2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

Optimal Bayesian Biomarker Selection for Gene Regulatory Networks under Regulatory Model Uncertainty

Mahdi Imani, Mohsen Imani and Seyede Fatemeh Ghoreishi

Abstract— **Gene regulatory networks (GRNs) are large and complex dynamical systems often monitored through RNA sequencing or microarray technologies. Genomics studies often focus on a small subset of genes and analyze only these genes due to the huge cost and time-limit constraints. Therefore, selecting a small subset of genes that carries the highest information about the underlying process of these complex systems is highly desired. The existing biomarker selection techniques rely on unrealistic assumptions such as direct observability of genes' states as well as the availability of perfect knowledge about the modeling process. To address the aforementioned issues, this paper models GRNs with uncertain regulatory models with the signal model of partially-observed Boolean dynamical systems (POBDS) and derives the optimal Bayesian biomarker selection framework given the noisy available gene-expression data. The proposed framework is built on the multiple-model adaptive estimation (MMAE) framework and the optimal minimum mean-square error (MMSE) state estimator for POBDS, called Boolean Kalman smoother (BKS). The proposed framework is an optimal solution relative to the uncertainty class, and its high performance is demonstrated using the mammalian cell-cycle Boolean network model and the p53-MDM2 negative feedback loop observed through gene-expression data.**

I. Introduction

Gene regulatory networks (GRNs) govern the functioning of key cellular processes, such as cell cycle, stress response, DNA repair, and more. Several mathematical models have been proposed to accurately capture the dynamical behavior of GRNs. These methods include Boolean networks [1]–[3], ordinary differential equations (ODE) [4], S-systems [5], and Bayesian networks [6].

The simplicity and interpretability of Boolean network models have made them one of the most successful frameworks for capturing the dynamical behavior of GRNs [7]. The evolution of Boolean network models has a long history, starting from the deterministic Boolean network model introduced by Kauffman and collaborators in 1969 [1]. In recent years, several sophisticated stochastic Boolean network models have been introduced, including Boolean Networks with perturbation (BNp), Probabilistic Boolean Networks (PBN), and Boolean Control Networks (BCN). Despite some fundamental differences between the aforementioned models, all rely on the assumption that the transcriptional states of genes are directly observable, meaning that 0 and 1 representing the inactivation and activation of each gene are perfectly distinguishable in the available data. However, this is never the case in practice, as modern transcriptional studies in genomics are based on technologies that produce noisy indirect measurements of gene activities, such as cDNA microarrays, RNA-seq, and cell imaging-based assays.

This paper employs the signal model of Partially-Observed Boolean Dynamical Systems (POBDS) [3], [8], [9], which unifies and generalizes most of the aforementioned Boolean network models and allows for indirect and incomplete observation of gene states, a realistic scenario encountered in practice. Several tools have been developed for the POBDS model in recent years, including the optimal filter and smoother based on the minimum mean square error (MMSE) criterion, called the Boolean Kalman filter (BKF) and Boolean Kalman smoother (BKS) [8], [10]–[12] respectively, fault detection [13]–[15], learning [16]–[18] and controllers [19]–[21].

Biomarker selection in GRNs has been the center of attention in recent years, primarily due to advances made in technology which have enabled acquiring much larger and richer genomics data [22]. While the information of a large number of genes is often embedded in genomics data, it has been shown that only a small subset of genes play the key role in the functionality and dynamics of these complex systems [7], [23], [24]. The small gene subsets are often used for various reasons such as analyzing healthy and unhealthy systems, early diagnosis or prognosis of chronic diseases, designing new/effective drugs/remedies that can alter these subsets and impact the systems, and developing targeted therapies to change the undesirable behavior of these systems. Therefore, identifying this small subset of genes is critical in enhancing the accuracy of analysis (e.g., cancer diagnosis, identifying genetic disorders, etc.) and reducing the huge cost associated with genomics studies.

Several mathematical techniques have been developed for the biomarker selection of GRNs. These include tools developed based on Bayesian networks, ordinary differential equations (ODE), Boolean network models. However, they often come short in practice due to reliance on full knowledge about the regulatory model or assumption regarding direct observability of genes' states.

We employ the signal model of POBDS for modeling GRNs with uncertain regulatory models and derive the optimal Bayesian biomarker selection framework using the combination of multiple model adaptive estimation (MMAE) framework [25], [26] and the optimal minimum mean square error (MMSE) [3]. The proposed framework offers several

Mahdi Imani is with the Department of Electrical and Computer Engineering at Northeastern University, Mohsen Imani is with the Department of Computer Science at University of California Irvine, and S. F. Ghoreishi is with Department of Civil and Environmental Engineering and Khoury College of Computer Sciences at Northeastern University. Emails: m.imani@northeastern.edu, m.imani@uci.edu, f.ghoreishi@northeastern.edu

978-1-6654-5196-3/$31.00 ©2022 AACC

benefits, including:

- *Optimality:* The proposed framework meets Bayesian optimality, which means the selected subset of genes is the best with respect to the posterior distribution of uncertain models obtained according to all available information.
- *Risk Consideration:* The proposed framework provides computation of the expected error of decisions, which is critical for risk consideration in sensitive domains.
- *Fast Computation:* The multiple model-based structure of the proposed framework allows parallelization of the process for fast decision making, leading to the applicability of the proposed framework to large GRNs.

In Section IV, the performance of the proposed framework is investigated using the mammalian cell cycle and p53-Mdm2 Boolean network models observed through noisy gene-expression data.

