JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 12, Number 2, 2005

Pp. 170–185

Explore Biological Pathways from Noisy Array Data by Directed Acyclic Boolean Networks

LEI M. LI[1] and HENRY HORNG-SHING LU[2]

ABSTRACT

We consider the structure of directed acyclic Boolean (DAB) networks as a tool for exploring biological pathways. In a DAB network, the basic objects are binary elements and their Boolean duals. A DAB is characterized by two kinds of pairwise relations: similarity and prerequisite. The latter is a partial order relation, namely, the on-status of one element is necessary for the on-status of another element. A DAB network is uniquely determined by the state space of its elements. We arrange samples from the state space of a DAB network in a binary array and introduce a random mechanism of measurement error. Our inference strategy consists of two stages. First, we consider each pair of elements and try to identify their most likely relation. In the meantime, we assign a score, s-p-score, to this relation. Second, we rank the s-p-scores obtained from the first stage. We expect that relations with smaller s-p-scores are more likely to be true, and those with larger s-p-scores are more likely to be false. The key idea is the definition of s-scores (referring to similarity), p-scores (referring to prerequisite), and s-p-scores. As with classical statistical tests, control of false negatives and false positives are our primary concerns. We illustrate the method by a simulated example, the classical arginine biosynthetic pathway, and show some exploratory results on a published microarray expression dataset of yeast *Saccharomyces cerevisiae* obtained from experiments with activation and genetic perturbation of the pheromone response MAPK pathway.

Key words: microarray, pathway, Boolean networks, measurement error, EM algorithm.

1. INTRODUCTION

ONE GREAT CHALLENGE OF POSTGENOMIC RESEARCH is to identify complex biological networks and pathways from genomewide data such as DNA sequences and expression profiles. This includes metabolic pathways, protein–protein interaction networks, gene regulatory pathways, etc. Along with biological methods such as phylogenetic profile and Rosetta Stone (see Eisenberg *et al.* [2000] and McGuire and Church [2000]), computational methods have been developed as powerful data-mining tools in the study of genomics.

[1]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089.
[2]Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan.

Clustering is one such important technique to group genes and samples from microarray data; see Eisen *et al.* (1998), Ben-Dor *et al.* (1999), Alon *et al.* (1999). A central component of a clustering algorithm is the definition of similarity scores, either from a biological perspective or from a statistical perspective. We note that the relation of similarity between two biological elements such as proteins or genes is symmetric in nature. On the other hand, a biological process may include a cascade of reactions to environmental factors and regulation of protein syntheses. Thus, concepts other than similarity are necessary for a complete description of pathways.

Data type is another consideration in the modeling of networks. In this article, we consider binary variables because we can always discretize continuous variables. In the presence of noise, careful discretization can even denoise to some degree. One such example can be found in Xing and Karp (2001). The use of Boolean networks has a long history in the literature. Kauffman (1977, 1979) considered a dynamic version of Boolean networks. A review of models of genetic regulatory systems including Boolean networks can be found in De Jong (2002). Based on the structure of Boolean networks, we introduce a new model for measurement error and propose a simple technique to infer pairwise relations between elements from noisy array data.

We note that the Bayesian networks model is a much more sophisticated and complete model to describe biological pathways than the method proposed in this article. For example, variables in a Bayesian networks model can be either discrete or continuous. Bayesian networks is a structure that contains directed relations among elements. It has been extensively studied in the last two decades; see Pearl (1988) and Jensen (1996). Its structure is characterized by two components. The first component is a directed acyclic graph whose vertices correspond to random variables. The second component describes a conditional distribution for each variable, given its parents in the graph. Murphy and Mian (1999) and Friedman *et al.* (2000) applied Bayesian network models to analyze microarray expression data. The family of Bayesian networks is fairly large, and the number of DAGs is superexponential. Although some algorithms searching for Bayesian networks have been developed (see Heckerman *et al.* [1995] and Spirtes *et al.* [2000b]), the learning of Bayesian networks is a challenging task without a priori knowledge. Also, to achieve high accuracy of estimation, sample sizes of several hundred are required even for relatively sparse graphs; see Spirtes *et al.* (2000a). The simple model considered in this article takes some aspects of Bayesian networks and serves as a tool of exploratory data analysis for array data.

Specifically, we consider the structure of directed acyclic Boolean (DAB) networks as a tool for exploring biological pathways. In a DAB networks model, the basic objects are binary elements and their Boolean duals. A DAB is characterized by two kinds of pairwise relations: similarity and prerequisite. The former represents a pair of elements with identical on–off states. The latter is a partial order relation, namely, the on-status of one element is prerequisite for the on-status of another element. A DAB networks model is uniquely determined by its state space: all possible on–off states subject to the pairwise relations. We arrange samples from the state space of a DAB network in a binary array and then introduce a random mechanism of measurement error. This results in a noisy array. Our goal is to reconstruct the DAB networks from the noisy array data.

Our inference strategy consists of two stages. First, we consider each pair of elements and try to identify their most likely relation. In the meantime, we assign a score, s-p-score, to this relation. Second, we rank the s-p-scores obtained from the first stage. We expect that those relations with smaller s-p-scores are more likely to be true, and those with larger s-p-scores are more likely to be false. The key idea is the definition of s-scores (referring to **similarity**), p-scores (referring to **prerequisite**), and s-p-scores (by model selection). As with classical statistical tests, control of false negatives and positives are our primary concerns.

The s-p-scoring method is one kind of exploratory data analysis and focuses on pairwise relations. After the ranking of pairwise relations, experts' knowledge may be incorporated. Depending on the data, we expect to reconstruct all or partial substructures of a network. If we set an upper bound to the number of E-M iterations involved, the computational complexity of the procedure is $O(m^2 \log m)$, where m is the number of elements in a network.

