7 Modeling genetic regulatory networks with probabilistic Boolean networks

Ilya Shmulevich and Edward R. Dougherty

7.1. Introduction

High-throughput genomic technologies such as microarrays are now allowing scientists to acquire extensive information on gene activities of thousands of genes in cells at any physiological state. It has long been known that genes and their products in cells are not independent in the sense that the activation of genes with subsequent production of proteins is typically jointly dependent on the products of other genes, which exist in a highly interactive and dynamic regulatory network composed of subnetworks and regulated by rules. However, discovering the network structure has thus far proved to be elusive either because we lack sufficient information on the components of the network or because we lack the necessary multidisciplinary approaches that integrate biology and engineering principles and computational sophistication in modeling. During the past several years a new mathematical rule-based model called probabilistic Boolean networks (PBN) has been developed to facilitate the construction of gene regulatory networks to assist scientists in revealing the intrinsic gene-gene relationships in cells and in exploring potential network-based strategies for therapeutic intervention (Shmulevich et al. [1, 2, 3, 4, 5, 6], Datta et al. [7, 8], Kim et al. [9], Zhou et al. [10], and Hashimoto et al. [11]). There is already evidence that PBN models can reveal biologically relevant gene regulatory networks and can be used to predict the effects of targeted gene intervention. A key goal of this chapter is to highlight some important research problems related to PBNs that remain to be solved, in hope that they will stimulate further research in the genomic signal processing and statistics community.

7.2. Background

Data comprised of gene expression (mRNA abundance) levels for multiple genes is typically generated by technologies such as the DNA microarray or chip. The role

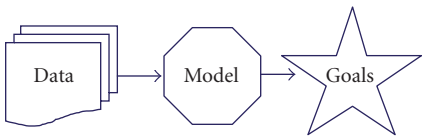


Figure 7.1

of the dynamical model or network is to simulate, via iteration of explicit rules, the dynamics of the underlying system presumed to be generating the observations. Such simulations can be useful for making predictions while the rules themselves, characterizing the relationships between the expressions of different genes, may hold important biological information. Time-course data, which are measurements taken at a number of time points, are often used for the inference of the model.

Before we discuss Probabilistic Boolean Networks, it may be worthwhile to pose several general but fundamental questions concerning modeling of genetic regulatory networks. The first and perhaps the most important question is the following.

7.2.1. What class of models should we choose?

We would like to argue that this choice must be made in view of (1) the data requirements and (2) the goals of modeling and analysis. Indeed, as shown in Figure 7.1 data is required to infer the model parameters from observations in the physical world, while the model itself must serve some purpose, in particular, prediction of certain aspects of the system under study. Simply put, the type and quantity of data that we can gather together with our prescribed purpose for using a model should be the main determining factors behind choosing a model class.

The choice of a model class involves a classical tradeoff. A *fine-scale* model with many parameters may be able to capture detailed *low-level* phenomena such as protein concentrations and reaction kinetics, but will require very large amounts of highly accurate data for its inference, in order to avert overfitting the model. In contrast, a *coarse-scale* model with lower complexity may succeed in capturing *high-level* phenomena, such as which genes are *on* or *off*, while requiring smaller amounts of more coarse-scale data. In the context of genetic regulatory systems, fine-scale models, typically involving systems of differential equations, can be applied to relatively small and isolated genetic circuits for which many types of accurate measurements can be made. On the other hand, coarse-scale models are more suited to global (genome-wide) measurements, such as those produced by microarrays. Such considerations should drive the selection of the model class. Needless to say, according to the principle of Ockham's razor, which underlies all scientific theory building, the model complexity should never be made higher than what is necessary to faithfully "explain the data" (Shmulevich [12]).

There is a rather wide spectrum of approaches for modeling gene regulatory networks, each with its own assumptions, data requirements, and goals, including linear models (van Someren et al. [13], D'haeseleer [14]), Bayesian networks (Murphy and Mian [15], Friedman et al. [16], Hartemink et al. [17], Moler et al. [18]), neural networks (Weaver et al. [19]), differential equations (Mestl et al. [20], Chen et al. [21], Goutsias and Kim [22]), as well as models including stochastic components on the molecular level (McAdams and Arkin [23]; Arkin et al. [24]) (see Smolen et al. [25], Hasty et al. [26], and de Jong [27] for reviews of general models).

The model system that has received, perhaps, the most attention is the Boolean network model originally introduced by Kauffman (Kauffman [28], Glass and Kauffman [29]). Good reviews can be found in Huang [30], Kauffman [31], Somogyi and Sniegoski [32], Aldana et al. [33]. In this model, the state of a gene is represented by a Boolean variable (on or off) and interactions between the genes are represented by Boolean functions, which determine the state of a gene on the basis of the states of some other genes.

One of the appealing properties of Boolean networks is that they are inherently simple, emphasizing generic principles rather than quantitative biochemical details, but are able to capture the complex dynamics of gene regulatory networks. Computational models that reveal these logical interrelations have been successfully constructed (Bodnar [34], Yuh et al. [35], Mendoza et al. [36], Huang and Ingber [37]). Let us now pose several questions related to this class of models.

7.2.2. To what extent do such models represent reality?

This question pertains more to modeling in general. All models only approximate reality by means of some formal representation. It is the degree to which we hope to approximate reality and, more importantly, our goals of modeling, namely, to acquire knowledge about some physical phenomenon that determines what class of models should be chosen. In the context of Boolean networks as models of genetic regulatory networks, the binary approximation of gene expression is only suitable to capture those aspects of regulation that possess a somewhat binary character. Even though most biological phenomena manifest themselves in the continuous domain, we often describe them in a binary logical language such as "on and off," "up-regulated and down-regulated," and "responsive and nonresponsive." Moreover, recent results suggest that gene regulation may indeed function "digitally" (Lahav et al. [38]). Before embarking on modeling gene regulatory networks with a Boolean formalism, it is prudent to test whether or not meaningful biological information can be extracted from gene expression data entirely in the binary domain. This question was taken up by Shmulevich and Zhang [39]. They reasoned that if the gene expression levels, when quantized to only two levels (1 or 0), would not be informative in separating known subclasses of tumors, then there would be little hope for Boolean modeling of realistic genetic networks based on gene expression data. Fortunately, the results

were very promising. By using binary gene expression data, generated via cDNA microarrays, and the Hamming distance as a similarity metric, they were able to show a clear separation between different subtypes of gliomas (a similar experiment was also performed for sarcomas), using multidimensional scaling. This seems to suggest that a good deal of meaningful biological information, to the extent that it is contained in the measured continuous-domain gene expression data, is retained when it is binarized. Zhou et al. [40] took a similar approach, but in the context of classification. The revealing aspect of their approach is that classification using binarized expressions proved to be only negligibly inferior to that using the original continuous expression values, the difference being that the genes derived via feature selection in the binary setting were different than the ones selected in the continuous. The expression values of those possessing binary-like behavior fell into bimodal distributions and these were naturally selected for classification.

7.2.3. Do we have the "right" type of data to infer these models?

With cDNA microarray data, it is widely recognized that reproducibility of measurements and between-slide variation is a major issue (Zhang et al. [41], Chen et al. [42], Kerr et al. [43]). Furthermore, genetic regulation exhibits considerable uncertainty on the biological level. Indeed, evidence suggests that this type of "noise" is in fact advantageous in some regulatory mechanisms (McAdams and Arkin [44]). Thus, from a practical standpoint, limited amounts of data and the noisy nature of the measurements can make useful quantitative inferences problematic, and a coarse-scale qualitative modeling approach seems to be justified. To put it another way, if our goals of modeling were to capture the genetic interactions with fine-scale quantitative biochemical details in a global large-scale fashion, then the data produced by currently available high-throughput genomic technologies would not be adequate for this purpose.

Besides the noise and lack of fine-scale data, another important concern is the design of dynamic networks using nondynamic data. If time-course data is available, then it is usually limited and the relation between the biological time-scale under which it has been observed and the transition routine of an inferred network is unknown. Moreover, most often the data being used to infer networks does not consist of time-course observations. In this situation, the usual assumption is that the data comes from the steady state of the system. There are inherent limitations to the design of dynamical systems from steady-state data. Steady-state behavior constrains the dynamical behavior, but does not determine it. Thus, while we might obtain good inference regarding the attractors, we may obtain poor inference relative to the steady-state distribution (see Section 7.4.2 for the definition of an attractor). Building a dynamical model from steady-state data is a kind of overfitting. It is for this reason that we view a designed network as providing a regulatory structure consistent with the observed steady-state behavior. If our main interest is in steady-state behavior, then it is reasonable to try to understand dynamical regulation corresponding to steady-state behavior.

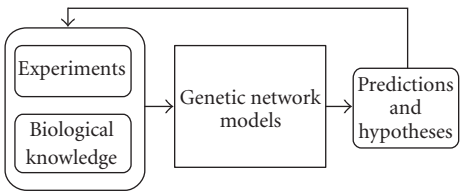


Figure 7.2

7.2.4. What do we hope to learn from these models?

Our last question is concerned with what type of knowledge we hope to acquire with the chosen models and the available data. As a first step, we may be interested in discovering qualitative relationships underlying genetic regulation and control. That is, we wish to emphasize fundamental generic coarse-grained properties of large networks rather than quantitative details, such as kinetic parameters of individual reactions (Huang [30]). Furthermore, we may wish to gain insight into the dynamical behavior of such networks and how it relates to underlying biological phenomena, such as cellular state dynamics, thus providing the potential for the discovery of novel targets for drugs. As an example, we may wish to predict the downstream effects of a targeted perturbation of a particular gene. Recent research indicates that many realistic biological questions may be answered within the seemingly simplistic Boolean formalism. Boolean networks are structurally simple, yet dynamically complex. They have yielded insights into the overall behavior of large genetic networks (Somogyi and Sniegoski [32], Szallasi and Liang [45], Wuensche [46], Thomas et al. [47]) and allowed the study of large data sets in a global fashion.

Besides the conceptual framework afforded by such models, a number of practical uses, such as the identification of suitable drug targets in cancer therapy, may be reaped by inferring the structure of the genetic models from experimental data, for example, from gene expression profiles (Huang [30]). To that end much recent work has gone into identifying the structure of gene regulatory networks from expression data (Liang et al. [48], Akutsu et al. [49, 50, 51], D'haeseleer et al. [52], Shmulevich et al. [53], Lähdesmäki et al. [54]). It is clear that "wet" lab experimental design and "dry" lab modeling and analysis must be tightly integrated and coordinated to generate, refine, validate, and interpret the biologically relevant models (see Figure 7.2).

Perhaps the most salient limitation of standard Boolean networks is their inherent determinism. From a conceptual point of view, it is likely that the regularity of genetic function and interaction known to exist is not due to "hard-wired" logical rules, but rather to the intrinsic self-organizing stability of the dynamical system, despite the existence of stochastic components in the cell. From an empirical point of view, there are two immediate reasons why the assumption of only one logical rule per gene may lead to incorrect conclusions when inferring these rules from gene expression measurements: (1) the measurements are typically noisy and the number of samples is small relative to the number of parameters to be inferred;

(2) the measurements may be taken under different conditions, and some rules may differ under these varying conditions.

7.3. Probabilistic Boolean networks

7.3.1. Background

The probabilistic Boolean network model was introduced by Shmulevich et al. [1]. These networks share the appealing properties of Boolean networks, but are able to cope with uncertainty, both in the data and the model selection. There are various reasons for utilizing a probabilistic network. A model incorporates only a partial description of a physical system. This means that a Boolean function giving the next state of a variable is likely to be only partially accurate. There will be conditions under which different Boolean functions may actually describe the transition, but these are outside the scope of the conventional Boolean model. If, consequently, we are uncertain as to which transition rule should be used, then a probabilistic Boolean network involving a set of possible Boolean functions for each variable may be more suitable than a network in which there is only a single function for each variable.

Even if one is fairly confident that a model is sufficiently robust that other variables can be ignored without significant impact, there remains the problem of inferring the Boolean functions from sample data. In the case of gene-expression microarrays, the data are severely limited relative to the number of variables in the system. Should it happen that a particular Boolean function has even a moderately large number of essential variables, its design from the data is likely to be imprecise because the number of possible input states will be too large for precise estimation. This situation is exacerbated if some essential variables are either unknown or unobservable (latent). As a consequence of the inability to observe sufficient examples to design the transition rule, it is necessary to restrict the number of variables over which a function is defined. For each subset of the full set of essential variables, there may be an optimal function, in the sense that the prediction error is minimized for that function, given the variables in the subset. These optimal functions must be designed from sample data. Owing to inherent imprecision in the design process, it may be prudent to allow a random selection between several functions, with the weight of selection based on a probabilistic measure of worth, such as the coefficient of determination (Dougherty et al. [55]).

The other basic issue regarding a probabilistic choice of transition rule is that in practice we are modeling an open system rather than a closed system. An open system has inputs (stimuli) that can affect regulation. Moreover, any model used to study a physical network as complex as a eukaryotic genome must inevitably omit the majority of genes. The states of those left out constitute external conditions to the model. System transition may depend on a particular external condition at a given moment of time. Such effects have been considered in the framework of using the coefficient of determination in the presence of external stresses (Kim et al. [56]). Under the assumption that the external stimuli occur asynchronously, it is prudent to allow uncertainty among the transition rules and

weight their likelihood accordingly. It may be that the probability of applying a Boolean function corresponding to an unlikely condition is low; however, system behavior might be seriously misunderstood if the possibility of such a transition is ignored.