II. POBDS Model of Gene Regulatory Networks Observed Through Noisy Measurements

A POBDS models a set of genes like a graph where genes are represented by nodes and interacting genes are connected by edges. The nature of the interaction is encoded by +1, −1, and 0, representing a positive, negative, and no interaction (more complex interactions can also be considered). The actual interactions of genes cannot easily be measured directly but must be inferred from noisy observations. In a partially-observed Boolean network, the unobserved state of the interactions is learned from the repeated measures of noisy indirect measurements. Assume that the system is described by a *state process* $\{\mathbf{X}_k; k = 0, 1, \ldots\}$, where $\mathbf{X}_k = [\mathbf{X}_k(1), ..., \mathbf{X}_k(d)]^T$ is a vector of size d (i.e., the number of genes). The ith state variable at any time k takes a real value from the set $\{0, 1\}$. The sequence of states is observed indirectly through the *measurement process* $\{\mathbf{Y}_k; k = 1, 2, \ldots\}$, where $\mathbf{Y}_k$ is a vector of measurements. The POBDS signal model can be represented by the following two processes [3]:

$$\begin{aligned} \mathbf{X}_k &= \mathbf{f}\left(\mathbf{X}_{k-1}, \mathbf{n}_k\right) \quad \text{(state process)} \\ \mathbf{Y}_k &= \mathbf{h}\left(\mathbf{X}_k, \mathbf{v}_k\right) \quad \text{(measurement process)} \end{aligned} \tag{1}$$

for $k = 1, 2, \ldots$; where $\mathbf{n}_k$ is the transition noise at time k, $\mathbf{f}$ is the *network function*, whereas $\mathbf{h}$ is a general function mapping the current state and observation noise $\mathbf{v}_k$ into the measurement space. The noise processes $\{\mathbf{n}_k, \mathbf{v}_k; k = 1, 2, \ldots\}$ are assumed to be "white" in the sense that the noises at distinct time points are uncorrelated random variables. It is also assumed that the noise processes are uncorrelated with each other and with the initial state $\mathbf{X}_0$; however, their distributions are arbitrary. The two components of this signal model are presented in this section.

A. GRN State Model

The state process for GRNs can be represented as $\mathbf{f}(\mathbf{X}_{k-1}, \mathbf{n}_k) = \mathbf{f}(\mathbf{X}_{k-1}) \oplus \mathbf{n}_k$, where "$\oplus$" indicates component-wise modulo-2 addition. The noise process $\mathbf{n}_k$ can be modeled by independent Bernoulli (p), where the parameter $p > 0$ models the noise "intensity" — the closer p is to 0.5, the more chaotic the system will be, while a value of p close to zero means that the state trajectories are nearly deterministic, being governed tightly by the network function

A well-known Boolean function which relies on pathway diagrams [27] can describe the behavior of GRNs. The components of Boolean function in this model are represented as $\mathbf{f} = (f_1, \ldots, f_d)$, where each component $f_i : \{0, 1\}^d \to \{0, 1\}$ is a Boolean function given by:

$$f_i(\mathbf{x}) = \begin{cases} 1, & \sum_{j=1}^{d} a_{ij}\mathbf{x}(j) + b_i > 0\,, \\ 0, & \sum_{j=1}^{d} a_{ij}\mathbf{x}(j) + b_i \le 0\,, \end{cases} \tag{2}$$

where a_{ij} and b_i are system parameters. The parameter a_{ij} can take three values: $a_{ij} = +1$ if there is a positive regulation (activation) from gene j to gene i; $a_{ij} = -1$ if there is a negative regulation (inhibition) from gene j to gene i; and $a_{ij} = 0$ if gene j is not an input to gene i. The parameter b_i specifies regulation *biases* and can take two values: $b_i = +1/2$ if gene i is positively biased, in the sense that an equal number of activation and inhibition inputs will produce activation; the reverse being the case if $b_i = -1/2$. The proposed network function is depicted in Fig. 1, where the threshold units are step functions that output 1 if the input is non-negative, and 0 otherwise. It should be noted that the model in (2) is nonlinear.

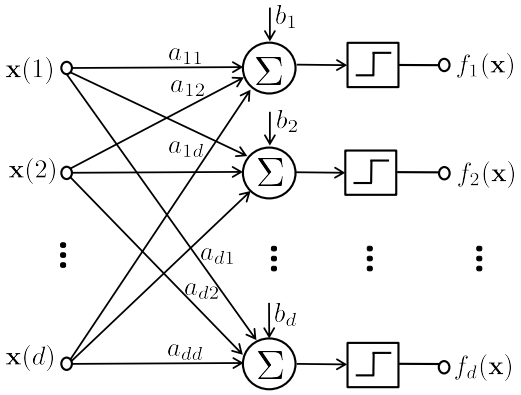


Fig. 1: Gene regulatory network model.

B. GRN Measurement Model

The genomics data are described by the observation process $\{\mathbf{Y}_k; k = 1, 2, \ldots\}$, where $\mathbf{Y}_k = (\mathbf{Y}_k(1), \ldots, \mathbf{Y}_k(d))$ is a vector containing the transcript abundance measurements at time k. The observation vector is entirely arbitrary; however, for simplicity, in this paper, we assume Gaussian gene-expression measurements at each time point. This is an appropriate model for many important gene-expression measurements technologies, such as cDNA microarrays and live cell imaging-based assays [28]–[30], in which gene expression measurements are continuous and unimodal (within a single population of interest). Furthermore, we assume conditional independence of the measurements given the states as:

$$P(\mathbf{Y}_k = \mathbf{y} \mid \mathbf{X}_k = \mathbf{x}) = \prod_{j=1}^{d} P(\mathbf{Y}_k(j) = \mathbf{y}(j) \mid \mathbf{X}_k(j) = \mathbf{x}(j))$$
$$= \frac{1}{\prod_{j=1}^{d}\sqrt{2\pi\sigma_j^2}} \exp\left(-\sum_{j=1}^{d} \frac{(\mathbf{y}(j) - \mu_j^0(1-\mathbf{x}(j)) - \mu_j^1 \mathbf{x}(j))^2}{2\sigma_j^2}\right), \tag{3}$$

where σ_j denotes the standard deviation of the abundance of transcript j, and the parameters μ_j^0 and μ_j^1 specify the means of the abundance of transcript j in the inactivated and activated states, respectively.

