Temporal Boolean Network Models of Genetic Networks and Their Inference from Gene Expression Time Series

Adrian Silvescu *

Department of Computer Science
Iowa State University
Ames, IA 50011, USA

Vasant Honavar †

Department of Computer Science and
Graduate Program in Bioinformatics and Computational Biology
Iowa State University
Ames, IA 50011, USA

Identification of genetic regulatory networks and genetic signal transduction pathways from gene expression data is one of key problems in computational molecular biology. Boolean networks offer a *discrete time Boolean* model of gene expression. In this model, each gene can be in one of two states (on or off) at any given time, and the expression of a given gene at time $t+1$ can be modeled by a *Boolean* function of the expression of at most k genes at time t. Typically $k << n$, where n is total number of genes under consideration. This paper motivates and introduces a generalization of the Boolean network model to address dependencies among activity of genes that span for more than one unit of time. The resulting model, called the *temporal Boolean network* or the $TBN(n,k,T)$ model, allows the expression of each gene to be controlled by a Boolean function of the expression levels of at most k genes at times in $\{t \cdots t-(T-1)\}$. We apply an adaptation of a popular machine learning algorithm for decision tree induction for inference of a $TBN(n,k,T)$ network from artificially generated gene expression data. Preliminary experiments with synthetic gene expression data generated from known $TBN(n,k,T)$ networks demonstrate the feasibility of this approach. We conclude with a discussion of some of the limitations of the proposed approach and some directions for further research.

*Electronic mail adress: silvescu@cs.iastate.edu
†Electronic mail adress: honavar@cs.iastate.edu

Complex Systems, **11** (1997) 1–1+;

1. Introduction

The central dogma of modern biology states that the functional state of an organism is determined largely by the pattern of expression of it's genes. Thus, the function of a cell, how well a cell performs its function, and even the determination of a cell type is controlled by the level at which genes are expressed in the cell. Many biological processes of interest (e.g., cellular differentiation during development, aging, disease) are controlled by complex interactions over time between hundreds of genes. Furthermore, each gene is involved in multiple functions. Therefore, understanding the nature of complex biological processes such as development, cellular differentiation, carcinogenesis, etc., requires determining the spatio-temporal expression patterns of thousands of genes, and, more importantly, seeking out the organizing principles that allow biological processes to function in a coherent manner under different environmental conditions. Given the fact that thousands of genes are involved in determining the functional state of an organism, the task of assigning functions to genes, identifying underlying genetic signalling pathways, genetic control and regulatory networks is a formidable task.

A number of emerging high-throughput technologies are revolutionizing the means by which genetic signalling pathways involving hundreds of genes can be studied. Many of these technologies exploit the power of multiplexed data acquisition from addressable solid-state arrays of biomolecules (BioArrays). Such technologies allow researchers to obtain, in a single experiment, significant amounts of biological information regarding thousands of genes. In order to determine gene expression levels, messenger RNA samples are collected from a population of cells under a given set of experimental conditions or at different times during the execution of a biological pathway or process (e.g., glycolysis [6], cell cycle [7], and development [8]). Using DNA microarrays, the levels of mRNA expression are measured. Similar methods for determining protein concentrations [9] exist but have a lower throughput.

These advances in data aquisiton make possible, at least in principle, the inference of models of genetic (regulatory and control) networks from gene expression data. However, the large number of genes involved, complexity of the pathways, and existence of pleiotropic and multigenic interactions [2] make this a challenging task. Some issues that arise in genetic network inference from gene expression data are discussed by Thieffry *et al.* [10] and Savageau [11]. Less sophisticated methods such as clustering [22, 23, 8] can also be used in order to analyze gene expression data but the information that can be obtained by these methods are mainly restricted to either positive or negative correlations among gene expression patterns.

A variety of formal models for capturing the interactions and functional dependencies in genetic networks have been proposed in the literature. These include: *electrical circuits* [12], *Boolean networks* [1, 2, 14, 3, 15], *Fourier coefficients* [4], *Bayesian Networks* [25], *differential equations* [16], *Petri nets* [17, 18], and *Weight matrices* [19]. Each of these approaches has its own strengths and limitations in terms of the following considerations: faithfulness or accuracy of the model relative to the biological phenomenon being modeled; transparency of the model (or equivalently, its explanatory value); experimental feasibility (in terms of data requirements) of model construction, and computational tractability of automated model inference from data.

The focus of this paper is on Boolean Network models of genetic networks wherein each gene can be in one of two states – ON (expressed) or OFF (not expressed). One of their main advantages is their comprehensibility due to the transparency of the representation. It is possible to extend such network models in order to allow for a discrete set of states for each gene, instead of just two states, or even to allow states that take on real (and hence an infinite) set of values.

Previous work on inference of Boolean models of genetic networks [3, 14] has focused on data that randomly sample the state transitions typically under the assumption that the state of a gene at a given time step is influenced by the states of a subset of genes in the network at the previous time step. Since many experiments involve obtaining gene expression data by monitoring the expression of genes involved in some biological process (e.g., cell or neural development) over a period of time, the resulting data is in the form of a *time series*. In such a setting, it is of interest to understand how the expression of a gene at some stage in the process is influenced by the expression levels of other genes during the stages of the process preceding it.

