Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Journal of Theoretical Biology 230 (2004) 289–299

Journal of Theoretical Biology

www.elsevier.com/locate/jtbi

Inferring models of gene expression dynamics

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

[a] McGill Centre for Bioinformatics, McGill University, 3775 University St. Montreal, Quebec, Canada H3A 2B4
[b] Department of Physiology, McGill University, 3655 Drummond Street, Montreal, Quebec, Canada H3G 1Y6

Received 8 December 2003; received in revised form 12 May 2004; accepted 13 May 2004

Available on line 3 August 2004

Abstract

We study the problem of identifying genetic networks in which expression dynamics are modeled by a differential equation that uses logical rules to specify time derivatives. We make three main contributions. First, we describe computationally efficient procedures for identifying the structure and dynamics of such networks from expression time series. Second, we derive predictions for the expected amount of data needed to identify randomly generated networks. Third, if expression values are available for only some of the genes, we show that the structure of the network for these "visible" genes can be identified and that the size and overall complexity of the network can be estimated. We validate these procedures and predictions using simulation experiments based on randomly generated networks with up to 30,000 genes and 17 distinct regulators per gene and on a network that models floral morphogenesis in *Arabidopsis thaliana*.

Keywords: Genetic network inference; Differential equations; Logic

1. Introduction

New high-throughput technologies for monitoring gene expression, microarrays, serial analysis of gene expression (SAGE), and others (Ding and Cantor, 2003; Kosman et al., 1998; Lockhart et al., 1996; Velculescu et al., 1995), offer an unprecedented capacity to view the complex regulatory machinery of cells. It is now possible to measure the activity of up to thousands of genes simultaneously, as they evolve over time or respond to different environmental, pharmaceutical, or genetic conditions. Such access to the states of cells gives rise to the hope of automatically inferring, on a large scale, the network of interactions that controls gene expression. However, a number of factors make this a difficult problem: noisy measurements, missing information, the apparent complexity and stochastic nature of gene regulation (McAdams and Shapiro, 1995; Yuh et al., 1998), and the sheer number of genes involved.

In this paper, we consider several questions: Is it possible to infer the regulatory relationships between genes from expression time series? If so, what inference algorithms are sufficient? How much data is required? Theoretical answers to these questions depend on how one models gene regulation and expression dynamics. Linear differential equation models can be inferred efficiently from expression data using least-squares regression techniques (Chen et al., 1999; D'Haeseleer et al., 1999; Gardner et al., 2003; Tegner et al., 2003; Yeung et al., 2002). However, linear models are not always flexible enough to model complex regulatory relationships. Nonlinear differential equations are potentially more realistic (Von Dassow et al., 2000; McAdams and Shapiro, 1995; Reinitz and Sharp, 1995; Yuh et al., 1998). Unfortunately, model-fitting is essentially a smooth nonlinear optimization problem that is typically impossible to solve exactly and that requires extensive computing power to solve even approximately (Von Dassow et al., 2000; Reinitz and Sharp, 1995). There has been some success in modeling gene networks using logical formalisms (Bodnar, 1997;

*Corresponding author. Tel.: +1-514-398-7071X09317; fax: +1-514-398-3387.

E-mail addresses: perkins@mcb.mcgill.ca (T.J. Perkins), hallett@mcb.mcgill.ca (M. Hallett), glass@cnd.mcgill.ca (L. Glass).

0022-5193/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.05.022

Kauffman, 1993; Mendoza and Alvarez-Buylla, 1998; Sanchez and Thieffry, 2001; Somogyi and Sniegoski, 1996; Thomas and D'Ari, 1990). However, logical models lose information by discretizing expression levels. Furthermore, fitting these models requires an exponential amount of computation (Akutsu et al., 1999; Liang et al., 1998).

We explore the problem of gene network inference using a formalism that combines nonlinear differential equations and logical approaches (De Jong et al., 2003; Glass and Kauffman, 1973; Glass, 1975; Mestl et al., 1995). While gene expression is real-valued and changes continuously in time, the rules that specify the time derivatives of expression take a logical form. These models can represent many of the complex, nonlinear regulatory phenomena observed in real gene networks. Despite this expressive power, we show that, under idealized observation conditions, these models can be efficiently inferred from expression time series. We provide analytical predictions for the amount of expression data needed as a function of the size and connectivity of the gene network. We test these predictions on simulated expression data from randomly generated networks and from a network that models floral morphogenesis in *Arabidopsis thaliana*. We also consider the case in which the expression time series includes data for only some of the genes in the network. We show that regulatory relationships between the observed genes can be inferred and that the overall size and connectivity of the network can be estimated, even if the detailed behavior of the unobserved genes cannot be deduced.

2. A model of gene expression

Many types of molecules are involved in gene expression. DNA, mRNAs, proteins and various small molecules all have roles in the regulation of protein production. By *genetic network* we mean a dynamical model of a finite set of interacting chemical species related to gene expression. In our model, each species i has a real-valued time-varying concentration, $x_i(t)$, which we assume to be normalized between zero and one. In addition, each species has a logical state, $X_i(t)$, which is 1 (high) if the $x_i(t)$ is greater than one half and 0 (low) otherwise. Concentration dynamics follow a production–decay model. The decay rate of species i is proportional to its concentration. The production rate of species i is controlled by a set of regulating species R_i. The production rate is always zero (no production) or one (maximum production), and depends on the logical states of the regulators. Thus, the dynamics for species i are

$$\dot{x}_i(t) = f_i(X_{R_i}(t)) - x_i(t), \qquad (1)$$

where $X_{R_i}(t)$ is a vector containing the logical states of the regulators of species i at time t, and f_i is a Boolean function (Glass, 1975). We call f_i the production rate function or regulation function of species i.