It has been shown that Markov chain theory could be used to analyze the dynamics of PBNs (Shmulevich et al. [1]). Also, the relationships to Bayesian networks have been established, and the notions of *influences* and *sensitivities* of genes defined. The latter have been used to study the dynamics of Boolean networks (Shmulevich and Kauffman [57]). The inference of networks from gene expression data has received great attention. It is important for the inferred network to be robust in the face of uncertainty. Much work in this direction has already been carried out specifically in the context of gene regulatory networks (Dougherty et al. [58], Kim et al. [56, 59], Shmulevich et al. [53], Lähdesmäki et al. [54], Zhou et al. [10, 60, 61]). Visualization tools for the inferred multivariate gene relationships in networks are described by Suh et al. [62].

A framework for constructing subnetworks, adjoining new genes to subnetworks, and mapping between networks in such a way that the network structure and parameters remain consistent with the data has been established by Dougherty and Shmulevich [5]. An algorithm for growing subnetworks from so-called "seed genes" has been developed by Hashimoto et al. [11] and applied to several datasets (glioma and melanoma). An important goal of PBN modeling is to study the long-run behavior of the genetic networks. This was studied by Shmulevich et al. [6], using Markov chain Monte Carlo (MCMC) methods, along with a detailed analysis of convergence. In particular, the effect of network mappings on long-run behavior is critically important and a preliminary study has been carried out regarding the effect of network compression, including the issue of how to compress a network to reduce complexity while at the same time maintain long-run behavior to the extent possible (Ivanov and Dougherty [63]).

A gene perturbation model was extensively studied by Shmulevich et al. [2]. This approach not only simplified the steady-state analysis, but also provided a theoretical framework for assessing the effects of single-gene perturbations on the global long-run network behavior. In addition, a methodology for determining which genes would be good potential candidates for intervention was developed. Intervention and perturbation were presented in a unified framework. In addition, another approach for intervention, based on structural control with the use of genetic algorithms, was presented by Shmulevich et al. [3]. Finally, intervention based on external control was considered by Datta et al. [7, 8]. In that work, given a PBN whose state transition probabilities depend on an external (control) variable, a dynamic programming-based procedure was developed by which one could choose the sequence of control actions that minimized a given performance index over a finite number of steps.

7.3.2. Biological significance

One of the main objectives of Boolean-based network modeling is to study generic coarse-grained properties of large genetic networks and the logical interactions of

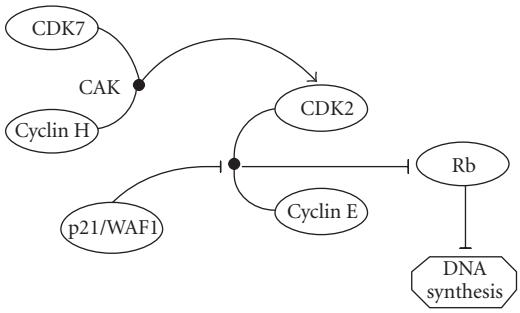


Figure 7.3. A diagram illustrating the cell cycle regulation example. Arrowed lines represent activation and lines with bars at the end represent inhibition.

genes, without knowing specific quantitative biochemical details, such as kinetic parameters of individual reactions. The biological basis for the development of Boolean networks as models of genetic regulatory networks lies in the fact that during regulation of functional states, the cell exhibits switch-like behavior, which is important for cells to move from one state to another in a normal cell growth process or in situations when cells need to respond to external signals, many of which are detrimental. Let us use cell cycle regulation as an example. Cells grow and divide. This process is highly regulated; failure to do so results in unregulated cell growth in diseases such as cancer. In order for cells to move from the G1 phase to the S phase, when the genetic material, DNA, is replicated for the daughter cells, a series of molecules such as cyclin E and cyclin-dependent kinase 2 (CDK2) work together to phosphorylate the retinoblastoma (Rb) protein and inactivate it, thus releasing cells into the S phase. CDK2/cyclin E is regulated by two switches: the positive switch complex called CDK activating kinase (CAK) and the negative switch p21/WAF1. The CAK complex can be composed of two gene products: cyclin H and CDK7. When cyclin H and CDK7 are present, the complex can activate CDK2/cyclin E. A negative regulator of CDK2/cyclin E is p21/WAF1, which in turn can be activated by p53. When p21/WAF1 binds to CDK2/cyclin E, the kinase complex is turned off (Gartel and Tyner [64]). Further, p53 can inhibit cyclin H, a positive regulator of cyclin E/CDK2 (Schneider et al. [65]). This negative regulation is an important defensive system in the cells. For example, when cells are exposed to mutagens, DNA damage occurs. It is to the benefit of cells to repair the damage before DNA replication so that the damaged genetic materials do not pass onto the next generation. Extensive amount of work has demonstrated that DNA damage triggers switches that turn on p53, which then turns on p21/WAF1. p21/WAF1 then inhibits CDK2/cyclin E, thus Rb becomes activated and DNA synthesis stops. As an extra measure, p53 also inhibits cyclin H, thus turning off the switch that turns on CDK2/cyclin E. Such delicate genetic switch networks in the cells are the basis for cellular homeostasis.

For purposes of illustration, let us consider a simplified diagram, shown in Figure 7.3, illustrating the effects of CDK7/cyclin H, CDK2/cyclin E, and p21/WAF1 on Rb. Thus, p53 and other known regulatory factors are not considered.

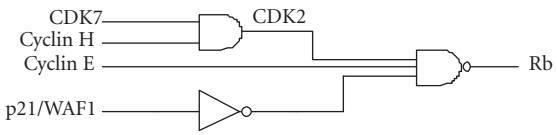


Figure 7.4. The logic diagram describing the activity of retinoblastoma (Rb) protein in terms of 4 inputs: CDK7, cyclin H, cyclin E, and p21. The gate with inputs CDK7 and cyclin H is an AND gate, the gate with input p21/WAF1 is a NOT gate, and the gate whose output is Rb is a NAND (negated AND) gate.

While this diagram represents the above relationships from a pathway perspective, we may also wish to represent the activity of Rb in terms of the other variables in a logic-based fashion. Figure 7.4 contains a logic circuit diagram of the activity of Rb (on or off) as a Boolean function of four input variables: CDK7, cyclin H, cyclin E, and p21/WAF1. Note that CDK2 is shown to be completely determined by the values of CDK7 and cyclin H using the AND operation and thus, CDK2 is not an independent input variable. Also, in Figure 7.3, p21/WAF1 is shown to have an inhibitive effect on the CDK2/cyclin E complex, which in turn regulates Rb, while in Figure 7.4, we see that from a logic-based perspective, the value of p21/WAF1 works together with CDK2 and cyclin E to determine the value of Rb. Such dual representations in the biological literature were pointed out by Rzhetsky et al. [66].

7.3.3. Definitions

Mathematically, a Boolean network $G(V,F)$ is defined by a set of nodes $V = \{x_1,\ldots,x_n\}$ and a list of Boolean functions $F = \{f_1,\ldots,f_n\}$. Each x_i represents the state (expression) of a gene i, where $x_i = 1$ represents the fact that gene i is expressed and $x_i = 0$ means it is not expressed. It is commonplace to refer to $x_1, x_2, \ldots, x_n$ as genes. The list of Boolean functions F represents the rules of regulatory interactions between genes. Each $x_i \in \{0,1\}$, $i = 1,\ldots,n$, is a binary variable and its value at time $t+1$ is completely determined by the values of some other genes $x_{j_1(i)}, x_{j_2(i)}, \ldots, x_{j_{k_i}(i)}$ at time t by means of a Boolean function $f_i \in F$. That is, there are k_i genes assigned to gene x_i and the set $\{x_{j_1(i)}, x_{j_2(i)}, \ldots, x_{j_{k_i}(i)}\}$ determines the "wiring" of gene x_i. Thus, we can write

$$x_i(t+1) = f_i\left(x_{j_1(i)}(t), x_{j_2(i)}(t), \ldots, x_{j_{k_i}(i)}(t)\right). \tag{7.1}$$

The *maximum connectivity* of a Boolean network is defined by $K = \max_i k_i$. All genes are assumed to update synchronously in accordance with the functions assigned to them and this process is then repeated. The artificial synchrony simplifies computation while preserving the qualitative, generic properties of global network dynamics (Huang [30], Kauffman [31], Wuensche [46]). It is clear that the dynamics of the network are completely deterministic.

The basic idea behind PBNs is to combine several promising Boolean functions, now called *predictors*, so that each can make a contribution to the prediction of a target gene. A natural approach is to allow a random selection of the

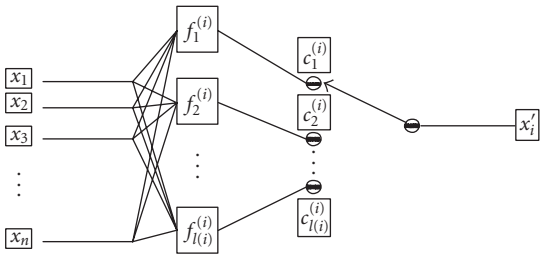


Figure 7.5. A basic building block of a probabilistic Boolean network. Although the "wiring" of the inputs to each function is shown to be quite general, in practice, each function (predictor) has only a few input variables.

predictors for a given target gene, with the selection probability being proportional to some measure of the predictor's determinative potential, such as the coefficient of determination, described later. At this point, it suffices for us to assume that each predictor has an associated probability of being selected. Given genes $V = \{x_1, \ldots, x_n\}$, we assign to each x_i a set $F_i = \{f_1^{(i)}, \ldots, f_{l(i)}^{(i)}\}$ of Boolean functions composed of the predictors for that target gene. Clearly, if $l(i) = 1$ for all $i = 1, \ldots, n$, then the PBN simply reduces to a standard Boolean network. The basic building block of a PBN is shown in Figure 7.5.

As first introduced in Shmulevich et al. [1], at each point in time or step of the network, a function $f_j^{(i)}$ is chosen with probability $c_j^{(i)}$ to predict gene x_i. Considering the network as a whole, a *realization* of the PBN at a given instant of time is determined by a vector of Boolean functions, where the ith element of that vector contains the predictor selected at that instant for gene x_i. If there are N possible realizations, then there are N vector functions, $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N$, of the form $f_k = (f_{k_1}^{(1)}, f_{k_2}^{(2)}, \ldots, f_{k_n}^{(n)})$, for $k = 1, 2, \ldots, N$, $1 \le k_i \le l(i)$, and where $f_{k_i}^{(i)} \in F_i$ $(i = 1, \ldots, n)$. In other words, the vector function $\mathbf{f}_k : \{0,1\}^n \to \{0,1\}^n$ acts as a transition function (mapping) representing a possible realization of the entire PBN. Such functions are commonly referred to as multiple-output Boolean functions. In the context of PBNs we refer to them as *network functions*. If we assume that the predictor for each gene is chosen independently of other predictors, then $N = \prod_{i=1}^{n} l(i)$. More complicated dependent selections are also possible. Each of the N possible realizations can be thought of as a standard Boolean network that operates for one time step. In other words, at every state $x(t) \in \{0,1\}^n$, one of the N Boolean networks is chosen and used to make the transition to the next state $x(t+1) \in \{0,1\}^n$. The probability P_i that the ith (Boolean) network or realization is selected can be easily expressed in terms of the individual selection probabilities $c_j^{(i)}$ (see Shmulevich et al. [1]).

The PBN model is generalized by assuming that the decision to select a new network realization is made with probability λ at every time step. In other words, at every time step, a coin is tossed with probability λ of falling on heads, and if it does, then a new network realization is selected as described above; otherwise, the current network realization is used for the next time step. The original PBN

definition as described above corresponds to the case $\lambda = 1$. We will refer to the model with $\lambda = 1$ as an *instantaneously random* PBN. The $\lambda < 1$ has a natural interpretation relative to external conditions. The Boolean network remains unchanged from moment to moment, except when its regulatory structure is altered by a change in an external condition. Any given set of conditions may be considered to correspond to a context of the cell. Hence, when $\lambda < 1$ we refer to the network as a *context-sensitive* PBN. Assuming conditions are stable, λ will tend to be quite small (Braga-Neto et al. [67], Zhou et al. [61], Brun et al. [68]).

Thus far, randomness has only been introduced relative to the functions (hence, implicitly, also the connectivity); however the model can be extended to incorporate transient gene perturbations. Suppose that a gene can get perturbed with (a small) 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. 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, and so forth. Thus, a network model should be able to capture this phenomenon.

Suppose that at every step of the network we have a realization of a random *perturbation vector* $\gamma \in \{0,1\}^n$. If the ith component of γ is equal to 1, then the ith gene is flipped, otherwise it is not. In general, γ need not be independent and identically distributed (i.i.d.), but will be assumed so for simplicity. Thus, we will suppose that $\Pr\{\gamma_i = 1\} = E[\gamma_i] = p$ for all $i = 1,\ldots,n$. Let $x(t) \in \{0,1\}^n$ be the state of the network at time t. Then, the next state $x(t+1)$ is given by

$$x(t+1) = \begin{cases} x(t) \oplus \gamma, & \text{with probability } 1-(1-p)^n, \\ \mathbf{f}_k\left(x_1(t),\ldots,x_n(t)\right), & \text{with probability } (1-p)^n, \end{cases} \tag{7.2}$$

where $\oplus$ is componentwise addition modulo 2 and $\mathbf{f}_k$, $k = 1,2,\ldots,N$, is the network function representing a possible realization of the entire PBN, as discussed above.

For a PBN with random perturbation, the following events can occur at any point of time: (1) the current network function is applied, the PBN transitions accordingly, and the network function remains the same for the next transition; (2) the current network function is applied, the PBN transitions accordingly, and a new network function is selected for the next transition; (3) there is a random perturbation and the network function remains the same for the next transition; (4) there is a random perturbation and a new network function is selected for the next transition.

7.4. Long-run behavior

7.4.1. Steady-state distribution

In the absence of random perturbations, the states of an instantaneously random PBN form a finite-state homogeneous Markov chain and that possesses a

stationary distribution, where the transition probabilities depend on the network functions. When random perturbations are incorporated into the model, the chain becomes ergodic and possesses a steady-state distribution. In Shmulevich et al. [2], an explicit formulation of the state-transition probabilities of the Markov chain associated with the PBN is derived in terms of the Boolean functions and the probability of perturbation p.