In order to model the temporal dependencies that span several time steps, we introduce in this paper, a generalization of Boolean Networks called Temporal Boolean Networks. This will basically transform the Boolean Networks from a Markov(1) to Markov(T) model where T is the length of the time window during which a gene can influence another gene. We will demonstrate how Temporal Boolean Networks can be efficiently inferred using an adaptation of a greedy decision tree learning algorithm [5]. The main contribution of this paper is in terms of computational techniques for genetic network inference from gene expression time series. The insights into the nature of functionally significant interactions among genes provided by the inferred genetic networks might also suggest novel ways to exploit artificial genetic networks to solve specific computational problems e.g., in evolutionary algorithms.

The rest of the paper is organized as follows: Section 2 provides a brief overview of gene expression data analysis. Section 3 motivates

and introduces the Temporal Boolean model for genetic networks. In Section 4 we present some theoretical results concerning data-driven inference of Temporal Boolean Networks from randomly sampled transitions. Section 5 describes our approach to inferring a Temporal Boolean Network from time series data. Section 6 presents experimental results that demonstrate the feasibility of the proposed approach. Section 7 concludes with a summary and discussion of some directions for further research.

2. Gene Expression Data Analysis

The widespread use of DNA microarray and related technologies have led to increased availability of gene expression data from plants and animals. Consequently, there is a growing need for sophisticated computational tools for extracting biologically significant information from gene expression data, assigning functions to genes, and identifying shared genetic signaling pathways and genetic regulatory networks. The data from a series of m microarray measurements involving n genes can be represented as an $m \times n$ gene expression table $\mathcal{E}$ where each of the n columns consist of m entries (numbers that correspond to the measured expression levels of a single gene across m measurements). Thus, the entry e_{ti} in row t and column i of the matrix $\mathcal{E}$ denotes the expression level of gene i in the tth measurement. The rows can represent expression values measured under different experimental conditions, or data obtained by monitoring the expression levels of genes at different times during a biological process (e.g., neural development). Given this data, a number of different types of analysis are possible [20, 21]. One of the simplest forms of analysis involves clustering of gene expression patterns (columns of the table) based on some predefined clustering criterion or distance measure, but no knowledge of the functional classes of the genes. If a subset of expression patterns form a tight cluster, it can be hypothesized that the genes in question are co-expressed, or even possibly co-regulated.

3. Temporal Boolean Networks

In order to model the dependence among the expression levels of the genes we will use a generalized version of Boolean Networks. The main advantage of these models is their high level of transparency. In this section we will examine the Boolean Networks model followed by the discussion of some issues that may arise with respect to this model. The need to address these issues provides the motivation for generalizing Boolean Networks to the Temporal Boolean Networks model, which will be defined at the end of this section.

3.1 Boolean Networks

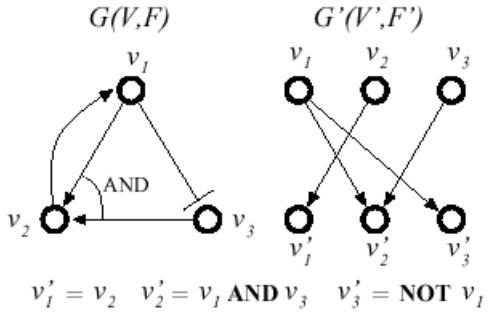
Boolean networks, introduced by Kaufmann [1] and explored in [2, 3, 14, 15] offer an attractive discrete time, boolean model for gene expression. In this model, each gene can be in one of two states either ON or OFF at any given time. The expression of a given gene at time $t+1$ is modeled by a *Boolean* function whose inputs are the expression levels of at most k genes at time t. Typically $k << n$, where n is total number of genes under consideration. We call this family of models $BN(n,k)$ networks. Some of the attractive features of this model include its conceptual simplicity and transparency as well as the availability of algorithms for automated inference of the model from expression data.

Following Akutsu *et al.* [3], we are going to give more precise definitions for the boolean networks. A Boolean Network is specified by pair $G(V,F)$, where the $V = \{v_1, ..., v_n\}$ is a set of nodes representing the genes of the network (one node for each gene) and a list $F = (f_1, ..., f_n)$ of Boolean functions. Each node v_i has an associated expression value that is either 1 (for expressed) or 0 (for not expressed) that will be denoted also by v_i. A boolean function $f_i(v_{i_1}, ..., v_{i_k}) \in F$ with inputs from the specified nodes $v_{i_1}, ..., v_{i_k}$ is assigned to each node v_i. The nodes $v_{i_1}, ..., v_{i_k}$ correspond to the genes that influence the expression of the gene associated with node v_i, and f_i represent the exact functional dependence of this influence. For example in Figure 1 the gene v_2 is influenced by genes v_1 and v_3 and the functional dependence is $v_2 = v_1\ AND\ v_3$. This functional dependence basically says that v_2 is expresses if and only if both v_1 and v_3 are expressed.