The regulator sets, R_i, define the *structure* of the network. The structure is often conceptualized as a directed graph in which vertices correspond to species and an arc from j to i means species j regulates species i (similar to the top of Fig. 1 and to Fig. 7). The regulator sets together with the regulation functions specify the dynamics. More general forms of Eq. (1) allow for more than two logical levels per species, separated by arbitrary real thresholds as well as production and decay rate constants other than zero and one (Glass, 1975; Mestl et al., 1995; De Jong et al., 2003). Partly for ease of exposition, we restrict attention to the simplest form of this class of models.

As an example of this framework, the synthetic "repressilator" gene network (Elowitz and Leibler, 2000) can be modeled as a cyclic network of three genes, each of which represses the next gene in the chain (see top of Fig. 1). This can be modeled by the regulator sets

$$R_1 = \{3\}, \quad R_2 = \{1\}, \quad R_3 = \{2\}$$

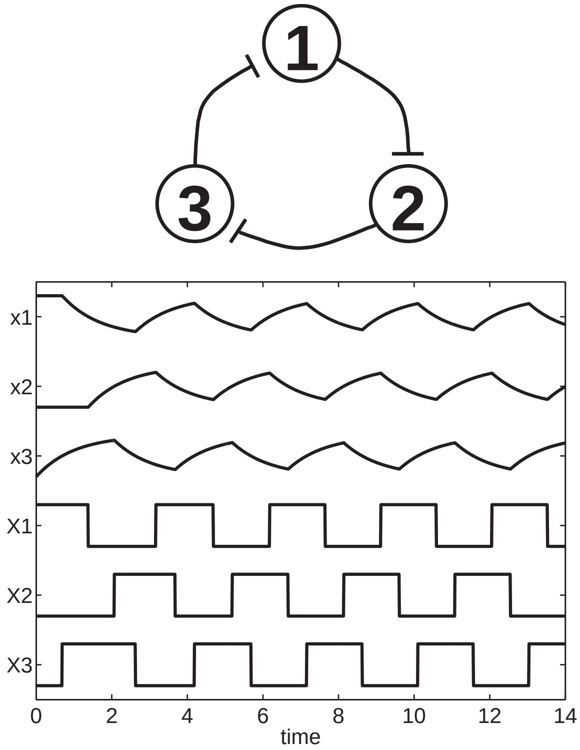

Fig. 1. Top: diagram of the three-gene repressilator network. Bottom: simulation from initial state $x_1 = 1$, $x_2 = 0$, $x_3 = 0$. The first three curves represent the concentrations of the three genes. The second three curves represent the logical states of the genes.

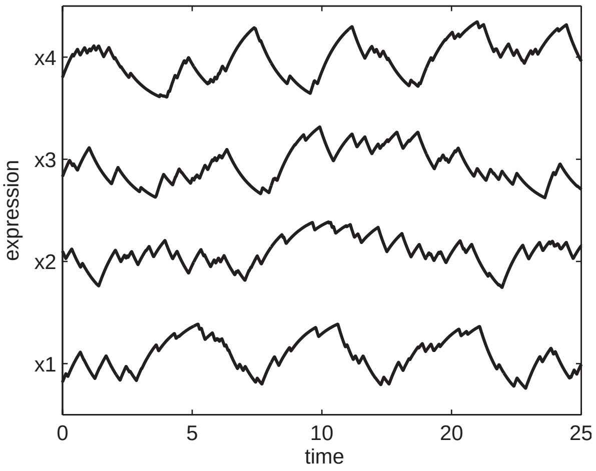

Fig. 2. Concentration of species one through four in a random 100-species network with 10 regulators per species.

and the regulation functions

$$f_1 = 1 - X_3, \quad f_2 = 1 - X_1, \quad f_3 = 1 - X_2.$$

The bottom of Fig. 1 shows the concentrations and logical states as a function of time from the initial condition $x_1 = 1$, $x_2 = 0$, $x_3 = 0$. Expression measurements for one of the genes in the real repressilator system showed a regular, periodic oscillation in activity (Elowitz and Leibler, 2000). This is in qualitative agreement with the predictions of this model. (The other two genes were presumed to be oscillating as well, but their expression was not measured.)

A time at which the logical state of any species changes is called a *switching time*. Between switching times, the logical states of all species are constant and so are production rates. Each concentration changes according to a simple linear differential equation $\dot{x}_i = -x_i$ or $1-x_i$. At a switching time, production rates can change, causing species to change from $\dot{x}_i = 1 - x_i$ dynamics to $1-x_i$ dynamics, or vice-versa. The piecewise linear equations (1) can exhibit complex behaviors not possible in linear differential equations. For example, Fig. 2 shows the concentrations of four species in a randomly generated network of 100 species. Each species has 10 regulators, chosen at random from the 99 other species in the network. The regulation function for each species was chosen randomly from all Boolean functions depending on 10 inputs. Typically, such randomly generated networks do not produce simple periodic behavior (Edwards and Glass, 2000; Mestl et al., 1997). However, their behavior exhibits statistical regularities that can be exploited to predict the amount of data needed to identify the network.

3. Network inference from fully observed continuous concentration time series

3.1. The inference problem and efficient solutions

We wish to determine the structure and regulation functions of an unknown system behaving according to Eq. (1), based on a set of concentration time series. In this section, we assume that each time series specifies the concentrations of all species over a continuous time interval. Access to concentration data over a continuous time interval implies that time derivatives can be inferred, as can the production rate of any species at any instant.

Our approach for inferring the structure of the network is based on the following observation. Suppose that in one instance the concentration of species i is rising (implying a production rate of one) and that in another instance its concentration is falling (implying no production). Suppose that in these two instances all species except one, species j, have the same logical state. Based on these two observations, it follows from the form of Eq. (1) that species j regulates species i. This suggests the following rule.