III. Optimal Bayesian Biomarker Selection for Partially-Observed GRNs

The complexity and scale of GRNs often pose significant uncertainty in the modeling process. This uncertainty often appears due to partial knowledge about the interactions between genes. In the model described in (2), the unknown interactions can be represented by unknown elements of connectivity matrix or bias units, i.e., $a_{ij} \in \{-1, 0, +1\}$ and $b_j \in \{-1/2, +1/2\}$. The process noise intensity p can also be unknown, which we assume can be quantized into a finite grid. Hence, the possible models for the dynamics of a GRN can be represented by a finite set Θ, in which one of these models is the correct model of the system. For instance, a single unknown regulation, a_{ij}, yields three possible models $\Theta = (\theta_1, \theta_2, \theta_3)$, in which each model differs in a single regulation.

While the network structure of GRNs provides valuable information about their dynamics, biologists are often interested in finding a small subset of genes that can properly predict these complex biological processes. The biologists benefit from these small gene subsets for various reasons such as analyzing healthy and unhealthy systems, early diagnosis or prognosis of chronic diseases, designing new/effective drugs/remedies that can alter these subsets and impact the systems, and developing targeted therapies to change the undesirable behavior of these systems.

Let $\mathbf{Y}_{1:T}$ be the sequence of observed measurements available through gene-expression data. The measurement at each time step is a vector containing the information of d genes as $\mathbf{Y}_k = (\mathbf{Y}_k(1), ..., \mathbf{Y}_k(d))$. The goal of biomarker selection is to pick a small subset of genes that carries the highest information about the underlying process. Let $\mathbb{S}$ denote the space of all possible subsets of genes. We use $\mathbf{Y}_k^s$ to refer to the sth subset of measurement vector $\mathbf{Y}_k$, for $s \in \mathbb{S}$.

The best subset in the set $\mathbb{S}$ is the subset that can perfectly capture the dynamics of the underlying process. This can be interpreted as the subset that leads to the minimum expected error of estimation throughout the whole system. The optimal minimum mean-square error (MMSE) state estimator for a system parametrized by θ given observation from the sth subset $\mathbf{Y}_{1:T}^s$ is given by:

$$\hat{\mathbf{X}}_{k|T,s}^{\mathrm{MS},\theta} = \underset{\hat{\mathbf{X}}_{k|T} \in \Psi}{\operatorname{argmin}} \mathbb{E}\left[\|\mathbf{X}_k - \hat{\mathbf{X}}_{k|T}\|_1^2 \mid \mathbf{Y}_{1:T}^s, \theta\right], \tag{4}$$

where $\|.\|$ denotes the usual vector norm and Ψ is the space of all Boolean estimators. The optimal solution to the above minimization problem is given by a Boolean Kalman smoother algorithm developed in [8] as:

$$\hat{\mathbf{X}}_{k|T,s}^{\mathrm{MS},\theta} = \overline{\mathbb{E}[\mathbf{X}_k \mid \mathbf{Y}_{1:T}^s, \theta]}, \tag{5}$$

for $k = 1, ..., T$; where $\overline{\mathbf{v}}$ maps the element of $\mathbf{v}$ smaller than 0.5 to 0 and others to 1. The sum of the expected error of the optimal estimator in this interval is given by:

$$C_s^{\mathrm{MS},\theta} = \sum_{k=1}^{T} \|\min\{\mathbb{E}[\mathbf{X}_k \mid \mathbf{Y}_{1:T}^s, \theta], \mathbf{1}_d - \mathbb{E}[\mathbf{X}_k \mid \mathbf{Y}_{1:T}^s, \theta]\}\|_1, \tag{6}$$

where "minimum" is applied component-wise. It can be shown that $0 \le C_s^{\mathrm{MS},\theta} \le Td/2$. The large values of $C_s^{\mathrm{MS},\theta}$ denote the high expected error of the optimal estimator for the sth subset of genes for a system parameterized by θ, whereas the small values denote low error (high confidence) of the optimal state estimator.

Due to the uncertainty in the regulatory network model, in this paper, we derive the *optimal Bayesian* (OB) biomarker selection policy, which can be obtained by solving the following minimization problem:

$$s^{\mathrm{OB}} = \underset{s \in \mathbb{S}}{\operatorname{argmin}}\, \mathbb{E}_{\theta|\mathbf{Y}_{1:T}}\left[C_s^{\mathrm{MS},\theta}\right] = \underset{s \in \mathbb{S}}{\operatorname{argmin}} \sum_{\theta \in \Theta} C_s^{\mathrm{MS},\theta}\, p_T(\theta), \tag{7}$$

where the expectation is taken with respect to the posterior distribution of regulatory models given all available information in the whole data, and $p_T(\theta) = P(\theta \mid \mathbf{Y}_{1:T})$ denotes the posterior distribution of the model parameterized by θ.