For a subset $U \subseteq V$, an *expression pattern* ψ of U is a function from U to $\{0,1\}$. An expression pattern of V is called a *state* of a boolean Network. That is ψ represents the states of nodes (genes), where each node is assumed to take either 0 (not-expressed) or 1 (expressed) as its state value. Typically, when the usage is clear from the context, we omit ψ. For example, we write $v_i = 1$ for denoting $\psi(v_i) = 1$. In a Boolean network, the expression pattern ψ_{t+1} at time $t+1$ is determined by the Boolean functions F from the expression pattern ψ_t at time t (i.e. $\psi_{t+1}(v_i) = f_i(v_{i_1}, ..., v_{i_k})$).

For better visualization, it is convenient to consider the wiring diagram [2, 14] $G'(V',F')$ of a Boolean network $G(V,F)$ [See Figure 1]. For each node v_i in V, let $v_{i_1}, ..., v_{i_k}$ be input nodes to v_i in $G(V,F)$. For each such node v_i, we consider an additional node v'_i, and we construct an edge from v_{i_j} to v'_i for each $1 \leq j \leq k$. Let $G' = (V',F')$ the network with nodes $v_1, ..., v_n, v'_1, ..., v'_n$ constructed in this way. Then the expression pattern of the set $\{v'_1, ..., v'_n\}$ is determined by $v'_i = f_i(v_{i_1}, ..., v_{i_k})$. That is, the expression of pattern $\{v_1, ..., v_n\}$ corresponds to one at a time t and the expression pattern of $\{v'_1, ..., v'_n\}$ corresponds to one at

time $t+1$. Moreover, it is convenient to consider the expression pattern of $\{v_1, ..., v_n\}$ as the INPUT, and the expression pattern of $\{v'_1, ..., v'_n\}$ as the OUTPUT.



INPUT			OUTPUT		
v_1	v_2	v_3	v'_1	v'_2	v'_3
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

Figure 1. A Boolean Network [from left to right: the network, its wiring diagram and the functional dependency table] (Akutsu *et al.* 1999)

The wiring diagram basically shows how the expression levels of genes at time t (INPUT) influence the the expression levels at time $t+1$ (OUTPUT). The exact trasition diagram for every combination of INPUT values can thus be given by a table as the one in Figure 1.

3.2 Motivations for Generalizing the Boolean Network Model

Our proposed generalization of the Boolean network model to a Temporal Boolean Network model is motivated by the following considerations:

- Boolean networks described above are incapable of modeling the existence of latency periods (lasting more than one unit of time) between the expression of a gene and the observation of its effect. For example a gene (say g_4) whose inhibitory effect (say on gene g_5) depends on an inducer (say g_3) first has to bind with this inducer in order to be able to bind to the inhibition site on g_5. Therefore there can be a significant delay between the expression of the inhibitor gene g_4 and its observed effect i.e. the inhibition of the gene g_5.

- Not all variables that can influence the expression level of a gene are necessarily observable. For instance, assume that genes $g_1, g_2, \cdots g_{1000}$ be the genes under study. Suppose that the expression of genes g_1 at time t might turn on a gene g_u at time $t+1$. It is quite possible that g_u is not among the genes that are being monitored in the experiment, or even among the genes that are currently known. Suppose gene g_u being on at time $t+1$ results in a gene g_2 being turned on at time $t+2$. Since the expression of g_u cannot be observed, the Boolean network

model described above will be unable to implicate g_1 in the control of g_2.

Note that even if all the genes are monitored in an experiment the unknown factor denoted by g_u may stand for a non-genetic environmental factor. Also long temporal delays observed in the our new model might indicate the presence of hidden factors, thus providing hints for their discovery and therefore contributing to the improvement of the quality of the data collected in future experiments.

In what follows, we introduce a generalization of the Boolean network model to address dependencies among the activity of genes that span for more than one unit of time. The resulting model, we will call the $TBN(n,k,T)$ model, allows the expression of each gene at time $t+1$ to be controlled by a Boolean function of the expression levels of at most k genes at times in $\{t \cdots t-(T-1)\}$.

3.3 Temporal Boolean Networks

In the Temporal Boolean Networks (TBN) we will allow the state of a gene at time $t+1$ to depend on the state of other genes at times $t, t-1, ..., t-(T-1)$, instead of only t. The representation of such a network, by analogy with the boolean networks is given in Figure 2.

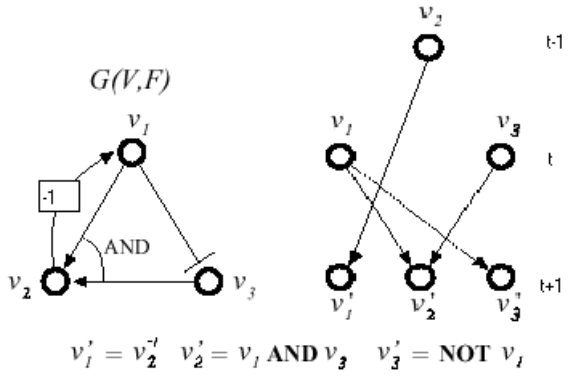


Figure 2. A Temporal Boolean Network

The only change from the Boolean Network model depicted in Figure 1 is that the expression level of gene v_2 at time $t+1$ depends on the value of the gene v_1 at time $t-1$ instead of time t. This is represented by a label of -1 label on the edge between v_1 and v_2. By default the edges are assumed to carry a value of 0, which means that the dependency they represent is from time t to time $t+1$. In general an edge labeled as $-k$ will represent a dependency between the values at

time $t-k$ and $t+1$. The wiring diagram is changed accordingly in order to be able to represent dependencies that represent the added functional $t-1$ dependencies. In general, each edge in the network can have a set of labels drawn from the set $\{0 \cdots T-1\}$. The functional dependency will be represented by a boolean table for each gene, but this time the inputs can be gene expression values at more than one time step in the past.