Rule 1: Infer that species j regulates species i if at two different times in the same or different time series, (1) all species except j have the same logical state, (2) the production rate of i differs.

We assume that each time series contains a finite number of switching times that divide the time series into intervals of constant logical states and production rates. Assuming that it is easy to find and iterate through these intervals, Rule 1 can be tested by looping over all pairs of intervals. The computation time required to apply this rule is proportional to the number of species and to the square of the number of switching times.

A special case of this rule is to look at pairs of times immediately before and after a switching time in one time series.

Rule 2: Infer that species j regulates species i if in one of the concentration time series, (1) there is a switching time at which j is the only species to switch logical state, (2) at that switching time, the production rate of i changes.

Given the same data, Rule 2 may identify fewer regulatory relationships than Rule 1. However, the computation time for Rule 2 is proportional to the number of switching times in the data set, rather than to its square. This advantage is important in the next subsection, where we infer very large networks from long time series. A second advantage is discussed in the section on partially observed time series.

To estimate the regulation function of species i given an estimate of its regulators, we store the observed production rates for species i, conditional on the logical states of its regulators, in a table. Suppose one of the above rules infers a set of regulators, $\hat{R}_i$, for species i, and let $X_{\hat{R}_i}$ denote a vector of logical states for the species in $\hat{R}_i$. Regulation function estimation is formalized by the rule below.

Rule 3: Infer $f(X_{\hat{R}_i})$

$$= \begin{cases} 1 & \text{if the production rate is one whenever the species in } \hat{R}_i \text{ are in logical states } X_{\hat{R}_i}, \\ 0 & \text{if the production rate is zero whenever the species in } \hat{R}_i \text{ are in logical states } X_{\hat{R}_i}, \\ ? & \text{otherwise.} \end{cases}$$

The last possibility, "?", can occur when the time series data does not contain any instance in which the species in $\hat{R}_i$ had the logical states specified in $X_{\hat{R}_i}$. In such a case, there is no basis for inferring the production rate of species i. The "?" inference can also occur if two different production rates are observed for the same pattern of logical regulator states, $X_{\hat{R}_i}$. This can only happen if at least one regulator of species i is not in $\hat{R}_i$, and the difference in production rate is due to a difference in the logical state of the unidentified regulator.

These rules are sufficient to recover the repressilator network from the time series depicted in Fig. 1. The first switch occurs when x_3 increases past one half, changing X_3 from zero to one. At the same time, x_1 begins to fall, revealing that gene 3 is a regulator of gene 1. Gene 1 is the second to switch logical state. The expression of gene 2 simultaneously begins to rise, indicating that gene 1 regulates gene 2. Finally, gene 2 switches logical state and the production rate of gene 3 changes, implying regulation of gene 3 by gene 2. By this time, after one period of the oscillator, the structure of the network and all regulation functions are determined.

If the concentration data is generated according to an equation of form (1), then Rules 1 and 2 posit that species j regulates species i only if that fact can be deduced from the data. Neither rule produces false positive regulatory relationships. False negatives are possible and unavoidable in general. There is no guarantee that a set of time series, however long, contains enough information to identify the network. This is a well-recognized pitfall in dynamical-system identification, and is not just a characteristic of systems with dynamics given by Eq. (1). Theoretically, this problem can be avoided by assuming multiple time series from independently random initial conditions or some other form of randomly sampled data (e.g., (Akutsu et al., 1999)). Another possibility is allowing the inference procedure to choose which samples are collected (Akutsu et al., 1998; Ideker et al., 2000). In the next section, we demonstrate the network inference procedures above on randomly generated networks. Although no amount of data is guaranteed to be sufficient for identification, statistical properties of these networks allow us to estimate the expected amount of data required for both structure and regulation function identification.

3.2. Analysis and simulation experiments based on randomly generated networks

We performed simulation experiments with randomly generated networks of N species and K regulators per species, for various N and K. "Randomly generated" means that for each species i, R_i was chosen randomly from all size-K sets of species excluding i. The regulation functions were chosen randomly from all Boolean functions depending on K inputs. To a first approximation, a time series from a randomly generated network typically displays three properties:

Property 1: A change in the logical state of any species has a 50% chance of changing the production rate of any species it regulates. This is due to the assumption that the regulation functions are randomly generated Boolean functions.

Property 2: At most one species switches logical state at a given time. This is because logical state switches occur at isolated, real-valued times.

Property 3: For any l, every species is equally likely to be the lth species to switch logical state, regardless of which species switched before. In the current work, we do not account for the possibility of correlations in the switching sequence (Mestl et al., 1997).

Suppose Rules 2 and 3 are applied to a single concentration time series obeying these three properties. How long must the time series be, in terms of the number of logical switches, before the network is identified? Consider a particular species i. Initially, none of its K regulators have been identified. To identify a regulator, two things must happen. The regulator must switch logical state and the production rate of species i must change. By Property 3, there is probability K/N that a particular logical switch will involve one of i's regulators. By Property 1, there is probability $\frac{1}{2}$ that i's production rate will change. Thus, there is probability $K/2N$ that any particular logical switch reveals one of the regulators of i. The expected number of switches until the first of K regulators is revealed is thus $2N/K$. Once the first regulator is identified, there is probability $(K-1)/2N$ that any subsequent switch will reveal one of i's other $K-1$ regulators. The expected additional switches until the second regulator is identified is thus $2N/(K-1)$. The expected number of logical switches until all of i's regulators are identified is

$$\frac{2N}{K} + \frac{2N}{K-1} + \cdots + \frac{2N}{1} = 2NH_K,$$

where $H_K = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{K}$ is the Kth harmonic number. H_K is approximately equal to the natural logarithm of K.