The computation of (7) requires the computation of the posterior distribution as well as the sum of the expected error of the optimal estimator associated with all subsets of genes and all possible models. To achieve this, we define the state conditional probability distribution vectors $\mathbf{\Pi}_{r|k}^{s,\theta}$, $\mathbf{\Pi}_{r|k}^{s,\theta}$ and $\mathbf{\Delta}_{k|T}^{s,\theta}$ of length 2^d via:

$$\begin{aligned} \mathbf{\Pi}_{r|k}^{\theta}(i) &= P\left(\mathbf{X}_r = \mathbf{x}^i \mid \mathbf{Y}_{1:k}, \theta\right), \\ \mathbf{\Pi}_{r|k}^{s,\theta}(i) &= P\left(\mathbf{X}_r = \mathbf{x}^i \mid \mathbf{Y}_{1:k}^s, \theta\right), \\ \mathbf{\Delta}_{k|T}^{s,\theta}(i) &= p\left(\mathbf{Y}_{k+1:T}^s \mid \mathbf{X}_k = \mathbf{x}^i, \theta\right), \end{aligned} \tag{8}$$

for $i = 1, ..., 2^d$, $1 \le k \le T, 1 \le r \le T$, $s \in \mathbb{S}$ and $\theta \in \Theta$. Let $T(\mathbf{Y}_k^s)$ be the updated matrix corresponding to sth subset of genes defined as:

$$\begin{aligned} \left(T(\mathbf{Y}_k^s)\right)_{ii} &= p(\mathbf{Y}_k^s \mid \mathbf{X}_k = \mathbf{x}^i) \\ &= \prod_{j \in s} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(\mathbf{Y}_k^s(j) - \mu_j^0(1-\mathbf{x}(j)) - \mu_j^1 \mathbf{x}(j))^2}{2\sigma_j^2}\right), \end{aligned} \tag{9}$$

for $i = 1, ..., 2^d$. Let the prediction matrix M_k of size $2^d \times 2^d$ be the transition matrix of the Markov chain corresponding to the model parameterized by $\theta \in \Theta$. This matrix can be defined as:

$$(M_k^{\theta})_{ij} = P(\mathbf{X}_k = \mathbf{x}^i \mid \mathbf{X}_{k-1} = \mathbf{x}^j, \theta), \text{ for } i, j = 1, ..., 2^d. \tag{10}$$

Posterior Computation: In this part, we describe how the posterior distribution of models can be computed. Let $p_0(\theta)$

denote the prior probability of model θ, where $\sum_{\theta\in\Theta} p_0(\theta) = 1$. In a case that there is no prior knowledge about the GRN available, one can simply use the uniform prior over the possible models: $p(\theta) = 1/|\Theta|$, for $\theta \in \Theta$.

The posterior probability of model θ can be computed using all available information as:

$$p_T(\theta) = \frac{p_\theta(\mathbf{Y}_{1:T})\, p_0(\theta)}{\sum_{\theta'\in\Theta} p_{\theta'}(\mathbf{Y}_{1:T})\, p_0(\theta')}, \tag{11}$$

where

$$p_\theta(\mathbf{Y}_{1:T}) = \prod_{k=1}^{T} ||T(\mathbf{Y}_k)\, M_k^\theta\, \mathbf{\Pi}_{k-1|k-1}^\theta||_1, \tag{12}$$

with $\mathbf{\Pi}_{k|k}^\theta = \frac{T(\mathbf{Y}_k)\, M_k^\theta \mathbf{\Pi}_{k-1|k-1}^\theta}{||T(\mathbf{Y}_k)\, M_k^\theta \mathbf{\Pi}_{k-1|k-1}^\theta||_1}$, and $\mathbf{\Pi}_{0|0}^\theta$ as initial distribution.

Computation of Expected Error of Optimal Estimator: For the computation of the expected error associated with each subset of genes and model, one needs to compute the smoothed distribution of states, $\mathbf{\Pi}_{k|T}^{s,\theta}$, at any given time k for all subsets of genes and regulatory models. Let A be a matrix of size $d \times 2^d$ containing all Boolean states of the system as $A = [\mathbf{x}^1 \cdots \mathbf{x}^{2^d}]$. It is easy to verify that $\mathbb{E}\left[\mathbf{X}_k \mid \mathbf{Y}_{1:T}^s, \theta\right] = A\mathbf{\Pi}_{k|T}^{s,\theta}$, so it follows from (5) that:

$$\hat{\mathbf{X}}_{k|T,s}^{\mathrm{MS}} = \overline{A\mathbf{\Pi}_{k|T}^{s,\theta}}\,. \tag{13}$$

with the sum of the expected MSE error in the interval as:

$$C_s^{\mathrm{MS},\theta} = \sum_{k=1}^{T} ||\min\{A\mathbf{\Pi}_{k|T}^{s,\theta}, \mathbf{1}_d - A\mathbf{\Pi}_{k|T}^{s,\theta}]\}||_1\,. \tag{14}$$

Proposed Algorithm: Upon describing the efficient ways of computation of the posterior distribution and expected error of the optimal estimator, the whole process of the proposed framework is described here. As it can be seen in the schematic diagram in Fig. 2, the proposed framework consists of a bank of Boolean Kalman smoothers (BKS) corresponding to different possible models for GRN, which are fed by different subsets of genes. Overall, there is $|\Theta| \times |\mathbb{S}|$ BKSs running in parallel, whose outcomes specify the expected errors of optimal estimators for all gene subsets and GRN models. The posterior distribution can be computed in parallel according to (11) using all available data. These computed values can be used for optimal Bayesian biomarker selection according to (7). The detailed procedure of the proposed framework is provided in Algorithm 1.

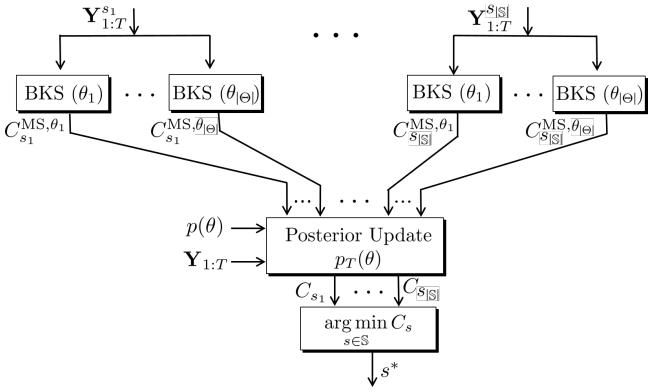


Fig. 2: The schematic diagram of the proposed framework.