It is straightforward to extend the proposed model to allow for multiple discrete levels of expression yielding Temporal Discrete Networks $TDN(n,k,T,D)$ wherein each gene can be expressed at levels $0,1,\cdots D-1$. Furthermore it is also possible to allow for continuous expression levels as well yielding $TCN(n,k,T,R)$ wherein the expression level of a gene g_i at time t is a real number in the interval $[-R,R]$.

4. Theoretical Results

Identification of a member of the $BN(n,k)$ family from noiseless as well as noisy data have been explored by Akutsu *et al.* [3, 13]. They assume that the genes in the (unknown) network to be inferred can be set to any arbitrary pattern of 0s and 1s and the resulting state at the next time step can be observed. Further, it is assumed that such "input-output" observations can be performed by uniformly randomly sampling the inputs. Under similar assumptions, we can state the corresponding results for $TBN(n,k,T)$ networks.

Theorem 1. *$O(2^{2k} \cdot (2k+\alpha) \cdot log\ nT)$ uniformly randomly sampled input patterns and the corresponding outputs of an (unknown) TBN $\tau \in TBN(n,k,T)$ are sufficient to guarantee exact inference of τ with probability at least $1-n^{-\alpha}$, where $\alpha > 1$ is any fixed constant.*

The proof of this theorem is a straightforward adaptation of similar results given in Akutsu *et al.* 1999 [3] in the case of Boolean networks. [See Appendix A for additional details].

Note that although the sample and time complexities are exponential in k (the degree of interaction) they are only logarithmic in nT (the product of the number of genes and the maximum time span of temporal dependence T). Since typically $k << nT$, such exact inference is feasible for small values of k. However, in general, it might be necessary to sacrifice exactness of inference in exchange for computational efficiency.

It is worth noting that these theoretical guarantees (like their counterparts given by Akutsu *et al.* [3, 13]) rely on the assumption that the input patterns constitute a uniformly random sample and that the output corresponding to each such input can be observed. Since many experiments involve obtaining gene expression data by monitoring the expression of genes involved in some biological process (e.g., cell or neural development) over a period of time, the resulting data is in the

form of a *time series*. Since the sequence of states in a time series are strongly temporally correlated, the assumption of uniform random sampling is no longer valid. Each such time series provides a trajectory through the state space of a genetic network. Each trajectory has a transient part which provides useful information about the underlying network and a steady state part that corresponds to an attractor of the network. Derivation of suitable bounds on the number of time series samples (as opposed to uniformly random samples) by identifying the necessary and sufficient constraints on the temporal structure of the time series data relative to the structure of the underlying network in this case is a topic of current research.

In a similar manner as in Akutsu [3] we can develop an information theoretic lower bound on the number of transitions needed to identify TBN(n,k,T) given by the following theorem:

Theorem 2. *At least* $\Omega(2^k + k \log nT)$. *INPUT/OUTPUT pairs are required in the worst case to identify a Temporal Boolean Network TBN(n,k,T).* [See Appendix A for details]. In what follows, we describe an approach to inference of Temporal Boolean Networks from expression data in the form of a time series.

5. Inference of Temporal Boolean Networks from Time Series Data

Temporal Boolean Networks model functional dependencies among genes using Boolean functions. Boolean functions have several alternative representations. The simplest (and the most explicit) representation is a truth table shown in Fig. 1. However, functions that correspond to descriptions of natural phenomena typically lend themselves to more compact representations in the Conjunctive Normal Form (CNF), Disjunctive Normal Form (DNF), or in the form of decision trees, decision lists, etc. We have chosen to represent the Boolean functions used to describe Temporal Boolean Networks in the form of Decision Trees.

For example, consider gene g_1 that is induced by a complex formed from the products of two genes g_2 and g_3 and these are the only gene that influence g_1. A decision tree that describes this dependence is given in Figure 3. Similarly, for each gene we can construct such a decision tree. The genes that control gene g_i's expression appear as nodes in its corresponding decision tree.

Since every k-input boolean function can be represented by a decision tree of depth at most k, we can ensures that the resulting representation is expressive enough to describe any member of the Temporal Boolean Network family $TBN(n, k, T)$. We use n decision trees, one for each gene, to collectively describe a Temporal Boolean Network model of a genetic network. The output of each decision tree represents the expression level (0 or 1) of the corresponding gene based on at most k expression levels over a time window of length at most T.