If we make the assumption that the processes of identifying regulators are statistically independent for the different species, then an inductive argument can be used to show that the expected number of logical switches before the structure of the network can be identified is no more than

$$2NH_NH_K. \tag{2}$$

That is, identifying the regulators of all species requires only about $H_N \approx \ln N$ times as much data as is needed to identify the regulators of a single species. We tested this prediction using computer simulations of networks with $N = 300$, 3000, and 30,000 species and K between five and 25. For each choice of N and K, we randomly generated 100 networks and initial concentrations. We simulated the dynamics of each network until the entire structure of the network was identified by Rule 2. Fig. 3 displays the mean number of logical switches before the structure was identified, along with the prediction of Eq. (2). Predictions and simulation results match well across a wide range of N and K. For most values of K, the expected number of switches required in the simulations was smaller than theoretically predicted. For the smallest values of K, however, the predictions were too low. The simulations display an increase in the number of switches needed for small K, which is absent in the theoretical predictions. We expect this is due to a failure of Property 1, according to which a change in the logical state of a regulator has a 50% chance to change the production rate of the regulated species. The fewer the regulators, the more likely it is that a randomly generated Boolean function will be mostly zero or mostly one for the different combinations of logical regulator states. In such cases, changes in the production rate are seen less often, and one expects structure identification would take longer.

How long must a time series be for the network structure and all of the regulation functions to be identified? Even if the regulators of species i are known, the regulation function cannot be identified until every combination of logical states of the regulators appears in the time series, 2^K combinations if there are K regulators. The sequence of logical regulator states in the time series can be viewed as a random walk on a K-dimensional hypercube. At each switching time, there is probability K/N that one of the regulators changes logical state, corresponding to a step to an adjacent vertex on the hypercube. The expected number of switches between steps on the hypercube is N/K. For a random walk on a K-dimensional hypercube, the expected number of steps until all vertices have been visited is $K2^K$ (Chandra et al., 1997). Thus, the expected number of switches until all combinations of regulator states have been observed is $(N/K)K2^K = N2^K$. If identifying regulation functions is assumed to be a statistically independent process for each species, then the expected number of switches before all of the regulation functions can be identified is no more than

$$NH_N2^K. \tag{3}$$

We tested this prediction using computer simulations of networks with $N = 300$, 3000, and 30,000 species and K between five and 12. For each choice of N and K, we used the same 100 networks and initial conditions as in the previous simulations, and simulated the dynamics until Rules 2 and 3 were sufficient to identify the structure of the network and all of the regulation functions. Fig. 4 displays the results of the simulations along with the theoretical predictions. As with structure identification, the networks are identified faster in simulation than predicted for higher values of K. For smaller values of K, the increasing number of switches

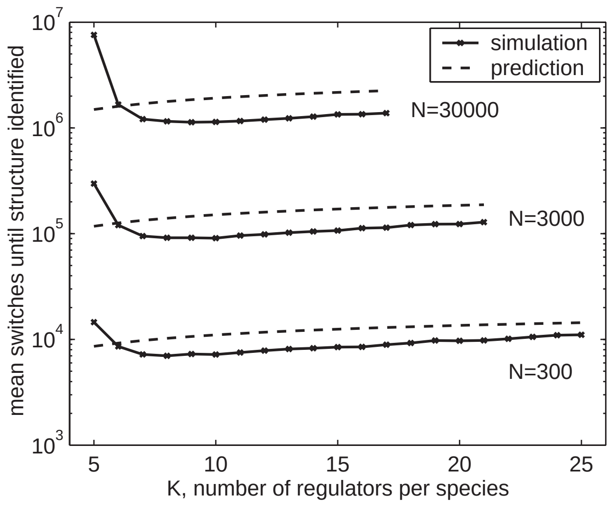

Fig. 3. Number of logical switches until network structure is identified—simulation results and theoretical predictions.

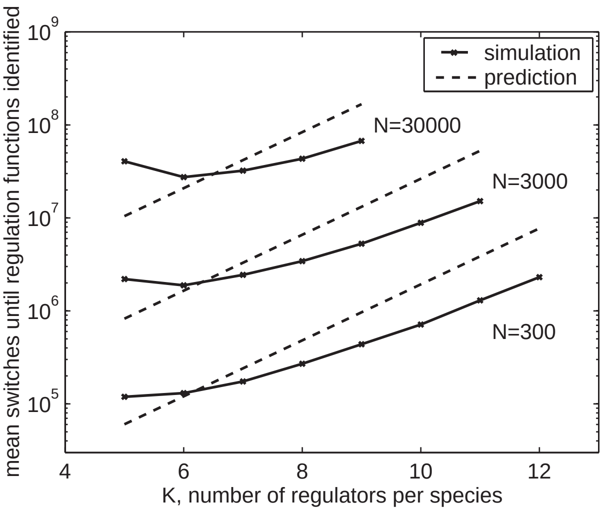

Fig. 4. Number of logical switches until all regulation functions are identified—simulation results and theoretical predictions.

needed in the simulations is not captured by the theoretical analysis.

These analyses and simulations rely on Rule 2 for structure identification. Analysis for Rule 1 is more difficult, but its data requirements can be no greater than those of Rule 2. We expect little difference for networks with many species, because there is little chance that at two arbitrary times there is only a single species in a different logical state. Most regulatory relationships should be revealed by single switching events. However, for smaller networks or networks whose dynamics do not obey the three properties we have assumed, the data requirements of the two rules may be quite different. For example, the structure of the *Arabidopsis thaliana* model below cannot be inferred entirely by Rule 2, because some of the regulatory relationships cannot be revealed by any single time series. Comparisons of multiple time series, and hence Rule 1, are necessary for inferring the complete structure.