Algorithm 1 The proposed Optimal Bayesian Biomarker Selection for POBDS

$\mathbf{\Pi}_{0|0}^\theta(i) = P\left(\mathbf{X}_0 = \mathbf{x}^i\right)$, for $\theta \in \Theta, i = 1, \ldots, 2^d$.
for $\theta \in \Theta$ **do**
  $L(\theta) = 0$.
  for $k = 1, \ldots, T$ **do**
    $\boldsymbol{\beta}_k^\theta = T_k(\mathbf{Y}_k) M_k^\theta\, \mathbf{\Pi}_{k-1|k-1}^\theta$.
    $L(\theta) = L(\theta) + \log ||\boldsymbol{\beta}_k^\theta||_1$.
    $\mathbf{\Pi}_{k|k}^\theta = \boldsymbol{\beta}_k^\theta / ||\boldsymbol{\beta}_k^\theta||_1$.
  end for
end for
$p_T(\theta) = \frac{\exp(L(\theta))\, p_0(\theta)}{\sum_{\theta'\in\Theta} \exp(L(\theta'))\, p_0(\theta')}$
$\mathbf{\Pi}_{0|0}^{s,\theta} = \mathbf{\Pi}_{0|0}^\theta$, for $s \in \mathbb{S}, \theta \in \Theta$.
$C_s = 0, s \in \mathbb{S}$.
for $s \in \mathbb{S}$ **do**
  for $\theta \in \Theta$ **do**
    for $k = 1, \ldots, T$ **do**
      $\mathbf{\Pi}_{k|k-1}^{s,\theta} = M_k^\theta\, \mathbf{\Pi}_{k-1|k-1}^{s,\theta}$.
      $\mathbf{\Pi}_{k|k}^{s,\theta} = T_k(\mathbf{Y}_k^{s,\theta}) \mathbf{\Pi}_{k|k-1} / ||T_k(\mathbf{Y}_k^{s,\theta}) \mathbf{\Pi}_{k|k-1}^{s,\theta}||_1$
    end for
    $\mathbf{\Delta}_{T|T} = \mathbf{1}_{2^d}$.
    for $k = T, T-1, \ldots, 1$ **do**
      $\mathbf{\Delta}_{k|k-1}^{s,\theta} = T_k(\mathbf{Y}_k^s)\, \mathbf{\Delta}_{k|k}^{s,\theta}$.
      $\mathbf{\Delta}_{k-1|k-1}^{s,\theta} = (M_k^\theta)^T\, \mathbf{\Delta}_{k|k-1}^{s,\theta}$.
    end for
    for $k = 1, \ldots, T$ **do**
      $\mathbf{\Pi}_{k|T}^{s,\theta} = (\mathbf{\Pi}_{k|k-1}^{s,\theta} \circ \mathbf{\Delta}_{k|k-1}^{s,\theta}) / ||\mathbf{\Pi}_{k|k-1}^{s,\theta} \circ \mathbf{\Delta}_{k|k-1}^{s,\theta}||_1$.
      $C_s = C_s + p_T(\theta) ||\min\{A\mathbf{\Pi}_{k|T}^{s,\theta}, \mathbf{1}_d - A\mathbf{\Pi}_{k|T}^{s,\theta}]\}||_1$.
    end for
  end for
end for
Optimal biomarker selection: $s^* = \mathrm{argmin}_{s\in\mathbb{S}} C_s$

IV. Numerical Experiments

A. Mammalian Cell-Cycle Boolean Network Model

We investigate the performance of the proposed Bayesian biomarker selection framework using the mammalian cell-cycle [7]. The mammalian cell cycle involves a sequence of events resulting in duplication and division of the cell. It occurs in response to growth factors, and under normal conditions, it is a tightly controlled process. A regulatory model for the mammalian cell cycle proposed in [7] is shown in Fig. 3. There are 10 genes involved in the process where the states of genes at time step k are denoted by a vector $\mathbf{X}_k$ = [CycD, Rb, p27, E2F, CycE, CycA, Cdc20, Cdh1, UbcH10]. These interactions can be expressed in terms of interaction parameters as $a_{21} = 1, a_{22} = 0, a_{23} = +1, a_{24} = 0, a_{25} = 1, a_{26} = -1, a_{27} = 0, a_{28} = 0, a_{29} = 0$ and $a_{2\,10} = 1$. All bias units are zero.

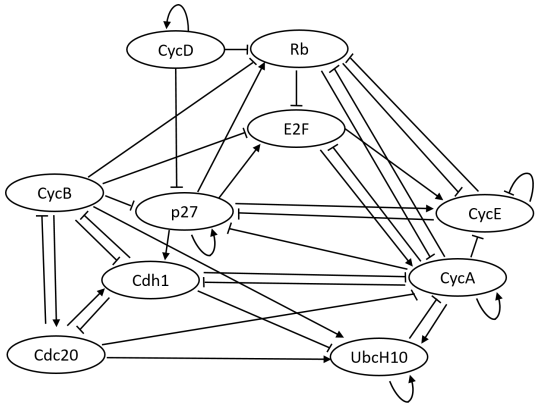


Fig. 3: Pathway diagrams of the mammalian cell cycle regulatory network models.