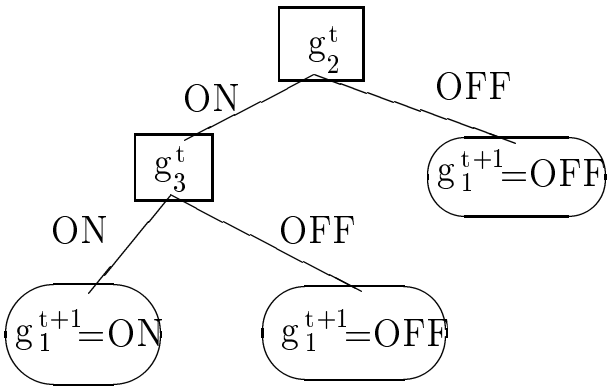


Figure 3. A decision tree that represents the fact that if genes g_2 and g_3 are expressed (turned ON) leads to the expression of gene g_1 (turned ON).

Several algorithms for inferring decision trees from data (in the form of sample input-output pairs) are available in the machine learning literature. The ID3[5] algorithm and its variants are based on a greedy search through the space of decision trees in order to identify a compact decision tree that adequately models the observed data and (under some reasonable assumptions) has high predictive accuracy on unobserved data. This search for a compact tree, guided by the entropy reduction (or information gain) criterion corresponds to a greedy version of the approach used in REVEAL [14]. Greedy search makes this approach computationally tractable for large genetic networks. A large body of empirical results in the machine learning literature suggest that the decision trees inferred by greedy search compare favorably with those inferred using exhaustive search in terms of predictive accuracy. Hence, we used a greedy search guided by information gain, over the space of depth k decision trees.

The training data for a decision tree modeling the functional dependency of the level the expression of gene g_i consists of observed input output pairs where each input is a nT bit boolean vector encoding the activities of each of the n genes at times $t, t-1, t-2 \cdots t-(T-1)$ and the corresponding output is the observed expression level of g_i at time $t+1$. For a given gene g_i, $m-T$ input–output samples (or training examples) are obtained by sliding a window of length T (where $T < m$) over the rows of a gene expression time series $\mathcal{E}$ (See section 2). Thus, an $m \times n$ gene expression matrix yields $m-T$ training samples for each of the n decision trees. Examples obtained from multiple time series are used for inferring a genetic network.

6. Experimental Setting and Results

The experiments described in this section were designed to explore the performance of the proposed approach to genetic network inference on randomly generated temporal boolean networks. In generating a network, we assume that the probability that the expression level of a gene at time $t+1$ depends on the expression levels of genes at time $t-\delta$ is proportional to ζ^{δ} $\forall t$ such that $0 \leq t \leq T-1$ for some choice of ζ where $0 < \zeta \leq 1$. For each network, we generated 20 time series of length 100, by setting the expression levels over a window of length T to random values and recording network's outputs over 100 time steps. Thus, each time series resulted in an $n \times (100+T)$ boolean matrix $\mathcal{E}$. The 20 time series collectively provided 100×20 training examples for each of the n genes. Each time point contains the expression levels for all genes at that moment of time. A decision tree was inferred for each gene.

Then we evaluated the results in terms of the *sensitivity*, *specificity* and *accuracy* of the inferred decision tree DT_i with respect to the corresponding temporal boolean network TBN_i for each gene g_i.

We denote the fact that the decision tree DT_i captures the dependence of the expression level of gene g_i at time $(t+1)$ on the expression level of the gene gene g_j at time $(t-\tau)$ where $\tau \in \{0, \cdots, (T-1)\}$, by writing $(\tau, j) \in DT_i$. Similarly, we denote the fact that the expression level of gene g_i at time $(t+1)$ depends on the expression level of gene g_j at time $(t-\tau)$ if gene g_i were controlled by the temporal boolean network TBN_i by writing $(\tau, j) \in TBN_i$. Let

$$DEP_{DT_i} = \{(\tau, j) | (\tau, j) \in DT_i\}$$

Let

$$DEP_{TBN_i} = \{(\tau, j) | (\tau, j) \in TBN_i\}$$

Then the *sensitivity* of the decision tree DT_i for gene g_i is defined as follows:

$$sensitivity(i) = \frac{|DEP_{TBN_i} \cap DEP_{DT_i}|}{|DEP_{TBN_i}|}$$

The *sensitivity* of the inferred decision tree measures the degree, to which this tree succeeds in capturing the dependency of gene g_i on other genes, with respect to the true Temporal Boolean Network.

The *sensitivity* of the inferred set of decision trees $DT = \{DT_i | i \in \{1, \cdots, n\}\}$ relative to the corresponding TBN is given by:

$$sensitivity(DT, TBN) = \frac{1}{n} \sum_{i=1}^{n} sensitivity(i)$$

The *specificity* of the decision tree inferred for gene g_i is defined as follows:

$$specificity(i) = \frac{|DEP_{TBN_i} \cap DEP_{DT_i}|}{|DEP_{DT_i}|}$$

Specificity of the inferred decision tree measures the degree to which it misleads us regarding the dependency of gene g_i on the various genes in the genetic network.

The specificity of the inferred set of decision trees $DT = \{DT_i | i \in \{1, \cdots, n\}\}$ relative to the corresponding TBN is given by:

$$specificity(DT, TBN) = \frac{1}{n} \sum_{i=1}^{n} specificity(i)$$

The *accuracy* of the decision tree inferred for gene g_i is defined to reflect the degree to which it correctly predicts the expression level of gene g_i (as estimated from a set of gene expression time series data).

Let Λ be a set of gene expression time series used to evaluate the accuracy of a decision tree inferred for gene g_i.

