Available online at www.sciencedirect.com
SCIENCE DIRECT
BioSystems 84 (2006) 115–123
Bio Systems
www.elsevier.com/locate/biosystems

Dynamical properties of model gene networks and implications for the inverse problem

Theodore J. Perkins [a,b,*], Mike Hallett [a], Leon Glass [b]

[a] *McGill Centre for Bioinformatics, 3775 University Street, Montreal, Que., Canada H3A 2B4*
[b] *Department of Physiology, McGill University, 3655 Prom. Sir William Osler, Montreal, Que., Canada H3G 1Y6*

Received 19 April 2005; received in revised form 24 August 2005; accepted 23 September 2005

Abstract

We study the inverse problem, or the "reverse-engineering" problem, for two abstract models of gene expression dynamics, discrete-time Boolean networks and continuous-time switching networks. Formally, the inverse problem is similar for both types of networks. For each gene, its regulators and its Boolean dynamics function must be identified. However, differences in the dynamical properties of these two types of networks affect the amount of data that is necessary for solving the inverse problem. We derive estimates for the average amounts of time series data required to solve the inverse problem for randomly generated Boolean and continuous-time switching networks. We also derive a lower bound on the amount of data needed that holds for both types of networks. We find that the amount of data required is logarithmic in the number of genes for Boolean networks, matching the general lower bound and previous theory, but are superlinear in the number of genes for continuous-time switching networks. We also find that the amount of data needed scales as 2^K, where K is the number of regulators per gene, rather than 2^{2K}, as previous theory suggests.

Keywords: Boolean network; Genetic network; Inverse problem; Dynamics; Sample complexity; Computational complexity

1. Introduction

Abstract models of genetic networks, such as Boolean networks and continuous-time switching networks, have been proposed as conceptual models for helping us understand the behavior of real genetic networks (Glass, 1975; Kauffman, 1969, 1993). These formalisms, or generalizations of them, have been used to model such systems as the sporulation network in *Bacillus subtilis* (de Jong et al., 2004a), the gap gene network of *Drosophila melanogaster* (Sanchez and Thieffry, 2001), and the segment polarity network in the same organism (Albert and Othmer, 2003). Although these models omit many of the details of the real chemical interactions, they serve as useful syntheses of the often-distributed biological knowledge about these networks, they allow testing of hypotheses about network behavior, and they sometimes lead to new biological hypotheses.

General properties of model networks, particularly randomly generated networks, have been studied in an effort to understand basic principles behind the functioning of real networks. For example, Kauffman has equated different cell types in real organisms with different fixed points or cycles in the dynamics of model networks (Kauffman, 1969, 1993). He and others have studied how the number and period of attractors in random net-

* Corresponding author. Tel.: +1 514 3987071x09317; fax: +1 514 3983387.
E-mail addresses: perkins@mcb.mcgill.ca (T.J. Perkins), hallett@mcb.mcgill.ca (M. Hallett), glass@cnd.mcgill.ca (L. Glass).

0303-2647/$ – see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2005.09.010

works depends on network size, topology, and how the dynamics functions are chosen (Bagley and Glass, 1996; Bastolla and Parisi, 1997; Bilke and Sjunnesson, 2001; Glass and Hill, 1998; Kauffman, 1969, 1993; Kauffman et al., 2003; Raeymaekers, 2002; Samuelsson and Troein, 2003; Shmulevich and Kauffman, 2004; Socolar and Kauffman, 2003).

The increasing availability of quantitative gene expression data has kindled the hope of automatically inferring regulatory relationships in real gene networks. There have been some successes (e.g. Jaeger et al., 2004a,b; Reinitz and Sharp, 1995), but there is not yet a standard methodology for doing so. Much remains to be understood about the problem. What is the computational complexity of the problem? What algorithms work best? How much data is needed? How should data be collected? Are there fundamental limits on what can be inferred from expression data alone?

Analyses of the inverse problem for Boolean and continuous-time switching networks have begun to provide theoretical answers to these questions. Liang et al. (1998) were the first to propose a solution to the inverse problem for Boolean networks. Later, Akutsu et al. (1999) and Ideker et al. (2000) described alternative solutions. Perkins et al. (2004) described solutions to the inverse problem for continuous-time switching networks. The approaches proposed for Boolean networks can also be applied to continuous-time switching networks, though the methods of Liang et al. (1998) and Akutsu et al. (1999) in particular require significantly more computation than the method described in Perkins et al. (2004).

We focus on the sample complexity of the inverse problem—that is, how much data is needed to identify the network? In particular, we study the problem for Boolean and continuous-time switching networks of N genes in which each gene has precisely K regulators. Akutsu et al. (1999) studied this problem for Boolean networks under the assumption that the data comprises uniformly randomly sampled states of the network. They proved that the amount of data needed scales as $\log N$ and as 2^{2K}. The 2^{2K} term is disheartening, because it suggests that it will take enormous amounts of data to identify densely connected networks. However, the $\log N$ dependence is encouraging because it suggests that network size per se is not a very important factor.

