



Gene perturbation and intervention in probabilistic Boolean networks

Ilya Shmulevich^{1,*}, Edward R. Dougherty² and Wei Zhang¹

¹Cancer Genomics Laboratory, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd, Box 85, Houston, TX 77030, USA and ²Department of Electrical Engineering, Texas A&M University, College Station, TX 77843, USA

Received on November 13, 2001; revised on March 14, 2002; accepted on March 21, 2002

ABSTRACT

Motivation: A major objective of gene regulatory network modeling, in addition to gaining a deeper understanding of genetic regulation and control, is the development of computational tools for the identification and discovery of potential targets for therapeutic intervention in diseases such as cancer. We consider the general question of the potential effect of individual genes on the global dynamical network behavior, both from the view of random gene perturbation as well as intervention in order to elicit desired network behavior.

Results: Using a recently introduced class of models, called Probabilistic Boolean Networks (PBNs), this paper develops a model for random gene perturbations and derives an explicit formula for the transition probabilities in the new PBN model. This result provides a building block for performing simulations and deriving other results concerning network dynamics. An example is provided to show how the gene perturbation model can be used to compute long-term influences of genes on other genes. Following this, the problem of intervention is addressed via the development of several computational tools based on first-passage times in Markov chains. The consequence is a methodology for finding the best gene with which to intervene in order to most likely achieve desirable network behavior. The ideas are illustrated with several examples in which the goal is to induce the network to transition into a desired state, or set of states. The corresponding issue of avoiding undesirable states is also addressed. Finally, the paper turns to the important problem of assessing the effect of gene perturbations on long-run network behavior. A bound on the steady-state probabilities is derived in terms of the perturbation probability. The result demonstrates that states of the network that are more 'easily reachable' from other states are more stable in the presence of gene perturbations. Consequently, these are hypothesized to correspond to cellular functional states.

Availability: A library of functions written in MATLAB for

simulating PBNs, constructing state-transition matrices, computing steady-state distributions, computing influences, modeling random gene perturbations, and finding optimal intervention targets, as described in this paper, is available on request from is@ieee.org

Contact: is@ieee.org

1 INTRODUCTION

The near-completion of the Human Genome Projects has revealed that there are 30–40 000 genes in the human genome (International Human Genome Sequencing Consortium, 2001; Hogenesch *et al.*, 2001). Genetic and molecular studies have shown that for many genes each is linked to other genes both at the level of transcription regulation and at the level of protein interaction. In this new era of genomic biology, single gene perspectives are becoming increasingly limited for gaining insight into biological processes. Global, systemic, or network perspectives are becoming increasingly important for making progress in our understanding of the biological processes and harnessing this understanding in educated intervention for correcting human diseases. The development of high throughput genomic and proteomic technologies is empowering researchers in the collection of broad-scope gene information. However, it remains a major challenge to digest the massive amounts of information and use it in an intelligent and comprehensive manner. The development of systematic approaches to finding genes for effective therapeutic intervention requires new models and powerful tools for understanding and managing complex genetic networks.

Boolean networks as models of gene regulatory networks have received much attention since they were first introduced approximately thirty years ago (Kauffman, 1969, 1993; Glass and Kauffman, 1973). In this model, gene expression is quantized to only two levels and the expression level (state) of each gene is functionally related to the expression states of some other genes using logical rules. The formalism of Boolean networks, which emphasize fundamental, generic principles rather than

*To whom correspondence should be addressed.

quantitative biochemical details, establishes a natural framework for capturing the dynamics of regulatory networks and regulation of cellular states, and provides the potential for the discovery of novel targets for anticancer drugs (Huang, 1999). Boolean networks have yielded insights into the overall behavior of large genetic networks (Somogyi and Sniegowski, 1996; Szallasi and Liang, 1998; Wuensche, 1998; Thomas *et al.*, 1995) and allow the study of large data sets in a global fashion. Perhaps part of the appeal of Boolean networks lies in the fact that they are structurally simple yet dynamically complex.

In Shmulevich *et al.* (2002), we introduced a new class of models called probabilistic Boolean networks (PBNs), which are probabilistic generalizations of the standard Boolean networks that offer a flexible and powerful modeling framework. PBNs share the appealing properties of Boolean networks in that they incorporate rule-based dependencies between genes and allow the systematic study of global network dynamics. However, because of their probabilistic nature, they are able to cope with uncertainty, which is intrinsic to biological systems. The dynamics of PBNs can be studied in the probabilistic context of Markov chains, with standard Boolean networks being special cases. Owing to this, the rich theory and numerous tools developed for Markov chains are applicable to the analysis of PBNs as well. PBNs also provide a natural way to quantify the relative influence and sensitivity of genes in their interactions with other genes.

A property of real gene regulatory networks is the existence of spontaneous emergence of ordered collective behavior of gene activity. Recent findings provide experimental evidence for the existence of these *attractors* in regulatory networks (Huang and Ingber, 2000). Boolean networks and PBNs also exhibit this behavior, the former with fixed point and limit cycle attractors (Kauffman, 1993), the latter with absorbing states and irreducible sets of states (Shmulevich *et al.*, 2002). There is abundant justification in the assertion that in real cells, functional states, such as growth or quiescence, correspond to these attractors (Huang, 1999; Huang and Ingber, 2000). Cancer is characterized by an imbalance between cellular states (attractors), such as proliferation and apoptosis (programmed cell death).

As supported by Boolean network simulations, attractors are quite stable under most gene perturbations (Kauffman, 1993), as are real cellular states. However, a characteristic property of dynamical systems such as PBNs (and Boolean networks) is that the activity of some genes may have a profound effect on the global behavior of the entire system. That is to say, a change of value of *certain* genes at *certain* states of the network may drastically affect the values of many other genes in the long-run and lead to different attractors. We should emphasize that the dependence on the current network state is crucial—a particular gene

may exert a significant impact on the network behavior at one time, but that same gene may be totally ineffectual in altering the network behavior at a later time.

One of the main goals of developing models such as PBNs is the identification of potential drug targets in cancer therapy. A random gene perturbation may cause the real regulatory network to transition into an undesirable cellular state, which itself will be stable under most subsequent gene perturbations. We are then faced with the challenge of determining which genes would be good potential candidates for intervention in order to reverse the effects or force the regulatory network to transition to another desirable stable state. Thus, it is important not only to study the effects of gene perturbation, especially on long-run network behavior, but also to develop tools for discovering intervention targets. While we distinguish between random gene perturbation and intentional gene intervention, in this paper, the PBN model class allows us to take a unified viewpoint. We will also make a distinction between so-called *transient* and *permanent* perturbation or intervention. The former type can be reversed by the network itself while the latter is unchangeable or fixed. Although for the most part, we focus on transient perturbation or intervention, this distinction will be discussed in Section 4.1.

Rather than going in depth to present PBNs, we give the necessary definitions and notation in Section 2 and refer the reader to Shmulevich *et al.* (2002) for a more detailed treatment. Section 3 is concerned with random gene perturbation in the context of PBNs. Specifically, it is shown how the underlying Markov chain reflects this phenomenon and an explicit state transition probability is derived in terms of the Boolean functions, their selection probabilities, and the probability of gene perturbation. An example showing an application to the computation of long-term influence of genes is presented for a small simulated network. Section 4 then discusses the notion of gene intervention and uses the theory of first passage times as a tool for deciding which genes are the best candidates for intervention. Several different strategies for selecting such genes are discussed and several examples are given. Finally, in Section 5, we address the question of sensitivity of the stationary distributions to random gene perturbations. We rely on some recent results from perturbation theory of stochastic matrices. Interestingly, these results relate back to Section 4 in that they are also given in terms of first passage times and as such reinforce the conceptual link between perturbation and intervention.