Then we will denote by $accuracy\ (i, \lambda)$ the accuracy of the tree for gene g_i on a time series $\lambda \in \Lambda$. $accuracy(i, \lambda)$ represents the fraction of expression levels of gene g_i from the time series λ that are correctly predicted by the inferred decision tree DT_i, relative to the total size of λ. Thus, if λ is a time series of length 100, and at 80 of the 100 time points, the expression level of gene g_i predicted by DT_i agrees with the corresponding values observed in λ, $accuracy(i, \lambda) = 0.8$. The *accuracy* of DT_i is estimated as follows:

$$accuracy(i) = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} accuracy(i, \lambda)$$

The accuracy of an inferred decision tree DT_i was estimated using an independently generated test set of 20 time series each of length 100 generated from TBN_i. The estimated accuracy of the inferred set of decision trees $DT = \{DT_i | i \in \{1, \cdots, n\}\}$ relative to the corresponding TBN is given by:

$$accuracy(DT, TBN) = \frac{1}{n} \sum_{i=1}^{n} accuracy(i)$$

Each experiment consisted of 10 independent runs of this procedure (each with a new randomly generated network) and the results presented show averages over these runs.

The first experiment presented in Figure 4a) is for a boolean network with $n = 16$ genes, with $T = 3$ and k varying from 2 to 10. Similar results were obtained for networks with 32 genes. The experiment shows that the sensitivity increases as the degree of interaction k increases.

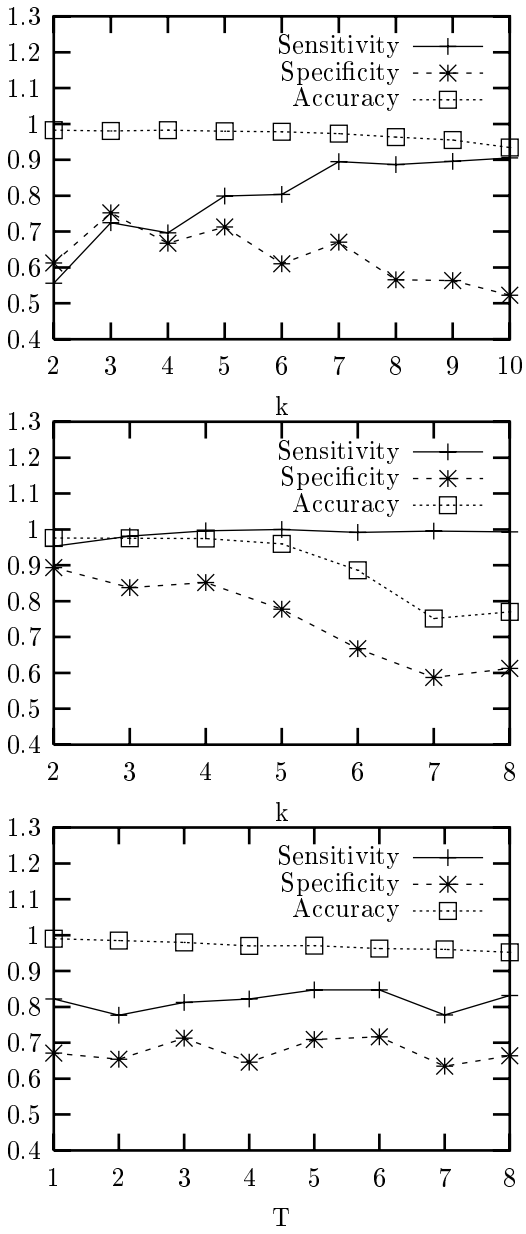


Figure 4. a)boolean net n=16, k=2-10, b)4-ary net n=16, k=2-8, c) boolean net n=16, k=6, T=1-8

The second experiment presented in Figure 4b) is for a 4-ary network with $n = 16$ genes, $T = 3$ and k varying from 2 to 10. This corresponds to a Temporal Discrete Network TDN (n=16, k=2-10, T=3, D=4). The goal of this experiment is to observe what is the effect of considering four levels of expression instead of only two. In this case we observe sensitivity and specificity increasing and accuracy decreasing with increase in k. These experiments show that as we increase the complexity of the model (in this case the number of expression levels) the sensitivity, which represents the probability that we will identify genes that represent the true dependency, increases (compare the Sensitivities in Figure 4a) and Figure 4b)). Further experiments revealed that the reduction in accuracy can be explained by the necessity for additional data as k increases, in the conditions of a 4-ary network were for each gene we have to disambiguate among 4^{4^k} 4-ary functions versus 2^{2^k} in the 2-ary (Boolean) case (once the dependencies are known qualitatively).

The third experiment presented in Figure 4c) is for a boolean network with $n = 16$ genes, $k = 6$ and T varying from 1 to 8. The goal of this experiment is to determine whether the variation of the time-span T of the dependency has any effect on the performance of the inference algorithm and it shows that this remains relatively constant as we change T.