We consider solving the inverse problem based on time series data, although, as we argue in Section 4, time series data from randomly generated Boolean networks behave in many respects as randomly sampled data. In Section 5, we show that, regardless of how the data is generated, solving the inverse problem for Boolean networks or continuous-time switching networks requires at least $\frac{1}{2}(2^K + K(\log_2(N-K) - \log_2 K))$ samples. We then derive new estimates for the expected amount of data required. It turns out that differences in the dynamical properties of these networks, examined in Section 4, have a significant impact. In Section 5, we estimate the expected sample complexity for Boolean networks as $O(K2^K \log N)$ and for continuous-time switching networks as $O(2^K N \log N)$. These estimates are supported by simulation experiments, reported in Section 6.

2. Boolean networks and continuous-time switching networks

Boolean networks, as introduced by Kauffman (1969, 1993), are a discrete-time model of gene expression dynamics. Each of N genes has a Boolean level of expression as a function of time, denoted by $X_i(t) \in \{0,1\}$, where $i \in \{1, 2, \ldots, N\}$ and $t \in \{0, 1, 2, \ldots\}$. Each gene i has K_i *regulators*, denoted $r_i^1, \ldots, r_i^{K_i}$. Each gene also has a *regulation function*, or dynamics function, $f_i : \{0,1\}^{K_i} \mapsto \{0,1\}$. The dynamics of gene i is given by

$$X_i(t+1) = f_i(X_{r_i^1}(t), \ldots, X_{r_i^{K_i}}(t)). \quad (1)$$

Our analysis and simulations focus on randomly generated Boolean networks in which each gene has the same number of regulators, K. The regulators of each gene are chosen uniformly at random, with autoregulation allowed. Usually the f_i are chosen uniformly randomly from all 2^{2^K} Boolean functions of K inputs. The random selection of a particular f_i can be implemented by constructing a truth table on K inputs and randomly assigning each of the 2^K rows of the table to an output value of 0 or 1 with equal probability. We use this notion of randomly choosing the f_i for most of our analyses. However, this method of choosing the f_i can be problematic when we study the inverse problem. It is possible that this procedure would choose, for example, an f_i that always outputs 1 regardless of the regulator states. In such a case, the "regulators" do not really regulate the target gene, and there is no way that a procedure for solving the inverse problem could detect these "regulators" from simulated expression data. For our simulations, we restrict attention to dynamics functions f_i, which truly depend on all inputs, in the sense that for any regulator r_i^j, there is an assignment of Boolean states to the other regulators such that changing the value of $X_{r_i^j}$ changes the value of f_i.

Continuous-time switching networks, introduced by Glass (1975), are a differential equation model of gene expression dynamics. Each gene i has a real-valued

expression level as a function of time, denoted $x_i(t) \in [0,1]$ for $t \in [0,\infty)$. Based on its real-valued expression, each gene is also associated with a Boolean "state"

$$X_i(t) = \begin{cases} 1 & \text{if } x_i(t) \geq \frac{1}{2} \\ 0 & \text{if } x_i(t) < \frac{1}{2} \end{cases}.$$

As in Boolean networks, each gene i has K_i regulators, $r_i^1, \ldots, r_i^{K_i}$ and a Boolean regulation function $f_i : \{0,1\}^{K_i} \mapsto \{0,1\}$. The dynamics of gene i is given by

$$\frac{dx_i(t)}{dt} = f_i\left(X_{r_i^1}(t), \ldots, X_{r_i^{K_i}}(t)\right) - x_i(t). \quad (2)$$

Because Eq. (2) has the form of a production-decay differential equation, $f_i(t)$ is sometimes called the production rate of gene i at time t, and f_i is referred to as the production rate function.

Fig. 1 shows an example of a continuous-time switching network time series. Whenever $f_i = 1$, x_i decays exponentially towards 1, and whenever $f_i = 0$, x_i decays exponentially towards 0. The "corners" in the curve for $x_i(t)$ correspond to times at which f_i changes. A change in f_i, of course, is due to a change in the Boolean state of one of the regulators of gene i. A time at which the Boolean state of any gene changes is called a *switching time*. Between switching times, all $X_i(t)$, and thus all $f_i(t)$, are constant.

Solutions to Eq. (2) are not always well defined because the (Boolean) production rates are a discontinuous function of the state of the system. In general, the method of Fillipov can be used to solve this problem (de Jong et al., 2004b). Such problems are uncommon, however, in randomly generated networks if autoregulation is ruled out. In our analyses based on randomly generated continuous-time switching networks, we will assume each gene has the same number of regulators, K, chosen uniformly at random from the other $N-1$ genes. Regulation functions are chosen as described above for Boolean networks.

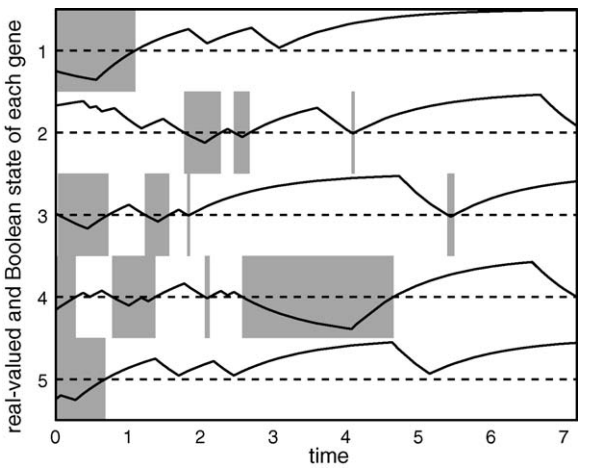

Fig. 1. Time series of genes one through five in a randomly generated continuous-time switching network of 20 genes with five regulators per gene. Each row corresponds to a different gene. The black curve gives $x_i(t)$ and the dotted line indicates $x_i = \frac{1}{2}$. Dark shading indicates periods of time when $X_i(t) = 0$.