The rest of the paper is organized as follows. In Section 2, we describe the structure of the model. In Section 3, we explain the s-p-scoring method. In Section 4, we illustrate the method by a simulated example, the classical arginine biosynthetic pathway, and show some exploratory results on the yeast *Saccharomyces cerevisiae* pheromone response MAPK pathway using an expression dataset obtained from experiments with activation and genetic perturbation. In Section 5, we discuss some relevant issues.

2. THE MODEL

The structure of directed acyclic Boolean (DAB) networks

Suppose we are concerned with m elements, $G_1, G_2, \ldots, G_m$, each taking two states: on and off. These elements are abstracts of biological objects such as genes, mRNAs, proteins, environmental conditions, or a mixture of them. If an element is measured on a continuous scale or has more than two expression levels, then we need to discretize it and encode it by binary variables. We will come back to this issue later. The theory of directed graphs is helpful for the description of our model; we refer readers to Brightwell (1997) for relevant results on this subject. We generate a graph with $2m$ vertices or nodes, $G_1, G_2, \ldots, G_m$, and their Boolean duals $\bar{G}_1, \bar{G}_2, \ldots, \bar{G}_m$, representing on-and-off states of the m elements, and this is referred to as the ground-set. We refer to a node A and its dual $\bar{A}$ as a Boolean pair.

We define a prerequisite relation between a pairs of elements A and B as follows: A is prerequisite for B if the on-status of A is necessary for the on-status of B, and we denote it by $A \prec B$. The prerequisite relation is a partial order. It is transitive on the ground-set, namely, $A \prec C$ and $C \prec B$ implies $A \prec B$. Also it is irreflexive in the sense that we never have $A \prec \bar{A}$. In addition, we assume that the dual of each partial order relation is also true; i.e., $\bar{B} \prec \bar{A}$ is true if and only $A \prec B$ is true. Similarly, we have the following three pairs of dual relations: $\bar{A} \prec \bar{B}$ with $B \prec A$; $A \prec \bar{B}$ with $B \prec \bar{A}$; and $\bar{A} \prec B$ with $\bar{B} \prec A$. We graphically represent a partial relation $A \prec B$ by drawing an arrow from the vertex A to B. It is not economical to include all the arcs in the directed graph due to the transitive property of partial orders. An ordered pair (A, B) is called a covering pair if there exists no vertex C such that $A \prec C$ and $C \prec B$. Thus, it suffices to represent all partial orders by arrows between covering pairs, and this is referred to as the diagram of the directed graph. It is well known that the diagram of a partial order is acyclic. In addition, no path exists to connect a Boolean pair in the diagram of a DAB because we never have $A \prec \bar{A}$.

Another relation between pairs of elements is similarity. Two elements A and B are *similar* if they are on and off simultaneously, and this is denoted by $A \sim B$. They are *negatively similar* if they are on and off in the opposite way, and this is denoted by $A \sim \bar{B}$. In the absence of measurement error, it is a trivial relation. But in practice, the presence of measurement error complicates the situation, and it needs to be inferred from the data.

We use "—" to connect two similar elements in the diagram. Figure 1 shows a directed acyclic Boolean network, which has seven elements with one similar and eleven prerequisite relations. Another way to identify a DAB is to consider the on–off states of its elements. There are in total $2^7 = 128$ states for a seven-element DAB. Only 13 of these states are compatible with the 12 pairwise relations in the above example. We enumerate them in Table 1, where "0" and "1" represent "off" and "on," respectively. It is a subset of the 128 states. In general, a directed acyclic Boolean network consisting of m elements corresponds to a unique subset of all 2^m states. Even though not every subset of the 2^m states corresponds to a directed acyclic Boolean network, the number of DABs, like the number of DAGs, is superexponential.

Consider n samples generated from a directed acyclic Boolean network; i.e., we sample with replacement from the state space compatible with the networks. Table 1 shows the compatible states for the above example. We arrange the data in a matrix (y_{ij}), where $i = 1, \ldots, n$, $j = 1, \ldots, m$, whose entries take

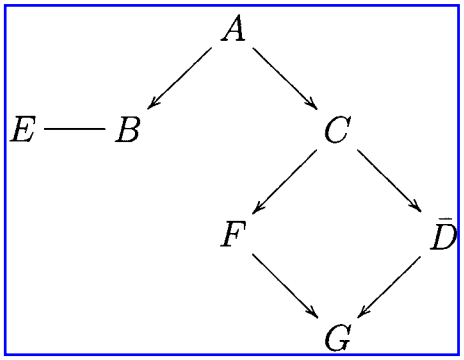

FIG. 1. Diagram of a directed acyclic Boolean network with seven elements and twelve pair relations. Only arrows between covering pairs are shown.

TABLE 1. THE TABLE OF STATES FOR DIRECTED ACYCLIC BOOLEAN NETWORK SHOWN IN FIG. 1

Case	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
A	0	1	1	1	1	1	1	1	1	1	1	1	1
B	0	0	1	1	1	1	1	1	0	0	0	0	0
C	0	0	0	1	1	1	1	1	1	1	1	1	1
D	1	1	1	1	0	1	0	0	1	0	1	0	0
E	0	0	1	1	1	1	1	1	0	0	0	0	0
F	0	0	0	0	0	1	1	1	0	0	1	1	1
G	0	0	0	0	0	0	0	1	0	0	0	0	1

TABLE 2. 2×2 TABLES FOR A PAIR OF ELEMENTS ASSUMING NO MEASUREMENT ERROR[a]

A / B	*0*	*1*	*A / B*	*0*	*1*
0	m_{00}	m_{01}	0	q_{00}	q_{01}
1	m_{10}	m_{11}	1	q_{10}	q_{11}

[a]The counts on the left are regarded as being generated from the multinomial distribution on the right.