2 PROBABILISTIC BOOLEAN NETWORKS: DEFINITIONS AND NOTATION

In this section, we give the basic definitions and notation for PBNs. The reader is referred to Shmulevich *et al.*

(2002) for more details. 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 (e.g. see Example 1). We will also refer to the functions $f_j^{(i)}$ as *predictors*. Thus, any given gene x_i transforms its inputs (regulatory factors that bind to it), using a rule $f_j^{(i)}$, into an output, which is the state or expression of gene x_i itself. All genes (nodes) are updated synchronously in accordance with the functions assigned to them and this process is then repeated. At any given 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 $f_{k_i}^{(i)} \in F_i$ ($i = 1, \dots, n$). In other words, 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 the realization \mathbf{f}_k . If the predictor for each gene is chosen independently of other predictors, then $N = \prod_{i=1}^n l(i)$. It should be noted that each biologically motivated predictor function $f_j^{(i)}$ typically has many fictitious, or 'don't care,' variables, which do not affect the output of the function. That is, although the domain of each predictor is $\{0, 1\}^n$, i.e. possible states of all genes, there are only a few input genes that actually regulate gene x_i at any given time, implying that each predictor is a relatively simple one. The biological and practical justifications for probabilistically choosing one of several simple predictors for each gene are discussed in Shmulevich *et al.* (2002).

Let $\mathbf{f} = (f^{(1)}, \dots, f^{(n)})$ be a random vector taking values in $F_1 \times \dots \times F_n$. That is, \mathbf{f} can take on all possible realizations of the PBN. Then, the probability that predictor $f_j^{(i)}$ is used to predict gene i ($1 \leq j \leq l(i)$) is equal to

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

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,b), is discussed in Shmulevich *et al.* (2002). The probability that a particular network realization is selected can be computed from (1) by defining an $N \times n$ matrix K such that entries in the j th column are integers between 1 and $l(j)$ and the rows are lexicographically ordered, each one corresponding to a possible network configuration. That is, row i corresponds to network realization i and the entry K_{ij} specifies that predictor $f_{K_{ij}}^{(j)}$ should be used for gene x_j . Then, the probability that network i is selected is

$$P_i = \Pr\{\text{Network } i \text{ is selected}\} = \prod_{j=1}^n c_{K_{ij}}^{(j)}. \quad (2)$$

In Shmulevich *et al.* (2002), it was shown that the dynamics of PBNs can be modeled by Markov chains, consisting of 2^n states, with the state transition matrix A given by

$$A(x, x') = \sum_{i: f_{K_{i1}}^{(1)}(x_1, \dots, x_n) = x'_1, f_{K_{i2}}^{(2)}(x_1, \dots, x_n) = x'_2, \dots, f_{K_{in}}^{(n)}(x_1, \dots, x_n) = x'_n} P_i, \quad (3)$$

where $A(x, x')$ is the probability of transitioning from $x = (x_1, \dots, x_n)$ to $x' = (x'_1, \dots, x'_n)$.

A method for quantifying the relative influence of genes on other genes, within the context of PBNs, was presented in Shmulevich *et al.* (2002). The *influence* $I_j(f)$ of the variable x_j on the function f , with respect to the probability distribution $D(x)$, $x \in \{0, 1\}^n$, over the n -dimensional hypercube, is defined as

$$I_j(f) = E_D \left[\frac{\partial f(x)}{\partial x_j} \right], \quad (4)$$

where $E_D[\cdot]$ is the expectation operator with respect to distribution D , $\frac{\partial f(x)}{\partial x_j} = f(x^{(j,0)}) \oplus f(x^{(j,1)})$ is the partial derivative of the Boolean function f , the symbol \oplus is addition modulo 2 (exclusive OR), and $x^{(j,k)} = (x_1, \dots, x_{j-1}, k, x_{j+1}, \dots, x_n)$, for $k = 0, 1$. In other words, (4) gives the influence as the probability (under the distribution $D(x)$) that a toggle of the j th variable changes the value of the function. In the context of PBNs, the influence of gene x_k on gene x_i is given by Shmulevich *et al.* (2002)

$$I_k(x_i) = \sum_{j=1}^{l(i)} I_k(f_j^{(i)}) \cdot c_j^{(i)}. \quad (5)$$

The *influence matrix* Γ contains information about influences between every pair of genes as $\Gamma_{ij} = I_i(x_j)$.

3 RANDOM GENE PERTURBATIONS

Suppose that any gene, out of n possible genes, can 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 and Levin, 1987; Kauffman, 1993) as well as in genetic algorithms and evolutionary computing (Goldberg, 1989; Altenberg, 1994). For Boolean networks, such random gene perturbations can be implemented with the popular DDLab software (Wuensche, 1996). This type of ‘randomization’, namely allowing genes to randomly flip value, is biologically meaningful. Since the genome is not a closed system, but rather has inputs from the outside, it is known that genes may become either activated or inhibited due to external stimuli, such as mutagens, heat stress, etc. Thus, a network model should be able to capture this phenomenon. If $p = 0$, then the model is reduced to the PBN described in Shmulevich *et al.* (2002). If $p > 0$, then we have the following situation. With probability $(1 - p)^n$, the transition from one state to another occurs as usual, by one of the randomly selected network realizations while with probability $1 - (1 - p)^n$, the state will change due to random bit perturbation(s).

We can frame the random gene perturbations 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 for now on for simplicity. The generalization to the non-i.i.d. case is conceptually straightforward. Thus, we will suppose that $\Pr\{\gamma_i = 1\} = E[\gamma_i] = p$ for all $i = 1, \dots, n$. Clearly,

$$\Pr\{\gamma = (0, \dots, 0)\} = (1 - p)^n.$$

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

$$x' = \begin{cases} x \oplus \gamma, & \text{with probability } 1 - (1 - p)^n \\ \mathbf{f}_k(x_1, \dots, x_n), & \text{with probability } (1 - p)^n \end{cases}, \quad (6)$$

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. In other words, Equation (6) states that if no genes are perturbed, the standard network transition function will be used, whereas if at least one perturbation does occur, then the next state will be determined according to the genes that are perturbed.

An important observation to make here is that for $p > 0$, any state of the network becomes in principle

accessible from any other state, due to the possibility of any combination of random gene perturbations. In fact, we can say the following.

PROPOSITION 1. *For $p > 0$, the Markov chain corresponding to the PBN is ergodic.*

PROOF. Since there are only a finite number of states, ergodicity is equivalent to the chain being aperiodic and irreducible. First, by virtue of Equation (6), we can note that the Markov state transition matrix has no zero entries, except possibly on the diagonal, the latter corresponding to the case when there does not exist a network transition function \mathbf{f}_k ($k = 1, 2, \dots, N$) such that $\mathbf{f}_k(x) = x$. This immediately implies that the chain is irreducible, since all states communicate. Indeed, let x be such a state (i.e. for which $\mathbf{f}_k(x) \neq x$ for all $k = 1, 2, \dots, N$) and let $y \neq x$ be any other state. The probability of transitioning from x to y is positive as is the probability of going from y back to x . Therefore, there is a positive probability that x is accessible from itself in just two steps. Using the same reasoning, the process may return to the same state after any number of steps, except possibly after one step, implying that the chain is also aperiodic.

The fact that the Markov chain is ergodic implies that it possesses a steady-state distribution equal to the stationary distribution, which can be estimated empirically simply by running the network for a sufficiently long time and by collecting information about the proportion of time the process spends in each state. The convergence rate, however, will surely depend on the parameter p . A simulation-based analysis of the network involving gene perturbation may require one to compute the transition probability $A(x, x') = \Pr\{(x_1, \dots, x_n) \rightarrow (x'_1, \dots, x'_n)\}$ between any two arbitrary states of the network. We turn to this next.