In the case of a context-sensitive PBN, the state vector of gene values at time t cannot be considered as a homogeneous Markov chain anymore because the transition probabilities depend on the function selected at time t. Instead of representing the states $\mathbf{x}$ of the PBN as the states of a Markov chain, we can represent the state-function pairs $(\mathbf{x}, \mathbf{f}_k)$ as states of a homogeneous Markov chain with transition probabilities

$$P_{\mathbf{y},\mathbf{f}_l}(\mathbf{x}, \mathbf{f}_k) = P(\mathbf{X}_t = \mathbf{x},\ \mathbf{F}_t = \mathbf{f}_k \mid \mathbf{X}_{t-1} = \mathbf{y},\ \mathbf{F}_{t-1} = \mathbf{f}_l) \tag{7.3}$$

for any time t. The chain must possess a stationary distribution, and if there are random perturbations, then it possesses a steady-state distribution. The probabilities $\pi(\mathbf{x})$ are the marginal probabilities of the steady-state distribution defined by

$$\pi(\mathbf{x}, \mathbf{f}_k) = \lim_{t\to\infty} P(\mathbf{X}_{t_0+t} = \mathbf{x},\ \mathbf{F}_{t_0+t} = \mathbf{f}_k \mid \mathbf{X}_{t_0} = \mathbf{y},\ \mathbf{F}_{t_0} = \mathbf{f}_l), \tag{7.4}$$

where t_0 is the initial time. These steady-state distributions for context-sensitive PBNs have been studied by Brun et al. [68].

7.4.2. Attractors

Owing to its deterministic and finite nature, if a Boolean network is initialized and then allowed to dynamically transition, it will return to a previously visited state within a bounded amount of time (based on the total number of genes). Once this occurs, it will cycle from that state through the same set of states and in the same order as it did after following the first visit to the state. The cycle of states is called an attractor cycle. Note that attractor cycles must be disjoint and either every state is a member of an attractor or it is transient, meaning it cannot be visited more than once. Each initialization leads to a unique attractor and the set of states leading to an attractor is called the basin of attraction for the attractor. A singleton attractor (absorbing state) has the property that once entered, the network cannot leave it.

The attractors of a Boolean network characterize its long-run behavior. If, however, we incorporate random perturbation, then the network can escape its attractors. In this case, full long-run behavior is characterized by its steady-state distribution. Nonetheless, if the probability of perturbation is very small, the network will lie in its attractor cycles for a large majority of the time, meaning that attractor states will carry most of the steady-state probability mass. The amount of time spent in any given attractor depends on its basin. Large basins tend to produce attractors possessing relatively large steady-state mass.

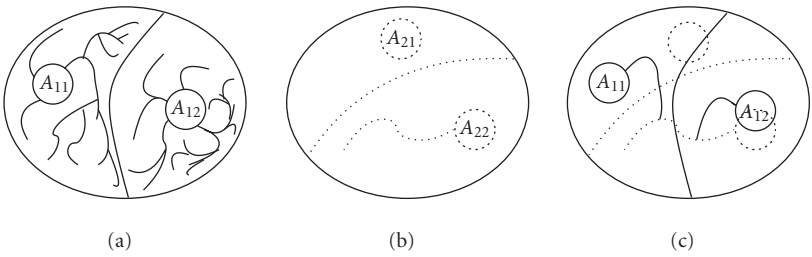


Figure 7.6. An illustration of the behavior of a context-sensitive PBN.

Let us now consider context-sensitive PBNs. So long as there is no switching, the current Boolean-network realization of the PBN characterizes the activity of the PBN and it will transition into one of its attractor cycles and remain there until a switch occurs. When a network switch does occur, the present state becomes an initialization for the new realization and the network will transition into the attractor cycle whose basin contains the state. It will remain there until another network switch. The attractor family of the PBN is defined to be the union of all the attractors in the constituent Boolean networks. Note that the attractors of a PBN need not be disjoint, although those corresponding to each constituent Boolean network must be disjoint.

Figure 7.6 shows an example of the behavior of a context-sensitive PBN relative to its attractors under a change of function. Part (a) shows the attractor cycles A_{11} and A_{12} for a network function $\mathbf{f}_1$, its corresponding basins, and some trajectories. Part (b) shows the attractor cycles A_{21} and A_{22} for a network function $\mathbf{f}_2$ and its corresponding basins. In part (c), we can see that if the system is using the function $\mathbf{f}_2$ and it makes a function change, to $\mathbf{f}_1$, then the future of the system depends on which part of the trajectory it is at the moment of the function change. In this example, for the particular trajectory shown with the dotted line toward the attractor A_{22}, the first part of the trajectory is in the basin corresponding to the attractor A_{11}, and the end of the trajectory is inside the basin corresponding to the attractor A_{12}. Therefore, if the change of function occurs before the system crosses the boundary between the basins, it will transition toward the attractor A_{11}. If the change of function occurs after it crosses the boundary, then it will transition toward the attractor A_{12}. In particular, we see that the attractor A_{22} lies completely inside the basin corresponding to the attractor A_{12}. In this case, if a change of function occurs when the system is inside the attractor A_{22}, it will always transition to the attractor A_{12}.

If one now incorporates perturbation into the PBN model, the stationary distribution characterizes the long-run behavior of the network. If both the switching and perturbation probabilities are very small, then the attractors still carry most of the steady-state probability mass. This property has been used to formulate analytic expressions of the probabilities of attractor states (Brun et al. [68]) and to validate network inference from data (Kim et al. [9], Zhou et al. [61]).

7.4.3. Monte-Carlo estimation of the steady-state distribution

Model-based simulations are invaluable for gaining insight into the underlying functioning of a genetic regulatory network. Simulation-supported decision making is essential in realistic analysis of complex dynamical systems. For example, one may wish to know the long-term joint behavior of a certain group of genes or the long-term effect of one gene on a group of others. After having robustly inferred the model structure and parameters, such questions can be answered by means of simulations. We have developed a methodology for analyzing steady-state (or long-run) behavior of PBNs using MCMC-type approaches (Shmulevich et al. [6]). By simulating the network until it converges to its steady-state distribution and monitoring the convergence by means of various diagnostics (Cowles and Carlin [69]), we can obtain the limiting probabilities of the genes of interest. Thus, the effects of permanent and transient interventions (e.g., turning a gene off) can be assessed on the long-run network behavior.

An approach found to be useful for determining the number of iterations necessary for convergence to the stationary distribution of the PBN is based on a method by Raftery and Lewis [70]. This method reduces the study of the convergence of the Markov chain corresponding to a PBN to the study of the convergence of a two-state Markov chain. Suppose that we are interested in knowing the steady-state probability of the event {Gene A is ON and Gene B is OFF}. Then, we can partition the state space into two disjoint subsets such that one subset contains all states on which the event occurs and the other subset contains the rest of the states. Consider the two meta-states corresponding to these two subsets. Although the sequence of these meta-states does not form a Markov chain in itself, it can be approximated by a first-order Markov chain if every k states from the original Markov chain is discarded (i.e., the chain is subsampled). It turns out in practice that k is usually equal to 1, meaning that nothing is discarded and the sequence of meta-states is treated as a homogeneous Markov chain (see Raftery and Lewis for details) with transition probabilities α and β between the two meta-states. Using standard results for two-state Markov chains, it can be shown that the burn-in period (the number of iterations necessary to achieve stationarity) m_0 satisfies

$$m_0 \geq \frac{\log\left(\varepsilon(\alpha+\beta)/\max(\alpha,\beta)\right)}{\log(1-\alpha-\beta)}. \tag{7.5}$$

We set $\varepsilon = 0.001$ in our experiments. In addition, it can be shown that the minimum total number of iterations N necessary to achieve a desired accuracy r (we used $r = 0.01$ in our experiments) is

$$N = \frac{\alpha\beta(2-\alpha-\beta)}{(\alpha+\beta)^3}\left(\frac{r}{\Phi\left((1/2)(1+s)\right)}\right)^{-2}, \tag{7.6}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and s is a parameter that we set to 0.95 in our experiments. For detailed explanations of the precision parameters ε, r, and s, see Raftery and Lewis [70]. The question becomes

Table 7.1. An example of the joint steady-state probabilities (in percentages) of several pairs of genes, computed from the network inferred from glioma gene expression data.

Tie-2	NFκB	%	Tie-2	TGFB3	%	TGFB3	NFκB	%
Off	Off	15.68	Off	Off	14.75	Off	Off	10.25
Off	On	41.58	Off	On	42.50	Off	On	12.47
On	Off	9.21	On	Off	7.96	On	Off	14.64
On	On	31.53	On	On	32.78	On	On	60.65

how to estimate the transition probabilities α and β, as these are unknown. The solution is to perform a test run from which α and β can be estimated and from which m_0 and N can be computed. Then, another run with the computed burn-in period m_0 and the total number of iterations N is performed and the parameters α and β are reestimated from which m_0 and N are recomputed. This can be done several times in an iterative manner until the estimates of m_0 and N are smaller than the number of iterations already achieved. We have used this method to determine the steady-state probabilities of some genes of interest from our gene expression data set, as described below.

We analyzed the joint steady-state probabilities of several combinations of two genes from a subnetwork generated from our glioma expression data: Tie-2 and NFκB, Tie-2 and TGFB3, and TGFB3 and NFκB. The steady-state probabilities for all pairs of considered genes are shown in Table 7.1 as percentages. Tie-2 is a receptor tyrosine kinase expressed on the endothelial cells. Its two ligands, angiopoietins 1 and 2, bind Tie-2 and regulate vasculogenesis (Sato et al. [71]), an important process in embryonic development and tumor development. Other related regulators for vasculogenesis are VEGF and VERF receptors, which are often overexpressed in the advanced stage of gliomas (Cheng et al. [72]). Although no experimental evidence supports a direct transcriptional regulation of those regulators by the transcriptional factor NFκB, which is also frequently activated in glioma progression (Hayashi et al. [73]) as predicted in this analysis, the results show that NFκB, at least indirectly, influence the expression of Tie-2 expression. Thus, it may not be surprising that when NFκB is on, Tie-2 is on about $31.53/(41.58 + 31.53) = 43\%$ of time. Because Tie-2 is only one of the regulators of vasculogenesis, which is important in glioma progression, it is consistent that our analysis of long-term (steady-state) gene expression activities shows that about 40% of the time Tie-2 is on. In contrast, NFκB is on 73% of the time, implying that fewer redundancies exist for NFκB activity.

Interestingly, a similar relationship exists between Tie-2 and TGFB3, as can be seen by comparing the percentages in columns 3 and 6 of Table 7.1. This suggests that TGFB3 and NFκB are more directly linked, which is also shown in the last three columns of the table (60% of the time, they are both on). This relationship is supported by the fact that TGFB1, a homologue of TGFB3, was shown to have a direct regulatory relationship with NFκB (Arsura et al. [74]) as well as by the recent work of Strauch et al. [75], who recently showed that NFκB activation indeed upregulates TGFB expression.

7.5. Inference of PBNs from gene expression data

Owing to several current limitations, such as availability of only transcriptional measurements (much regulation occurs on the protein level), cell population asynchrony and possible heterogeneity (different cell types exhibiting different gene activities), and latent factors (environmental conditions, genes that we are not measuring, etc.), it is prudent to strive to extract higher-level information or knowledge about the relationships between measurements of gene transcript levels. If we can discover such relationships, then we can potentially learn something new about the underlying mechanisms responsible for generating these observed quantities. The burden of discovery remains in the wet lab, where potentially interesting relationships must be examined experimentally. We now discuss some approaches to the inference problem. These approaches have already been used in a number of different studies. At this point it still remains to be known which aspects of the data tend to be emphasized by which approach and whether one approach reflects the actual regulatory relationships more faithfully than another. This constitutes an important research problem.

7.5.1. Coefficient of determination

A basic building block of a rule-based network is a *predictor*. In a probabilistic network, several good predictors are probabilistically synthesized to determine the activity of a particular gene. A predictor is designed from data, which means that it is an approximation of the predictor whose action one would actually like to model. The precision of the approximation depends on the design procedure and the sample size.

Even in the context of limited data, modest approaches can be taken. One general statistical approach is to discover associations between the expression patterns of genes via the *coefficient of determination* (CoD) (Dougherty et al. [55, 58], Kim et al. [56, 59]). This coefficient measures the degree to which the transcriptional levels of an observed gene set can be used to improve the prediction of the transcriptional state of a target gene relative to the best possible prediction in the absence of observations. Let Y be a target variable, $\mathbf{X}$ a set of variables, and f the function such that $f(\mathbf{X})$ is the optimal predictor of Y relative to minimum mean-square error, $\varepsilon(Y, f(\mathbf{X}))$. The CoD for Y relative to $\mathbf{X}$ is defined by

$$\theta_{\mathbf{X}}(Y) = \frac{\varepsilon_{\bullet}(Y) - \varepsilon\big(Y, f(\mathbf{X})\big)}{\varepsilon_{\bullet}(Y)}, \tag{7.7}$$

where $\varepsilon_{\bullet}(Y)$ is the error of the best constant estimate of Y in the absence of any conditional variables. The CoD is between 0 and 1.

The method allows incorporation of knowledge of other conditions relevant to the prediction, such as the application of particular stimuli, or the presence of inactivating gene mutations, as predictive elements affecting the expression level of a given gene. Using the coefficient of determination, one can find sets of genes

related multivariately to a given target gene. The CoD has been used in gene-expression studies involving genotoxic stress (Kim et al. [56]), melanoma (Kim et al. [9]), glioma (Hashimoto et al. [11]), and atherogenesis (Johnson et al. [76]).