3. The inverse problem

Given one or more time series generated by a Boolean network or a continuous-time switching network, the inverse problem is to identify the network generating the data. For simplicity of exposition we assume, in the Boolean network case, a single time series of length $T+1$. Thus, the data is a sequence, $\{X(0), X(1), \ldots, X(T)\}$, where $X(t)$ is a vector of the Boolean states of the genes at time t. We say this sequence comprises T samples of the network dynamics, because it includes T transitions from one network state to the next.

For continuous-time switching networks, we assume a single continuous-time time series, $x(t)$ for $t \in [0,T]$. For solving the inverse problem, all we need to know is the Boolean states and production rates of the genes as a function of time, $X(t)$ and $f(t)$. These can be deduced from $x(t)$. Furthermore, because switching times divide the time series into intervals of constant $X(t)$ and $f(t)$, it suffices to know the sequence of Boolean network states and production rates. Suppose there are Z intervals between switching times in the data $x(t)$. Then the data for the inverse problem is the sequence $\{(X(1), f(1)), (X(2), f(2)), \ldots, (X(Z), f(Z))\}$, where $X(z)$ and $f(z)$, respectively, denote the Boolean state of the network and the vector of production rates during the zth interval of time between switching times. We say this time series has Z samples of the network dynamics.

The inverse problem cannot be solved with certainty unless all 2^N possible Boolean network states appear in the data. For large N, this would be an unrealistic amount of data, hence we can give up on identifying the network with certainty. Instead, we seek the most parsimonious hypothesis that explains the data. Specifically, for each gene i we seek a minimal-size set of candidate regulators, $\hat{r}_i^1, \ldots, \hat{r}_i^{\hat{K}_i}$, and a candidate regulation function, $\hat{f}_i$, that are consistent with the data. For Boolean networks, a candidate solution is consistent if for all $i \in \{1, 2, \ldots, N\}$ and all $t \in \{0, 1, \ldots, T-1\}$,

$$X_i(t+1) = \hat{f}_i\left(X_{\hat{r}_i^1}(t), \ldots, X_{\hat{r}_i^{\hat{K}_i}}(t)\right).$$

For continuous-time switching networks, a candidate solution is consistent if for all $i \in \{1, 2, \ldots, N\}$ and all

$z \in \{1, 2, \ldots, Z\}$,

$$f_i(z) = \hat{f}_i\left(X_{\hat{r}_i^1}(z), \ldots, X_{\hat{r}_i^{\hat{K}_i}}(z)\right).$$

A time series generated by a Boolean or continuous-time switching network is *sufficient for solving the inverse problem* if, for every gene, (1) there is a unique minimal-size set of candidate regulators and a unique candidate regulator function that are consistent with the data, and (2) the candidate regulators and regulation function are correct. If all 2^N possible Boolean states appear in the time series, then it is sufficient for solving the inverse problem. But it is possible for much shorter time series to be sufficient. In the next section, we study statistical properties of time series generated by Boolean and continuous-time switching networks, and in the following section, we use these properties to estimate the expected length that a time series must be to be sufficient for solving the inverse problem.

4. Dynamical properties of Boolean networks and continuous-time switching networks

Consider a Boolean network of N genes and K regulators per gene, generated randomly as described in Section 2. A Boolean network is a deterministic dynamical model with a finite number of possible states. Thus, the asymptotic behavior of the network is to reach a fixed point of the dynamics or to reach a repeating cycle of states, where a cycle can be between 2 and 2^N states long. It has been observed that when $K = 1$or 2, a typical network rapidly reaches a fixed point or short-period cycle (Kauffman, 1969, 1993). However, when $K \geq 3$ and N is "not too small", typical time series are "complex" and cycles can be very long (Bastolla and Parisi, 1997; Kauffman, 1969, 1993; Raeymaekers, 2002). For example, Fig. 2(A) displays the first 100 steps of a time series from a randomly generated network of 20 genes, each having 10 regulators.

The simplest possible model of such a time series is to assume that the state of each gene is independently randomly zero or one on every time step, and this is the model we assume for the expected sample complexity analysis in Section 5. To justify this model, first consider a more general model in which we assume every gene changes state independently randomly with some probability p on each step. Does p depend on N and K, and if so, how?

For a gene to change state, one of its regulators must change state on the previous time step. The probability that at least one of the K regulators of the gene changes state is $1-(1-p)^K$. But even if a regulator changes state, f_i may not change. Assuming that f_i is chosen randomly from all 2^{2^K} Boolean functions of K inputs, a change in the state of the regulators has a $\frac{1}{2}$ chance of changing the output value of f_i. So the probability that a gene changes on every time step, p, should satisfy $p = \frac{1}{2}(1-(1-p)^k)$, or $(1-p)^K + 2p - 1 = 0$. For any K, $p = 0$ is a solution to this equation and corresponds to a network at a fixed point—no genes change state. For $K \geq 3$, there is another root, less than $\frac{1}{2}$ and approaching $\frac{1}{2}$ for increasing K. For $K = 5$, for example, the root is approximately 0.48121. This analysis suggests that genes in randomly generated networks should change state on each time step with a probability that is independent of N and that approaches $\frac{1}{2}$ with increasing K.