THEOREM 2. *Given a PBN $G(V, F)$ with genes $V = \{x_1, \dots, x_n\}$ and a list $F = (F_1, \dots, F_n)$ of sets $F_i = \{f_1^{(i)}, \dots, f_{l(i)}^{(i)}\}$ of Boolean predictors, as well as a gene perturbation probability $p > 0$,*

$$A(x, x') = \left(\sum_{i=1}^N P_i \left[\prod_{j=1}^n (1 - |f_{K_{ij}}^{(j)}(x_1, \dots, x_n) - x'_j|) \right] \right) \times (1 - p)^n + p^{\eta(x, x')} \times (1 - p)^{n - \eta(x, x')} \times 1_{[x \neq x']},$$

where $\eta(x, x') = \sum_{i=1}^n (x_i \oplus x'_i)$ is the Hamming distance between vectors x and x' , P_i is given in (2), and $1_{[x \neq x']}$ is an indicator function that is equal to 1 only when $x \neq x'$.

PROOF. The two terms in Theorem 2 essentially correspond to the two cases in Equation (6). First, consider the case when no gene is perturbed or equivalently, $\gamma =$

$(0, \dots, 0)$. This occurs with probability $(1-p)^n$. Thus, the next state is determined via the Boolean functions selected at that time step. The probability of transitioning from $x = (x_1, \dots, x_n)$ to $x' = (x'_1, \dots, x'_n)$, then, is equal to the sum of the probabilities of all network realizations \mathbf{f}_k such that $\mathbf{f}_k(x_1, \dots, x_n) = (x'_1, \dots, x'_n)$, $k = 1, 2, \dots, N$. Thus, given that no perturbation occurred,

$$A(x, x') = \sum_{i: \mathbf{f}_i(x) = x'} P_i,$$

which, in terms of the individual Boolean functions, can be expressed as

$$\sum_{i=1}^N P_i \left[\prod_{j=1}^n (1 - |f_{K_{ij}}^{(j)}(x_1, \dots, x_n) - x'_j|) \right],$$

where we treat binary values as real values (cf. Equation (3)). This is in fact the transition probability when $p = 0$, as shown in Shmulevich *et al.* (2002).

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 $\eta(x, x')$, which is 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 $1_{[x \neq x']}$.

If the perturbation vector γ is not identically distributed (i.e. some genes are more likely to get flipped), then the above transition probabilities become slightly more complicated, requiring products of individual probabilities $\Pr\{\gamma_i = 1\}$. It can be seen from Theorem 2 that the transition probability between two different states cannot be zero so long as $p > 0$.

A practical benefit of the randomization afforded by gene perturbation is that it empirically simplifies various computations involving PBNs. For example, consider the computation of influence $I_k(f_j^{(i)})$ of gene x_k on the predictor function $f_j^{(i)}$, as given in Equation (4). The computation of influence of a gene on the predictor entails computing the joint distribution $D(x)$ of all the genes used by that predictor, in order to compute the expectation of the partial derivatives of the predictors. This distribution, however, should be consistent with the model itself. For example, if we wish to quantify ‘long-term’ influence, we need to obtain the stationary distribution of the Markov chain corresponding to the PBN. Obtaining these long-run probabilities, however, may be problematic from an

empirical point of view, since the Markov chain may consist of a number of irreducible subchains and these probabilities will depend on the initial starting state. In other words, depending on where we start the process, we may end up in different irreducible subchains. Obtaining long-run behavior directly from the state-transition matrix A may also be impractical even for moderate sizes of PBNs, thus requiring simulation-based analysis.

The assumption of random gene perturbation, described above, solves this problem by ridding us of the dependence on the initial starting state. Since all states communicate, according to Proposition 1, the steady-state distribution is the same as the stationary distribution and by letting the process run for a sufficiently long time, we can empirically compute the distribution $D(x)$ simply by keeping track of the proportion of time each combination of values of the genes in the domain of the predictor occurs. For instance, if the predictor is a function of some given three variables, then we simply have to tabulate the frequency of appearance of each of the eight combinations of these three variables to obtain the necessary distribution in order to compute the influence on that predictor. Let us illustrate these ideas with an example.

EXAMPLE 1. Consider a PBN consisting of three genes $V = (x_1, x_2, x_3)$ and the function sets $F = (F_1, F_2, F_3)$, where $F_1 = \{f_1^{(1)}, f_2^{(1)}\}$, $F_2 = \{f_1^{(2)}\}$, and $F_3 = \{f_1^{(3)}, f_2^{(3)}\}$. The function truth tables as well as selection probabilities $c_j^{(i)}$ are given in Example 1 in Shmulevich *et al.* (2002). Let us assume that the initial (starting) distribution of the Markov chain is the uniform distribution, that is, $D(x) = 1/8$ for all $x \in \{0, 1\}^3$. Using this distribution, we can compute the influence matrix Γ (see Example 2 in Shmulevich *et al.* 2002. At the next time step, however, the distribution of all the states is no longer uniform. In general, the distribution at a given time step can be obtained simply by multiplying the distribution at the previous time step by the state-transition matrix A . Therefore, if we would like to compute the influence matrix at an arbitrary time point, we must have the distribution vector corresponding to that time point. Similarly, if we would like to compute the long-term influence (i.e. influence after the network has reached equilibrium), we must have the stationary distribution vector. Let us suppose that the perturbation probability is equal to $p = 0.01$ and see how the influence matrix changes over time. That is, for every step of the network, we will recompute the influence matrix. Let us focus on the influence of gene x_2 on the other three genes (i.e. row 2 of the influence matrix). Figure 1 shows the trajectories for these three influences for the first 100 time steps. First, it can be seen that the influences indeed converge to their asymptotic values. Second, it is worthwhile noting

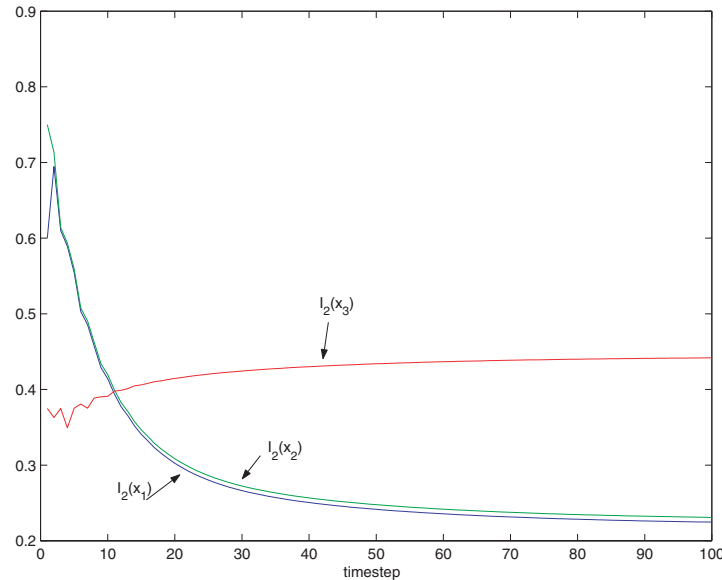


Fig. 1. The trajectories of the influences $I_2(x_i)$ for $i = 1, 2, 3$, plotted as a function of the time-steps taken by the PBN given in Example 1. The gene perturbation probability is equal to $p = 0.01$.

that the ‘transient’ influence (e.g. first 10 time steps in this example) can be very different from the long-term influence. For example, the influence of x_2 on x_3 is the lowest at the beginning and is the highest at the end. The important thing to note here is that the long-term influences are guaranteed to be independent of the initial starting state or distribution because a non-zero gene perturbation probability was used.