4. Network inference from partially observed expression time series

For the partially observed inference problem, we again assume the concentration data is in the form of a finite set of concentration time series, each over a continuous time interval. However, we assume that concentration values are given for only $M<N$ of the species. The concentrations of the other species are unknown, as is the total number of species in the network, N. We call species 1 through M visible species, and we call species j a visible regulator of species i if j is visible and if j regulates i.

Determining the entire structure and regulation functions of the network from such data is not possible in general. However, it is possible to determine the visible regulators of each visible species, which we call the visible structure of the network. In the partially observed problem, using Rule 1 to identify regulators may produce false positives. If two production rates for species i are observed while a single visible species j is in a different logical state, it does not follow that j regulates i. The difference in i's production rate may result from regulation by an unobserved species. Rule 2 can also produce false positives, but only if an unobserved species and an observed species change logical states at the same time. If simultaneous changes are ruled out (Property 2), then Rule 2 does not produce false positives.

How much data is required to identify the visible structure of a randomly generated network? We assume a single time series obeying the same three properties described in the previous section. For a species i with K_i visible regulators, we expect $2NH_{K_i}$ logical switches until all the visible regulators are identified by Rule 2. (This counts all switches, not just the visible ones. The number of visible switches would be $2MH_{K_i}$.) Because each species may have a different number of visible regulators, the number of logical switches until the entire visible structure is determined can be estimated as

$$2NH_M \sum_{l=0}^{K} Pr(K_i = l)H_l, \qquad (4)$$

where $Pr(K_i = l) = \binom{M-1}{l}\binom{N-M}{K-l}/\binom{N-1}{K}$ is the probability that species i has l visible regulators in a randomly generated network.

Fig. 5 compares this prediction with the results of simulations for randomly generated networks of $N =$ 300, 3000, and 30,000 species with $K = 10$ regulators per species. The fraction of species visible was varied between 10 and 90 percent. The predictions are slightly low, but scale correctly with network size and with the fraction of the species visible.

Although it is not possible to make detailed inferences about the unobserved part of the network, one can estimate the total network size, N, and the number of regulators per gene, K, for randomly generated networks. Two types of data are useful for making these estimates: the visible structure of the network, once it has been determined, and the frequency with which changes in the production rate of a visible species can be attributed to a visible species versus an unknown, invisible species. Suppose the structure identification procedure has determined that species i has K_i visible regulators. Suppose the data contains T_i changes in the production rate of species i, and V_i of these can be attributed to one of the visible regulators. For a randomly generated network, a regulator of a species

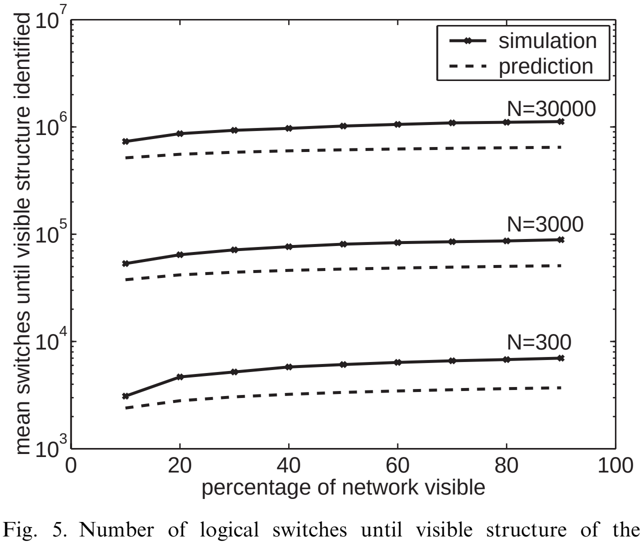

Fig. 5. Number of logical switches until visible structure of the network is identified—simulation results and theoretical predictions.

has probability M/N of being a visible regulator. Each change in the production rate of species i has probability K_i/K of being due to a visible regulator. Thus, the probability of a particular set of K_i's, T_i's and V_i's is approximately

$$\prod_{i=1}^{M}\binom{K}{K_i}\left(\frac{M}{N}\right)^{K_i}\left(1-\frac{M}{N}\right)^{K-K_i} \times\binom{T_i}{V_i}\left(\frac{K_i}{K}\right)^{V_i}\left(1-\frac{K_i}{K}\right)^{T_i-V_i}. \qquad (5)$$

The first three terms approximate the probability that species i has K_i visible regulators, and the last three terms express the probability of the observed number of production rate switches of species i due to visible and invisible regulators.

The principle of maximum likelihood, which states that the best hypothesis is the one under which the observed data is most likely, can be used to derive estimates for N and K. It is readily shown that for any fixed K, the choice $N = M^2K/\left(\sum_{i=1}^{M} K_i\right)$ maximizes Eq. (5). The optimal choice of K can be found simply by evaluating different choices of K over a reasonable range. We tested this procedure on the same networks as above, using for data the visible structure discovered and the switching data up until the time at which the visible structure was fully identified. The results are presented in Fig. 6. Estimates improve with increasing N and with increasing percentage of the network visible. In 100 independent runs, the estimates were always correct to within a factor of two even with networks of 300 species of which only 10% of the species were observed.