TABLE 3. COUNT PATTERNS FOR THE SIX PAIRWISE RELATIONS ASSUMING EXHAUSTIVE SAMPLING AND NO MEASUREMENT ERROR

$A \sim B$			$A \sim \bar{B}$			$A \prec B,\ \bar{B} \prec \bar{A}$		
A / B	*0*	*1*	*A / B*	*0*	*1*	*A / B*	*0*	*1*
0	+	0	0	0	+	0	+	0
1	0	+	1	+	0	1	+	+
$\bar{A} \prec \bar{B},\ B \prec A$			$A \prec \bar{B},\ B \prec \bar{A}$			$\bar{A} \prec B,\ \bar{B} \prec A$		
A / B	*0*	*1*	*A / B*	*0*	*1*	*A / B*	*0*	*1*
0	+	+	0	0	+	0	+	+
1	0	+	1	+	+	1	+	0

values of either 0 or 1. Table 1 is the transpose of (y_{ij}), and each row corresponds to an element and each column corresponds to a sample.

Without measurement error, we can reconstruct the directed acyclic Boolean network in Fig. 1 from Table 1 by identifying all the pairs with prerequisite or similar relations. This is carried out by the following procedure. For each pair of elements, say, A and B, we count the four incidences of (A, B) being $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$ from the corresponding columns of (y_{ij}) and arrange them in a 2×2 table; see the left of Table 2. We mark a cell "+" if the count is positive and mark it "0" otherwise. Consequently, the six relations are characterized by the count patterns in Table 3.

Next, we consider the issue of selection bias. In practice, we sample from all the possible states compatible with a directed acyclic Boolean network. In the above example, we have only 13 cases. When m is large, this number could be large, and possibly only a fraction of them are sampled. Then the issue of estimableness arises. If we cannot have an exhaustive sample, i.e., some compatible states are missed in observation, then the count strategy described above may lead to false positive pairwise relations, either prerequisite or similarity. For example, if case 3 in Table 1 is missed from observations, then the count strategy indicates $C \prec B$, which is not consistent with the truth. Nevertheless, this strategy will not lead to any false negatives in the absence of measurement error.

Measurement error

Next we introduce a mechanism of measurement error to the data sampled from a directed acyclic Boolean network. This results in a more practical model for many biological data, such as expression levels. We assume that each entry in (y_{ij}) is switched to its opposite value according to a misclassification probability p, independently with one another; i.e.,

$$x_{ij} = \begin{cases} y_{ij} & \text{with probability } 1-p\,, \\ 1-y_{ij} & \text{with probability } p\,. \end{cases}$$

This creates the noisy array (x_{ij}), which are the observations.

Problem and pairwise structure

Our goal is to reconstruct the directed acyclic Boolean network from the array of binary data (x_{ij}). It is clear that the problem is equivalent to identifying all the pairs of elements with estimable similarity or prerequisite relations.

3. METHOD

Our inference strategy consists of two stages. First, we consider each pair of elements and try to find their most likely relation. In the meantime, we assign a score, s-p-score, to this relation. Second, we rank the s-p-scores obtained from the first stage. We expect that those relations with smaller s-p-scores are more likely to be true, and those with larger s-p-scores are more likely to be false.

Probabilistic models for 2 by 2 tables

To deal with measurement error, we resort to probabilistic models. Instead of a full model including every element, we consider pairwise models in the first stage. The count data in the 2×2 table on the left of Table 2 can be thought of as being generated from a multinomial distribution with four cells whose probabilities are q_{00}, q_{01}, q_{10}, q_{11}, respectively, as shown on the right of Table 2, where $q_{00}+q_{01}+q_{10}+q_{11}=1$. Then the six types of relations between elements A and B are reformulated as hypotheses on the probability patterns; see Table 4. Please notice that $(q_{00}, q_{01}, q_{10}, q_{11})$ depend on both the structure of the DAB network and the sampling scheme.

Similarly to (y_{ij}), we extract the data in (x_{ij}) for each pair of elements, say, A and B, and arrange them on the left of Table 5. Now the counts n_{00}, n_{01}, n_{10}, n_{11} are not generated from the multinomial $(q_{00}, q_{01}, q_{10}, q_{11})$, but from another multinomial $(r_{00}, r_{01}, r_{10}, r_{11})$ as shown on the right of Table 5, where $r_{00}+r_{01}+r_{10}+r_{11}=1$.

Missing data structure

With measurement error, a part of m_{00} may leak to the other three cells. We denote the redistributed counts from m_{00} to the four cells by $m_{00,00}$, $m_{00,01}$, $m_{00,10}$, $m_{00,11}$. Analogous notation is defined for m_{01},

TABLE 4. THE SIX PAIRWISE RELATIONS, THEIR CORRESPONDING PROBABILISTIC HYPOTHESES AND S-SCORES, P-SCORES

	Relation	*Hypothesis*	*Scores*
Diagonal	$A \sim B$	$q_{01}=q_{10}=0$	$s_{A\sim B}$
Similarity	$\bar{A} \sim B$	$q_{00}=q_{11}=0$	$s_{\bar{A}\sim B}$
Triangular	$A \prec B$	$q_{01}=0$	$p_{A\prec B}$
Prerequisite	$\bar{A} \prec \bar{B}$	$q_{10}=0$	$p_{\bar{A}\prec\bar{B}}$
	$A \prec \bar{B}$	$q_{00}=0$	$p_{A\prec\bar{B}}$
	$\bar{A} \prec B$	$q_{11}=0$	$p_{\bar{A}\prec B}$

TABLE 5. THE 2×2 COUNT TABLE FOR A PAIR OF ELEMENTS AND THEIR GENERATING PROBABILITIES IN THE PRESENCE OF MEASUREMENT ERROR

A / B	*0*	*1*	*A / B*	*0*	*1*
0	n_{00}	n_{01}	0	r_{00}	r_{01}
1	n_{10}	n_{11}	1	r_{10}	r_{11}

TABLE 6. SPLITTING COUNTS CAUSED BY MISCLASSIFICATION ERROR

A/B	*0*		*1*	
0	$m_{00,00}$	$m_{00,01}$	$m_{01,00}$	$m_{01,01}$
	$m_{00,10}$	$m_{00,11}$	$m_{01,10}$	$m_{01,11}$
1	$m_{10,00}$	$m_{10,01}$	$m_{11,00}$	$m_{11,01}$
	$m_{10,10}$	$m_{10,11}$	$m_{11,10}$	$m_{11,11}$