7. Summary and Discussion

This paper introduced the Temporal Boolean Networks (and Temporal Discrete Networks) which generalize the Boolean Network model in order to cope with dependencies that span over more than one unit of time. Some bounds on the size of data needed to infer Temporal Boolean Networks from time series data under uniformly random sampling assumptions are stated. We also showed how the problem of Temporal Boolean Network inference can be translated into a problem of inferring a set of decision trees. We demonstrated, through a series of experiments using artificially generated networks, the effectiveness of a simple and fast decision tree inference algorithm for inferring Temporal Boolean Networks and Temporal Discrete Networks from time series data.

The main hindrance against applying the TBN inference algorithm on real data is the lack of sufficiently large data sets. This hindrance is likely to become less serious as additional data are gathered.

The approaches to genetic network inference from gene expression data rely on the assumption that only the expression of a gene is likely to be controlled by a relatively small number (say k) of genes. Biologically meaningful value of k is currently unknown, but is believed to be much smaller than the total number of genes n [1]. It is clear

from the sample complexity results presented in this paper that purely data-driven approaches to inference of temporal boolean networks will be computationally infeasible unless $k << n$.

Work in progress is aimed at evaluating the effectiveness of the described approach for inferring genetic networks from biological gene expression time series data.

Some directions for future work include investigation of variations of the model to accommodate probabilistic Boolean and Discrete valued functional dependencies, and continuous valued expression levels as well as alternative (e.g., event-based and interval-based) representations of time. Investigation of techniques for incorporating prior knowledge (in the form of known biological constraints) into the inference algorithm represents another direction for futire research. Also of interest are active learning approaches wherein the learning algorithm helps identify promising experiments as opposed to the purely data-driven passive learning approach examined in this paper.

Acknowledgments

This research was supported in part by grants from the National Science Foundation (awards 9982341 and 9972653), the Carver Foundation, and Pioneer Hi-Bred Inc. This research has benefited from interactions with Dr. Drena Dobbs, Dr. Pat Schnable, Dr. Xun Gu, Dr. Gavin Naylor, and Dr. Steve Rodermel of the Iowa State University Bioinformatics and Computational Biology Program.

Appendix

A. Theoretical Results

Theorem 1. *$O(2^{2k} \cdot (2k+\alpha) \cdot log\ nT)$ uniformly randomly sampled input patterns and the corresponding outputs of an (unknown) TBN $\tau \in TBN(n,k,T)$ are sufficient to guarantee exact inference of τ with probability at least $1 - n^{-\alpha}$, where $\alpha > 1$ is any fixed constant.*

Proof Sketch: The proof of this theorem is an adaptation of the corresponding theorem proved in [3] in the case of Boolean Networks. The proof in [3] is based on a brute-force algorithm for identifying a boolean function from the set of all boolean functions with less than k inputs chosen from a set of n possibilities (i.e., roughly $2^{2^k} \cdot n^k$ functions). A precise characterisation can be given to a minimal set of INPUT/OUTPUT pairs that will allow the algorithm to do the identification of a Boolean Network exactly. Then the bound stated by the theorem (i.e., $O(2^{2k} \cdot (2k+\alpha) \cdot log\ nT)$) is derived as the number of uniformly random sampled INPUT/OUTPUT pairs needed, in order

to be sure that our sample will contain the minimal set with probability at least $1 - n^{-\alpha}$.

The only difference in the case of temporal boolean networks is that we have to do the identification from the set of all boolean functions with less than k inputs chosen from a set of nT possibilities (i.e., roughly $2^{2^k} \cdot (nT)^k$ functions). Aside from this the proof follows along the same lines as the one in [3].

Theorem 2. *At least $\Omega(2^k + k \log nT)$ INPUT/OUTPUT pairs are required in the worst case to identify a Temporal Boolean Network TBN(n,k,T).*

Proof: The proof for this theorem is a typical information theoretic argument.

The general problem in the information theoretic setting is the following: "We want to identify an element e, from a set S of finite cardinality $|S|$, by asking D-ary questions (i.e. that have D possible answers) about where in the set S the element e to be found. And the problem is to find what is the minimum number of D-ary questions that we need in the worst case in order to identify the element e." The solution to this problem is as follows: "We need at least $\Omega(\log_D |S|)$ questions in order to identify the element e. This is because a D-ary question splits the set S into D subsets and at least one of these sets has a size of at least $|S|/D$. If we iterate this procedure for l succesive questions then we get that the size of at least one of the subsets has to be at least $|S|/D^l$. By setting this size equal to 1 (in order to obtain a 100% identification of e) and solving the equation we obtain a minimum of $\Omega(\log_D |S|)$ questions needed for the identification of e".

Returning to our problem, if we are to recast it in information theoretic terms, the element that we want to identify is a TBN from the set of all possible TBNs of the type (n, k, T) (i.e., having n genes, dependency of at most k, and having a dependency timespan of at most T). The questions in our case are the INPUTs (which are $n \times T$ matrices that represent the levels of expression for all n genes during the previous T timesteps) and the answers to these questions are represented by the OUTPUTs (which are the levels of expression for all of the n genes at the current moment of time). Therefore in our case an INPUT/OUTPUT is equivalent to an answer to a 2^n-ary question (because there are 2^n possible OUTPUT patterns, hence 2^n possible answers). The number of TBNs of the type (n, k, T) is given by all the possible combinations of n(one for each gene) boolean functions with k inputs chosen from $n \cdot T$ possibilities. Since there are 2^{2^k} possible Boolean functions with k inputs and $\Omega((nT)^k)$ ways to chose those k inputs from $n \cdot T$ possibilities, it follows that there are $\Omega((2^{2^k} \cdot (n \cdot T)^k)^n)$ possible TBNs of type (n,k,T). Now applying the information theoretic argument it follows that we need $\Omega(\log_{2^n}(2^{2^k} \cdot (n \cdot T)^k)^n)$ questions. But

this is the same as $\Omega(2^k + k \log_2 nT)$ (because $\log_{a^n} b^n = \log_a b$). Which completes the proof.