We consider three cases with 2, 4 and 6 unknown regulations (M). The unknown regultions are chosen randomly among +1 or -1 interactions, and for the sake of simplicity, we assume the unknown interactions take in $\{-1, +1\}$. These unknown connections result in 2^M different Boolean network models for the system, differing in one or more of these uncertain regulations. Let $\Theta = \{\theta_1, ..., \theta_{2^M}\}$ be the uncertainty class of these network models, the prior distribution over Boolean network models has been generated from Dirichlet distribution as:

$$p(\theta) \sim f(p(\theta); \phi) = \frac{\Gamma\left(\phi 2^M\right)}{\Gamma(\phi)^{2^M}} \prod_{j=1}^{2^M} p(\theta_j)^{\phi-1}, \qquad (15)$$

where Γ is the gamma function and $\phi > 0$ is the parameter of the symmetric Dirichlet distribution. ϕ specifies the variability of the initial distributions; the smaller ϕ is, the more the initial distributions deviate from the uniform distribution. $\phi = 1$ and $\phi = 10$ are considered in our numerical experiments. The goal is to find the best 3 genes among 10, which refers to the genes' subset: $\mathbb{S} = (s_1, ..., s_{120})$, where $\binom{10}{3} = 120$. For a test time series of length T', the normalized average mean squared error (MSE) per time step for a given subset of genes is defined as:

$$\text{Nor. Ave. MSE per Time } = \frac{1}{T'd} \sum_{k=1}^{T'} ||\hat{\mathbf{X}}^s_{k|T'} - \mathbf{X}_k||_1^2, \quad (16)$$

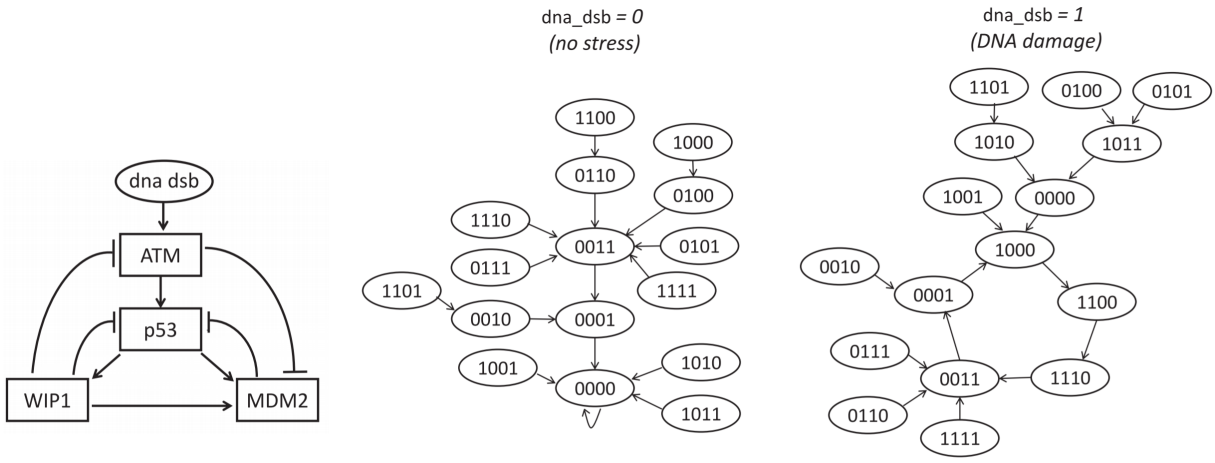
where $\hat{\mathbf{X}}^s_{k|T'} = \overline{\sum_{\theta \in \Theta} (A\mathbf{\Pi}^{s,\theta}_{k|T'}) P(\theta \mid \mathbf{Y}^s_{1:T'})}$.

Table I represents the normalized average mean square errors per time-step for the proposed method in the following cases: the best subset containing 3 genes among 10 genes are sought by the proposed framework and random policy; no selection is made, and all genes are used for the estimation process, and the BKS is tuned to the "true" model. It should be noted that the results of the BKS tuned to the true model is the best result that can be achieved, as all genes are assumed to be observed, and no uncertainty in the modeling process is considered. It can be seen that the proposed framework has significantly outperformed the random policy and its normalized average MSE per time-step is close to the case when all genes are used for the estimation process. This clearly shows that the proposed framework is capable of selecting the best 3 genes among 10 that can capture the dynamical behavior of the mammalian cell-cycle network. Meanwhile, the result of the proposed framework is closer to the baseline solution (BKS tuned to the true model) for more informative prior knowledge ($\phi = 10$) over the unknown regulations and a lower number of unknown regulations. The reason is that the number of possible models for the system increases exponentially with the number of unknown regulations (M), and the posterior probability becomes less informative given a small available gene-expression data. Finally, as expected, the estimation error is higher for larger measurement noise, σ.

B. P53-Mdm2 Boolean Network Model

In this part of numerical experiments, we describe an application of the proposed framework in a well-known *p53-MDM2* negative-feedback gene regulatory network [23]. The pathway diagram for this network is presented in the left plot of Fig. 4. The *p53* gene codes for the tumor suppressor protein p53 in humans, and its activation plays a critical role in cellular responses to various stress signals that might cause genome instability. The gene regulatory network consists of four genes: *ATM*, *p53*, *Wip1*, and *MDM2*, and the input "dna_dsb" which indicates the presence of DNA double-strand breaks. A single unknown connection is assumed to be the activation from "p53" to "WIP1", which poses three possible models in Θ. Four subsets of genes are considered as: $\mathbb{S}$ =("ATM","p53","WIP1","MDM2"). We can see that 0000 is a singleton attractor state under no-stress, while the other states are transient; on the other hand, under DNA damage, there is a cyclic attractor, corresponding to an oscillation of p53, along with the other proteins in its regulatory pathway. This reproduces the known biological behavior described previously.