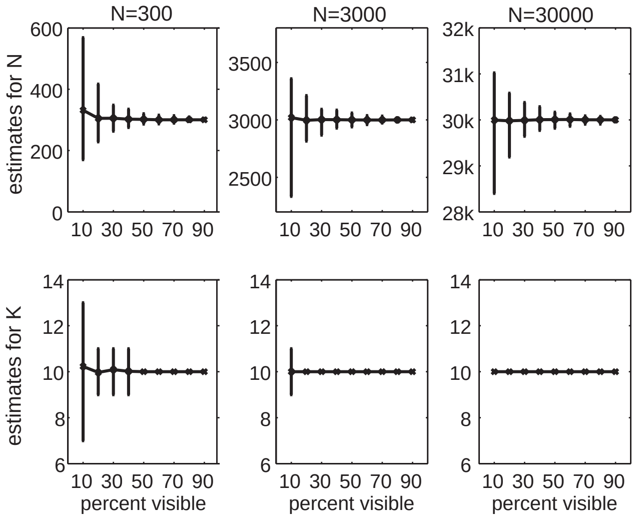

Fig. 6. Minimum, mean, and maximum estimates for N and K for networks of varying sizes, over 100 independent simulations.

5. Demonstration on a model of floral morphogenesis in *Arabidopsis thaliana*

In this section, we demonstrate the network inference procedures on a model of floral morphogenesis in *Arabidopsis thaliana*. Mendoza and Alvarez-Buylla (1998) proposed a Boolean network model that qualitatively describes differentiation of the four floral organs: sepals, petals, stamens, and carpels. We adapt this model to equations of form (1). The model contains twelve interacting chemical species, depicted in Fig. 7. Eleven are proteins, designated EMF1, TFL1, LFY, AP1, CAL, LUG, UFO, AG, AP3, PI and SUP. The twelfth species, BFU, is a dimer of the AP3 and PI proteins. The Mendoza–Alvarez-Buylla model specified conditions for expression of the different species by means of linear threshold functions. We translated these into the logical rules below.

$$f_{\mathrm{LUG}} = 0$$

$$f_{\mathrm{UFO}} = 0$$

$$f_{\mathrm{LFY}} = 0$$

$$f_{\mathrm{SUP}} = 0$$

$$f_{\mathrm{EMF1}} = X_{\mathrm{EMF1}}$$

$$f_{\mathrm{CAL}} = X_{\mathrm{LFY}}$$

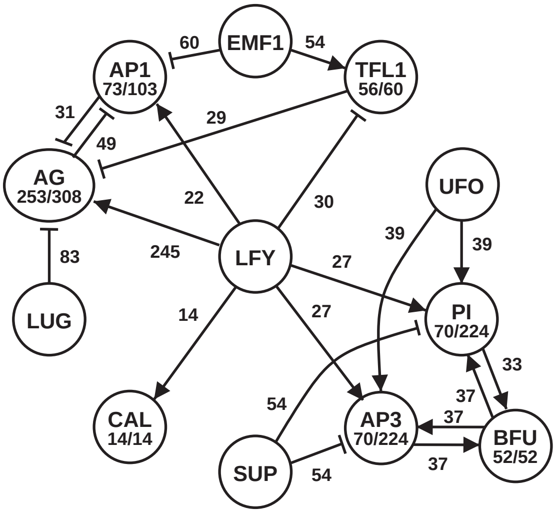

Fig. 7. Diagram of the network modeling floral morphogenesis in *Arabidopsis thaliana* (Mendoza and Alvarez-Buylla, 1998). Activation indicated by →, repression by ⊣. Regulation functions are given in the text. (The original Mendoza–Alvarez-Buylla model includes regulators for LFY, but the regulation function they propose does not allow activation under any conditions. We have removed those links.) Numbers on edges indicate the mean number of switches before that regulatory link was discovered. The two numbers in circles indicate the mean number of switches before all regulators of the species are identified and the mean number of switches before all regulators and the regulation function are determined.

$f_{\mathrm{TFL1}} = X_{\mathrm{EMF1}}$ and not X_{LFY}

$f_{\mathrm{AP1}} = X_{\mathrm{LFY}}$ or (not X_{EMF1} and not X_{AG})

$f_{\mathrm{BFU}} = X_{\mathrm{AP3}}$ and X_{PI}

$f_{\mathrm{AG}}=$ not X_{TFL1} and not X_{AP1} and (X_{LFY} or not X_{LUG})

$f_{\mathrm{AP3}} = X_{\mathrm{LFY}}$ or (X_{UFO} and X_{BFU}) or
(not X_{SUP} and (X_{UFO} or X_{BFU}))

$f_{\mathrm{PI}} = X_{\mathrm{LFY}}$ or (X_{UFO} and X_{BFU}) or
(not X_{SUP} and (X_{UFO} or X_{BFU}))

Five of the proteins have no regulators or are autoregulating and act as “inputs” to the network. The first four, LUG, UFO, LFY, and SUP, always decay to zero from their initial concentration. EMF1 is autoregulating. If its initial concentration is high (greater than one-half), then it stays high, approaching one. If its initial concentration is low, then it decays to zero. Depending on initial conditions, the system tends to one of six attractors in which each concentration is zero or one. Four of these correspond to the four types of floral organs. One corresponds to a non-floral state, and one corresponds to a floral state that has not been observed experimentally, perhaps because unmodeled factors prevent its occurrence or because the initial conditions that lead to this attractor do not occur naturally (Mendoza and Alvarez-Buylla, 1998). Fig. 8 shows an example time series for the network.

To test the inference procedures, we simulated the model to produce time series data. A single time series was generated by setting the initial concentrations uniformly randomly between zero and one and simulating until the logical states of the species reached one of the six attractors. For a single test of the amount of data needed to infer the network, we generated a sequence of such time series until Rules 1 and 3 were sufficient to infer the network structure and all regulation functions. This test was repeated 100 times, to estimate average data requirements. It was necessary to use Rule 1 for inferring structure because EMF1 never switches logical state during a time series. Only by comparing across time series can one identify species regulated by EMF1. Fortunately, the simulated data sets required to infer the network were small enough that applying Rule 1 was not a computational burden.