TABLE 7. SPLITTING PROBABILITIES CAUSED BY MISCLASSIFICATION ERROR

A/B	*0*		*1*	
0	$q_{00,00} = (1-p)^2 q_{00}$	$q_{00,01} = p(1-p)\,q_{00}$	$q_{01,00} = p(1-p)\,q_{01}$	$q_{01,01} = (1-p)^2\,q_{01}$
	$q_{00,10} = p(1-p)\,q_{00}$	$q_{00,11} = p^2\,q_{00}$	$q_{01,10} = p^2\,q_{01}$	$q_{01,11} = p(1-p)\,q_{01}$
1	$q_{10,00} = p(1-p)\,q_{10}$	$q_{10,01} = p^2\,q_{10}$	$q_{11,00} = p^2\,q_{11}$	$q_{11,01} = p(1-p)\,q_{11}$
	$q_{10,10} = (1-p)^2\,q_{10}$	$q_{10,11} = p(1-p)\,q_{10}$	$q_{11,10} = p(1-p)\,q_{11}$	$q_{11,11} = (1-p)^2\,q_{11}$

m_{10}, and m_{11}. This splitting pattern is shown in Table 6. Correspondingly, their generating probabilities (q_{00}, q_{01}, q_{10}, q_{11}) are redistributed as shown in Table 7, where we adopt the notation $q_{ij,kl}$ analogous to $m_{ij,kl}$. The two sets of counts and probabilities are linked as follows:

$$\begin{cases} n_{ij} = \displaystyle\sum_{k,l=0,1} m_{kl,ij}\,, \\ r_{ij} = \displaystyle\sum_{k,l=0,1} q_{kl,ij}\,, \end{cases} \tag{1}$$

and

$$\begin{cases} m_{kl} = \displaystyle\sum_{i,j=0,1} m_{kl,ij}\,, \\ q_{kl} = \displaystyle\sum_{i,j=0,1} q_{kl,ij}\,. \end{cases}$$

MLE and the E-M algorithm

The log-likelihood of the data is given, up to a constant, by the following

$$L = \sum_{i,j=0,1} n_{ij} \log r_{ij}\,, \tag{2}$$

where the probabilities r_{ij}'s are computed according to (1) and Table 7. Later we define s-scores and p-scores via maximum likelihood estimates (MLE). Except for a constant, the log-likelihood of the full data $\{m_{ij,kl}\}$ is given by

$$\sum_{i,j,k,l=0,1} m_{ij,kl} \log q_{ij,kl} , \tag{3}$$

where $q_{ij,kl}$ are those splitting probabilities in Table 7.

To estimate the MLE, the celebrated E-M algorithm maximizes the likelihood of full data (3) rather than that in (2); see Dempster *et al.* (1977) and McLachlan and Krishnan (1997). In the E-step, we impute the splitting counts by their conditional expectations calculated at the current value of the parameter by the formula

$$E_{(p,q_{00},q_{01},q_{10},q_{11})}(m_{ij,kl}|n_{kl}) = \frac{n_{kl}\, q_{ij,kl}}{\sum_{i',j'=0,1} q_{i'j',kl}} , \tag{4}$$

where $i, j, k, l = 0, 1$. Under different hypotheses specified in Table 4, one or two probabilities of q_{00}, q_{01}, q_{10}, and q_{11} are zero. In the M-step, we update the value of the parameter by maximizing the conditional expectation of the log-likelihood for the full data; See Li and Lu (2001) for details.

Pairwise scores

We first consider a problem simpler than reconstructing a DAB network: what is the most likely relation for a pair of elements?

Definition 1. *For a pair of elements A and B,*

- *the* s-scores $s_{A\sim B}$ *and* $s_{A\sim \bar{B}}$ *are, respectively, the maximum likelihood estimates of p under the diagonal model:* $q_{01} = q_{10} = 0$ *and* $q_{00} = q_{11} = 0$*;*
- *the* p-scores $p_{A\prec B}$, $p_{\bar{A}\prec \bar{B}}$, $p_{A\prec \bar{B}}$, *and* $p_{\bar{A}\prec B}$ *are, respectively, the maximum likelihood estimates of p under the triangular model:* $q_{01} = 0$, $q_{10} = 0$, $q_{00} = 0$, *and* $q_{11} = 0$*; cf. Table 4.*

We compute s-scores and p-scores by the E-M algorithm described earlier. The heuristic of the definition is that we use the MLE $\hat{p}$ to measure the goodness of fit of each hypothesis: the smaller the score, the more support to the corresponding hypothesis.

Next we need to choose one score out of the two s-scores and four p-scores for a pair of elements. In other words, we need to select the hypothesis that is most consistent with the data. This is a problem of model selection; see Schwarz (1978).

Definition 2. *For a pair of elements A and B,*

- *between the two diagonal models, select the one that achieves the smaller s-score;*
- *among the four triangular models, select the one that achieves the smallest p-score;*
- *for the diagonal model corresponding to the smaller s-score and the triangular model corresponding to the smallest p-score, we compare their corresponding BIC values, namely, the penalized log-likelihoods as follows:*

$$BIC = -\log likelihood + \frac{d \log n}{2} ,$$

where n is the sample size and d is the number of parameters. This number is two for a diagonal model and is three for a triangular model. We choose the model with the smaller BIC value as the most likely relation for the pair A and B, and define their s-p-score *to be the score corresponding to the most likely relation.*

Please notice that s-p-score is one of the s-scores and p-scores and BIC values are used only to choose the hypothesis. It is easy to understand why we select the smallest s-score and p-score. Notice that each diagonal model is nested in two triangular models. To make the choice between a diagonal and a triangular model, we need to take into account model complexity. We here adopt the technique of BIC for model selection.

