

Supplementary Material for Steady-State Analysis of Genetic Regulatory Networks Modeled by Probabilistic Boolean Networks

Ilya Shmulevich, Ilya Gluhovsky, Ronaldo Hashimoto,
Edward R. Dougherty, Wei Zhang

1 Definitions and Preliminary Results

We provide the basic definitions and notations for PBNs and refer the reader to (Shmulevich *et al.*, 2002a) for more details and examples. A PBN $G(V, F)$ is defined by a set of binary-valued nodes $V = \{x_1, \dots, x_n\}$ and a list $F = (F_1, \dots, F_n)$ of sets $F_i = \{f_1^{(i)}, \dots, f_{l(i)}^{(i)}\}$ of Boolean functions. Each node $x_i \in \{0, 1\}$ represents the state (expression) of gene i , where $x_i = 1$ means that gene i is expressed and $x_i = 0$ means it is not expressed. The set F_i represents the possible rules of regulatory interactions for gene x_i . That is, each $f_j^{(i)} : \{0, 1\}^n \rightarrow \{0, 1\}$ is a possible Boolean function determining the value of gene x_i in terms of some other genes and $l(i)$ is the number of possible functions for gene x_i . The functions $f_j^{(i)}$ are called *predictors*. Any given gene transforms its inputs (regulatory factors that bind to it) into an output, which is the state or expression of the gene itself. All genes (nodes) are updated synchronously in accordance with the functions assigned to them and this process is then repeated. At every time step, one of the predictors for gene x_i is selected randomly from the set F_i , according to a predefined probability distribution, discussed below.

A *realization* of the PBN at a given instant of time is determined by a vector of Boolean functions. If there are N possible realizations, then there are N vector functions, $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_N$ of the form $\mathbf{f}_k = (f_{k_1}^{(1)}, f_{k_2}^{(2)}, \dots, f_{k_n}^{(n)})$, for $k = 1, 2, \dots, N$, $1 \leq k_i \leq l(i)$ and where $f_{k_i}^{(i)} \in F_i$ ($i = 1, \dots, n$). The vector function (also called multiple-output function) $\mathbf{f}_k : \{0, 1\}^n \rightarrow \{0, 1\}^n$ acts as a transition function (mapping) representing a possible realization of the entire PBN. Thus, given the values of all genes (x_1, \dots, x_n) , $\mathbf{f}_k(x_1, \dots, x_n) = (x'_1, \dots, x'_n)$ gives us the state of the genes after one step of the network given by \mathbf{f}_k . If the predictor for each gene is chosen independently of other predictors, then $N = \prod_{i=1}^n l(i)$. Each predictor function $f_j^{(i)}$ typically has many fictitious variables, which means

that although the domain of each predictor is $\{0,1\}^n$, there are only a few input genes that actually regulate gene x_i at any given time. The biological and practical justifications for probabilistically choosing one of several simple predictors for each gene are discussed in (Shmulevich *et al.*, 2002a).

Stochastically, the multiple-output function is a random vector $\mathbf{f} = (f^{(1)}, \dots, f^{(n)})$ taking values in $F_1 \times \dots \times F_n$, meaning that $\mathbf{f} = \mathbf{f}_k$ for some k . The *selection probability* that the predictor $f_j^{(i)}$ is used to determine gene i ($1 \leq j \leq l(i)$) is given by

$$c_j^{(i)} = \Pr \left\{ f^{(i)} = f_j^{(i)} \right\} = \sum_{k: f_{k_i}^{(i)} = f_j^{(i)}} \Pr \{ \mathbf{f} = \mathbf{f}_k \}. \quad (1)$$

In general there needs to be no assumption that $f^{(1)}, \dots, f^{(n)}$ are selected independently; however, here we make that assumption. Hence,

$$\Pr \{ \mathbf{f} = \mathbf{f}_k \} = \prod_{j=1}^n \Pr \left\{ f^{(j)} = f_{k_j}^{(j)} \right\} = \prod_{j=1}^n c_{k_j}^{(j)}. \quad (2)$$

An approach for obtaining the probabilities $c_j^{(i)}$ from gene expression data, using the coefficient of determination (Dougherty *et al.*, 2000; Kim *et al.*, 2000a; Kim *et al.*, 2000b), is discussed in (Shmulevich *et al.*, 2002a).