The coefficient of determination is defined in terms of the population distribution. However, in practice, we use the sample-based version; much like the sample mean (average) is the estimate of the population mean. An important research goal related to the CoD is to study and characterize the behavior of its estimator in the context of robustness. That is, it is important to understand to what extent the presence of outliers influences the estimate of the CoD. Various tools to analyze robustness, used in the nonlinear signal processing community (e.g., Shmulevich et al. [77]), may be applicable in this context.

7.5.2. Best-fit extensions

Most recent work with Boolean networks has focused on identifying the structure of the underlying gene regulatory network from gene expression data (Liang et al. [48], Akutsu et al. [49, 50], Ideker et al. [78], Karp et al. [79], Maki et al. [80], Noda et al. [81], Shmulevich et al. [53]). A related issue is to find a network that is consistent with the given observations or determine whether such a network exists at all. Much work in this direction has been traditionally focused on the so-called *consistency problem*, namely, the problem of determining whether or not there exists a network that is consistent with the observations.

The consistency problem represents a search for a rule from examples. That is, given some sets T and F of true and false vectors, respectively, the aim is to discover a Boolean function f that takes on the value 1 for all vectors in T and the value 0 for all vectors in F. It is also commonly assumed that the target function f is chosen from some class of possible target functions. In the context of Boolean networks, such a class could be the class of canalizing functions (discussed later) or functions with a limited number of essential variables. More formally, let $T(f) = \{v \in \{0,1\}^n : f(v) = 1\}$ be called the *on-set* of function f and let $F(f) = \{v \in \{0,1\}^n : f(v) = 0\}$ be the *off-set* of f. The sets $T, F \subseteq \{0,1\}^n$, $T \cap F = \emptyset$, define a *partially defined* Boolean function $g_{T,F}$ as

$$g_{T,F}(v) = \begin{cases} 1, & v \in T \\ 0, & v \in F \\ *, & \text{otherwise.} \end{cases} \tag{7.8}$$

A function f is called an *extension* of $g_{T,F}$ if $T \subseteq T(f)$ and $F \subseteq F(f)$. The consistency problem (also called the extension problem) can be posed as follows: given a class C of functions and two sets T and F, is there an extension $f \in C$ of $g_{T,F}$?

While this problem is important in computational learning theory, since it can be used to prove the hardness of learning for various function classes (e.g., Shmulevich et al. [82]), it may not be applicable in realistic situations containing

noisy observations, as is the case with microarrays. That is, due to the complex measurement process, ranging from hybridization conditions to image processing techniques, as well as actual biological variability, expression patterns exhibit uncertainty.

A learning paradigm that can incorporate such inconsistencies is called the best-fit extension problem. Its goal is to establish a network that would make as few misclassifications as possible. The problem is formulated as follows. Suppose we are given positive weights $w(x)$ for all vectors $x \in T \cup F$ and define $w(S) = \sum_{x \in S} w(x)$ for a subset $S \subseteq T \cup F$. Then, the *error size* of function f is defined as

$$\varepsilon(f) = w(T \cap F(f)) + w(F \cap T(f)). \tag{7.9}$$

If $w(x) = 1$ for all $x \in T \cup F$, then the error size is just the number of misclassifications. The goal is then to output subsets T^* and F^* such that $T^* \cap F^* = \varnothing$ and $T^* \cup F^* = T \cup F$ for which the partially defined Boolean function g_{T^*,F^*} has an extension in some class of functions C and so that $w(T^* \cap F) + w(F^* \cap T)$ is minimum. Consequently, any extension $f \in C$ of g_{T^*,F^*} has minimum error size. A crucial consideration is computational complexity of the learning algorithms. In order for an inferential algorithm to be useful, it must be computationally tractable. It is clear that the best-fit extension problem is computationally more difficult than the consistency problem, since the latter is a special case of the former, that is, when $\varepsilon(f) = 0$. Shmulevich et al. [53] showed that, for many function classes, including the class of all Boolean functions, the best-fit extension problem is polynomial-time solvable in the number of genes and observations, implying its practical applicability to real data analysis. Also, fast optimized and scalable search algorithms for best-fit extensions were developed by Lähdesmäki et al. [54].

The best-fit method is very versatile in the sense that one can specify the relative cost of making an error in the inference for various states of gene expression. There are a number of available quality measurements (Chen et al. [83]) which could be used in this fashion. Thus, instead of discarding low-quality measurements, one may be able to control their relative influence by down-weighting them in the best-fit extension inference method, in proportion to their quality measure. This is a useful topic to explore in future research.

7.5.3. Bayesian connectivity-based design

A recently proposed Bayesian method for constructing PBNs (that applies to a more general class of networks) is based on the network connectivity scheme (Zhou et al. [61]). Using a reversible-jump MCMC technique, the procedure finds possible regulatory gene sets for each gene, the corresponding predictors, and the associated probabilities based on a neural network with a very simple hidden layer. An MCMC method is used to search the network configurations to find those with the highest Bayesian scores from which to construct the PBN. We briefly outline the method, leaving the details to the original paper.

Consider a Boolean network $G(V,F)$ as previously defined, V being the genes and F the predictors. Construction is approached in a Bayesian framework relative to network topology by searching for networks with the highest a posteriori probabilities

$$P(V|D) \propto P(D|V)P(V), \tag{7.10}$$

where D is the data set and $P(V)$, the prior probability for the network, is assumed to satisfy a uniform distribution over all topologies. Note that $P(V|D)$ is given by

$$P(V|D) = \int p(D|V,F)\,p(F)dF, \tag{7.11}$$

where p denotes the density. The computation of this integral is virtually intractable and therefore it is approximated (Zhou et al. [61]).

In the context of this Bayesian framework, a PBN is constructed by searching the space of network topologies and selecting those with the highest Bayesian scores $P(V|D)$ to form the PBN. The algorithm proceeds in the following general manner: generate an initial graph $V^{(0)}$; compute $P(V|D)$; for $j = 1, 2, \ldots$, calculate the predictors $F^{(j)}$ corresponding to $G^{(j)}$; compute the Bayesian $P(V|D)$ score; and choose $G^{(j+1)}$ via an MCMC step.

7.5.4. Plausible classes of genetic interactions

While the focus in computational learning theory has mostly been on the complexity of learning, very similar types of problems have been studied in nonlinear signal processing, specifically, in optimal filter design (Coyle and Lin [84], Coyle et al. [85], Yang et al. [86], Dougherty and Loce [87], Dougherty and Chen [88]). This typically involves designing an estimator from some predefined class of estimators that minimizes the error of estimation among all estimators in the class. An important role in filter design is played by these predefined classes or constraints. For example, the so-called stack filters are represented by the class of monotone Boolean functions. Although it would seem that imposing such constraints can only result in a degradation of the performance (larger error) relative to the optimal filter with no imposed constraints, constraining may confer certain advantages. These include prior knowledge of the degradation process (or in the case of gene regulatory networks, knowledge of the likely class of functions, such as canalizing functions), tractability of the filter design, and precision of the estimation procedure by which the optimal filter is estimated from observations. For example, we often know that a certain class of filters will provide a very good suboptimal filter, while considerably lessening the data requirements for its estimation. It is with the issue of filter complexity versus sample size that the design of nonlinear filters intersects with the theory of classification (Dougherty and Barrera [89]). We now discuss several promising constraints for inferring Boolean predictors for PBNs.

An important class of functions, known to play an important role in regulatory networks (Kauffman [31, 90], Harris et al. [91]), is the class of *canalizing functions*. Canalizing functions constitute a special type of Boolean function in which at least one of the input variables is able to determine the value of the output of the function. For example, the function $f(x_1, x_2, x_3) = x_1 + x_2x_3$, where the addition symbol stands for disjunction and the multiplication for conjunction, is a canalizing function, since setting x_1 to 1 guarantees that the value of the function becomes 1 regardless of the value of x_2 or x_3. Although their defining property is quite simple, canalizing functions have been implicated in a number of phenomena related to discrete dynamical systems as well as nonlinear digital filters (see references in Shmulevich et al. [92]).

Canalizing functions, when used as regulatory control rules, are one of the few known mechanisms capable of preventing chaotic behavior in Boolean networks (Kauffman [31]). In fact, there is overwhelming evidence that canalizing functions are abundantly utilized in higher vertebrate gene regulatory systems. Indeed, a recent large-scale study of the literature on transcriptional regulation in eukaryotes demonstrated an overwhelming bias towards canalizing rules (Harris et al. [91]).

Recently, Shmulevich et al. [93] have shown that certain *Post classes*, which are classes of Boolean functions that are closed under superposition (Post [94]), also represent plausible evolutionarily selected candidates for regulatory rules in genetic networks. These classes have also been studied in the context of synthesis (Nechiporuk [95]) and reliability (Muchnik and Gindikin [96]) of control systems—a field that bears an important relationship to genetic control networks. The Post classes considered by Shmulevich et al. [93] play an important role in the emergence of order in Boolean networks. The closure property mentioned above implies that any gene at any number of steps in the future is guaranteed to be governed by a function from the same class. It was demonstrated that networks constructed from functions belonging to these classes have a tendency toward ordered behavior and are not overly sensitive to initial conditions, moreover and damage does not readily spread throughout the network. In addition, the considered classes are significantly larger than the class of canalizing functions, as the number of inputs per Boolean function increases. Additionally, functions from this class have a natural way to ensure robustness against noise and perturbations, thus representing plausible evolutionarily selected candidates for regulatory rules in genetic networks. Efficient spectral algorithms for testing membership of functions in these classes as well as the class of canalizing functions have been developed by Shmulevich et al. [92].

An important research goal is to determine whether the constraints described above are plausible not only from the point of view of evolution, noise resilience, and network dynamical behavior, but also in light of experimental data. Tools from model selection theory can be used to answer this question. Should this prove to be the case, by having prior knowledge of the plausible rules of genetic interaction, one can significantly improve model inference by reducing data requirements and increasing accuracy and robustness.

7.6. Subnetworks

It is likely that genetic regulatory networks function in what might be called a *multiscale* manner. One of the basic principles in multiscale modeling is that meaningful and useful information about a system or object exists on several different "levels" simultaneously. In the context of genetic networks, this would imply that genes form small groups (or clusters) wherein genes have close interactions. Some of these clusters are functionally linked forming larger "metaclusters" and these metaclusters have interactions as well. This process may continue on several different scales. This type of clustering effect has been observed in many other types of networks, such as social networks (Newman et al. [97]), the power grid of the western United States, and neural networks (Watts and Strogatz [98]). Interestingly, dynamical systems that have this property exhibit enhanced signal-propagation speed and computational power.

7.6.1. Growing subnetworks from seed genes

An important goal is to discover relatively small subnetworks, out of the larger overall network, that function more or less independently of the rest of the network. Such a small subnetwork would require little or sometimes even no information from the "outside." We can proceed by starting with a "seed" consisting of one or more genes that are believed to participate in such a subnetwork. Then, we iteratively adjoin new genes to this subnetwork such that we maintain the aforementioned "autonomy" of the subnetwork as much as possible, using the notions of gene influence (Shmulevich et al. [1]) or the coefficient of determination. Such an algorithm for growing subnetworks from seed genes has been developed by Hashimoto et al. [11].

Subnetwork construction proceeds in a way that enhances a strong collective strength of connections among the genes within the subnetwork and also limits the collective strength of the connections from outside the subnetwork. Consider Figure 7.7. Suppose we have a subnetwork $\mathbf{S}$ and are considering the candidate gene Y for inclusion in this subnetwork. We would like the collective strength (to be defined in a moment) of the genes in $\mathbf{S}$ on the candidate gene Y as well as the strength of gene Y on the genes in $\mathbf{S}$ to be high. In other words, the genes in $\mathbf{S}$ and Y should be tightly interdependent. At the same time, other genes from outside of the subnetwork should have little impact on Y if we are to maintain the subnetwork autonomy or "self determinacy." Thus, their collective strength on Y should be low. At each step, the subnetwork grows by one new gene so as to ensure maximal autonomy. An overall measure of subnetwork autonomy, which serves as an objective function in the subnetwork growing algorithm, is a combination of the three types of strength just described (Hashimoto et al. [11]). Finally, the strength itself can be naturally captured either by the coefficient of determination or by the influence, which we now discuss.

The *influence* of a variable relative to a Boolean function for which it is one among several Boolean variables is defined via the partial derivative of a Boolean

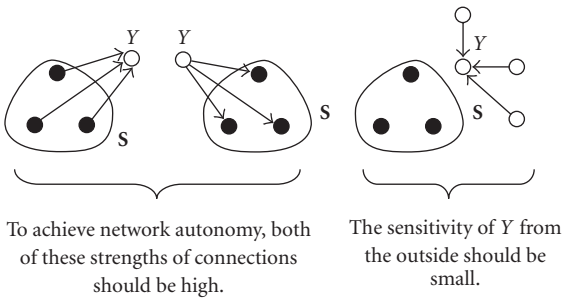


Figure 7.7. In order to maintain the self-autonomy of subnetwork **S**, the collective strength of the genes in **S** on gene Y—a candidate for inclusion in the subnetwork—should be high. The strength of Y on **S** should be high as well, thus maintaining high interdependency between the genes in the subnetwork. At the same time, the strength of genes outside the subnetwork on gene Y should be low.