The basic idea of most powerful statistical tests is to minimize the chance of type II error (false positive) subject to a constraint on the chance of type I error (false negative); see Lehmann (1986). Even though the classical theory of hypothesis testing does not directly apply to our situation, its rationale remains our guide. For each hypothesis in Table 4, we expect that the s-score or p-score has the following property: it is a good estimate of the parameter p when the hypothesis is true; whereas it is considerably biased upward when the hypothesis is false.

Accuracy of estimation and control of false negative

We next consider the statistical behavior of the s-scores and p-scores under the null hypothesis. Without loss of generality, we take the hypothesis $q_{01} = 0$, for example. Notice that this is a composite hypothesis. In general, the maximum likelihood estimate in a regular setting is both consistent and efficient; see Bickel and Doksum (1977).

Proposition 1. *Suppose that the hypothesis $A \prec B$, i.e., $q_{01} = 0$ holds. Then, except for the singular point at $q_{00} = q_{11} = 0$, the maximum likelihood estimate of p has the property of asymptotical normality, i.e.,*

$$\sqrt{n}\,[\hat{p} - p, \hat{q}_{00} - q_{00}, \hat{q}_{10} - q_{10}, \hat{q}_{11} - q_{11}] \longrightarrow N(0, I^{-1}),$$

where I is the Fisher information matrix,

$$I = -\begin{pmatrix} E\left[\frac{\partial^2 logL}{\partial p^2}\right] & E\left[\frac{\partial^2 logL}{\partial p \partial q_{00}}\right] & E\left[\frac{\partial^2 logL}{\partial p \partial q_{10}}\right] & E\left[\frac{\partial^2 logL}{\partial p \partial q_{11}}\right] \\ E\left[\frac{\partial^2 logL}{\partial p \partial q_{00}}\right] & E\left[\frac{\partial^2 logL}{\partial q_{00}^2}\right] & E\left[\frac{\partial^2 logL}{\partial q_{00} \partial q_{10}}\right] & E\left[\frac{\partial^2 logL}{\partial q_{00} \partial q_{11}}\right] \\ E\left[\frac{\partial^2 logL}{\partial p \partial q_{10}}\right] & E\left[\frac{\partial^2 logL}{\partial q_{00} \partial q_{10}}\right] & E\left[\frac{\partial^2 logL}{\partial q_{10}^2}\right] & E\left[\frac{\partial^2 logL}{\partial q_{10} \partial q_{11}}\right] \\ E\left[\frac{\partial^2 logL}{\partial p \partial q_{11}}\right] & E\left[\frac{\partial^2 logL}{\partial q_{00} \partial q_{11}}\right] & E\left[\frac{\partial^2 logL}{\partial q_{00} \partial q_{10}}\right] & E\left[\frac{\partial^2 logL}{\partial q_{11}^2}\right] \end{pmatrix}.$$

It will take more than 10 pages to write down the expression of I^{-1}. In fact, the computation was carried out by the symbolic calculation in MAPLE. Here we choose to give only the term corresponding to the parameter p as follows:

$$\frac{p(1-p)(3p^2q_{00} + 3p^2q_{11} - p^2q_{10} - 3pq_{00} - 3pq_{11} + pq_{10} + q_{11} + q_{00})}{n(4p^2q_{11}^2 + 4p^2q_{00}^2 + 8p^2q_{00}q_{11} - 4pq_{11}^2 - 4pq_{00}^2 - 8q_{00}pq_{11} + 2q_{00}q_{11} + q_{11}^2 + q_{00}^2)}. \tag{5}$$

In Fig. 2, we plot the element of I^{-1} corresponding to p as a function of q_{00} and q_{01} in which p is fixed to be 0.05. The only singularity point occurs at $q_{10} = 1$ and $q_{00} = q_{11} = q_{01} = 0$. In this case, one element is house-keeping (on all the time), and the other one is silent (off all the time). By filtering out silent and house-keeping elements, we can eliminate this kind of singularity for the sake of inference. Consequently, we can find a bound on the inverse of the Fisher information matrix, and this means that the p-score will be around p within an order $1/\sqrt{n}$ radius asymptotically.

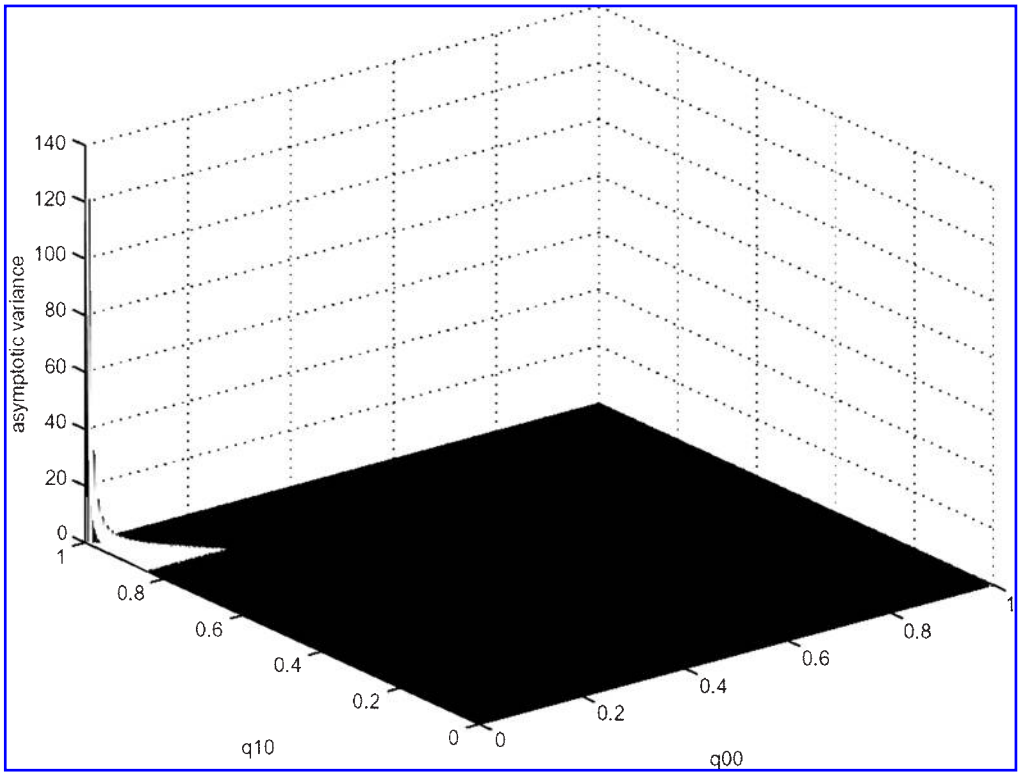

FIG. 2. The asymptotic variance of the MLE of p when $p = 0.05$. One singularity point occurs at $q_{10} = 1$ and $q_{00} = q_{11} = q_{01} = 0$.