As shown in (Shmulevich *et al.*, 2002a), a PBN is a homogeneous Markov chain consisting of 2^n states. A further extension to the PBN model permitting so-called gene *perturbations* was proposed in (Shmulevich *et al.*, 2002b), allowing any out of n possible genes to get perturbed with probability p , independently of other genes. In the Boolean setting, this is represented by a flip of value from 1 to 0 or vice versa and directly corresponds to the bit-flipping mutation operator in *NK* Landscapes (Kauffman, 1993) as well as in genetic algorithms and evolutionary computing (Goldberg, 1989). This situation is modeled as follows. Suppose that at every step of the network, we have a realization of a so-called random *perturbation vector* $\gamma \in \{0,1\}^n$. If the i -th component of γ is equal to 1, then the i -th gene is flipped, otherwise it is not. In general, γ need not be independent and identically distributed (i.i.d.), but we will assume this from now on for simplicity. Thus, we will suppose that $\Pr \{ \gamma_i = 1 \} = p$ for all $i = 1, \dots, n$. Let $x = (x_1, \dots, x_n)$ be the state of the network (i.e. values of all the genes) at some given time. Then, the next state x' is given by

$$x' = \begin{cases} x \oplus \gamma, & \text{with probability } 1 - (1-p)^n \\ \mathbf{f}_k(x_1, \dots, x_n), & \text{with probability } (1-p)^n \end{cases}, \quad (3)$$

where \oplus is component-wise addition modulo 2 and $\mathbf{f}_k(x_1, \dots, x_n)$, $k = 1, 2, \dots, N$, is the transition function representing a possible realization of the entire PBN. We should note that according to this perturbation model, one model time step is used by a perturbation should one happen to occur. An alternative formulation would be to use $\mathbf{f}_k(x_1, \dots, x_n) \oplus \gamma$ instead of $x \oplus \gamma$, which would signify the fact that the network transition is immediately followed by

a perturbation and so, only takes one time step. Since it is not known how discrete time steps relate to actual physical time, the choice of the perturbation model is arbitrary. The relevant issue is that either model will be suitable for removing the dependence on the initial conditions and obtaining a steady-state distribution.

Taking gene perturbation into account, the state transition matrix A of the Markov chain can be expressed as

$$A(x, x') = \left(\sum_{k=1}^N \Pr\{\mathbf{f} = \mathbf{f}_k\} \times 1_{[\mathbf{f}_k(x)=x']} \right) \times (1-p)^n + p^{\eta(x, x')} \times (1-p)^{n-\eta(x, x')} \times 1_{[x \neq x']}, \quad (4)$$

where $\eta(x, x') = \sum_{i=1}^n (x_i \oplus x'_i)$ is the Hamming distance between vectors x and x' and $1_{[A]}$ is an indicator function that is equal to 1 only when event A holds true (Shmulevich *et al.*, 2002b). Briefly, the two terms in equation (4) essentially correspond to the two cases in equation (3). With probability $(1-p)^n$, no gene is perturbed and the next state is determined via the Boolean functions selected at that time step. If at least one gene is perturbed, then the transition probability depends on the number of perturbed genes. Given that a perturbation did occur, causing a transition from state x to state x' , we can conclude that the number of perturbed genes was $\eta(x, x')$ – the Hamming distance between x and x' . Because $\gamma \in \{0, 1\}^n$ is i.i.d. with $E[\gamma_i] = p$, $i = 1, \dots, n$, the probability that x got changed to x' is equal to $p^{\eta(x, x')} \times (1-p)^{n-\eta(x, x')}$. It is clear that the fact that at least one perturbation occurred implies that x and x' cannot be equal and so this expression must be multiplied by $1_{[x \neq x']}$. More details are available in (Shmulevich *et al.*, 2002b).

It is relatively straightforward to see that for $p > 0$, the Markov chain corresponding to the PBN is ergodic. Consequently, this fact simplifies the analysis of long-term behavior of the network. For example, it is possible to assess the sensitivity of the stationary distribution to such random gene perturbations and study the long-term effects of gene intervention (Shmulevich *et al.*, 2002b).

2 Minorization condition

A Markov chain with transition probability matrix A on state space $\{0, 1\}^n$ satisfies the minorization condition if there is a probability measure $Q(\cdot)$ on $\{0, 1\}^n$ and $\varepsilon > 0$, such that

$$A^{m_0}(x, x') \geq \varepsilon Q(x'), \quad \forall x, x' \in \{0, 1\}^n, \quad (5)$$

where $A(x, x')$ is given in (4). It is shown in (Rosenthal, 1995) that if the minorization condition is satisfied for a Markov chain with stationary distribution π , then for any initial distribution,

$$\left\| \pi^{(k)} - \pi \right\|_{\text{TV}} \leq (1 - \varepsilon)^{\lfloor k/m_0 \rfloor}, \quad (6)$$

where $\pi^{(k)}$ is the distribution at time k and $\|\mu - \pi\|_{\text{TV}} = \frac{1}{2} \sum_i |\mu_i - \pi_i|$ is the *total variation* distance between μ and π . Let us consider this result in the framework of PBNs. In order for us to apply equation (6), we must select suitable m_0 , ε and $Q(\cdot)$ in equation (5). Define $r(x') = \min_x A^{m_0}(x, x')$. For any choice of m_0 , we then have to choose the largest ε such that $\varepsilon \leq r(x')/Q(x')$. Thus, the optimal $\hat{Q}(x') \propto r(x')$ and $\hat{\varepsilon} = \sum r(x')$ by virtue of the fact that $\sum Q(x') = 1$.