The normalized average MSE per time-step obtained by the proposed framework over 10 time steps for various selected genes are presented in Table II. One can see that the average errors are larger for larger process noise. This can be justified because larger process noise pulls the system out of attractors more often, making the estimation process more challenging. One can see that the error is larger for higher measurement noise. Furthermore, it can be seen that



(a) (b) (c)

Fig. 4: Activation/repression pathway diagram and state transition diagrams corresponding to a constant input dna_dsb = 0 (no-stress) and dna_dsb = 1 (DNA-damage) for the P53-MDM2 negative feedback loop Boolean network model with negative regulation biases.

TABLE I: Results of mammalian cell-cycle Boolean network.

σ	ϕ	M	Prop. Method			All-Genes			Random			BKS(θ^*)
			T = 10	T = 50	T = 100	T = 10	T = 50	T = 100	T = 10	T = 50	T = 100	
5	1	2	0.16	0.11	0.08	0.11	0.09	0.05	0.32	0.29	0.24	0.04
		4	0.20	0.18	0.13	0.16	0.11	0.08	0.36	0.33	0.28	
		6	0.31	0.27	0.22	0.23	0.19	0.12	0.41	0.37	0.33	
	10	2	0.12	0.09	0.06	0.09	0.07	0.04	0.28	0.24	0.21	
		4	0.16	0.14	0.10	0.13	0.10	0.06	0.30	0.28	0.25	
		6	0.24	0.20	0.16	0.20	0.13	0.08	0.33	0.30	0.28	
10	1	2	0.23	0.18	0.12	0.15	0.12	0.09	0.41	0.37	0.31	0.07
		4	0.27	0.26	0.20	0.20	0.15	0.13	0.40	0.35	0.32	
		6	0.38	0.34	0.30	0.29	0.21	0.15	0.43	0.40	0.37	
	10	2	0.20	0.16	0.15	0.15	0.12	0.10	0.33	0.29	0.27	
		4	0.23	0.22	0.18	0.16	0.12	0.09	0.36	0.32	0.30	
		6	0.32	0.29	0.23	0.24	0.21	0.16	0.40	0.37	0.35	

TABLE II: Results of p53-Mdm2 Boolean network.

ϕ	p	Initial Distribution	No stress (dna_dsb = 0)				DNA damage (dna_dsb = 1)			
			ATM	p53	Wip1	Mdm2	ATM	p53	Wip1	Mdm2
10	0.01	$[\frac{1}{16},...,\frac{1}{16}]^T$	0.0968	0.0940	0.0694	**0.0547**	0.3967	0.4020	**0.3888**	0.4613
		$[0,...,1]^T$	0.1220	0.1205	0.1121	**0.0996**	**0.4175**	0.4192	0.4205	0.4554
	0.1	$[\frac{1}{16},...,\frac{1}{16}]^T$	0.2221	0.2143	0.1698	**0.1427**	**0.3325**	0.3709	0.3413	0.3980
		$[0,...,1]^T$	0.2394	0.2332	0.1916	**0.1586**	**0.3433**	0.3768	0.3490	0.3858
1	0.01	$[\frac{1}{16},...,\frac{1}{16}]^T$	0.0971	0.0947	0.0749	**0.0637**	0.4236	0.4147	**0.4081**	0.4698
		$[0,...,1]^T$	0.1223	0.1210	0.1159	**0.1095**	0.4486	**0.4406**	0.4427	0.4663
	0.1	$[\frac{1}{16},...,\frac{1}{16}]^T$	0.2254	0.2191	0.1828	**0.1630**	0.3659	0.3831	**0.3636**	0.4100
		$[0,...,1]^T$	0.2428	0.2377	0.2078	**0.1851**	0.3750	0.3890	**0.3738**	0.4055

the average error is larger in the case of an active dna_dsb input in comparison to an inactive one. This can be explained by the attractor structure of the p53-MDM2 Boolean network in the presence and absence of external input, in which the system has a singleton and cyclic attractor in the absence and presence of DNA damage, respectively.

The optimal Bayesian genes selected by the proposed framework are specified by bold numbers in Table II. One

can see that in the case of an inactive dna_dsb input, MDM2 is the best choice in all cases. With an active dna_dsb input, either ATM, p53, or Wip1 genes are the best choices, depending on the system's parameters. For example, the choices of the optimal gene in the case of small process and observation noise are different for two initial distribution vectors. A similar trend can be seen in the case of large measurement noise. From the results of Table II, one can clearly understand the importance of the biomarker selection process and its dependency on the initial distribution, the values of noise, and input to the system.

V. Conclusion

This paper proposed an optimal Bayesian biomarker selection framework for selecting a subset of genes that carries the highest information about the underlying process of gene regulatory networks. The partial observability of the states of genes as well as the imperfect knowledge about the regulatory model is accounted for by the use of partially-observed Boolean dynamical systems (POBDS) signal model. The proposed framework consists of multiple Boolean Kalman smoothers (BKSs) running in parallel, each tuned to a set of possible models and a subset of genes. We derived the exact Bayesian solution, which is the optimal solution with respect to the posterior distribution of the possible models. The high performance of the proposed framework is demonstrated through the biomarker selection process of the mammalian cell-cycle regulatory model and the p53-Mdm2 negative feedback loop Boolean network model.

Acknowledgment

The authors acknowledge the support of the National Science Foundation through the NSF awards IIS-2202395 and ENG-2127780, and ARMY Research Office through the award W911NF2110299 and Office of Naval Research awards N00014-21-1-2225 and N00014-22-1-2067.

References

[1] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of theoretical biology*, vol. 22, no. 3, pp. 437–467, 1969.