Control of false positives

Next we look at how the p-score $p_{A \prec B}$ behaves under the alternatives: $q_{01} > 0$ versus the null $q_{01} = 0$. We study the asymptotic bias of the MLE.

Proposition 2. *Let the parameters in the true model be $(p, q_{00}, q_{01}, q_{10}, q_{11})$, where $q_{01} > 0$. As the sample size $n \to \infty$, the MLE $(\tilde{p}, \tilde{q}_{00}, \tilde{q}_{01}, \tilde{q}_{10}, \tilde{q}_{11})$ subject to $\tilde{q}_{01} = 0$ is given by the value that minimizes the Kullback–Leibler divergence between the null and alternative:*

$$D[\{p, q_{00}, q_{01}, q_{10}, q_{11}\} || \{\tilde{p}, \tilde{q}_{00}, \tilde{q}_{01} = 0, \tilde{q}_{10}, \tilde{q}_{11})\}] = D[\{r_{ij}\} || \{\tilde{r}_{ij}\}] = \sum_{i,j=0,1} [-r_{ij} \log \tilde{r}_{ij} + r_{ij} \log r_{ij}],$$

where $\{r_{ij}\}$ and $\{\tilde{r}_{ij}\}$ are respectively defined by $\{p, q_{00}, q_{01}, q_{10}, q_{11}\}$ and $\{\tilde{p}, \tilde{q}_{00}, \tilde{q}_{01} = 0, \tilde{q}_{10}, \tilde{q}_{11}\}$ via (1) and Table 7.

Proof. The concept of Kullback–Leibler divergence can be found in Cover and Thomas (1991). The proof lies in the connection between likelihood and Kullback–Leibler divergence. When $n \longrightarrow \infty$, $n_{ij}/n \longrightarrow r_{ij}$, and maximizing the quantity in (2) becomes maximizing the following:

$$\sum_{i,j=0,1} n\, r_{ij} \log \tilde{r}_{ij}\,,$$

over $\{\tilde{r}_{ij}\}$. This is equivalent to minimizing

$$\sum_{i,j=0,1} [-r_{ij} \log \tilde{r}_{ij} + r_{ij} \log r_{ij}]\,,$$

which is $D[\{r_{ij}\} || \{\tilde{r}_{ij}\}]$. Thus we complete the proof. ■

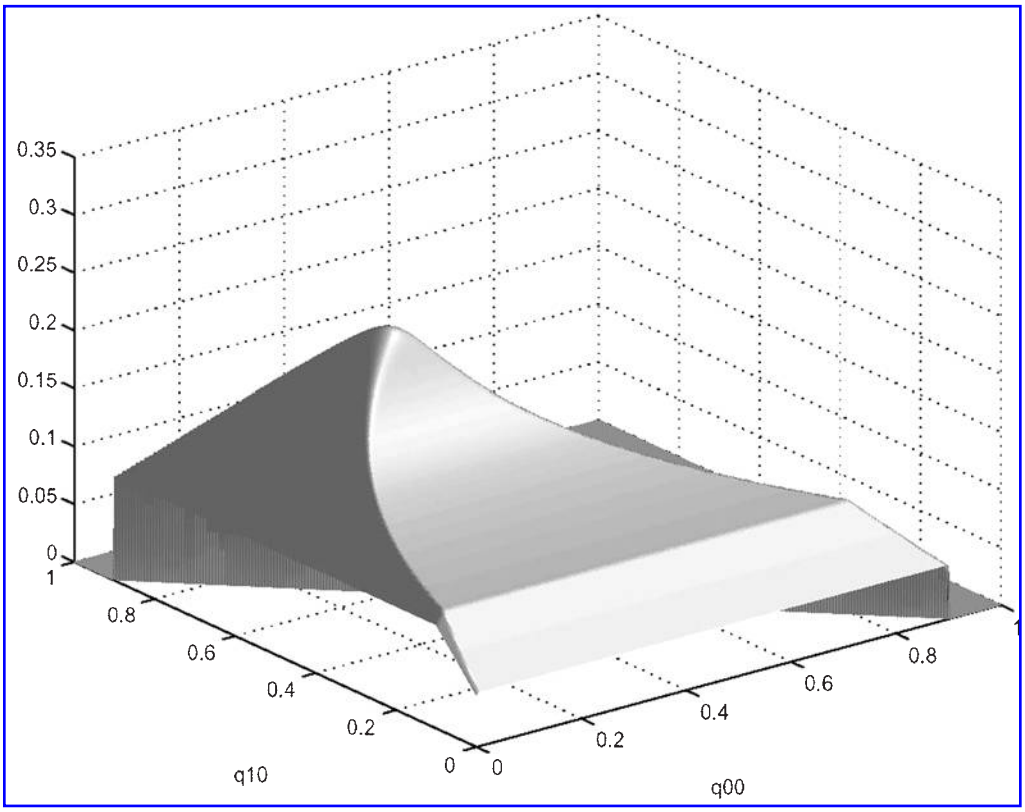

FIG. 3. $p_{A\prec B} - p$, where $p = 0.05$ and $q_{01} = 0.1$. It confirms that $p_{A\prec B}$ is larger than p when $q_{01} > 0$.