Example. Suppose there exists state x_0 such that for any state x , $A^{m_0}(x, x') \geq uA^{m_0}(x_0, x')$ for some $u \leq 1$. That is, $uA^{m_0}(x_0, x')$ is a lower bound for all rows of $A^{m_0}(x, x')$. Then $\hat{\varepsilon} = \sum r(x') \geq u \sum A^{m_0}(x_0, x') = u$.

Ideally, we should look for m_0 for which the sum of column minima is greatest. However, tracking powers of A is computationally difficult. Note that we cannot use $m_0 = 1$ because it may be impossible for the PBN to stay in the same state; thus, $A(x, x)$ may be zero for some x , resulting in zero column minima. Therefore, we try $m_0 = 2$. We find by only using perturbation transitions (since we do not make assumptions about the network itself) that

$$A^2(x, x') \geq [2p(1-p)]^{\eta(x, x')} [p^2 + (1-p)^2]^{n-\eta(x, x')} - (1 + 1_{\{x \neq x'\}})(1-p)^n (1-p)^{n-\eta(x, x')} p^{\eta(x, x')}$$

because there are $n - \eta(x, x')$ genes whose value must change as a result of two transitions, $\eta(x, x')$ genes whose value remains the same, and we have to subtract the transitions that involve a no-move step. For small p , this is approximately

$$A^2(x, x') \geq (2p)^{\eta(x, x')} (1-2p)^{n-\eta(x, x')}. \quad (7)$$

Thus, the column minima $r(x') \geq (2p)^n$ and $\hat{\varepsilon} = (4p)^n$ and we get

$$\left\| \pi^{(k)} - \pi \right\|_{\text{TV}} \leq (1 - (4p)^n)^{\lfloor k/2 \rfloor}. \quad (8)$$

Consider the sets of states that correspond to irreducible subchains when $p = 0$. Such sets of states were called *implicitly irreducible* in (Shmulevich *et al.*, 2002b) and are hypothesized to correspond to functional cellular states of the organism being modeled. Thus, the state space is divided into blocks and it is only through small perturbation probabilities that jumps between different blocks are possible. It, therefore, may be reasonable to assume that a column minimum is achieved when x and x' are in different blocks, as the transitions within blocks, being furnished by the Boolean functions in the PBN, are more probable than perturbations. Thus, if it so happens that x and x' with large $\eta(x, x')$ are in the same block, $r(x')$ for that column improves. We make the following comments to take this into account.

We will assume that states belong to blocks at random and will compute the expected maximum Hamming deviation from x' : $x_0 = \arg \min_{x \not\parallel x'} A^2(x, x') = \arg \max_{x \not\parallel x'} \eta(x, x')$ by (7), where by $x \not\parallel x'$ we mean that the two are not in

the same subchain (block). We will then use $(2p)^{\eta(x_0, x')} (1 - 2p)^{n - \eta(x_0, x')}$ as an estimate of $r(x')$.

Suppose that the size of the block that includes x' is $k + 1$ and let $m = 2^n - (k + 1)$ be the size of the rest of the state space. Let k_L be the number of states that are Hamming distance greater than L from x' . $k_L = \sum_{i > L} \binom{n}{i}$ since we choose i genes to flip to get i units away from x' . Now $\eta(x_0, x') \leq L$ if and only if all states with distance greater than L are in the same block as x' . The number of such arrangements of states is $\binom{k}{k_L} k_L! (m + k - k_L)!$ because we place the k_L farthest states together with x' and the rest of the states are unconstrained. Thus,

$$P\{\eta(x_0, x') \leq L\} = \binom{k}{k_L} k_L! (m + k - k_L)! / (m + k)! = \binom{k}{k_L} / \binom{m + k}{k_L}$$

and

$$E\eta(x_0, x') = \sum_{L \geq 0} P\{\eta(x_0, x') > L\} = \sum_{L \geq 0} \left(1 - \binom{k}{k_L} / \binom{m + k}{k_L} \right).$$

In particular, we can see that $\eta(x_0, x')$ is always at least $\min\{L : k_L \leq k\}$. So if $k = n/2$, namely, the block that contains x' is about half the state space, even under the most favorable arrangement for that column, $\eta(x_0, x') = n/2$ and therefore $r(x') \geq (2p)^{n/2} (1 - 2p)^{n/2}$, which is not that different from $(2p)^n$ above. This brief analysis indicates that making any assumptions about the relative magnitudes of the probabilities of perturbation and transition via the selected Boolean functions is not likely to significantly improve the bound on convergence. It is only by having some information about the structure of the PBN itself that the bound in (8) can be lowered.