4 INTERVENTION

In Section 3, we considered the effects of random gene perturbations. In a similar vein, one can consider the effects of deliberately affecting a particular gene by means of intervention. One of the key goals of PBN modeling is the determination of possible intervention targets (genes) such that the network can be ‘persuaded’ if not forced to transition into a desired state or set of states. Whereas in Boolean networks, attractors are hypothesized to correspond to functional cellular states (Huang, 1999), in PBNs, this role is played by irreducible subchains. When the probability of perturbation, p , is equal to zero, a PBN is unable to escape from an irreducible subchain, implying that the cellular state cannot be altered. When p becomes positive, there is a chance that the current cellular state may switch to another cellular state by means of a random gene perturbation. Clearly, perturbation of certain genes is more likely to achieve the desired result than that of some other genes. Our goal, then, is to discover which genes are the best potential ‘lever points,’ to borrow the terminology from Huang (1999), in the sense of having

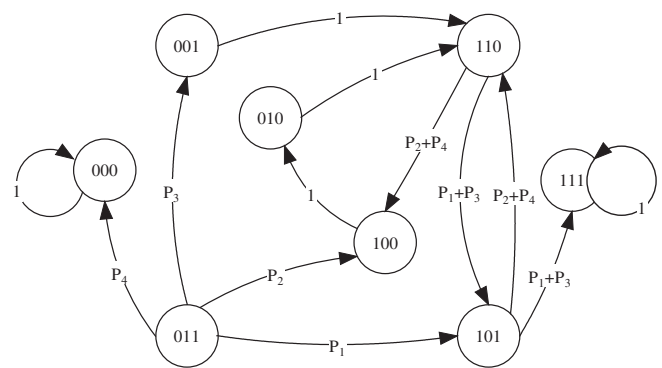


Fig. 2. State transition diagram corresponding to the PBN in Example 1.

the greatest possible impact on desired network behavior so that we can intervene with them by changing their value (1 or 0) as needed. In addition, we wish to be able to intervene with as few genes as possible in order to achieve our goals. To motivate the discussion, let us illustrate the idea with an example. We will use the PBN given in Example 1, which is also used in Shmulevich *et al.* (2002).

Suppose the state transition diagram of the Markov chain corresponding to the PBN in Example 1 is shown in Figure 2. For the predictor probabilities given in Example 1, the probabilities of the four possible network realizations are: $P_1 = 0.3$, $P_2 = 0.3$, $P_3 = 0.2$, and $P_4 = 0.2$. Suppose that we are currently in state (111)

and wish to eventually transition to state (000). Finally, let us assume, for the moment, that the probability of random perturbation is zero ($p = 0$). The question is, with which of the three genes, x_1 , x_2 , or x_3 , should we intervene such that the probability is greatest that we will end up in (000). By direct inspection of the diagram in Figure 2, we can see that if we make $x_1 = 0$, then with probability $P_4 = 0.2$, we will transition into (000) whereas if we make $x_2 = 0$ or $x_3 = 0$, then it will be impossible for us to end up in (000) and with probability 1, we will eventually come back to (111), where we started. In other words, the network will be resistant to perturbations of the second or third genes and will eventually maintain the same state. Thus, the answer to our question in this rather simple example is that only by intervening with gene x_1 do we have a chance of achieving our goal. In order for us to be able to answer such questions in general, we need to develop several tools.

When $p > 0$, by Proposition 1, the entire Markov chain is ergodic and thus, every state will eventually be visited. Thus, the question of intervention should be posed in the sense of *reaching a desired state as early as possible*. For instance, in the example considered above, if p is very small and we are in state (111), then it will be a long time until we reach (000) and setting $x_1 = 0$ is much more likely to get us there faster. We are, therefore, interested in the probability $F_k(x, y)$ that, starting in state x , the first time the PBN will reach some given state y will be at time k . This is often referred to as the *first passage time* from state x to state y . A related measure of interest is the mean first passage time from state x to state y , defined as

$$M(x, y) = \sum_k k F_k(x, y). \quad (7)$$

This measure tells us how long, on the average, it will take to get from state x to state y .

It is easy to see that for $k = 1$, $F_k(x, y) = A(x, y)$, which is just the transition probability from x to y . For $k \geq 2$, it is also straightforward to show (e.g. Çinlar, 1997) that $F_k(x, y)$ satisfies

$$F_k(x, y) = \sum_{z \in \{0,1\}^n - \{y\}} A(x, z) F_{k-1}(z, y). \quad (8)$$

Every required entry of the matrix A can be computed directly using Theorem 2. Let us illustrate this computation with the same example given above.

Suppose, as before, that $p = 0.01$. Then, the steady-state distribution equals [0.0752 0.0028 0.0371 0.0076 0.0367 0.0424 0.0672 0.7310], where the leftmost element corresponds to (000) and the rightmost to (111). As expected, the PBN spends much more time in state (111) than in any other state. In fact, more than 70% of the time is spent in that state. Let our starting state x be (111) and

the destination state y be (000), as before. The question with which we concern ourselves is whether we should intervene with gene x_1 , x_2 , or x_3 . In other words, we would like to compute $F_k((011), (000))$, $F_k((101), (000))$, and $F_k((110), (000))$, where the states are written in their binary representations. We can then assess our results by plotting

$$H_{K_0}(x, y) = \sum_{k=1}^{K_0} F_k(x, y)$$

for the states x of interest and for a sufficiently large K_0 . The intuition behind this approach is the following. Since the events {the first passage time from x to y will be at time k } are disjoint for different values of k , the sum of their probabilities for $k = 1, \dots, K_0$ is equal to the probability that the network, starting in state x , will visit state y before time K_0 . As a special case, when $K_0 = \infty$, this is equal to the probability that the chain *ever* visits state y , starting at state x , which of course is equal to 1, since our chains are ergodic if $p > 0$. Figure 3 shows the plots of $H_{K_0}(x, y)$ for $K_0 = 1, \dots, 20$ and for the three states of interest, namely, (011), (101), and (110).

The plots indicate that if we start with state (011), we are much more likely to enter state (000) sooner than if we start with states (110) or (101). For example, during the first 20 steps, we have an almost 25% chance of entering (000) if we start with (011), whereas if we start with (110) or (101), we only have about a 5% chance. This, in turn, indicates that we should intervene with gene x_1 rather than with gene x_2 or x_3 . Of course, in this rather simple example, we could have discerned this by visual inspection of Figure 2, but for larger networks, this method provides a tool for answering these kinds of questions.

In biology, there are numerous examples when the (in)activation of one gene or protein can lead much quicker (or with a higher probability) to a certain cellular functional state or phenotype than the (in)activation of another gene or protein. For instance, let's use a stable cancer cell line as an example. Without any intervention, the cells will keep proliferating. Let us assume that the goal of the intervention is to push the cell into programmed cell death (apoptosis). Let us further assume that we will achieve this intervention with two gene candidates: p53 and telomerase. The p53 gene is the most well-known tumor suppressor gene, encoding a protein that regulates the expression of several genes such as Bax and Fas/APO1 that function to promote apoptosis (Miyashita and Reed, 1995; Owen-Schaub *et al.*, 1995) and p21/WAF1 that functions to inhibit cell growth (El-Deiry *et al.*, 1993). The telomerase gene encodes telomerase, which maintains the integrity of the end of chromosomes (telomeres) in our germ cells, which are responsible for propagating our complete genetic material to the following generation, as well as progenitor cells,

