A. Choudhary, and E.R. Dougherty, "Control approaches for probabilistic gene regulatory networks-what approaches have been developed for addreassinig the issue of intervention?," *Signal Processing Magazine, IEEE*, vol. 24, no. 1, pp. 54–63, 2007.

[3] Aniruddha Datta, Ashish Choudhary, Michael L. Bittner, and Edward R. Dougherty, "External control in markovian genetic regulatory networks," *Machine Learning*, vol. 52, pp. 169–191, 2003.

[4] O Abul, R Alhaj, and F Polat, "Markov decision processes based optimal control policies for probabilistic Boolean networks," in *IEEE Symposium on Bioinformatics and Bioengineering*, May 2004, pp. 337 – 344.

[5] R. Pal, A. Datta, and E. Dougherty, "Optimal infinite horizon control for probabilistic Boolean networks," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2375–2387, 2006.

[6] T Akutsua, M Hayashida, W-K Ching, and M K Ng, "Control of Boolean networks: Hardness results and algorithms for tree structured networks," *Journal of Theoretical Biology*, vol. 244, no. 4, pp. 670–679, February 2007.

[7] Babak Faryabi, Golnaz Vahedi, Jean-Francois Chamberland, Aniruddha Datta, and Edward R. Dougherty, "Optimal constrained stationary intervention in gene regulatory networks," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2008, May 2008.

[8] W-K Ching, S-Q Zhang, Y Jiao, T Akutsu, N-K Tsing, and A S Wong, "Optimal control policy for probabilistic Boolean networks with hard constraints," *IET Systems Biology*, vol. 3, no. 2, pp. 90–99, March 2009.

[9] C Yang, C Wai-Ki, T Nam-Kiu, and L Ho-Yin, "On finite-horizon control of genetic regulatory networks with multiple hard-constraints," *BMC Systems Biology*, vol. 4, no. Suppl 2, 2010.

[10] N. Ghaffari, I. Ivanov, X. Quian, and E. Dougherty, "A CoD-based reduction algorithm preserving the effects of stationary control policies for Boolean networks," *Bioinformatics*, vol. 26, no. 12, pp. 1556–1563, 2010.

[11] Mehmet Tana, Reda Alhajj, and Faruk Polata, "Scalable approach for effective control of gene regulatory networks," *Artificial Intelligence in Medicine*, vol. 48, no. 1, pp. 51–59, January 2010.

[12] X. Qian, N. Ghaffari, I. Ivanov, and E.R. Dougherty, "State reduction for network intervention with probabilistic Boolean networks," *Bioinformatics*, vol. 26, no. 24, pp. 3098–3104, 2010.

[13] M Tan, R Alhajj, and F Polat, "Automated large-scale control of gene regulatory networks," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 2, pp. 286 – 297, April 2010.

[14] G. Vahedi, B. Faryabi, J.F. Chamberland, A. Datta, and E.R. Dougherty, "Intervention in gene regulatory networks via a stationary mean-first-passage-time control policy," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 10, pp. 2319 – 2331, October 2008.

[15] Xiaoning Qian, Ivan Ivanov, Noushin Ghaffari, and Edward R Dougherty, "Intervention in gene regulatory networks via greedy control policies based on long-run behavior," *BMC Systems Biology*, vol. 3, no. 61, pp. 1–16, June 2009.

[16] S. Huang, "Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery," *Journal of Molecular Medicine*, vol. 77, no. 6, pp. 469–80, June 1999.

[17] N. Bouaynaya, R. Shterenberg, and D. Schonfeld, "Inverse perturbation for optimal intervention in gene regulatory networks," *Bioinformatics*, vol. 27, no. 1, pp. 103–110, 2011.

[18] N. Bouaynaya, M. Rasheed, R. Shterenberg, and D. Schonfeld, "Intervention in general topology gene regulatory networks," in *IEEE International Workshop on Genomic Signal Processing and Statistics*, 2011, pp. 222–225.

[19] S. Kim, H. Li, E. R. Dougherty, N. Chao, Y. Chen, M. L. Bittner, and E. B. Suh, "Can Markov chain models mimic biological regulation," *Biological Systems*, vol. 10, no. 4, pp. 447–458, 2002.

[20] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, no. 6795, pp. 536–540, 2000.

[21] Xiaoning Qian and Edward R Dougherty, "Effect of function perturbation on the steady-state distribution of genetic regulatory networks: Optimal structural intervention," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 4966–4976, October 2008.

[22] R. Pal, I. Ivanov, A. Datta, and E. R. Dougherty, "Generating Boolean networks with a prescribed attractor structure," *Bioinformatics*, vol. 21, pp. 4021 – 4025, 2005.