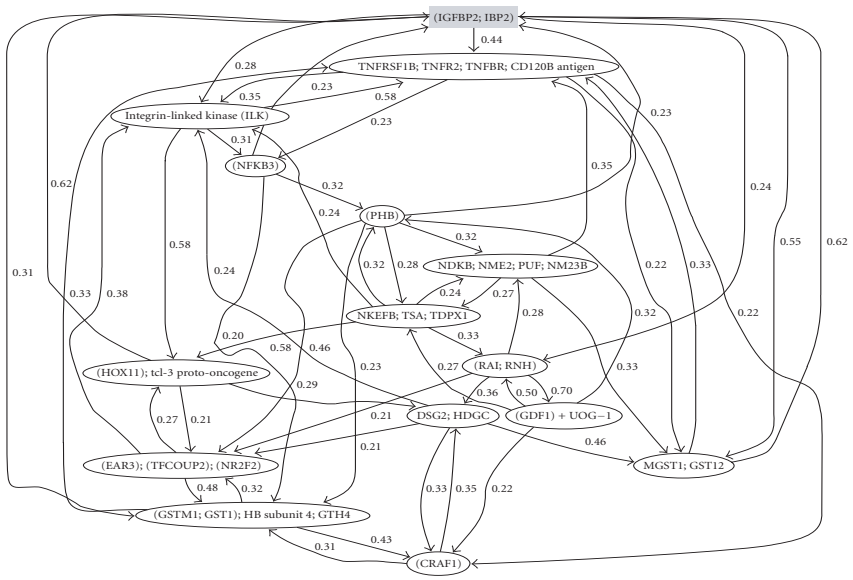
function. One can define the partial derivative of a Boolean function in several equivalent ways; however, for our purposes here, we simply note that the partial derivative of f with respect to the variable x_j is 0 if toggling the value of variable x_j does not change the value of the function, and it is 1 otherwise. The influence of x_j on f is the expectation of the partial derivative with respect to the distribution of the variables. In the context of a probabilistic Boolean network, there are a number of predictor functions associated with each gene, and each of these functions has associated with it a selection probability (Shmulevich et al. [1]). The influence of gene x_k on gene x_j is the sum of the influences of gene x_k on x_j relative to the family of predictor functions for x_j, weighted by the selection probabilities for these x_j-predicting functions.

Examples of subnetworks with IGFBP2 or VEGF as seeds are shown in Figure 7.8. In both glioma subnetworks, we used the influence as the strength of connection. The numbers over the arrows represent influences.

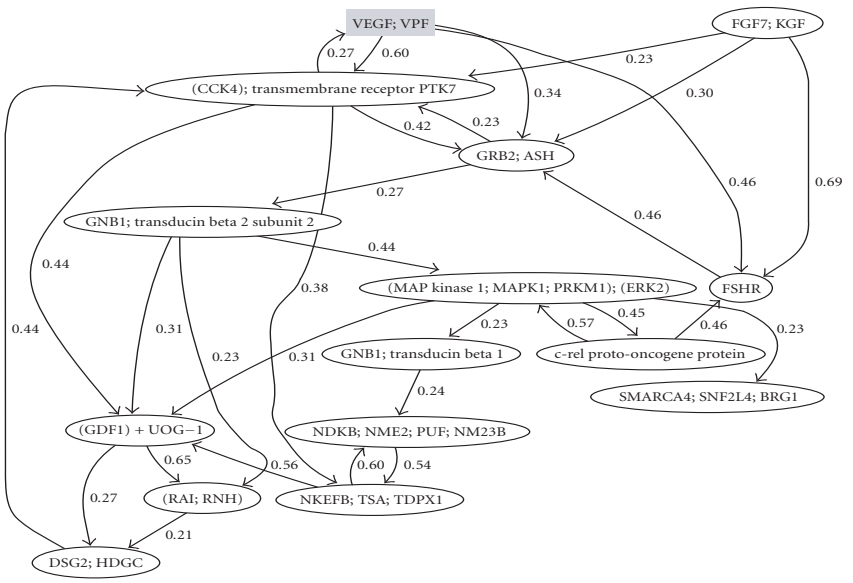
The subnetworks generated thus far have been constructed using the "seed growing" algorithm, starting from a large inferred network. All predictor functions in the starting network, consisting of almost 600 genes, were inferred using the CoD. A useful research aim is to repeat the inference using the best-fit extension method and reconstruct the subnetworks again. It is quite possible that the resultant subnetworks may exhibit some differences from those that have already been constructed. This possibility would furnish one with two opportunities. Firstly, it may reveal other genetic interactions that were not made apparent from the CoD-based inference method, in turn providing new hypotheses to be experimentally tested. Secondly, relationships consistent in both inference methods would strengthen one's confidence in the model-based results.

7.6.2. Interpretation and validation of subnetworks with prior biological knowledge and experimental results

Having constructed subnetworks in Figure 7.8 from expression via the seed-based growing algorithm, we would like to interpret and validate (in so far as that is



(a)



(b)

Figure 7.8. Two subnetworks generated from PBN modeling applied to a set of human glioma transcriptome data generated in our laboratory. (a) The subnetwork has been "grown" from the IGFBP2 (insulin-like growth factor binding protein 2) "seed." (b) The subnetwork has been grown from the VEGF (vascular endothelial growth factor) seed.

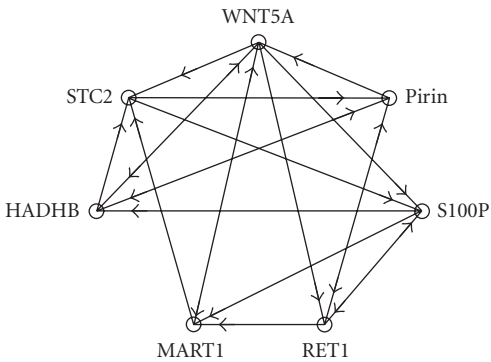


Figure 7.9. The seven-gene WNT5A network.

possible) the constructions using prior knowledge and independent experimental results. IGFBP2 and VEGF are genes that have been extensively studied and well characterized in the literature. It is known that IGFBP2 and VEGF are overexpressed in high-grade gliomas, glioblastoma multiforme (GBM)—the most advanced stage of tumor (Kleihues and Cavenee, WHO [99])—as compared to other types of glioma (Fuller et al. [100]). This finding was confirmed by two independent studies (Sallinen et al. [101], Elmlinger et al. [102]). Ongoing functional studies in the Cancer Genomics Laboratory (MD Anderson Cancer Center) using cell lines showed that when IGFBP2 is overexpressed, the cells become more invasive.

Studies that were completely independent of the PBN modeling work showed that NFκB activity is activated in cells stably overexpressing IGFBP2. This was done by using a luciferase reporter gene linked to a promoter element that contains an NFκB binding site. An analysis of the IGFBP2 promoter sequence showed that there are several NFκB binding sites, suggesting that NFκB transcriptionally regulates IGFBP2. A review of the literature revealed that Cazals et al. [103] indeed demonstrated that NFκB activated the IGFBP2-promoter in lung alveolar epithelial cells. Interestingly, in the IGFBP2 network (Figure 7.8a), we see an arrow linking NFκB3 to IGFBP2, and we see a two-step link from IGFBP2 to NFκB through TNF receptor 2 (TNFR2) and integrin-linked kinase (ILK). This parallels what was observed in the Cancer Genomics Laboratory. The presence of NFκB binding sites on the IGFBP2 promoter implies a direct influence of NFκB on IGFBP2. Although higher NFκB activity in IGFBP2 overexpressing cells was found, a transient transfection of IGFBP2 expressing vector together with NFκB promoter reporter gene construct did not lead to increased NFκB activity, suggesting an indirect effect of IGFBP2 on NFκB that will require time to take place. In fact, because of this indirect effect, this observation was not pursued for a year until the PBN-based subnetwork was linked to the laboratory experiments. IGFBP2 also contains an RGD domain, implying its interaction with integrin molecules. Integrin-linked kinase is in the integrin signal transduction pathway. The second subnetwork starting with VEGF (Figure 7.8b) offers even more compelling insight and supporting evidence.

Gliomas, like other cancers, are highly angiogenic, reflecting the need of cancer tissues for nutrients. To satisfy this need, expression of VEGF or vascular endothelial growth factor gene is often elevated. VEGF protein is secreted outside the cells and then binds to its receptor on the endothelial cells to promote their growth (Folkman [104]). Blockage of the VEGF pathway has been an intensive research area for cancer therapeutics (Bikfalvi and Bicknell [105]). A scrutiny of the VEGF network (Figure 7.8b) revealed some very interesting insight, which is highly consistent with prior biological knowledge derived from biochemical and molecular biology experiments. Let us elaborate. From the graph, VEGF, FGF7, FSHR, and PTK7 all influence Grb2. FGF7 is a member of fibroblast growth factor family (Rubin et al. [106]). FSHR is a follicle-stimulating hormone receptor. PTK7 is another tyrosine kinase receptor (Banga et al. [107]). The protein products of all four genes are part of signal transduction pathways that involve surface tyrosine kinase receptors. Those receptors, when activated, recruit a number of adaptor proteins to relay the signal to downstream molecules. Grb2 is one of the most crucial adaptors that have been identified (Stoletov et al. [108]). We should note that Grb2 is a target for cancer intervention (Wei et al. [109]) because of its link to multiple growth factor signal transduction pathways including VEGF, EGF, FGF, PDGF. Thus, the gene transcript relationships among the above five genes in the VEGF subnetwork appear to reflect their known or likely functional and physical relationship in cells. Molecular studies reported in the literature have further demonstrated that activation of protein tyrosine kinase receptor-Grb-2 complex in turn activates ras-MAP kinase- NFκB pathway to complete the signal relay from outside the cells to the nucleus of the cells (Bancroft et al. [110]). Although *ras* is not present on our VEGF network, a *ras* family member, GNB2, or transducing beta 2, is directly influenced by Grb2; GNB2 then influences MAP kinase 1 or ERK2, which in turn influences NFκB component c-rel (Pearson et al. [111]).

In the VEGF subnetwork shown in Figure 7.8, we also observe some potential feedback loop relationships. For example, c-rel influences FSHR, which influences Grb2-GNB2-MAPK1, and then influences c-rel itself. This may be a feedback regulation, a crucial feature of biological regulatory system in cells to maintain homeostasis. Other feedback regulation may also exist. RAI, or rel-A (another NFκB component) associated inhibitor (Yang et al. [112]), influences GNB2, which is two steps away from c-rel. RAI is further linked to PTK7 through GDF1, reflecting potentially another feedback regulatory mechanism. Whether those relationships are true negative feedback control mechanisms will need to be validated experimentally in the future. In this regard, the networks built from these models provide valuable theoretical guidance to experiments.

7.7. Perturbation and intervention

A major goal in gene network modeling is the ability to predict downstream effects on the gene network when a node is perturbed. This is very important for therapeutics. If we can predict the effect of such a perturbation, we can evaluate the virtue of a potential target when the effect on the entire system is considered.

The mathematical framework for performing this type of analysis has been developed by Shmulevich et al. [2]. Although this methodology has already been used for steady-state prediction, we have not attempted the prediction of downstream effects of specific perturbations. This, along with laboratory-based experimental verification, constitutes a valuable research goal.

A property of real gene regulatory networks is the existence of spontaneous emergence of ordered collective behavior of gene activity—that is, the evolution of networks into attractors. There is experimental evidence for the existence of attractors in real regulatory networks (Huang and Ingber [37]). As previously discussed, Boolean networks and PBNs also exhibit this behavior, the former with fixed-point and limit-cycle attractors (Kauffman [31]), the latter with absorbing states and irreducible sets of states (Shmulevich et al. [1, 53]). There is abundant justification in the assertion that in real cells, functional states, such as growth or quiescence, correspond to these attractors (Huang [30], Huang and Ingber [37]). Cancer is characterized by an imbalance between cellular states (attractors), such as proliferation and apoptosis (programmed cell death) resulting in loss of homeostasis.

As supported by Boolean network simulations, attractors are quite stable under most gene perturbations (Kauffman [31]). The same is true for 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.

A detailed perturbation analysis, including the long-range effect of perturbations, has been carried out by Shmulevich et al. [2]. It was demonstrated that states of the network that are more "easily reachable" from other states (in terms of mean first-passage times) are more stable in the presence of gene perturbations. Consequently, these sets of states are those that correspond to cellular functional states and represent the probabilistic version of homeostatic stability of attractors in the PBN model.

Suppose, on the other hand, that we wish to elicit certain long-run behavior from the network. What genes would make the best candidates for intervention so as to increase the likelihood of this behavior? That is, suppose that the network is operating in a certain "undesirable" set of states and we wish to "persuade" it to transition into a "desirable" set of states by perturbing some gene. For practical reasons, we may wish to be able to intervene with as few genes as possible in order to achieve our goals. Such an approach can expedite the systematic search and identification of potential drug targets in cancer therapy.

This question was taken up by Shmulevich et al. in [2], where several methods for finding the best candidate genes for intervention, based on first-passage times, were developed. The first-passage times provide a natural way to capture the goals

of intervention in the sense that we wish to transition to certain states (or avoid certain states, if that is our goal) "as quickly as possible," or, alternatively, by maximizing the probability of reaching such states before a certain time. Suppose, for example, that we wish to persuade the network to flow into a set of states (irreducible subchain—the counterpart of an attractor) representing apoptosis (programmed cell death). This could be very useful, for example, in the case of cancer cells, which may keep proliferating. We may be able to achieve this action via the perturbation (intervention) of several different genes, but some of them may be better in the sense that the mean first-passage time to enter apoptosis is shorter.

The type of intervention described above—one that allows us to intervene with a gene—can be useful for modulating the dynamics of the network, but it is not able to alter the underlying structure of the network. Accordingly, the steady-state distribution remains unchanged. However, a lack of balance between certain sets of states, which is characteristic of neoplasia in view of gene regulatory networks, can be caused by mutations of the "wiring" of certain genes, thus permanently altering the state-transition structure and, consequently, the long-run behavior of the network (Huang [30]).