We tested this prediction with simulation studies using 50 randomly generated networks for each $N \in \{25, 50, 100, 200\}$ and $K \in \{3, 4, \ldots, 10\}$. We simulated each

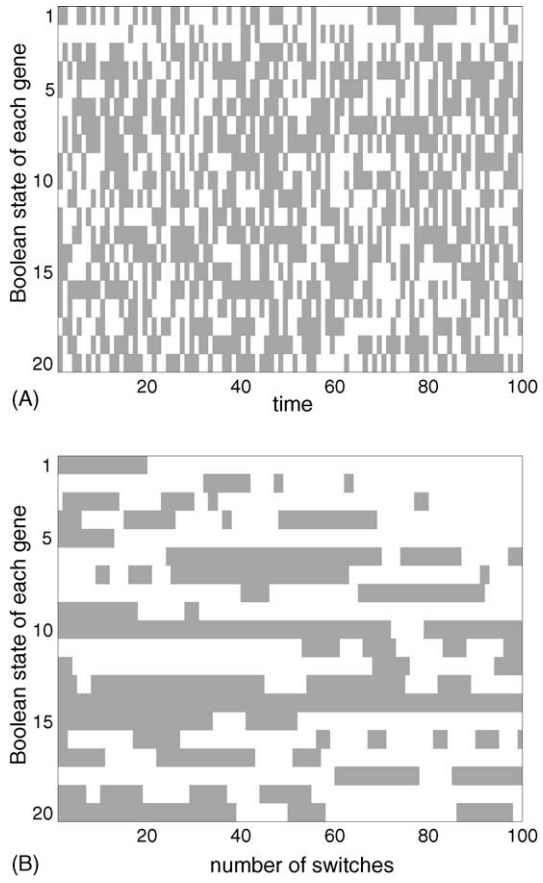

Fig. 2. (A) The first 100 steps of a time series from a randomly generated Boolean network with 20 genes and 10 regulators per gene. Each row corresponds to a gene, with a white box representing the high state and a shaded box representing the low state. (B) The sequence of Boolean gene states for a time series from a continuous-time switching network with the same regulators, regulation functions, and initial condition.

network from a random initial condition for 200 time steps and computed the mean switching probability over all genes during the second 100 steps. We discarded runs that ended at a fixed point, as these correspond to the root $p = 0$. We averaged the mean switching probabilities for the remaining runs. The results are plotted in Fig. 3. At $N = 25$ and $K = 3$, the predicted and empirical switching probabilities differ less than 0.03; at higher values of N and/or K, they match more closely. Contrary to the prediction, N does appear to have an effect on the switching probability, though its effect is smaller than that of K. According to predictions and simulations, as N and K grow large, the switching probability goes to $\frac{1}{2}$. That is, $X_i(t)$ and $X_i(t+1)$ can be considered independent random variables.

Continuous-time switching networks are also deterministic dynamical models, but they have infinite state spaces. Their dynamics can have fixed points and stable periodic orbits, but they can also have aperiodic orbits and can be chaotic (Glass and Hill, 1998; Mestl et al., 1997). Like Boolean networks, convergence to a fixed point or short-period attractor is common when N or K are "small", but when N and K are "not too small", complex long-period or aperiodic time series are typical.

Because switches in the Boolean states of genes occur at isolated real-valued times, we generally do not expect any two genes to switch at the exact same time. It is possible to construct networks and initial conditions for which this happens, but it is not likely for randomly generated networks. As a result, we expect that changes in the Boolean state of the network involve only a single gene and that the sequence of Boolean network states looks significantly different than the state sequence for a Boolean network. For example, Fig. 2(B) shows the first 100 Boolean network states in a time series generated by a continuous-time switching network. The regulators, regulation functions, and initial condition are the same as for the Boolean network whose time series appears in the figure. It is certainly not the case that $X_i(z)$ and $X_i(z+1)$ can be considered independent random variables. Genes typically keep the same Boolean state through many switching times.

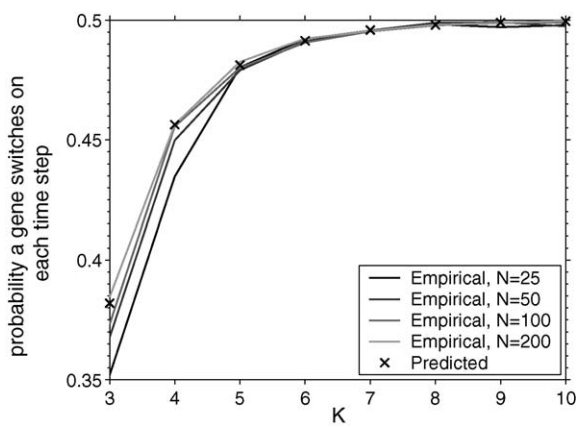

Fig. 3. Predicted and empirical probability that a gene changes state on each time step.