Fig. 7 displays the mean number of logical switches observed before inferring each part of the network. On average, it took 264 switches to determine the entire structure of the network and 343 switches to determine the structure and all the regulation functions. Although this network was not generated randomly, one can compare these numbers to the predictions for random networks. There are 12 species and approximately 2 regulators per species on average. This suggests $2 * 12 * H_{12} * H_2 = 112$ switches would be needed to identify the structure, and $12 * H_{12} * 2^2 = 148$ switches would be needed to identify all the regulation functions. Both predictions are low by approximately the same amount, $\frac{112}{264} = 42\%$ and $\frac{148}{343} = 43\%$. The link that required the most data to identify was LFY→AG. In part, this is because LFY can influence AG only if TFL1 and AP1 are high and LUG is low. Furthermore, LFY activates AG only when LFY is high. This is comparatively rare, as LFY converges to zero in all trajectories.

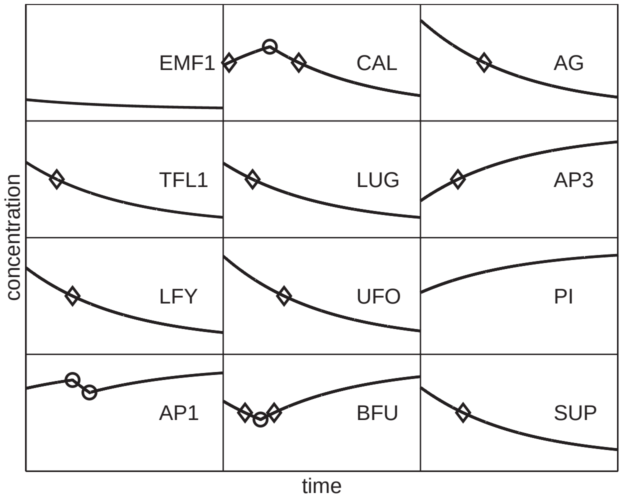

Fig. 8. A simulated time series of the *Arabidopsis thaliana* model. The curves represent the concentrations of the 12 species over time. Diamonds mark changes in the logical state of a species. Circles mark changes in the production rate of a species. This time series approaches the attractor 000100010110, which corresponds to a petal cell.

6. Discussion

We have examined the problem of inferring dynamical models of gene expression in which the time derivatives of chemical concentrations are expressed by logical rules. Such models are able to capture complex, combinatorial relationships between the concentration or rate of production of a gene’s products and the concentrations of the regulators of the gene (McAdams and Shapiro, 1995; Yuh et al., 1998).

One of the contributions of this paper is to describe computationally efficient algorithms for network inference. A network of N chemical species, each regulated by K species, can be connected in $\binom{N}{K}^N \approx N^{NK}$ different ways. Many algorithms that explicitly extract regulatory relationships search directly in this exponentially large space of network structures, either exhaustively or using local search heuristics (Akutsu et al., 1999; Ideker et al., 2000; Liang et al., 1998). Running these algorithms can be a significant computational burden.

Our algorithms are predicated upon a number of simplifying assumptions. Clearly, data in which there is noise, time delays, an unknown number of regulators for each species, and variable and unknown thresholds would be more difficult or impossible to analyse using the methods outlined here. For example, if the data is noisy, no single pair of observations is sufficient to definitively infer a regulatory relationship. However, if the same regulatory link is suggested by many independent pairs of observations, one begins to have confidence that the link is real. We used the assumption that concentrations are observed continuously in two ways: (i) to estimate time derivatives of expression, and hence production rates, and (ii) to pinpoint switches in the logical state of a gene. The latter is most important for inference by Rule 2; Rule 1 does not require that all switches be observed. However, both rules need the time derivative information, which no expression monitoring technology can report directly. Worse, expression experiments are often designed with samples spread evenly over some time interval, in order to maximize coverage. This makes it difficult to estimate time derivatives. The importance of time derivative information suggests an alternative strategy for expression experiment design. At least some expression samples should be packed closely in time so that expression levels and their time derivatives could be estimated well. A key assumption of Eq. (1) is that there are only two (or more generally, just a few) possible production rates for each gene. For some genes, such as developmental genes, which show clear patterns of turning "on" or "off", this may be a tenable assumption. Eq. (1) is not a good model for genes which exhibit graded responses. These issues would need to be addressed before applying the inference algorithms we propose to real gene expression data.

A number of features of real genetic control networks might simplify the inference problem. For example, there is growing recognition that genetic control might be modular (Alon, 2003; Von Dassow et al., 2000), and this would restrict the set of possible network structures. Kauffman and colleagues have suggested that logical functions controlling transcription tend to be a small subset, called canalyzing functions, of the possible logical functions (Harris et al., 2002). Finally, methods based on time series analysis can be supplemented by data from other sources, such as observations of the network under perturbed environmental or genetic conditions, genomic data or localization (chip-on-chip) data (Akutsu et al., 1998; Ideker et al., 2000; Segal et al., 2002).

A second contribution of the paper is that we have analysed the data requirements of our inference procedures. We analytically derived predictions for randomly generated networks based on their size and connectivity, and performed simulation experiments on randomly generated networks as well as on a more realistic network that models floral morphogenesis in *Arabidopsis thaliana*. A surprising result of our analytical predictions, verified by simulation, is that for randomly generated networks, the amount of data needed to infer the structure of the network scales only logarithmically with the number of regulators per species. For example, to identify all the regulators of a species with 100 regulators takes only twice as much data as to identify all the regulators of a species with 10 regulators. In contrast, theoretical analysis and simulations of Boolean network inference suggest that inferring the structure requires an amount of data that is exponential in the number of regulators (Akutsu et al., 1999). We made two major assumptions that differ from the analysis of Akutsu et al. We assumed time series data in which one gene at a time changes logical state, and we assumed that the regulation functions are chosen randomly from the set of all Boolean functions. It is not yet clear if both assumption are necessary or if only one of the two results in such a dramatic difference in the data requirements.