Therefore, it is prudent to develop a methodology for altering the steady-state probabilities of certain states or sets of states with minimal modifications to the rule-based structure. The motivation is that these states may represent different phenotypes or cellular functional states, such as cell invasion and quiescence, and we would like to decrease the probability that the whole network will end up in an undesirable set of states and increase the probability that it will end up in a desirable set of states. One mechanism by which we can accomplish this consists of altering some Boolean functions (predictors) in the PBN. For practical reasons, as above, we may wish to alter as few functions as possible. Such alterations to the rules of regulation may be possible by the introduction of a factor or drug that alters the extant behavior.

Shmulevich et al. [3] developed a methodology for altering the steady-state probabilities of certain states or sets of states, with minimal modifications to the underlying rule-based structure. This approach was framed as an optimization problem that can be solved using genetic algorithms, which are well suited for capturing the underlying structure of PBNs and are able to locate the optimal solution in a highly efficient manner. For example, in some computer simulations that were performed, the genetic algorithm was able to locate the optimal solution (structural alteration) in only 200 steps (evaluations of the fitness function), out of a total of 21 billion possibilities, which is the number of steps a brute-force approach would have to take. The reason for such high efficiency of the genetic algorithm is due to the embedded structure in the PBN that can be exploited.

7.8. External control

The aforementioned intervention methods do not provide effective "knobs" that could be used to externally guide the time evolution of the network towards more desirable states. By considering possible external interventions as control inputs,

and given a finite treatment horizon, ideas from optimal control theory can be applied to develop a general optimal intervention theory for Markovian gene regulatory networks, in particular, for PBNs. This strategy makes use of dynamic programming. The costs and benefits of using interventions are incorporated into a single performance index, which also penalizes the state where the network ends up following the intervention. The use of auxiliary variables makes sense from a biological perspective. For instance, in the case of diseases like cancer, auxiliary treatment inputs such as radiation, chemo-therapy, and so forth may be employed to move the state probability distribution vector away from one which is associated with uncontrolled cell proliferation or markedly reduced apoptosis. The auxiliary variables can include genes which serve as external master-regulators for all the genes in the network. To be consistent with the binary nature of the expression status of individual genes in the PBN, we will assume that the auxiliary variables (*control inputs*) can take on only the binary values zero or one. The values of the individual control inputs can be changed from one time step to the other in an effort to make the network behave in a desirable fashion. Interventions using full information (Datta et al. [7]) and partial information (Datta et al. [8]) have been considered for instantaneously random PBNs, for which the states of the Markov chain are the states of the PBN. Following Datta et al. [7], we summarize the full-information case here.

7.8.1. The optimal control problem

To develop the control strategy, let $\mathbf{x}(k) = [x_1(k), x_2(k), \ldots, x_n(k)]$ denote the state vector (gene activity profile) at step k for the n genes in the network. The state vector $\mathbf{x}(k)$ at any time step k is essentially an n-digit binary number whose decimal equivalent is given by

$$z(k) = 1 + \sum_{j=1}^{n} 2^{n-1} x_j(k). \tag{7.12}$$

As $\mathbf{x}(k)$ ranges from $000\cdots0$ to $111\cdots1$, $z(k)$ takes on all values from 1 to 2^n. The map from $\mathbf{x}(k)$ to $z(k)$ is one-to-one, onto, and hence invertible. Instead of the binary representation $\mathbf{x}(k)$ for the state vector, we can equivalently work with the decimal representation $z(k)$.

Suppose that the PBN has m control inputs, $u_1, u_2, \ldots, u_m$. Then at any given time step k, the row vector $\mathbf{u}(k) = [u_1(k), u_2(k), \ldots, u_m(k)]$ describes the complete status of all the control inputs. Clearly, $\mathbf{u}(k)$ can take on all binary values from $000\cdots0$ to $111\cdots1$. An equivalent decimal representation of the control input is given by

$$v(k) = 1 + \sum_{i=1}^{m} 2^{m-1} u_i(k). \tag{7.13}$$

As $\mathbf{u}(k)$ takes on binary values from $000\cdots0$ to $111\cdots1$, $\nu(k)$ takes on all values from 1 to 2^m. We can equivalently use $\nu(k)$ as an indicator of the complete control input status of the PBN at time step k.

As shown by Datta et al. [7], the one-step evolution of the probability distribution vector in the case of such a PBN with control inputs takes place according to the equation

$$\mathbf{w}(k+1) = \mathbf{w}(k)\mathbf{A}\big(\nu(k)\big), \tag{7.14}$$

where $\mathbf{w}(k)$ is the 2^n-dimensional state probability distribution vector and $\mathbf{A}(\nu(k))$ is the $2^n \times 2^n$ matrix of control-dependent transition probabilities. Since the transition probability matrix is a function of the control inputs, the evolution of the probability distribution vector of the PBN with control depends not only on the initial distribution vector but also on the values of the control inputs at different time steps.

In the control literature, (7.14) is referred to as a *controlled Markov chain* (Bertsekas [113]). Given a controlled Markov chain, the objective is to come up with a sequence of control inputs, usually referred to as a *control strategy*, such that an appropriate cost function is minimized over the entire class of allowable control strategies. To arrive at a meaningful solution, the cost function must capture the costs and benefits of using any control. The design of a good cost function is application dependent and likely to require considerable expert knowledge. In the case of diseases like cancer, treatment is typically applied over a finite time horizon. For instance, in the case of radiation treatment, the patient may be treated with radiation over a fixed interval of time, following which the treatment is suspended for some time as the effects are evaluated. After that, the treatment may be applied again but the important point to note is that the treatment window at each stage is usually finite. Thus we will be interested in a finite horizon problem where the control is applied only over a finite number of steps.

Suppose that the number of steps over which the control input is to be applied is M and we are interested in controlling the behavior of the PBN over the interval $k = 0, 1, 2, \ldots, M-1$. We can define a cost $C_k(z(k), \nu(k))$ as being the cost of applying the control input $\nu(k)$ when the state is $z(k)$. The expected cost of control over the entire treatment horizon is

$$E\left[\sum_{k=0}^{M-1} C_k\big(z(k), \nu(k)\big) \,|\, z(0)\right]. \tag{7.15}$$

Even if the network starts from a given (deterministic) initial state $z(0)$, the subsequent states will be random because of the stochastic nature of the evolution in (7.14). Consequently, the cost in (7.15) must be defined using an expectation. Expression (7.15) gives us one component of the finite horizon cost, namely the cost of control.

Regarding the second component of the cost, the net result of the control actions $v(0), v(1), \ldots, v(M-1)$ is that the state of the PBN will transition according to (7.14) and will end up in some state $z(M)$. Owing to the stochastic nature of the evolution, the terminal state $\mathbf{z}(M)$ is a random variable that can potentially take on any of the values $1, 2, \ldots, 2^n$. We assign a penalty, or terminal cost, $C_M(z(M))$ to each possible state. To do this, divide the states into different categories depending on their desirability and assign higher terminal costs to the undesirable states. For instance, a state associated with rapid cell proliferation leading to cancer should be associated with a high terminal penalty while a state associated with normal behavior should be assigned a low terminal penalty. For our purposes here, we will assume that the assignment of terminal penalties has been carried out and we have a terminal penalty $C_M(z(M))$ which is a function of the terminal state. This is the second component of our cost function. $C_M(z(M))$ is a random variable and so we must take its expectation while defining the cost function to be minimized. In view of (7.15), the finite horizon cost to be minimized is given by

$$E\left[\sum_{k=0}^{M-1} C_k(z(k), v(k)) + C_M(z(M)) \mid z(0)\right]. \tag{7.16}$$

To proceed further, let us assume that at time k, the control input $v(k)$ is a function of the current state $z(k)$, namely, $v(k) = \mu_k(z(k))$. The optimal control problem can now be stated: given an initial state $z(0)$, find a control law $\pi = [\mu_0, \mu_1, \ldots, \mu_{M-1}]$ that minimizes the cost functional

$$J_\pi(z(0)) = E\left[\sum_{k=0}^{M-1} C_k(z(k), \mu_k(z(k))) + C_M(z(M))\right] \tag{7.17}$$

subject to the probability constraint

$$P[z(k+1) = j \mid z(k) = i] = a_{ij}(v(k)), \tag{7.18}$$

where $a_{ij}(v(k))$ is the ith row, jth column entry of the matrix $\mathbf{A}(v(k))$. Optimal control problems of the type described by the preceding two equations can be solved by using *dynamic programming*, a technique pioneered by Bellman in the 1960s. We will not pursue the solution here, instead referring the reader to Datta et al. [7] for the complete solution. We will, however, follow Datta et al. [7] in providing an application.

7.8.2. Control of WNT5A in metastatic melanoma

In expression profiling studies concerning metastatic melanoma, the abundance of mRNA for the gene WNT5A was found to be a highly discriminating difference between cells with properties typically associated with high metastatic competence

versus those with low metastatic competence (Bittner et al. [114]; Weeraratna et al. [115]). In this study, experimentally increasing the levels of the WNT5A protein secreted by a melanoma cell line via genetic engineering methods directly altered the metastatic competence of that cell as measured by the standard in vitro assays for metastasis. A further finding of interest was that an intervention that blocked the WNT5A protein from activating its receptor, the use of an antibody that binds WNT5A protein, could substantially reduce WNT5A's ability to induce a metastatic phenotype. This suggests a study of control based on interventions that alter the contribution of the WNT5A gene's action to biological regulation, since the available data suggests that disruption of this influence could reduce the chance of a melanoma metastasizing, a desirable outcome.

The methods for choosing the 10 genes involved in a small local network that includes the activity of the WNT5A gene and the rules of interaction have been described by Kim et al. [9]. The expression status of each gene was quantized to one of three possible levels: -1 (down-regulated), 0 (unchanged), and 1 (up-regulated). Although the network is ternary valued instead of binary valued, the PBN formulation extends directly, with the terminology "probabilistic gene regulatory network" being applied instead of probabilistic Boolean network (Zhou et al. [10, 61]). The control theory also extends directly. Indeed, to apply the control algorithm of (Datta et al. [7]), it is not necessary to actually construct a PBN; all that is required are the transition probabilities between the different states under the different controls. For this study, the number of genes was reduced from 10 to 7 by using CoD analysis. The resulting genes along with their multivariate relationships are shown in Figure 7.9.

The control objective for this seven-gene network is to externally down-regulate the WNT5A gene. The reason is that it is biologically known that WNT5A ceasing to be down-regulated is strongly predictive of the onset of metastasis. For each gene in this network, its two best two-gene predictors were determined, along with their corresponding CoDs. Using the procedure by Shmulevich et al. in [1], the CoD information was used to determine the seven-by-seven matrix of transition probabilities for the Markov chain corresponding to the dynamic evolution of the seven-gene network.

The optimal control problem can now be completely specified by choosing (i) the treatment/intervention window, (ii) the terminal penalty, and (iii) the types of controls and the costs associated with them. For the treatment window, a window of length 5 was arbitrarily chosen, that is, control inputs would be applied only at time steps 0, 1, 2, 3, and 4. The terminal penalty at time step 5 was chosen as follows. Since the objective is to ensure that WNT5A is down regulated, a penalty of zero was assigned to all states for which WNT5A equals -1, a penalty of 3 to all states for which WNT5A equals 0, and a penalty of 6 to all states for which WNT5A equals 1. Here the choice of the numbers 3 and 6 is arbitrary but they do reflect our attempt to capture the intuitive notion that states where WNT5A equals 1 are less desirable than those where WNT5A equals 0. Two types of possible controls were considered by Datta et al. [7]; here only one of them was considered, where WNT5A is controlled via pirin.

The control objective is to keep WNT5A down-regulated. The control action consists of either forcing pirin to -1 or letting it remain wherever it is. A control cost of 1 is incurred if and only if pirin has to be forcibly reset to -1 at that time step. Using the resulting optimal controls, the evolution of the state probability distribution vectors has been studied with and without control. For every possible initial state, the resulting simulations have indicated that, at the final state, the probability of WNT5A being equal to -1 is higher with control than that without control; however, the probability of WNT5A being equal to -1 at the final time point is not, in general, equal to 1. This is not surprising given that one is trying to control the expression status of WNT5A using another gene and the control horizon of length 5 simply may not be adequate for achieving the desired objective with such a high probability. Nevertheless, even in this case, if the network starts from the state corresponding to STC2 $= -1$, HADHB $= 0$, MART-1 $= 0$, RET-1 $= 0$, S100P $= -1$, pirin $= 1$, WNT5A $= 1$ and evolves under optimal control, then the probability of WNT5A $= -1$ at the final time point equals 0.673521. This is quite good in view of the fact that the same probability would have been equal to 0 in the absence of any control action.

Bibliography

[1] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[2] I. Shmulevich, E. R. Dougherty, and W. Zhang, "Gene perturbation and intervention in probabilistic Boolean networks," *Bioinformatics*, vol. 18, no. 10, pp. 1319–1331, 2002.

[3] I. Shmulevich, E. R. Dougherty, and W. Zhang, "Control of stationary behavior in probabilistic Boolean networks by means of structural intervention," *Journal of Biological Systems*, vol. 10, no. 4, pp. 431–445, 2002.

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

[5] E. R. Dougherty and I. Shmulevich, "Mappings between probabilistic Boolean networks," *Signal Processing*, vol. 83, no. 4, pp. 799–809, 2003.

[6] I. Shmulevich, I. Gluhovsky, R. Hashimoto, E. R. Dougherty, and W. Zhang, "Steady-state analysis of probabilistic Boolean networks," *Comparative and Functional Genomics*, vol. 4, no. 6, pp. 601–608, 2003.

[7] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks," *Machine Learning*, vol. 52, no. 1-2, pp. 169–191, 2003.

[8] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks: the imperfect information case," *Bioinformatics*, vol. 20, no. 6, pp. 924–930, 2004.

[9] S. Kim, H. Li, E. R. Dougherty, et al., "Can Markov chain models mimic biological regulation?," *Journal of Biological Systems*, vol. 10, no. 4, pp. 337–357, 2002.