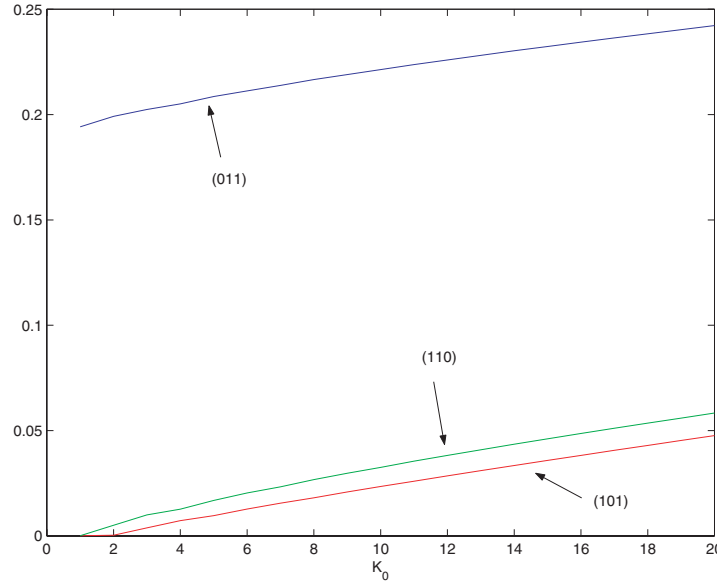


Fig. 3. $H_{K_0}(x^{(i)}, y)$ for $K_0 = 1, \dots, 20$, for starting states (011), (101), and (110), corresponding to perturbations of first, second, and third genes, respectively.

which are responsible for replenishing our cells during the normal cell turnover (homeostasis). In somatic cells, the telomerase gene is turned off, resulting in telomere shortening each time the cell divides—a key reason for the limited life span of our normal cells (Harley, 1991). In the majority of tumor cells, telomerase is activated, which is believed to contribute to the prolonged life-span of the tumor cells (Kim *et al.*, 1994) and worsened prognosis for the cancer patients (Hiyama *et al.*, 1995; Zhang *et al.*, 1996). Extensive experimental results indicate that when p53 is activated in the cells, for example, in response to radiation, the cells undergo rapid growth inhibition and apoptosis in as short as a few hours (Lowe *et al.*, 1993; Kobayashi *et al.*, 1998). In contrast, inhibition of the telomerase gene also leads to cell growth inhibition, differentiation, and cell death, but only after cells go through a number of cell divisions (allowing telomere shortening), which takes a longer time to occur than via p53.

Another valuable computational tool is the mean first passage times given in Equation (7). Intuitively, the best candidate gene for intervention should be the one that results in the smallest mean first passage time to the destination state. Using the same example as above, we have computed the three mean first passage times corresponding to the perturbation of genes x_1 , x_2 , and x_3 . These are equal to 337.51, 424.14, and 419.20, respectively. Since the first one is the smallest, this again supports that gene x_1 is the best candidate for intervention.

To summarize, we simply generate different states

$x^{(i)} = x \oplus e_i$, $i = 1, \dots, n$, where e_i is the unit binary vector with a 1 in the i th coordinate, by perturbing each of the n genes and compute $H_{K_0}(x^{(i)}, y)$ for some desired destination state y and constant K_0 . Then, the best gene for intervention is the one for which $H_{K_0}(x^{(i)}, y)$ is maximum. That is, given a fixed K_0 , the optimal gene $x_{i_{\text{opt}}}$ satisfies

$$i_{\text{opt}} = \arg \max_i H_{K_0}(x^{(i)}, y). \quad (9)$$

Alternatively, by minimizing the mean first passage times, the optimal gene satisfies

$$i_{\text{opt}} = \arg \min_i M(x^{(i)}, y). \quad (10)$$

Another related approach to the one in (9) might be to first fix a probability h_0 and wait until one of the $H_{K_0}(x^{(i)}, y)$ reaches it first. Note that due to ergodicity, for every state $x^{(i)}$, there will always be a $K_0^{(i)}$ large enough such that $H_{K_0^{(i)}}(x^{(i)}, y) > h_0$. In that sense, the optimal gene for intervention $x_{i_{\text{opt}}}$ is one for which

$$i_{\text{opt}} = \arg \min_i \min_{K_0^{(i)}} \{K_0^{(i)} : H_{K_0^{(i)}}(x^{(i)}, y) > h_0\}. \quad (11)$$

At first glance, it might appear as if both approaches, (9) and (11), will yield the same answer, since Figure 3 seems to suggest that the plots do not intersect and that if one of them is maximum for a given K_0 , it will be the first to reach any fixed h_0 thereafter. While it is true that for

sufficiently large K_0 , the plots will not intersect, this is not in general true for smaller values of K_0 .

The criteria imbedded in Equations (9) and (11) have underlying different interpretations. The first aims to *maximize the probability* of reaching a particular state before a certain fixed time while the second aims to *minimize the time* needed to reach a certain state with a given fixed probability. These two approaches are complementary and may be used in conjunction. Finally, the approach in (10) based on minimizing mean first passage times is another simple alternative. We will come back to mean first passage times in Section 5, when we discuss sensitivity analysis of PBNs.

4.1 Sets of states, avoidance of states, and permanent intervention

So far, we have discussed the notion of intervention in terms of a single starting state and a single destination state. However, we may often be more interested in the same types of questions, but concerning *sets* of states. For example, two different sets of states may correspond to different functional cellular states, such as proliferation or quiescence, much in the same way attractors play this role in standard Boolean networks (Huang, 1999). In PBNs, this role is typically played by irreducible subchains when no perturbations can occur ($p = 0$). In other words, once the network enters an irreducible subchain (cf. attractor), it can't escape. When the perturbation probability is positive, there are no longer any irreducible subchains (see Proposition 1), but the sets of states that correspond to these irreducible subchains when $p = 0$ still represent the functional states of the organism that is being modeled—there is now simply a probability of escaping due to random perturbations. Those sets of states that correspond to irreducible subchains when $p = 0$ could be referred to as *implicitly irreducible* subchains. They are essentially 'islands' of states and the probability of perturbation controls the amount of 'bridges' between these islands. When $p = 0$, there are no bridges, and when p becomes larger, it becomes easier to 'travel' between the islands.

Going back to the question of intervention, we may be interested in posing it as follows. Given that we are in a set of states X , what gene is the best candidate for intervention if we want to end up in the set of states Y ? The question may be posed in the sense of either (9), (10), or (11). Fortunately, the mathematical framework does not really change when we talk about sets of states. For example, if $X = \{x\}$ consists of just one state, but Y is a set comprised of many states, then the first passage probabilities $F_k(x, y)$ may simply be summed over all states $y \in Y$ and we can define $F_k(x, Y) = \sum_{y \in Y} F_k(x, y)$. Then, the same approaches as discussed above to find the best gene for intervention can be used.

The situation when X is comprised of a number of states is conceptually a bit more complicated, since now, the starting set of states X , rather than just one starting state x , represents a type of uncertainty in our knowledge of the current state of the network. That is, we may not know exactly in what state the network is in at a particular time, but we may know that it is in a certain set of states. This may be relevant not only from an experimental perspective, as it may be difficult to determine precisely the current state at a given time, but perhaps more importantly, we may not be interested in restricting ourselves just to one state, but rather consider a whole set of states X that is believed to correspond to the current functional cellular state.

Consequently, a gene that may be the best candidate for intervention for one of the starting states in X may not be the best for another state in X . Therefore, the best we can do in such a case is to combine the individual results for all states $x \in X$, but weigh them by their respective probabilities of occurrence. The latter is furnished by the steady-state probabilities π_x . In other words, we can define

$$F_k(X, Y) = \frac{\sum_{x \in X} \sum_{y \in Y} F_k(x, y) \cdot \pi_x}{\sum_{x \in X} \pi_x} \quad (12)$$

to be the first passage probability from a set X to a set Y .