We expect that $\tilde{p} - p > 0$ when $q_{01} > 0$. We have confirmed this result numerically. In the range of $0 < p < 0.45$, $0 < q_{01} < 0.5$, we set up a mesh and calculate $\tilde{p} - p = p_{A\prec B} - p$. Figure 3 shows the result when $p = 0.05$ and $q_{01} = 0.1$.

Now we explain why we rather take $\hat{p}$ than the likelihood ratio as the statistics to test the hypothesis.

Proposition 3. *Suppose* $(\tilde{p}, \tilde{q}_{00}, \tilde{q}_{01} = 0, \tilde{q}_{10}, \tilde{q}_{11})$ *and* $(p, q_{00}, q_{01} > 0, q_{10}, q_{11})$ *are respectively the null and alternative hypotheses. Denote the significance level by* α*, and the chance of type II error of the optimal test by* β_n*, where* n *is the sample size. Then*

$$\lim_{\alpha\to 0}\lim_{n\to\infty}\frac{1}{n}\log\beta_n = -D[(\tilde{p}, \tilde{q}_{00}, \tilde{q}_{01} = 0, \tilde{q}_{10}, \tilde{q}_{11})||(p, q_{00}, q_{01}, q_{10}, q_{11})]\,.$$

This result is a direct application of the Stein's lemma; see Chapter 12 of Cover and Thomas (1991). It says that the chance of type II error (false positive) is characterized by the Kullback–Leibler divergence between the two hypotheses. We plot the Kullback–Leibler divergence for the case $p = 0.05$, $q_{00} = q_{11} = q_{10}$ in Fig. 4. It remains zero until q_{01} reaches 0.25. This indicates that the likelihood ratio test cannot give good protection against false positives. In comparison, we plot $\tilde{p} - p = p_{A\prec B} - p$ against q_{01} for the case $p = 0.05$, $q_{00} = q_{11} = q_{10}$ in Fig. 5. It can be seen that the score immediately goes up as q_{01} moves away from zero. Thus we rather adopt p-scores to play the role of test statistic.

Reconstruction of directed acyclic Boolean networks

The s-p-scores are more meaningful if they are generated from a directed acyclic Boolean network because we may discover significant pairwise relations by ranking the scores in the ascending order. We collect those pairwise relations whose s-p-scores are smaller than a threshold and put them in a watch list. Known biological results are helpful for the determination of threshold. For example, if we know the relation $A \prec B$ is true, then those s-p-scores smaller than $p_{A\prec B}$ should be in our watch list. Please

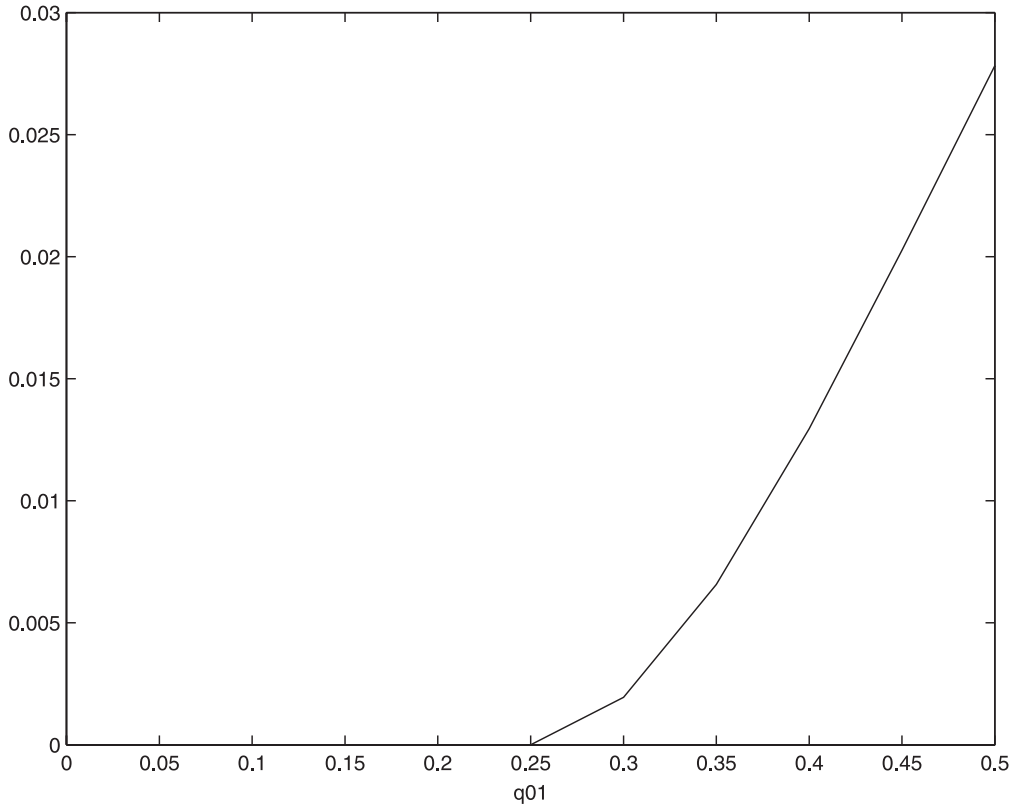

FIG. 4. The Kullback–Leibler divergence between the full model $q_{01} > 0$ and the triangular model $q_{01} = 0$ against q_{01}, where $p = 0.05$, $q_{00} = q_{11} = q_{10}$.