[10] X. Zhou, X. Wang, and E. R. Dougherty, "Construction of genomic networks using mutual-information clustering and reversible-jump Markov-Chain-Monte-Carlo predictor design," *Signal Processing*, vol. 83, no. 4, pp. 745–761, 2003.

[11] R. F. Hashimoto, S. Kim, I. Shmulevich, W. Zhang, M. L. Bittner, and E. R. Dougherty, "Growing genetic regulatory networks from seed genes," *Bioinformatics*, vol. 20, no. 8, pp. 1241–1247, 2004.

[12] I. Shmulevich, "Model selection in genomics," *Environ. Health Perspect.*, vol. 111, no. 6, pp. A328–A329, 2003.

[13] E. P. van Someren, L. F. A. Wessels, and M. J. T. Reinders, "Linear modeling of genetic networks from experimental data," in *Proc. 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00)*, pp. 355–366, San Diego, Calif, USA, August 2000.

[14] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear modeling of mRNA expression levels during CNS development and injury," in *Pac. Symp. Biocomput. (PSB '99)*, vol. 4, pp. 41–52, Hawaii, USA, January 1999.

[15] K. Murphy and S. Mian, "Modelling gene expression data using dynamic Bayesian networks," Tech. Rep., University of California, Berkeley, Calif, USA, 1999.

[16] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian network to analyze expression data," *J. Computational Biology*, vol. 7, pp. 601–620, 2000.

[17] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks," in *Pac. Symp. Biocomput. (PSB '01)*, pp. 422–433, Hawaii, USA, January 2001.

[18] E. J. Moler, D. C. Radisky, and I. S. Mian, "Integrating naive Bayes models and external knowledge to examine copper and iron homeostasis in S. cerevisiae," *Physiol. Genomics*, vol. 4, no. 2, pp. 127–135, 2000.

[19] D. C. Weaver, C. T. Workman, and G. D. Stormo, "Modeling regulatory networks with weight matrices," in *Pac. Symp. Biocomput. (PSB '99)*, vol. 4, pp. 112–123, Hawaii, USA, January 1999.

[20] T. Mestl, E. Plahte, and S. W. Omholt, "A mathematical framework for describing and analysing gene regulatory networks," *J. Theor. Biol.*, vol. 176, no. 2, pp. 291–300, 1995.

[21] T. Chen, H. L. He, and G. M. Church, "Modeling gene expression with differential equations," in *Pac. Symp. Biocomput. (PSB '99)*, vol. 4, pp. 29–40, Hawaii, USA, January 1999.

[22] J. Goutsias and S. Kim, "A nonlinear discrete dynamical model for transcriptional regulation: construction and properties," *Biophys. J.*, vol. 86, no. 4, pp. 1922–1945, 2004.

[23] H. H. McAdams and A. Arkin, "Stochastic mechanisms in gene expression," *Proc. Natl. Acad. Sci. USA*, vol. 94, no. 3, pp. 814–819, 1997.

[24] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells," *Genetics*, vol. 149, no. 4, pp. 1633–1648, 1998.

[25] P. Smolen, D. A. Baxter, and J. H. Byrne, "Mathematical modeling of gene networks," *Neuron*, vol. 26, no. 3, pp. 567–580, 2000.

[26] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins, "Computational studies of gene regulatory networks: in numero molecular biology," *Nat. Rev. Genet.*, vol. 2, no. 4, pp. 268–279, 2001.

[27] H. de Jong, "Modeling and simulation of genetic regulatory systems: a literature review," *J. Comput. Biol.*, vol. 9, no. 1, pp. 69–103, 2002.

[28] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *J. Theor. Biol.*, vol. 22, no. 3, pp. 437–467, 1969.

[29] L. Glass and S. A. Kauffman, "The logical analysis of continuous, non-linear biochemical control networks," *J. Theor. Biol.*, vol. 39, no. 1, pp. 103–129, 1973.

[30] S. Huang, "Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery," *J. Mol. Med.*, vol. 77, no. 6, pp. 469–480, 1999.

[31] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, New York, NY, USA, 1993.

[32] R. Somogyi and C. Sniegoski, "Modeling the complexity of gene networks: understanding multigenic and pleiotropic regulation," *Complexity*, vol. 1, pp. 45–63, 1996.

[33] M. Aldana, S. Coppersmith, and L. P. Kadanoff, "Boolean dynamics with random couplings," in *Perspectives and Problems in Nonlinear Science*, E. Kaplan, J. E. Marsden, and K. R. Sreenivasan, Eds., Applied Mathematical Sciences Series, pp. 23–89, Springer-Verlag, New York, NY, USA, 2003.

[34] J. W. Bodnar, "Programming the Drosophila embryo," *J. Theor. Biol.*, vol. 188, no. 4, pp. 391–445, 1997.

[35] C. H. Yuh, H. Bolouri, and E. H. Davidson, "Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene," *Science*, vol. 279, no. 5358, pp. 1896–1902, 1998.

[36] L. Mendoza, D. Thieffry, and E. R. Alvarez-Buylla, "Genetic control of flower morphogenesis in Arabidopsis thaliana: a logical analysis," *Bioinformatics*, vol. 15, no. 7-8, pp. 593–606, 1999.

[37] S. Huang and D. E. Ingber, "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," to appear in InterJournal Genetics, http://www.interjournal.org.

[38] G. Lahav, N. Rosenfeld, A. Sigal, et al., "Dynamics of the p53-Mdm2 feedback loop in individual cells," *Nat. Genet.*, vol. 36, no. 2, pp. 147–150, 2004.

[39] I. Shmulevich and W. Zhang, "Binary analysis and optimization-based normalization of gene expression data," *Bioinformatics*, vol. 18, no. 4, pp. 555–565, 2002.

[40] X. Zhou, X. Wang, and E. R. Dougherty, "Binarization of microarray data on the basis of a mixture model," *Mol. Cancer Ther.*, vol. 2, no. 7, pp. 679–684, 2003.

[41] W. Zhang, I. Shmulevich, and J. Astola, *Microarray Quality Control*, John Wiley & Sons, New York, NY, USA, 2004.

[42] Y. Chen, E. R. Dougherty, and M. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *Biomedical Optics*, vol. 2, no. 4, pp. 364–374, 1997.

[43] M. K. Kerr, E. H. Leiter, L. Picard, and G. A. Churchill, "Sources of variation in microarray experiments," in *Computational and Statistical Approaches to Genomics*, W. Zhang and I. Shmulevich, Eds., Kluwer Academic Publishers, Boston, Mass, USA, 2002.

[44] H. H. McAdams and A. Arkin, "It's a noisy business! Genetic regulation at the nanomolar scale," *Trends Genet.*, vol. 15, no. 2, pp. 65–69, 1999.

[45] Z. Szallasi and S. Liang, "Modeling the normal and neoplastic cell cycle with "realistic Boolean genetic networks": their application for understanding carcinogenesis and assessing therapeutic strategies," in *Pac. Symp. Biocomput. (PSB '98)*, vol. 3, pp. 66–76, Hawaii, USA, January 1998.

[46] A. Wuensche, "Genomic regulation modeled as a network with basins of attraction," in *Pac. Symp. Biocomput. (PSB '98)*, vol. 3, pp. 89–102, Hawaii, USA, January 1998.

[47] R. Thomas, D. Thieffry, and M. Kaufman, "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.*, vol. 57, no. 2, pp. 247–276, 1995.

[48] S. Liang, S. Fuhrman, and R. Somogyi, "Reveal, a general reverse engineering algorithm for inference of genetic network architectures," in *Pac. Symp. Biocomput. (PSB '98)*, vol. 3, pp. 18–29, Hawaii, USA, January 1998.

[49] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano, "Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions," in *Proc. 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '98)*, pp. 695–702, San Francisco, Calif, USA, January 1998.

[50] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model," *Pac. Symp. Biocomput.*, vol. 4, pp. 17–28, 1999.

[51] T. Akutsu, S. Miyano, and S. Kuhara, "Inferring qualitative relations in genetic networks and metabolic pathways," *Bioinformatics*, vol. 16, no. 8, pp. 727–734, 2000.

[52] P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co-expression clustering to reverse engineering," *Bioinformatics*, vol. 16, no. 8, pp. 707–726, 2000.

[53] I. Shmulevich, A. Saarinen, O. Yli-Harja, and J. Astola, "Inference of genetic regulatory networks via best-fit extensions," in *Computational and Statistical Approaches to Genomics*, W. Zhang and I. Shmulevich, Eds., Kluwer Academic Publishers, Boston, Mass, USA, 2002.

[54] H. Lähdesmäki, I. Shmulevich, and O. Yli-Harja, "On learning gene regulatory networks under the Boolean network model," *Machine Learning*, vol. 52, no. 1-2, pp. 147–167, 2003.

[55] E. R. Dougherty, S. Kim, and Y. Chen, "Coefficient of determination in nonlinear signal processing," *Signal Processing*, vol. 80, no. 10, pp. 2219–2235, 2000.

[56] S. Kim, E. R. Dougherty, Y. Chen, et al., "Multivariate measurement of gene expression relationships," *Genomics*, vol. 67, no. 2, pp. 201–209, 2000.

[57] I. Shmulevich and S. A. Kauffman, "Activities and sensitivities in Boolean network models," *Phys. Rev. Lett.*, vol. 93, no. 4, p. 048701, 2004.

[58] E. R. Dougherty, M. Bittner, Y. Chen, et al., "Nonlinear filters in genomic control," in *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP '99)*, Antalya, Turkey, June 1999.

[59] S. Kim, E. R. Dougherty, M. L. Bittner, et al., "General nonlinear framework for the analysis of gene interaction via multivariate expression arrays," *J. Biomed. Opt.*, vol. 5, no. 4, pp. 411–424, 2000.

[60] X. Zhou, X. Wang, and E. R. Dougherty, "Gene prediction using multinomial probit regression with Bayesian gene selection," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 1, pp. 115–124, 2004.

[61] X. Zhou, X. Wang, R. Pal, I. Ivanov, M. L. Bittner, and E. R. Dougherty, "A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks," *Bioinformatics*, vol. 20, no. 17, pp. 2918–2927, 2004.

[62] E. B. Suh, E. R. Dougherty, S. Kim, et al., "Parallel computation and visualization tools for code-termination analysis of multivariate gene-expression relations," in *Computational and Statistical Approaches to Genomics*, W. Zhang and I. Shmulevich, Eds., Kluwer Academic Publishers, Boston, Mass, USA, 2002.

[63] I. Ivanov and E. R. Dougherty, "Reduction mappings between probabilistic Boolean networks," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 1, pp. 125–131, 2004.

[64] A. L. Gartel and A. L. Tyner, "Transcriptional regulation of the p21(WAF1/CIP1) gene," *Exp. Cell Res.*, vol. 246, no. 2, pp. 280–289, 1999.

[65] E. Schneider, M. Montenarh, and P. Wagner, "Regulation of CAK kinase activity by p53," *Oncogene*, vol. 17, no. 21, pp. 2733–2741, 1998.

[66] A. Rzhetsky, T. Koike, S. Kalachikov, et al., "A knowledge model for analysis and simulation of regulatory networks," *Bioinformatics*, vol. 16, no. 12, pp. 1120–1128, 2000.

[67] U. Braga-Neto, R. Hashimoto, E. R. Dougherty, D. V. Nguyen, and R. J. Carroll, "Is cross-validation better than resubstitution for ranking genes?," *Bioinformatics*, vol. 20, no. 2, pp. 253–258, 2004.

[68] M. Brun, E. R. Dougherty, and I. Shmulevich, "Attractors in probabilistic Boolean networks: steady-state probabilities and classification," to appear in Signal Process.

[69] M. K. Cowles and B. P. Carlin, "Markov Chain Monte Carlo convergence diagnostics: a comparative study," *Journal of the American Statistical Association*, vol. 91, pp. 883–904, 1996.

[70] A. E. Raftery and S. Lewis, "How many iterations in the Gibbs sampler?" in *Bayesian Statistics*, J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith, Eds., vol. 4, pp. 763–773, Oxford University Press, Oxford, UK, 1992.

[71] T. N. Sato, Y. Qin, C. A. Kozak, and K. L. Audus, "Tie-1 and tie-2 define another class of putative receptor tyrosine kinase genes expressed in early embryonic vascular system," *Proc. Natl. Acad. Sci. USA*, vol. 90, no. 20, pp. 9355–9358, 1993.

[72] S. Y. Cheng, H. J. Huang, M. Nagane, et al., "Suppression of glioblastoma angiogenicity and tumorigenicity by inhibition of endogenous expression of vascular endothelial growth factor," *Proc. Natl. Acad. Sci. USA*, vol. 93, no. 16, pp. 8502–8507, 1996.

[73] S. Hayashi, M. Yamamoto, Y. Ueno, et al., "Expression of nuclear factor-kappa B, tumor necrosis factor receptor type 1, and c-Myc in human astrocytomas," *Neurol. Med. Chir. (Tokyo)*, vol. 41, no. 4, pp. 187–195, 2001.

[74] M. Arsura, M. Wu, and G. E. Sonenshein, "TGF beta 1 inhibits NF-kappa B/Rel activity inducing apoptosis of B cells: transcriptional activation of I kappa B alpha," *Immunity*, vol. 5, no. 1, pp. 31–40, 1996.

[75] E. D. Strauch, J. Yamaguchi, B. L. Bass, and J. Y. Wang, "Bile salts regulate intestinal epithelial cell migration by nuclear factor-kappa B-induced expression of transforming growth factor-beta," *J. Am. Coll. Surg.*, vol. 197, no. 6, pp. 974–984, 2003.

[76] C. D. Johnson, Y. Balagurunathan, K. P. Lu, et al., "Genomic profiles and predictive biological networks in oxidant-induced atherogenesis," *Physiol. Genomics*, vol. 13, no. 3, pp. 263–275, 2003.