In addition to reaching a desired state or set of states, we may also be interested in *avoiding* a state or set of states. This is quite natural in terms of inducing a network not to enter into some sets of states corresponding to unwanted functional cellular states (e.g. proliferation). This goal is in a sense complementary to what has been described above in terms of reaching a desired state either as soon as possible with a given probability or with as high probability as possible, before a given time. For example, in Equation (10), our goal was to minimize the mean first passage time to a destination state. In order to avoid a destination state, we simply have to maximize the mean first passage time to that state. So, the underlying mechanism is quite the same and we will not give a separate example illustrating the avoidance of states. We would like to point out, however, that it may be possible that performing no intervention whatsoever is the best option, regardless of whether we want to reach or avoid a state or set of states. In other words, depending on the network as well as on the starting and destination states or sets of states, it may be the case that not intervening with any gene is optimal in terms of the criteria given in (9), (10), or (11).

In our model, the interventions and perturbations that we have considered up to this point are in the gene's expression state, which is generally a transient phenomenon. Thus, it could be termed *transient* intervention or perturbation. That is, the effect on a gene, whether by random

perturbation or forced intervention, is applied at only one time point and the network itself is responsible for determining the values of that gene thereafter. It could be said that the effect has the potential to be reversed by the network itself. For example, in Figure 2, if we are in state (111) and the second gene changes value, resulting in (101), at the next time step, regardless of where the network transitions, (110) or (111), the second gene will always get changed back to 1 again. Since in that example (111) is an absorbing state, the network will eventually return to it, and the perturbation or intervention—whatever the means was of changing the second gene—will have been ‘compensated’ by the network itself. This inherent resistance to perturbations is a key factor for stability and robustness of PBNs.

We can also consider a *permanent* intervention or perturbation. In this scenario, a gene changes value and remains at that value forever. From a genetic perspective, permanent intervention is achieved through removing a gene or ‘transplanting’ a gene, as done in gene therapy. From a network perspective, the permanent intervention (or perturbation) of a gene essentially reduces the state space by half, since all the states in which that gene is not equal to the fixed value cannot appear. The rest of the genes are predicted as usual, via the Boolean functions and their selection probabilities $c_j^{(i)}$ remain unaltered. The Boolean function corresponding to the fixed gene is the identity function (0 or 1) with selection probability 1.

Permanent intervention by gene manipulation is used by both nature and humans. It is an efficient way to generate mutations and also hoped to be an efficient way for correcting mutations (therapy). Perhaps the best example for the first scenario is viral infection. Let us use Simian Virus 40 (SV40) as an example. SV40 virus was discovered in the 1950s during the development of vaccine for poliovirus (Carbone *et al.*, 1997). It was found that SV40 could transform monkey kidney cells and develop tumors when injected into rodents (Abrahams and Van der Eb, 1975). SV40 was not believed to cause tumor in human cells, however, SV40 DNA was found in some human brain tumors in recent years (Kouhata *et al.*, 2001) suggesting that SV40 may have a tumorigenic effect in humans too, although with a long latent period.

Extensive research has been carried out to elucidate how SV40 causes cancer in mouse cells. Though SV40 does not have a big genome, one of the most important proteins encoded by SV40 is large T-antigen. Large T-antigen interacts with host cell molecules and triggers a series of events that are beneficial for the viral replication and bad for the host cells. For example, T-antigen inactivates the functions of p53 (Zhu *et al.*, 1991; Bargonetti *et al.*, 1992), which may be the key mechanism for the tumorigenic effect of SV40 T-antigen. We should point out

that SV40 T-antigen also interacts with other molecules such as retinoblastoma (Rb DeCaprio *et al.*, 1988)—an important protein the activation of which inhibits DNA synthesis. From a network perspective, the permanent mutation caused by SV40 T-antigen may permanently alter the dynamics of the network, causing it to shift into a set of states associated with tumorigenesis. To further prove that T-antigen itself is sufficient to cause this effect, T-antigen was ‘transplanted’ into the mouse brain using a tissue-specific transgenic mouse model (second scenario, man-made event). As expected, brain tumors were found in many of the transgenic mice (Brinster *et al.*, 1984). Since SV40 DNA was detected in some human brain tumors, one cannot help but to speculate that SV40 may be causing human brain tumors too.

From the point of view of man-made intervention, it may be that permanent rather than transient intervention is the only way to reach a desired set of states. That is, it may be the case that the network is so resistant to transient intervention of any gene, that it will be extremely unlikely for the network to ever reach (or avoid) the desired states. Permanent intervention, though less desirable as it introduces permanent changes to the network, may be the only alternative to reach a set of states with a sufficiently high probability. The question, as before, is what genes are the most likely ‘lever points’ for controlling the global behavior of the network.

For example, based on what is known, p53 is one such gene. This is clearly demonstrated by the fact that p53 gene deletion or mutation (permanent perturbation) is one of the most frequent genetic changes in cancers (Hollstein *et al.*, 1991). Removing p53 genes from mouse through embryonic stem cell gene knock-out technology, researchers generated the p53 null mice. The mouse can be born normally and develop into adult normally, but develop cancers in most of the mice at 4.5 months (Donehower *et al.*, 1992). So p53 may be an important lever gene for regulation of homeostasis—a delicate balance between cell growth and cell death. Thus, it may not be surprising that p53 is often selected as a therapeutic target for permanent intervention. In cultured cells, the introduction of p53 back to p53-null cells leads to cell growth inhibition or cell death (El-Deiry *et al.*, 1993). Thus one properly chosen lever gene has the potential to lead the network into a specific implicitly irreducible subchain (cf. attractor in standard Boolean networks). p53 gene is also being used in gene therapy, where the target gene (p53 in this case) is cloned into a viral vector (adenovirus vector is a common one). The modified virus serves as a vehicle to transport p53 gene into the tumor cells to generate a permanent intervention (Swisher *et al.*, 1999; Bouvet *et al.*, 1998).

5 SENSITIVITY OF STATIONARY DISTRIBUTIONS TO GENE PERTURBATIONS

In this section, we briefly address the question of sensitivity of the stationary distributions to random gene perturbations, as discussed in Section 3. This is an important issue because it characterizes the effect of perturbations on long-term network behavior. It is clear that whatever is meant by sensitivity, it will no doubt depend on the probability of random perturbation, p . The general question is: if we perturb the transition probabilities, how much will the stationary distributions, or equivalently, the limiting probabilities change? This question has generally been addressed in the area known as perturbation theory of stochastic matrices and dates back to Schweitzer (1968). If A and $\tilde{A} = A - E$ are the original and ‘perturbed’ Markov matrices, where E represents the perturbation, and π and $\tilde{\pi}$ are their respective stationary distributions, then most results are of the form

$$\|\tilde{\pi} - \pi\| \leq \kappa \|E\|, \text{ or } \left| \frac{\pi_j - \tilde{\pi}_j}{\pi_j} \right| \leq \kappa_j \|E\|,$$

for some matrix norm $\|\cdot\|$, and κ, κ_j are called condition numbers and are used as measures of sensitivity. Recently, a new approach to measure the sensitivity of the Markov chain to perturbations, in terms of mean first passage times, has been proposed by Cho and Meyer (2000). This approach has the advantage in that it does not require computing or estimating the condition numbers. The result is given in the following Theorem.