In the next section, the model we assume for continuous-time switching networks is that the identity of the zth gene to switch is independently random and uniform over all genes for all z. In other words, at each switching time, a single gene changes Boolean state and each gene has chance $\frac{1}{N}$ of being the gene to change. This assumption has been examined in detail by Mestl et al. (1997), who found some deviations between theoretical predictions based on the assumption and the results of simulation studies. Nevertheless, it is a simple and convenient assumption, which we make for the sake of our sample complexity analysis.

5. The amount of data needed to solve the inverse problem

How much data is needed to solve the inverse problem for a Boolean or continuous-time switching network of N genes, each having K regulators? A simple lower bound can be derived based on the number of possible networks. There are

$$\binom{N}{K}$$

possible sets of regulators for each gene in a Boolean network, and

$$\binom{N-1}{K}$$

for a continuous-time switching network. In either case, this is more than $\left(\frac{N-K}{K}\right)^K$. There are also 2^{2^K} possible dynamics functions for each gene. Thus, the total number of networks on N genes with K regulators per gene is at least

$$\left(2^{2^K}\left(\frac{N-K}{K}\right)^K\right)^N.$$

Each sample of a Boolean or continuous-time switching network's dynamics contains $2N$ bits—$X(t)$ and $X(t+1)$ for a Boolean network, and $X(z)$ and $f(z)$ for a continuous-time switching network. To solve the inverse problem, the data must contain at least enough bits to distinguish among all possible networks of the same N and K. Thus,

the minimal number of samples needed to solve the inverse problem is

$$\frac{\log_2\left(2^{2^K}\left(\frac{N-K}{K}\right)^K\right)^N}{2N} = \frac{1}{2}(2^K + K(\log_2(N-K) - \log_2 K)). \quad (3)$$

Next, we derive estimates for the expected amount of data needed. Our analyses rely repeatedly on the following Lemma, which is proved in Appendix A.

Lemma 1. *If there are A different events, each of which occurs with probability at least p in any block of τ time steps, then the expected number of time steps until all A events have occurred is* $O\left(\frac{\tau}{p}\log A\right)$.

Consider Boolean networks first. Following Section 4, we assume that the data is a single time series $\{X(0), X(1), \ldots, X(T)\}$ which can be treated as if each $X(t)$ is an independent random Boolean vector. Two things must happen for the network to be identified. First, the correct regulators for each gene must be identified. Second, the regulation functions must be identified. To identify the correct regulators, all incorrect regulator sets must be ruled out. There are

$$\binom{N}{K} - 1$$

incorrect regulator sets of size K for each gene, or no more than N^{K+1} incorrect regulator sets total. One way that an incorrect regulator set can be ruled out is if, from one time step to the next, none of the regulators change Boolean state but, on the following time step, the target gene does change Boolean state. The probability of the former on any time step is, by assumption, 2^{-K}, and the probability of the latter is 2^{-1}. Thus, there is probability at least $2^{-(K+1)}$ on any time step of ruling out an incorrect regulator set. By Lemma 1, the expected number of samples until all incorrect regulator sets is ruled out is $O(2^{K+1}\log N^{K+1})$ or $O(K2^K \log N)$.

To identify the regulation functions, every combination of regulator states for every gene must occur in the time series, along with the next network state to which it leads. There are 2^K combinations for each of N genes, and each occurs with probability 2^{-K} on each time step. Thus, the total number of samples until all have appeared is $O(2^K \log N2^K) = O(2^K(K + \log N))$. This is of lower order than the number of samples needed to identify the correct regulators, so the expected number of samples needed for Boolean network identification is

$$O(K2^K \log N). \quad (4)$$

Next, we consider continuous-time switching networks. Following Section 4, we assume that the data is a single time series $\{(X(1), f(1)), (X(2), f(2)), \ldots, (X(Z), f(Z))\}$ which can be treated as if each $X(z+1)$ differs from $X(z)$ in one uniformly randomly chosen position. Again, to solve the inverse problem, the regulators and regulation functions of every gene must be determined. There are NK regulatory relationships to be uncovered. Suppose gene j regulates gene i. One way this can be determined is if, for some z, $X_j(z) \neq X_j(z+1)$ and $f_i(z) \neq f_i(z+1)$; the observed change in i's production rate must be attributed to a change in the Boolean state of some regulator, and j is the only gene whose Boolean state changes. Gene j has chance $\frac{1}{N}$ of being the one that changes Boolean state at any particular switching time. Assuming f_i is chosen randomly, there is a $\frac{1}{2}$ chance that a change in the state of a regulator changes the output value of f_i. Thus, there is chance $\frac{1}{2N}$ that a sample reveals that j regulates i. By Lemma 1, the expected number of samples until all regulators are identified is $O(2N\log(NK)) = O(N\log N)$.