[2] I. Shmulevich, E. R. Dougherty, and W. Zhang, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks," *Proceedings of the IEEE*, vol. 90, no. 11, pp. 1778–1792, 2002.

[3] M. Imani and U. M. Braga-Neto, "Maximum-likelihood adaptive filter for partially observed Boolean dynamical systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 359–371, 2017.

[4] S. Chen, A. Shojaie, and D. M. Witten, "Network reconstruction from high-dimensional ordinary differential equations," *Journal of the American Statistical Association*, vol. 112, no. 520, pp. 1697–1707, 2017.

[5] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita, "Dynamic modeling of genetic networks using genetic algorithm and S-system," *Bioinformatics*, vol. 19, no. 5, pp. 643–650, 2003.

[6] F. Liu, S.-W. Zhang, W.-F. Guo, Z.-G. Wei, and L. Chen, "Inference of gene regulatory network based on local Bayesian networks," *PLoS computational biology*, vol. 12, no. 8, p. e1005024, 2016.

[7] A. Fauré, A. Naldi, C. Chaouiya, and D. Thieffry, "Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle," *Bioinformatics*, vol. 22, no. 14, pp. e124–e131, 2006.

[8] M. Imani, E. Dougherty, and U. Braga-Neto, "Boolean Kalman filter and smoother under model uncertainty," *Automatica*, vol. 111, p. 108609, 2020.

[9] M. Imani and U. Braga-Neto, "Particle filters for partially-observed Boolean dynamical systems," *Automatica*, vol. 87, pp. 238–250, 2018.

[10] M. Imani and U. Braga-Neto, "Gene regulatory network state estimation from arbitrary correlated measurements," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, p. 22, 2018.

[11] M. Imani and b. y. o. Ghoreishi, Seyede Fatemeh, "Adaptive real-time filter for partially-observed Boolean dynamical systems,"

[12] L. D. McClenny, M. Imani, and U. Braga-Neto, "Boolfilter package vignette," 2017.

[13] S. F. Ghoreishi and M. Imani, "Offline fault detection in gene regulatory networks using next-generation sequencing data," in *2019 53rd Asilomar Conference on Signals, Systems and Computers*, IEEE, 2019.

[14] M. Imani and U. M. Braga-Neto, "Optimal finite-horizon sensor selection for Boolean Kalman filter," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pp. 1481–1485, IEEE.

[15] A. Bahadorinejad, M. Imani, and U. Braga-Neto, "Adaptive particle filtering for fault detection in partially-observed Boolean dynamical systems," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 17, no. 4, pp. 1105–1114, 2018.

[16] M. Imani and j. y. p. Ghoreishi, Seyede Fatemeh, "Optimal finite-horizon perturbation policy for inference of gene regulatory networks,"

[17] M. Imani and U. Braga-Neto, "Control of gene regulatory networks using Bayesian inverse reinforcement learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1250–1261, 2019.

[18] M. Imani and U. Braga-Neto, "Optimal control of gene regulatory networks with unknown cost function," in *Proceedings of the 2018 American Control Conference (ACC 2018)*, pp. 3939–3944, IEEE, 2018.

[19] M. Imani and U. M. Braga-Neto, "Point-based methodology to monitor and control gene regulatory networks via noisy measurements," *IEEE Transactions on Control Systems Technology*, vol. 27, pp. 1023 – 1035, 2019.

[20] M. Imani and U. M. Braga-Neto, "Control of gene regulatory networks with noisy measurements and uncertain inputs," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 2, pp. 760–769, 2018.

[21] M. Imani and U. M. Braga-Neto, "Finite-horizon LQR controller for partially-observed Boolean dynamical systems," *Automatica*, vol. 95, pp. 172–179, 2018.

[22] L. Qian, H. Wang, and E. R. Dougherty, "Inference of noisy nonlinear differential equation models for gene regulatory networks using genetic programming and Kalman filtering," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3327–3339, 2008.

[23] E. Batchelor, A. Loewer, and G. Lahav, "The ups and downs of p53: understanding protein dynamics in single cells," *Nature Reviews Cancer*, vol. 9, no. 5, p. 371, 2009.

[24] M. Imani and S. F. Ghoreishi, "Partially-observed discrete dynamical systems," in *Proceedings of the 2021 American Control Conference (ACC 2021)*, IEEE, 2021.

[25] P. S. Maybeck and P. D. Hanlon, "Performance enhancement of a multiple model adaptive estimator," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 31, no. 4, pp. 1240–1254, 1995.

[26] M. Imani and U. Braga-Neto, "Multiple model adaptive controller for partially-observed boolean dynamical systems," in *2017 American Control Conference (ACC)*, pp. 1103–1108, IEEE, 2017.

[27] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang, "The yeast cell-cycle network is robustly designed," *Proceedings of the National Academy of Sciences*, vol. 101, no. 14, pp. 4781–4786, 2004.

[28] Y. Chen, V. Kamat, E. R. Dougherty, M. L. Bittner, P. S. Meltzer, and J. M. Trent, "Ratio statistics of gene expression levels and applications to microarray data analysis," *Bioinformatics*, vol. 18, no. 9, pp. 1207–1215, 2002.

[29] J. Hua, C. Sima, M. Cypert, G. C. Gooden, S. Shack, L. Alla, E. A. Smith, J. M. Trent, E. R. Dougherty, and M. L. Bittner, "Dynamical analysis of drug efficacy and mechanism of action using GFP reporters," *Journal of Biological Systems*, vol. 20, no. 04, pp. 403–422, 2012.

[30] E. Hajiramezanali, M. Imani, U. Braga-Neto, X. Qian, and E. R. Dougherty, "Scalable optimal Bayesian classification of single-cell trajectories under regulatory model uncertainty," *BMC genomics*, 2019.