Note: The previous theorem can be generalized also for the case when we have D-ary Temporal Networks by replacing 2 in all the places with D (including the logarithm).

References

[1] Kauffmann, S.A., *The Origins of Order, Self-Organization and Selection in Evolution*. Oxford University Press, (1993).

[2] Somogyi R., Sniegoski C.A., "Modeling the complexity of genetic Networks: Understanding Multigenic and Pleitropic Regulation". *Complexity* **1**(6):45-63, (1996).

[3] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of Genetic Networks from a Small Number of Gene Expression Patterns Under the Boolean Network Model", *Pacific Symposium on Biocomputing*, 4:17-28, (1999).

[4] Kargupta, H. "A striking property of genetic code-like transformations", Technical Report EECS-99-004, Department of Electrical Engineering and Computer Science, Washington State University, (1999).

[5] Quinlan J.R., "Rule induction with statistical data - a comparison with multiple regression," *Journal of the Operational Research Society*, **38**, 347-352, (1987).

[6] J. L. DeRisi, V. R. Iyer, P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale", *Science*, **278**(5338), 680-6, (1997).

[7] P. T. Spellman, et al., "Comprehensive Identification of Cell-Cycle regulated Genes of the Yeast Saccharomyces Cervisiae by MicroArray Hybridization", *Molecular Biology of the Cell*, **9**, 3273-97, (1998).

[8] X. Wen, et al., "Large-scale temporal gene expression mapping of central nervous system development", *Proceedings of the National Academy of Science*,

95, 334-9, (1998).

[9] S. P. Gygi, Y. Rochon, B. R. Franza, R. Aebersold, "Correlation between protein and mRNA abundance in yeast", *Mollecular Biology of the Cell*, **19**, 1720-30, (1999).

[10] D. Thieffry and R. Thomas; "Qualitative Analysis of Gene Networks", Pacific Symposium on Biocomputing, 3:77-88, (1998).

[11] M.A. Savageau; "Rules for the Evolution of Gene Circuitry" Pacific Symposium on Biocomputing, 3:54-65, (1998).

[12] McAdams H.H., Shapiro L., "Circuit Simulation of Genetic Networks", *Science*, **269**, 4 August : pp. 650-656, (1995).

[13] T. Akutsu, S. Miyano, and S.Kuhara; "Algorithms for Inferring Qualitative Models of Biological Networks", *Pacific Symposium on Biocomputing*, 5:290-301, (2000).

[14] S. Liang, S. Fuhrman and R. Somogyi; "REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures", *Pacific Symposium on Biocomputing*, 3:18-29, (1998).

[15] T.E. Ideker, V. Thorsson, and R.M. Karp; "Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design", *Pacific Symposium on Biocomputing*, 5:302-313, (2000).

[16] T. Chen, H. L. He, and G.M. Church; "Modeling Gene Expression with Differential Equations", *Pacific Symposium on Biocomputing*, 4:29-40, (1999).

[17] Goss P.J.E., Peccoud J., "Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri Nets", *PNAS*, 95, 6750-6755, (1998).

[18] H. Matsuno, A. Doi, M. Nagasaki, and S. Miyano; "Hybrid Petri Net Representation of Gene Regulatory Network", *Pacific Symposium on Biocomputing*, 5:338-349, (2000).

[19] D.C. Weaver, C.T. Workman, G.D. Stormo; "Modeling Regulatory Networks with Weight Matrices", *Pacific Symposium on Biocomputing*, 4:112-123, (1999).

[20] R. Somogyi, H. Kitano, S. Miyano, and Q. Zheng, "Molecular Network Modelling and Data Analysis", Session Introduction - *Pacific Symposium on Biocomputing*, 5:288-289, (2000).

[21] Stormo, G., "Identification of Coordinated Gene Expression and Regulatory Sequences:, Session Introduction - *Pacific Symposium on Biocomputing*, 5:413-414, (2000).

[22] Eisen, M., Spellman, P., Brown, P., and Botstein, D., "Cluster analysis and display of genome-wide expression patterns", *Proceedings of the National Academy of Science*, **95**:14863-14868, (1998).

[23] Ben-Dor, A., Yakhini, Z., "Clustering Gene Expression Patterns", Journal of Computational Biology, **6**:281-297, (1999).

[24] G.S. Michaels, D.B. Carr, M. Askenazi, S. Fuhrman, X. Wen and R. Somogyi, "Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data", Pacific Symposium on Biocomputing, 3:42-53, (1998).

[25] Murphy, K. and Mian, S., "Modelling gene expression data using dynamic Bayesian networks", Technical report, Computer Science Division, University of California, Berkeley, CA, (1999).