To identify the regulation functions, all 2^K Boolean regulator state combinations must occur for all N genes. Consider a particular gene i and its K regulators. For each successive sample in the data, there is chance $\frac{K}{N}$ that precisely one of the regulators changes Boolean state, and otherwise none of them change. The sequence of changes in the Boolean state of i's regulators can be viewed as a random walk on a K-dimensional hypercube. For an ordinary walk on the hypercube, the expected number of steps for all states to be visited is no more than $cK2^K$ for some constant c and all K. The present case is slightly different, because there is only probability $\frac{K}{N}$ of moving to an adjacent vertex of the hypercube on each step. But that just means there is expected time $\frac{N}{K}$ between steps on the hypercube, so the expected number of steps for all states to be visited is no more than $\frac{N}{K}cK2^K = cN2^K$ (Chandra et al., 1997). By the Markov inequality, there is probability at least $\frac{1}{2}$ that all states are visited in any period of $2cN2^K$ time. By Lemma 1, the expected number of samples until all combinations of regulators states appear for every gene is thus $O(4cN2^K \log N) = O(2^K N \log N)$. The expected number of samples needed for continuous-time switching network identification is thus

$$O(2^K N \log N). \quad (5)$$

6. Simulation experiments

We performed simulation experiments to test the bound and estimates of the previous section. In the

first experiment, we tested the sample complexity of the inverse problem for Boolean and continuous-time switching networks with $K = 5$ regulators per gene and number of genes $N \in \{10, 15, 20, \ldots, 50\}$. For each choice of N we randomly generated 10 networks, each comprising regulator sets and regulation functions for each gene. Each network was simulated as a Boolean network and as a continuous-time switching network. We simulated the networks to produce time series data until the data was sufficient to solve the inverse problem. (Because we generated the networks, we can check if a particular data set uniquely identifies the network.) Both Boolean and continuous-time switching networks are prone to reaching fixed points in the dynamics or periodic attractors. To protect against this, the data we use is actually a set of time series. Each time series began from a random initial state and was ended whenever it reached any Boolean network state for a second time.

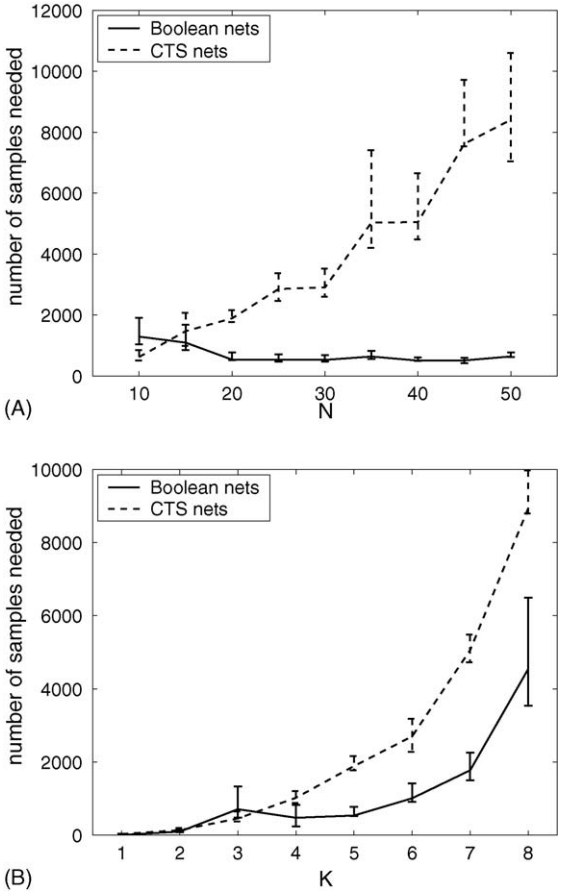

Fig. 4. Mean and quartiles of the empirical number of samples needed to solve the inverse problem for Boolean and continuous-time switching networks. (A) $K = 5$ and N varies. (B) $N = 20$ and K varies.

Fig. 4(A) shows the results of the experiment. The sample complexity for continuous-time switching networks appears to grow more than linearly with N, consistent with the $N \log N$ behavior predicted by Eq. (5). Surprisingly, the sample complexity for Boolean networks appears independent of N over the range tested, except for a slight rise at the lowest values of N. The rise is easy to explain. For such small networks, the approximation that the time series is an independent random sequence is poor. Often, such networks contain short periodic cycles, and trajectories from different initial conditions quickly reach the same cycle. Thus, after the first few trajectories, it takes longer to reach new Boolean states than the theory predicts. The flatness of the sample complexity curve for larger values of N is harder to explain. It may be that the approximation of independent randomness continues to improve with larger N, offsetting the expected rise in sample complexity, which is only proportional to $\log N$.

In a second experiment, we held N constant at 20, while varying K between 1 and 8. Fig. 4(B) presents the results. At least for $K \geq 5$, the sample complexity is approximately doubling for each increase in K, as predicted both by the lower bound of Eq. (3) and the estimates for expected sample complexity in Eqs. (4) and (5).

7. Discussion

We have observed that the dynamics of randomly generated Boolean and continuous-time switching networks have much different statistical characteristics, which lead to different estimates for the amount of data needed to solve the inverse problem. As did Akutsu et al. (1999), we observed that the number of samples needed for Boolean network identification scales as $\log N$, where N is the number of genes in the network. The $\log N$ dependence comes from the assumption that the data is independently randomly sampled, which, we have argued, is a good model of time series data from randomly generated Boolean networks. However, independent random sampling is not a good description of real gene expression time series, as has been noted by others (e.g. Szallasi and Liang, 1998). Thus, it is unclear how well these results will apply to real genetic network inference problems. For continuous-time switching networks, we estimated an $N \log N$ dependence, which may be more realistic.