[77] I. Shmulevich, O. Yli-Harja, J. Astola, and A. Korshunov, "On the robustness of the class of stack filters," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1640–1649, 2002.

[78] T. E. Ideker, V. Thorsson, and R. M. Karp, "Discovery of regulatory interactions through perturbation: inference and experimental design," in *Pac. Symp. Biocomput. (PSB '00)*, vol. 5, pp. 302–313, Hawaii, USA, January 2000.

[79] R. M. Karp, R. Stoughton, and K. Y. Yeung, "Algorithms for choosing differential gene expression experiments," in *Proc. 3rd Annual International Conference on Computational Molecular Biology (RECOMB '99)*, pp. 208–217, Lyon, France, 1999.

[80] Y. Maki, D. Tominaga, M. Okamoto, S. Watanabe, and Y. Eguchi, "Development of a system for the inference of large scale genetic networks," in *Pac. Symp. Biocomput. (PSB '01)*, vol. 6, pp. 446–458, Hawaii, USA, January 2001.

[81] K. Noda, A. Shinohara, M. Takeda, S. Matsumoto, S. Miyano, and S. Kuhara, "Finding genetic network from experiments by weighted network model," *Genome Inform.*, vol. 9, pp. 141–150, 1998.

[82] I. Shmulevich, M. Gabbouj, and J. Astola, "Complexity of the consistency problem for certain Post classes," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 31, no. 2, pp. 251–253, 2001.

[83] 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.

[84] E. J. Coyle and J. H. Lin, "Stack filters and the mean absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 1244–1254, 1988.

[85] E. J. Coyle, J. H. Lin, and M. Gabbouj, "Optimal stack filtering and the estimation and structural approaches to image processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 2037–2066, 1989.

[86] R. Yang, L. Yin, M. Gabbouj, J. Astola, and Y. Neuvo, "Optimal weighted median filtering under structural constraints," *IEEE Trans. on Signal Processing*, vol. 43, no. 3, pp. 591–604, 1995.

[87] E. R. Dougherty and R. P. Loce, "Precision of morphological-representation estimators for translation-invariant binary filters: increasing and nonincreasing," *Signal Processing*, vol. 40, no. 2-3, pp. 129–154, 1994.

[88] E. R. Dougherty and Y. Chen, "Optimal and adaptive design of logical granulometric filters," *Adv. Imaging Electron Phys.*, vol. 117, pp. 1–71, 2001.

[89] E. R. Dougherty and J. Barrera, "Pattern recognition theory in nonlinear signal processing," *J. Math. Imaging Vision*, vol. 16, no. 3, pp. 181–197, 2002.

[90] S. A. Kauffman, "The large scale structure and dynamics of gene control circuits: an ensemble approach," *J. Theor. Biol.*, vol. 44, no. 1, pp. 167–190, 1974.

[91] S. E. Harris, B. K. Sawhill, A. Wuensche, and S. Kauffman, "A model of transcriptional regulatory networks based on biases in the observed regulation rules," *Complexity*, vol. 7, no. 4, pp. 23–40, 2002.

[92] I. Shmulevich, H. Lähdesmäki, and K. Egiazarian, "Spectral methods for testing membership in certain Post classes and the class of forcing functions," *IEEE Signal Processing Lett.*, vol. 11, no. 2, pp. 289–292, 2004.

[93] I. Shmulevich, H. Lähdesmäki, E. R. Dougherty, J. Astola, and W. Zhang, "The role of certain Post classes in Boolean network models of genetic networks," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 19, pp. 10734–10739, 2003.

[94] E. Post, "Introduction to a general theory of elementary propositions," *Amer. J. Math.*, vol. 43, pp. 163–185, 1921.

[95] E. I. Nechiporuk, "On the complexity of circuits in some bases, containing non-trivial elements with zero weights," *Prob. Cybern.*, , no. 8, 1962.

[96] A. A. Muchnik and S. G. Gindikin, "On the completeness of systems of unreliable elements which realize functions of the algebra of logic," *Dokl. Akad. Nauk SSSR*, vol. 144, no. 5, 1962.

[97] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. Suppl 1, pp. 2566–2572, 2002.

[98] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[99] P. Kleihues and W. K. Cavenee, Eds., *World Health Organization Classification of Tumours: Pathology and Genetics of Tumours of Nervous System*, Oxford University Press, Oxford, UK, 2000.

[100] G. N. Fuller, C. H. Rhee, K. R. Hess, et al., "Reactivation of insulin-like growth factor binding protein 2 expression in glioblastoma multiforme: a revelation by parallel gene expression profiling," *Cancer Res.*, vol. 59, no. 17, pp. 4228–4232, 1999.

[101] S. L. Sallinen, P. K. Sallinen, H. K. Haapasalo, et al., "Identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques," *Cancer Res.*, vol. 60, no. 23, pp. 6617–6622, 2000.

[102] M. W. Elmlinger, M. H. Deininger, B. S. Schuett, et al., "In vivo expression of insulin-like growth factor-binding protein-2 in human gliomas increases with the tumor grade," *Endocrinology*, vol. 142, no. 4, pp. 1652–1658, 2001.

[103] V. Cazals, E. Nabeyrat, S. Corroyer, Y. de Keyzer, and A. Clement, "Role for NF-kappa B in mediating the effects of hyperoxia on IGF-binding protein 2 promoter activity in lung alveolar epithelial cells," *Biochim. Biophys. Acta.*, vol. 1448, no. 3, pp. 349–362, 1999.

[104] J. Folkman, "Angiogenesis in cancer, vascular, rheumatoid and other disease," *Nat. Med.*, vol. 1, no. 1, pp. 27–31, 1995.

[105] A. Bikfalvi and R. Bicknell, "Recent advances in angiogenesis, anti-angiogenesis and vascular targeting," *Trends Pharmacol. Sci.*, vol. 23, no. 12, pp. 576–582, 2002.

[106] J. S. Rubin, H. Osada, P. W. Finch, W. G. Taylor, S. Rudikoff, and S. A. Aaronson, "Purification and characterization of a newly identified growth factor specific for epithelial cells," *Proc. Natl. Acad. Sci. USA*, vol. 86, no. 3, pp. 802–806, 1989.

[107] S. S. Banga, H. L. Ozer, and S. T. Park, S. K. Lee, "Assignment of PTK7 encoding a receptor protein tyrosine kinase-like molecule to human chromosome 6p21.1-p12.2 by fluorescence in situ hybridization," *Cytogenet. Cell Genet.*, vol. 76, no. 1-2, pp. 43–44, 1997.

[108] K. V. Stoletov, K. E. Ratcliffe, and B. I. Terman, "Fibroblast growth factor receptor substrate 2 participates in vascular endothelial growth factor-induced signaling," *FASEB J.*, vol. 16, no. 10, pp. 1283–1285, 2002.

[109] C. Q. Wei, Y. Gao, K. Lee, et al., "Macrocyclization in the design of Grb2 SH2 domain-binding ligands exhibiting high potency in whole-cell systems," *J. Med. Chem.*, vol. 46, no. 2, pp. 244–254, 2003.

[110] C. C. Bancroft, Z. Chen, G. Dong, et al., "Coexpression of proangiogenic factors IL-8 and VEGF by human head and neck squamous cell carcinoma involves coactivation by MEK-MAPK and IKK-NF-kappaB signal pathways," *Clin. Cancer Res.*, vol. 7, no. 2, pp. 435–442, 2001.

[111] G. Pearson, J. M. English, M. A. White, and M. H. Cobb, "ERK5 and ERK2 cooperate to regulate NF-kappaB and cell transformation," *J. Biol. Chem.*, vol. 276, no. 11, pp. 7927–7931, 2001.

[112] J.-P. Yang, M. Hori, T. Sanda, and T. Okamoto, "Identification of a novel inhibitor of nuclear factor-kappa-B, RelA-associated inhibitor," *J. Biol. Chem.*, vol. 274, no. 22, pp. 15662–15670, 1999.

[113] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*, Academic Press, Orlando, Fla, USA, 1976.

[114] M. Bittner, P. Meltzer, and Y. Chen, "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, no. 6795, pp. 536–540, 2000.

[115] A. T. Weeraratna, Y. Jiang, G. Hostetter, et al., "Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma," *Cancer Cell*, vol. 1, no. 3, pp. 279–288, 2002.

[116] S. Agaian, J. Astola, and K. Egiazarian, *Binary Polynomial Transforms and Nonlinear Digital Filters*, Marcel Dekker, New York, NY, USA, 1995.

[117] O. Coudert, "Doing two-level logic minimization 100 times faster," in *Proc. 6th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '98)*, pp. 112–121, Society for Industrial and Applied Mathematics, San Francisco, Calif, USA, January 1995.

[118] S. Dedhar, B. Williams, and G. Hannigan, "Integrin-linked kinase (ILK): a regulator of integrin and growth-factor signalling," *Trends Cell Biol.*, vol. 9, no. 8, pp. 319–323, 1999.

[119] E. R. Dougherty and S. N. Attoor, "Design issues and comparison of methods for microarray-based classification," in *Computational and Statistical Approaches to Genomics*, W. Zhang and I. Shmulevich, Eds., Kluwer Academic Publishers, Boston, Mass, USA, 2002.

[120] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York, NY, USA, 1993.

[121] C. J. Geyer and E. A. Thompson, "Annealing Markov Chain Monte Carlo with applications to ancestral inference," *Journal of the American Statistical Asssociation*, vol. 90, pp. 909–920, 1995.

[122] G. D. Hachtel, E. Macii, A. Pardo, and F. Somenzi, "Markovian analysis of large finite state machines," *IEEE Trans. Computer-Aided Design*, vol. 15, no. 12, pp. 1479–1493, 1996.

[123] P. L. Hammer, A. Kogan, and U. G. Rothblum, "Evaluation, strength, and relevance of variables of Boolean functions," *SIAM Journal on Discrete Mathematics*, vol. 13, no. 3, pp. 302–312, 2000.

[124] W. K. Hastings, "Monte Carlo sampling methods using Markov Chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[125] S. Kim, E. R. Dougherty, I. Shmulevich, et al., "Identification of combination gene sets for glioma classification," *Mol. Cancer Ther.*, vol. 1, no. 13, pp. 1229–1236, 2002.

[126] L. M. Loew and J. C. Schaff, "The virtual cell: a software environment for computational cell biology," *Trends Biotechnol.*, vol. 19, no. 10, pp. 401–406, 2001.

[127] D. A. McAllester, "PAC-Bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[128] P. Mendes, W. Sha, and K. Ye, "Artificial gene networks for objective comparison of analysis algorithms," *Bioinformatics*, vol. 19, no. Suppl 2, pp. II122–II129, 2003.

[129] C. Mircean, I. Tabus, J. Astola, et al., "Quantization and similarity measure selection for discrimination of lymphoma subtypes under k-nearest neighbor classification," in *Proc. SPIE Photonics West, Biomedical Optics*, San Jose, Calif, USA, January 2004.

[130] M. Mitchell, J. P. Crutchfield, and P. T. Hraber, "Evolving cellular automata to perform computations: mechanisms and impediments," *Physica D.*, vol. 75, no. 1–3, pp. 361–391, 1994.

[131] S. R. Neves and R. Iyengar, "Modeling of signaling networks," *Bioessays*, vol. 24, no. 12, pp. 1110–1117, 2002.

[132] S. N. Nikolopoulos and C. E. Turner, "Integrin-linked kinase (ILK) binding to paxillin LD1 motif regulates ILK localization to focal adhesions," *J. Biol. Chem.*, vol. 276, no. 26, pp. 23499–23505, 2001.

[133] B. Plateau and K. Atif, "Stochastic automata network of modeling parallel systems," *IEEE Trans. Software Eng.*, vol. 17, no. 10, pp. 1093–1108, 1991.

[134] J. S. Rosenthal, "Minorization conditions and convergence rates for Markov chain Monte Carlo," *Journal of the American Statistical Asssociation*, vol. 90, no. 430, pp. 558–566, 1995.

[135] I. Shmulevich, O. Yli-Harja, K. Egiazarian, and J. Astola, "Output distributions of recursive stack filters," *IEEE Signal Processing Lett.*, vol. 6, no. 7, pp. 175–178, 1999.

[136] P. T. Spellman, G. Sherlock, M. Q. Zhang, et al., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol. Biol. Cell.*, vol. 9, no. 12, pp. 3273–3297, 1998.

[137] I. Tabus and J. Astola, "On the use of MDL principle in gene expression prediction," *EURASIP J. Appl. Signal Process.*, vol. 4, no. 4, pp. 297–303, 2001.

[138] A. Wagner, "How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps," *Bioinformatics*, vol. 17, no. 12, pp. 1183–1197, 2001.

[139] H. Wang, H. Wang, W. Shen, et al., "Insulin-like growth factor binding protein 2 enhances glioblastoma invasion by activating invasion-enhancing genes," *Cancer Res.*, vol. 63, no. 15, pp. 4315–4321, 2003.

[140] D. E. Zak, F. J. Doyle, G. E. Gonye, and J. S. Schwaber, "Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data," in *Proc. 2nd International Conference on Systems Biology (ICSB '01)*, pp. 231–238, Pasadena, Calif, USA, November 2001.

[141] D. E. Zak, R. K. Pearson, R. Vadigepalli, G. E. Gonye, J. S. Schwaber, and F. J. Doyle III, "Continuous-time identification of gene expression models," *OMICS*, vol. 7, no. 4, pp. 373–386, 2003.

Ilya Shmulevich: Cancer Genomics Laboratory, Department of Pathology, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Email: is@ieee.org

Edward R. Dougherty: Department of Electrical Engineering, Texas A&M University, TX 77843-3128, USA; Cancer Genomics Laboratory, Department of Pathology, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

Email: e_dougherty@tamu.edu