Our analysis relied on statistical properties of randomly generated networks in which every gene has K regulators for some fixed K. Other models may better capture the structure of real genetic networks. For example, there is some evidence that transcriptional regulatory networks show a scale-free degree distribution (Lee et al., 2002). The structure of scale-free networks can result in much different dynamical properties than seen with fixed-indegree networks (Aldana and Cluzel, 2003; Oosawa and Savageau, 2002). We performed a small number of simulation experiments to determine how the conclusions of the current work might change for scale-free networks. We found some evidence that the expected amount of data needed for structure identification scales as in Eq. (2), where K is taken to be the mean number of regulators per gene. However, identifying regulation functions takes at least $2^{K_{max}}$ logical switches, where K_{max} is the maximum number of regulators for any gene. For scale-free networks, this can be much greater than the mean number of regulators. Determining a succinct set of parameters which characterize, for any type of network structure, the amount of data required for inference remains an open problem.

No gene expression-monitoring technology simultaneously measures the concentrations of all chemical species related to gene expression (mRNAs, proteins, small molecules, etc.). While this fact is well-recognized, many methods being proposed for network inference make no allowance for the influence of unobserved factors. As a third contribution, we have introduced a "partially observed" version of the network inference problem, in which concentration values are reported for only a subset of the species in the network. We showed that the structure of the visible portion of the network can be identified. We also showed that the total size of the network (observed and unobserved species) as well as the number of regulators per species can be inferred from the observable data. We did this for a particular kind of randomly generated network in which each species has exactly K regulators chosen uniformly at random. It remains to be seen whether similar inferences could be made for other types of randomly generated networks, such as scale-free networks.

Despite reservations about the applicability of these computations to real biological networks, the current work demonstrates that it is possible to infer the dynamics of very large model networks and to estimate the amount of data required to do so. These results underscore the importance and value of collecting data in which gene expression is monitored over time, and show how this data might be used to relate the patterns of activity of genes to the underlying logical structure of the network controlling gene activity.

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.

References

Akutsu, T., Kuhara, S., Maruyama, O., Miyano, S., 1998. Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. In: Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms, pp. 695–702.

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 Pacific Symposium Biocomputing (PSB'99), pp. 17–28.

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

Alon, U., 2003. Biological networks: the tinkerer as an engineer. Science 301, 1866–1867.

Bodnar, J., 1997. Programming the *Drosophila* embryo. J. Theor. Biol. 188, 391–445.

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. Compl. 6 (4), 312–340.

Chen, T., He, H.L., Church, G.M., 1999. Modeling gene expression with differential equations. In Pacific Symposium on Biocomputing, pp. 29–40.

De Jong, H., Geiselmann, J., Hernandez, C., Page, M., 2003. Genetic network analyzer: qualitative simulation of genetic regulatory networks. Bioinformatics 19 (3), 336–344.

D'Haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R., 1999. Linear modeling of mRNA expression levels during CNS development and injury. In Pacific Symposium on Biocomputing, pp. 41–52.

Ding, C., Cantor, C.R., 2003. A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. Proc. Natl Acad. Sci. USA 100 (6), 3059–3064.

Edwards, R., Glass, L., 2000. Combinatorial explosion in model gene networks. Chaos 10, 691–704.

Elowitz, M.B., Leibler, S., 2000. A synthetic oscillatory network of transcriptional regulators. Nature 403, 335–338.

Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J., 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301, 102–105.

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

Glass, L., Kauffman, S.A., 1973. The logical analysis of continuous, non-linear biochemical control networks. J. Theor. Biol. 39, 103–129.

Harris, S.E., Sawhill, B.K., Wuensche, A., Kauffman, S., 2002. A model of transcriptional regulatory networks based on biases in the observed regulation rules. Complexity 7 (4), 23–40.

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

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

Kosman, D., Reinitz, J., Sharp, D.H., 1998. Automated assay of gene expression at cellular resolution. In Pacific Symposium on Biocomputing, pp. 6–17.

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., 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 Pacific Symposium on Biocomputing, pp. 18–29.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E.L., 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnol. 14 (12), 1675–1680.

McAdams, H.H., Shapiro, L., 1995. Circuit simulation of genetic networks. Science 269, 650–656.

Mendoza, L., Alvarez-Buylla, E.R., 1998. Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. J. Theor. Biol. 193, 307–319.

Mestl, T., Plahte, E., Ornholt, S.W., 1995. A mathematical framework for describing and analysing gene regulatory networks. J. Theor. Biol. 176, 291–300.

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.

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

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

Segal, E., Barash, Y., Simon, I., Friedman, N., Koller, D., 2002. From promoter sequence to expression: a probabilistic framework. In: Proceedings of the Sixth Annual International Conference on Computational Molecular Biology (RECOMB).

Somogyi, R., Sniegoski, C.A., 1996. Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. Complexity 1, 45–63.

Tegner, J., Yeung, M.K.S., Hasty, J., Collins, J.J., 2003. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proc. Natl Acad. Sci. 100, 5944–5949.

Thomas, R., D'Ari, R., 1990. Biological Feedback. CRC, Boca Raton, FL.

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., 1995. Serial analysis of gene expression. Science 270, 484–487.

Von Dassow, G., Meir, E., Munro, E.M., Odell, G.M., 2000. The segment polarity network is a robust developmental module. Nature 406, 188–192.

Yeung, M.K.S., Tegner, J., Collins, J.J., 2002. Reverse engineering gene networks using singular value decomposition and robust regression. Proc. Natl Acad. Sci. 99, 6163–6168.

Yuh, C.-H., Bolouri, H., Davidson, E.H., 1998. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279, 1896–1902.