We also studied how the number of regulators per gene, K, affects the amount of data needed to solve the inverse problem. In our lower bound, Eq. (3), and in our estimates for both Boolean and continuous-time switching networks, we found that the amount of data needed

scales exponentially with K. Specifically, it scales as 2^K or $K2^K$. This is significantly smaller than the 2^{2K} bound produced by Akutsu et al. (1999), but still suggests that inferring networks in which there are many regulators per gene will require large amounts of data. However, even the 2^K estimate may be pessimistic, because it is based on the assumption that the expression level or rate of production of a gene can be an arbitrary Boolean function of the states of its regulators. If there are biological reasons why certain types of dynamics functions are more likely, such as "canalyzing" functions (Kauffman et al., 2003), then the sample complexity may be yet lower.

Our analysis is limited by the assumptions that each gene is regulated by precisely K others and that the regulation functions and inputs are chosen uniformly at random from all 2^{2^K} Boolean functions on K inputs. Such networks tend to exhibit complex long-period, aperiodic or chaotic dynamics, which nevertheless have very regular statistical properties. These properties allow us to derive predictions for the amount of data needed to solve the inverse problem. Recent work has demonstrated that many of the assumptions of our model do not accurately reflect the structure of genetic networks. Thus, genetic networks may exhibit scale free connectivity with a range of different values of in degree (Lee et al., 2002), characteristic network motifs of local connectivity (Shen-Orr et al., 2002), and canalyzing, rather than random, Boolean functions regulating genetic activity (Kauffman et al., 2003). These factors can dramatically affect the dynamical properties of the network (Aldana and Cluzel, 2003; Glass and Hill, 1998; Kauffman et al., 2003; Oosawa and Savageau, 2002). Consequently, the estimates for the sample complexity in the current paper may not apply to networks with different dynamical properties. However, the general approach employed here, of first characterizing the statistics of the time series arising from a particular class of networks and then using these statistics to derive sample complexity estimates, should still apply.

In a similar vein, while we have derived different sample complexity estimates for Boolean and continuous-time switching networks, and supported these differences with simulation experiments, it is important to realize that these differences are not inherent to Boolean and continuous-time switching networks. Rather, these results stem from statistical differences in the dynamics of randomly generated networks of these two types. Looking at restricted classes of these networks or sampling the data in a different way may change these results. For example, if we restrict attention to the class of Boolean networks in which only a single gene changes state from one time step to the next, then we would expect the sample complexity to be more similar to that of randomly generated continuous-time switching networks. Likewise, if we imagine sampling the dynamics of a continuous-time switching network periodically, with enough time passing in between samples that many genes change Boolean state, then we would expect a sample complexity more like that for Boolean networks. We are presently investigating the sample complexity for continuous-time switching networks as a function of the sampling period.

In implementing algorithms to carry out the inverse problem, it is often not necessary to specify K (Liang et al., 1998; Ideker et al., 2000; Perkins et al., 2004) or to have a fixed number of regulators for each gene, or to require a particular type of dynamics—although, if K is unknown, the algorithms cannot determine with certainty whether all regulators have been detected until all 2^N Boolean network states have appeared in the data. However, it appears that the statistical properties of the dynamics can have a profound affect on the computational complexity of solving the inverse problem. The algorithms of Liang et al. (1998) and Akutsu et al. (1999) require $O(N^K)$ search effort to find minimal-size candidate regulator sets for each gene. The algorithms described in Perkins et al. (2004), tailored for continuous-time switching networks, requires an amount of computation that is only polynomial in the size of the data. Thus, while the sample complexity of the inverse problem may be better for Boolean networks than for continuous-time switching networks, the computational complexity may be worse. Formally establishing this apparent "trade-off" between sample complexity and computational complexity, and studying how other types of dynamics affect the sample and computational complexity of the inverse problem are important directions for future work.

Acknowledgements

This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. This material is based upon work supported by the National Science Foundation under a grant awarded in 2002. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Proof sketch for Lemma 1

Recall that there are A events, each of which occurs with probability at least p in any block of τ time steps.

Let t_a be the random variable denoting the time at which event a occurs. The expected time until all A events occur is

$$E \max_a t_a \leq \sum_{i=1}^{\infty} \tau \,\text{Prob}(\text{at least one } t_a \geq \tau i)$$
$$\leq \sum_{i=1}^{\infty} \tau \min(1,\, A\,\text{Prob}(\text{a particular } t_a \geq \tau i))$$
$$\leq \sum_{i=1}^{\infty} \tau \min(1,\, A(1-p)^i)$$

Let I be the smallest integer such that $A(1-p)^I \leq 1$. Then we have

$$E \max_a t_a \leq \tau \left(I + \sum_{i=I}^{\infty} A(1-p)^i \right)$$
$$= t \left(I + A\frac{(1-p)^I}{p} \right)$$

Because $A(1-p)^I \leq 1$, the second term inside the parenthesis is no more than $\frac{1}{p}$. Furthermore, $A(1-p)^I \leq 1 \Leftrightarrow \log A + I\log(1-p) \leq 0 \Leftrightarrow I \geq \frac{-\log A}{\log(1-p)}$. It can be shown that $\frac{-\log A}{\log(1-p)} = O\left(\frac{\log A}{p}\right)$. Thus, the expected time until all A events occur is $O\left(\frac{\tau}{p}\log A\right)$.