THEOREM 3 (CHO AND MEYER, 2000). *Let A and $\tilde{A} = A - E$ be transition probability matrices for two irreducible Markov chains with respective stationary distributions π and $\tilde{\pi}$. Denote by $\|E\|_\infty$ the infinity-norm of E , which is the maximum over the row sums $\sum_j |E(i, j)|$. Let $M(x, y) = \sum_k k F_k(x, y)$ denote the mean first passage time from state x to state y in the chain corresponding to A . Then, the relative change in the limiting probability for state y is*

$$\left| \frac{\pi_y - \tilde{\pi}_y}{\pi_y} \right| \leq \frac{1}{2} \|E\|_\infty \max_{x \neq y} M(x, y).$$

Cho and Meyer (2000) also showed that their bound is tight in the sense that there always exists a perturbation E that attains the bound. Let us now consider this result in the context of random gene perturbations.

THEOREM 4. *Given a PBN $G(V, F)$ with an existing steady-state distribution, let π_y be a limiting probability of state y when $p = 0$ (no perturbations) and let $\tilde{\pi}_y$ be the*

limiting probability of the same state when $0 < p < 1/2$. Then,

$$\left| \frac{\pi_y - \tilde{\pi}_y}{\pi_y} \right| \leq (1 - (1 - p)^n) \max_{x \neq y} M(x, y).$$

PROOF. The perturbation matrix E from Theorem 3 can be expressed directly from Theorem 2 as follows. Let $E(x, x')$ be the entry in E corresponding to the transition probability from x to x' , for $x, x' \in \{0, 1\}^n$. Also, let

$$A(x, x') = \sum_{i=1}^N P_i \left[\prod_{j=1}^n (1 - |f_{K_{ij}}^{(j)}(x_1, \dots, x_n) - x'_j|) \right]$$

denote the transition matrix when $p = 0$ (see Shmulevich *et al.* 2002) and $\tilde{A}(x, x')$ denote the transition matrix given in Theorem 2, where a non-zero perturbation probability is assumed. In other words,

$$\tilde{A}(x, x') = A(x, x') \times (1 - p)^n + p^{\eta(x, x')} \times (1 - p)^{n - \eta(x, x')} \times 1_{[x \neq x']}. \quad (13)$$

Then, $E(x, x') = A(x, x') - \tilde{A}(x, x')$ and for each row of E , we have

$$\begin{aligned} \sum_{x'} |E(x, x')| &= \sum_{x'} |A(x, x') \times (1 - (1 - p)^n) \\ &\quad - p^{\eta(x, x')} \times (1 - p)^{n - \eta(x, x')} \times 1_{[x \neq x']}| \\ &\leq \sum_{x'} (|A(x, x') \times (1 - (1 - p)^n)| \\ &\quad + |p^{\eta(x, x')} \times (1 - p)^{n - \eta(x, x')} \times 1_{[x \neq x']}|). \end{aligned} \quad (14)$$

First, we observe that since $\sum_{x'} A(x, x') = 1$, the first term of the summation in (14) is simply equal to $(1 - (1 - p)^n)$. Next, we have

$$\begin{aligned} \sum_{x'} |p^{\eta(x, x')} \times (1 - p)^{n - \eta(x, x')} \times 1_{[x \neq x']}| \\ = \sum_{x' \neq x} p^{\eta(x, x')} \times (1 - p)^{n - \eta(x, x')}, \end{aligned} \quad (15)$$

where we can remove the absolute value symbols since each summand is positive. Since the summation in Equation (15) is taken over all possible values of x' except $x' = x$, the Hamming distance $\eta(x, x')$ ranges from 1 to n . As there are $\binom{n}{k}$ states x' that are Hamming distance k from x (i.e. $|\{x' : \eta(x, x') = k\}| = \binom{n}{k}$), Equation (15) can be rewritten as

$$\begin{aligned} \sum_{x' \neq x} p^{\eta(x, x')} \times (1 - p)^{n - \eta(x, x')} &= \sum_{k=1}^n \binom{n}{k} p^k (1 - p)^{n - k} \\ &= 1 - (1 - p)^n. \end{aligned} \quad (16)$$

Thus, every row of E satisfies

$$\sum_{x'} |E(x, x')| \leq 2(1 - (1 - p)^n)$$

and so

$$\|E\|_{\infty} \leq 2(1 - (1 - p)^n) \quad (17)$$

as well.

Using (17) together with Theorem 3 gives the desired result.

Theorem 4 allows us to bound the sensitivity of the limiting probabilities of any state of the PBN, relative to the probability of random gene perturbation. The mean first passage times $M(x, y)$ can be computed in a straightforward way by using the recursive formula in (8). The same type of analysis as above may be conducted between two PBNs with different perturbation probabilities $p_1 < p_2$ and the relative sensitivity of the limiting probabilities can be expressed in terms of p_1 , p_2 , and the mean first passage times. One important implication of Theorem 4 is that if a particular state of a PBN can be ‘easily reached’ from other states, meaning that the mean first passage times are small, then its steady-state probability will be relatively unaffected by perturbations. Such sets of states, if we hypothesize them to correspond to some functional cellular states, are thus relatively insensitive to random gene perturbations.

6 CONCLUSION

The complex interplay of the elements in a genetic regulatory network implies that any individual element or group of elements exerts an effect on the entire network. The extent of this effect depends on the nature of the relationships between the elements as well as on the state of the network. This paper is concerned with two related questions. Given the possibility of a random gene perturbation with a certain probability, to what extent do such perturbations affect the long-term behavior of the entire network? Alternately, given a desire to elicit certain behavior from the network, what genes would make the best candidates for intervention so as to increase the likelihood of this behavior?

The first question has been addressed by constructing an explicit formulation of the state-transition probabilities in terms of the Boolean functions and the probability of perturbation, and then deriving a bound on the steady-state probabilities, given in Theorem 4. In concordance with intuition, an interesting implication of this theorem is that the steady-state probabilities of those states of the network to which it is easy to transition from other states, in terms of mean first-passage times, are more resilient to random gene perturbations. The first passage times provide a conceptual link with the second question in that

they furnish the means by which we develop the tools for finding the best candidate genes for intervention.

The problem of capturing long-run network behavior for large-size networks is difficult owing to the exponential increase of the state space. Matrix-based methods quickly become prohibitive. We plan to focus on effective strategies for obtaining steady-state behavior through simulation and efficient data structures. Alternately, it may be important to be able to select relatively small sub-networks, out of a large network, that function more or less independently of the rest of the network. Such a small sub-network would require little or no information from the outside. Algorithms for efficiently finding such sub-networks, inferred from real gene-expression data, along with a formal representation for performing such reductions, will be part of future work.

ACKNOWLEDGEMENTS

The authors are grateful to the Referees for their careful reading of the manuscript and many helpful suggestions. Many thanks go to Professor Ioan Tabus for stimulating and helpful discussions on long-term influence of genes and to Dr Ilya Gluhovsky for his careful reading and suggestions on the first draft of this paper. This work was partially supported by the Tobacco Settlement Funds as appropriated by the Texas State Legislature, by a generous donation from the Michael and Betty Kadoorie Foundation, and by a grant from the RGK Foundation.

REFERENCES

- Abrahams,P.J. and Van der Eb,A.J. (1975) *In vitro* transformation of rat and mouse cells by DNA from simian virus 40. *J. Virol.*, **16**, 206–209.
- Altenberg,L. (1994) The evolution of evolvability in genetic programming. In Kinnear,K.E. (ed.), *Advances in Genetic Programming*. MIT Press, Cambridge, MA, pp. 47–74.
- Bargonetti,J., Reynisdottir,I., Friedman,P. and Prives,C. (1992) Site-specific binding of wild-type p53 to cellular DNA is inhibited by SV40 T antigen and mutant p53. *Genes Dev.*, **6**, 1886–1898.
- Bouvet,M., Bold,R.J., Lee,J., Evans,D.B., Abbruzzese,J.L., Chiao,P.J., McConkey,D.J., Chandra,J., Chada,S., Fang,B. and Roth,J.A. (1998) Adenovirus-mediated wild-type p53 tumor suppressor gene therapy induces apoptosis and suppresses growth of human pancreatic cancer. *Ann. Surg. Oncol.*, **5**, 681–688.
- Brinster,R.L., Chen,H.Y., Messing,A., van Dyke,T., Levine,A.J. and Palmiter,R.D. (1984) Transgenic mice harboring SV40 T-antigen genes develop characteristic brain tumors. *Cell*, **37**, 367–379.
- Carbone,M., Rizzo,P. and Pass,H.I. (1997) Simian virus 40, poliovaccines and human tumors: a review of recent developments. *Oncogene*, **15**, 1877–1888.
- Cho,G.E. and Meyer,C.D. (2000) Markov chain sensitivity measured by mean first passage times. *Linear Algebra and Its Applications*, **316**, 21–28.