References

Akutsu, T., Miyano, S., Kuhara, S., 1999. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Proceedings of the Pacific Symposium on Biocomputing, pp. 17–28.

Albert, R., Othmer, H.G., 2003. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. J. Theor. Biol. 223, 1–18.

Aldana, M., Cluzel, P., 2003. A natural class of robust networks. Proc. Natl. Acad. Sci. U.S.A. 100 (15), 8710–8714.

Bagley, R.J., Glass, L., 1996. Counting and classifying attractors in high dimensional dynamical systems. J. Theor. Biol. 183, 269–284.

Bastolla, U., Parisi, G., 1997. A numerical study of the critical line of Kauffman networks. J. Theor. Biol. 187, 117–133.

Bilke, S., Sjunnesson, F., 2001. Stability of the Kauffman model. Phys. Rev. E 65.

Chandra, A.K., Raghavan, P., Ruzzo, W.L., Smolensky, R., Tiwari, P., 1997. The electrical resistance of a graph captures its commute and cover times. Comput. Complexity 6 (4), 312–340.

de Jong, H., Geiselmann, J., Batt, G., Hernandez, C., Page, M., 2004a. Qualitative simulation of the initiation of sporulation in *Bacillus subtilis*. Bull. Math. Biol. 66 (2), 261–300.

de Jong, H., Gouze, J.-L., Hernandez, C., Page, M., Sari, T., Geiselmann, J., 2004b. Qualitative simulation of genetic regulatory networks using piecewise-linear models. Bull. Math. Biol. 66 (2), 301–340.

Glass, L., 1975. Combinatorial and topological methods in nonlinear chemical kinetics. J. Chem. Phys. 63 (4), 1325–1335.

Glass, L., Hill, C., 1998. Ordered and disordered dynamics in random networks. Europhys. Lett. 41, 599–604.

Ideker, T.E., Thorsson, V., Karp, R.M., 2000. Discovery of regulatory interactions through perturbation: inference and experimental design. In: Proceedings of the Pacific Symposium on Biocomputing, pp. 302–313.

Jaeger, J., Blagov, M., Kosman, D., Kozlov, K.N., Manu, Myasnikova, E., Surkova, S., Vanario-Alonso, C.E., Samsonova, M., Sharp, D.H., Reinitz, J., 2004a. Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster*. Genetics 167, 1721–1737.

Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K.N., Manu, Myasnikova, E., Vanario-Alonso, C.E., Samsonova, M., Sharp, D.H., Reinitz, J., 2004b. Dynamic control of positional information in the early *Drosophila* embryo. Nature 430, 368–371.

Kauffman, S., Peterson, C., Samuelsson, B., Troein, C., 2003. Random Boolean network models and the yeast transcriptional network. Proc. Natl. Acad. Sci. U.S.A. 100 (25), 14796–14799.

Kauffman, S.A., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theor. Biol. 22, 437–467.

Kauffman, S.A., 1993. The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, New York.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.-B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A., 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298, 799–804.

Liang, S., Fuhrman, S., Somogyi, R., 1998. REVEAL, a general reverse-engineering algorithm for inference of genetic network architectures. In: Proceedings of the Pacific Symposium on Biocomputing, pp. 18–29.

Mestl, T., Bagley, R.J., Glass, L., 1997. Common chaos in arbitrarily complex feedback networks. Phys. Rev. Lett. 79 (4), 653–656.

Oosawa, C., Savageau, M.A., 2002. Effects of alternative connectivity on behavior of randomly constructed Boolean networks. Physica D 170, 143–161.

Perkins, T.J., Hallett, M.T., Glass, L., 2004. Inferring models of gene expression dynamics. J. Theor. Biol. 230, 289–299.

Raeymaekers, L., 2002. Dynamics of Boolean networks controlled by biologically meaningful functions. J. Theor. Biol. 218, 331–341.

Reinitz, J., Sharp, D.H., 1995. Mechanism of *eve* stripe formation. Mech. Dev. 49, 133–158.

Samuelsson, B., Troein, C., 2003. Superpolynomial growth in the number of attractors in Kauffman networks. Phys. Rev. Lett. 90 (9).

Sanchez, L., Thieffry, D., 2001. A logical analysis of the gap gene system. J. Theor. Biol. 211, 115–141.

Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U., 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat. Genet. 31, 64–68.

Shmulevich, I., Kauffman, S.A., 2004. Activities and sensitivies in Boolean network models. Phys. Rev. Lett. 93 (4).

Socolar, J.E.S., Kauffman, S.A., 2003. Scaling in ordered and critical random Boolean networks. Phys. Rev. Lett. 90 (6).

Szallasi, Z., Liang, S., 1998. Modeling the normal and neoplastic cell cycle with "realistic Boolean genetic networks": their application for understanding carcinogenesis and assessing therapeutic strategies. In: Proceedings of the Pacific Symposium on Biocomputing, pp. 66–76.