- Çınlar, E. (1997) *Introduction to Stochastic Processes*. Prentice Hall, Englewood, NJ.
- DeCaprio, J.A., Ludlow, J.W., Figge, J., Shew, J.Y., Huang, C.M., Lee, W.H., Marsilio, E., Paucha, E. and Livingston, D.M. (1988) SV40 large tumor antigen forms a specific complex with the product of the retinoblastoma susceptibility gene. *Cell*, **54**, 275–283.
- Donehower, L.A., Harvey, M., Slagle, B.L., McArthur, M.J., Montgomery, Jr, C.A., Butel, J.S. and Bradley, A. (1992) Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature*, **356**, 215–221.
- Dougherty, E.R., Kim, S. and Chen, Y. (2000) Coefficient of determination in nonlinear signal processing. *Signal Process.*, **80**, 2219–2235.
- El-Deiry, W.S., Tokino, T., Velculescu, V.E., Levy, D.B., Parsons, R., Trent, J.M., Lin, D., Mercer, W.E., Kinzler, K.W. and Vogelstein, B. (1993) WAF1, a potential mediator of p53 tumor suppression. *Cell*, **75**, 817–825.
- Glass, K. and Kauffman, S.A. (1973) The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.*, **39**, 103–129.
- Goldberg, D. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA.
- Harley, C.B. (1991) Telomere loss: mitotic clock or genetic time bomb? *Mutation Res.*, **256**, 271–282.
- Hiyama, E., Hiyama, K., Yokoyama, T., Matsuura, Y., Piatyszek, M.A. and Shay, J.W. (1995) Correlating telomerase activity levels with human neuroblastoma outcomes. *Nat. Med.*, **1**, 249–255.
- Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R., Zhou, Y., Kay, S.A., Schultz, P.G. and Cooke, M.P. (2001) A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell*, **106**, 413–415.
- Hollstein, M., Sidransky, D., Vogelstein, B. and Harris, C.C. (1991) p53 mutations in human cancers. *Science*, **253**, 49–53.
- Huang, S. (1999) Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.*, **77**, 469–480.
- Huang, S. and Ingber, D.E. (2000) Regulation of cell cycle and gene activity patterns by cell shape: evidence for attractors in real regulatory networks and the selective mode of cellular control. *InterJournal Genetics*, **MS: 238**, <http://www.interjournal.org>
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Kauffman, S.A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, **22**, 437–467.
- Kauffman, S.A. (1993) *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, Oxford.
- Kauffman, S.A. and Levin, S. (1987) Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, **128**, 11–45.
- Kim, N.W., Piatyszek, M.A., Prowse, K.R., Harley, C.B., West, M.D., Ho, P.L.C., Coviello, G.M., Wright, W.E., Weinrich, S.L. and Shay, J.W. (1994) Specific association of human telomerase activity with immortal cells and cancer. *Science*, **266**, 2011–2015.
- Kim, S., Dougherty, E.R., Chen, Y., Sivakumar, K., Meltzer, P., Trent, J.M. and Bittner, M. (2000a) Multivariate measurement of gene expression relationships. *Genomics*, **67**, 201–209.
- Kim, S., Dougherty, E.R., Bittner, M.L., Chen, Y., Sivakumar, K., Meltzer, P. and Trent, J.M. (2000b) General nonlinear framework for the analysis of gene interaction via multivariate expression arrays. *J. Biomed. Optics*, **5**, 411–424.
- Kobayashi, T., Ruan, S.-B., Jabbur, J.R., Consoli, U., Clodi, K., Shiku, H., Owen-Schaub, L., Andreeff, M., Reed, J. and Zhang, W. (1999) Differential p53 phosphorylation and activation of apoptosis-promoting genes Bax and Fas/APO-1 by radiation and ara-C treatment. *Cell Death and Differentiation*, **85**, 584–591.
- Kouhata, T., Fukuyama, K., Hagihara, N. and Tabuchi, K. (2001) Detection of simian virus 40 DNA sequence in human primary glioblastomas multiforme. *J. Neurosurg.*, **95**, 96–101.
- Lowe, S.W., Schmitt, E.M., Smith, S.W., Osborne, B.A. and Jacks, T. (1993) p53 is required for radiation-induced apoptosis in mouse thymocytes. *Nature*, **362**, 847–849.
- Miyashita, T. and Reed, J.C. (1995) Tumor suppressor p53 is a direct transcriptional activator of the human bax gene. *Cell*, **80**, 293–299.
- Owen-Schaub, L.B., Zhang, W., Cusack, J., Angelo, L.S., Santee, S.M., Fujiwara, T., Roth, J.A., Deisseroth, A.B., Zhang, W.-W., Kruzel, E. and Radinsky, R. (1995) Wild-type and a temperature-sensitive mutant of human p53 induce Fas/APO-1 expression. *Mol. Cell. Biol.*, **15**, 3032–3040.
- Schweitzer, P.J. (1968) Perturbation theory and finite Markov chains. *J. Appl. Probability*, **5**, 401–413.
- Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W. (2002) Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**, 261–274.
- Somogyi, R. and Sniegowski, C. (1996) Modeling the complexity of gene networks: understanding multigenic and pleiotropic regulation. *Complexity*, **1**, 45–63.
- Swisher, S.G., Roth, J.A., Nemunaitis, J., Lawrence, D.D., Kemp, B.L., Carrasco, C.H., Connors, D.G., El-Naggar, A.K., Fossella, F., Glisson, B.S. et al. (1999) Adenovirus-mediated p53 gene transfer in advanced non-small-cell lung cancer. *J. Natl Cancer Institute*, **91**, 763–771.
- Szallasi, Z. and Liang, S. (1998) Modeling the normal and neoplastic cell cycle with realistic boolean genetic networks: their application for understanding carcinogenesis and assessing therapeutic strategies. *Pac. Symp. Biocomput.*, **3**, 66–76.
- Thomas, R., Thieffry, D. and Kaufman, M. (1995) Dynamical behaviour of biological regulatory networks—I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.*, **57**, 247–276.
- Wuensche, A. (1996) Discrete Dynamics Lab (DDLab), <http://www.santafe.edu/~wuensch/ddlab.html>.
- Wuensche, A. (1998) Genomic regulation modeled as a network with basins of attraction. *Pac. Symp. Biocomput.*, **3**, 89–102.
- Zhang, W., Piatyszek, M.A., Kobayashi, T., Estey, E., Andreeff, M., Deisseroth, A.B. and Shay, J.W. (1996) Detection of telomerase activity in human acute myelogenous leukemia and modulation of the activity by differentiation-inducing agents. *Clin. Cancer Res.*, **2**, 799–803.
- Zhu, J., Abate, M., Rice, P.W. and Cole, C. (1991) The ability of simian virus 40 large T antigen to immortalize primary mouse embryo fibroblasts cosegregates with its ability to bind p53. *J. Virol.*, **65**, 6872–6880.