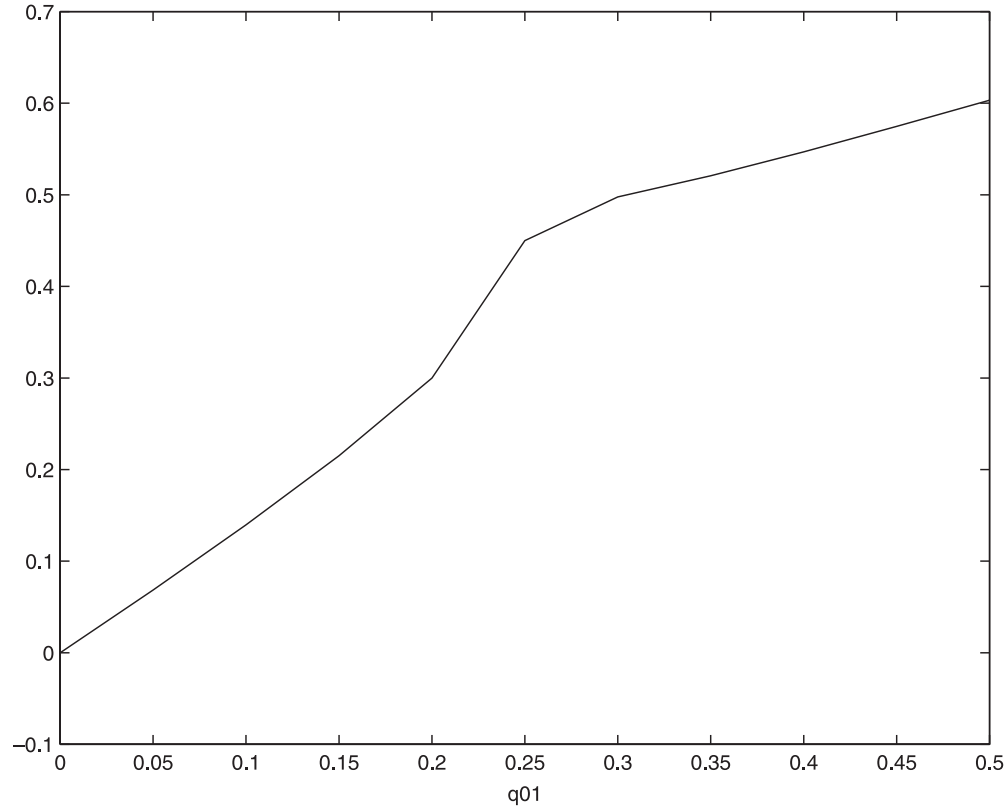

FIG. 5. $p_{A \prec B} - p$ against q_{01}, where $p = 0.05$, $q_{00} = q_{11} = q_{10}$.

notice that as more pairwise relations are included in the watch list, the more likely we are to observe incompatible ones. In this case, no DAB network exists to explain all the relations. We here mention one strategy, namely, the *maximum compatibility criterion*: choose the maximum threshold value so that the selected pairwise relations contain no conflict. Next we illustrate the method by some examples.

4. EXAMPLES

Simulated example

For the DAB example consisting of seven elements in Fig. 1, we simulate a data set of 76 samples with misclassification probability $p = 0.05$. The data can be arranged in an array similar to that obtained from microarray. Namely, each row in this array corresponds to an element, and each column corresponds to a sample. We compute the 21 s-p-scores and sort them in Table 8. For each pair of elements, we show the counts of $n_{i,j}$ in the last four columns, two s-scores, and four p-scores in the middle. The sorted s-p-scores and their corresponding hypotheses are shown in the first two columns. The true relations and false relations (in parentheses) cross each other by only one case.

Arginine biosynthetic pathway

Boolean logic is a useful tool for the study of pathways. We here revisit the analysis of the experiment concerning the biochemical pathway for the synthesis of the amino acid arginine in *Neurospora crassa*. It is a standard example to illustrate the one gene–one enzyme hypothesis; see Russell (1995). The pathway is shown in Fig. 6. Using genetic crosses and complementation tests, we know the process involves four genes, which are designated $argE^+$, $argF^+$, $argG^+$, and $argH^+$ in a wild-type cell. The experiments generated growth pattern of the mutant strains on media supplemented with presumed arginine precursors. These intermediates are ornithine, citrulline, and argininosuccinate.

TABLE 8. FOR THE DAB IN FIG. 1, WE GENERATE 76 SAMPLES, AND TAKE $p = 0.05$[a]

Ranking		*Hypotheses*						*Counts in cells*			
Relation	*s-p-score*	$q_{01} = q_{10} = 0$	$q_{00} = q_{11} = 0$	$q_{01} = 0$	$q_{10} = 0$	$q_{00} = 0$	$q_{11} = 0$	n_{00}	n_{01}	n_{10}	n_{11}
$C \prec G$	0.000	0.441	0.250	0.000	0.441	0.250	0.197	23	0	38	15
$A \prec G$	0.000	0.441	0.138	0.000	0.441	0.079	0.138	6	0	55	15
$A \prec C$	0.017	0.146	0.388	0.017	0.146	0.079	0.388	5	1	18	52
$A \prec \bar{D}$	0.028	0.250	0.329	0.079	0.250	0.028	0.329	1	5	31	39
$A \prec E$	0.030	0.342	0.237	0.030	0.342	0.079	0.237	5	1	41	29
$B \sim E$	0.041	0.041	0.498	0.028	0.041	0.605	0.395	42	2	4	28
$A \prec F$	0.054	0.309	0.270	0.054	0.309	0.079	0.270	4	2	37	33
$F \prec G$	0.058	0.219	0.368	0.058	0.219	0.368	0.197	38	3	23	12
$C \prec \bar{D}$	0.059	0.362	0.231	0.303	0.362	0.059	0.231	3	20	29	24
$A \prec B$	0.060	0.329	0.250	0.060	0.329	0.079	0.250	4	2	40	30
$C \prec F$	0.099	0.244	0.382	0.099	0.244	0.303	0.382	18	5	23	30
$(C \prec E)$	0.112	0.319	0.349	0.112	0.319	0.303	0.349	18	5	28	25
$\bar{D} \prec G$	0.120	0.388	0.257	0.197	0.388	0.257	0.120	23	9	38	6
$(C \prec B)$	0.134	0.319	0.362	0.319	0.134	0.303	0.362	17	27	6	26
$(\bar{E} \prec G)$	0.148	0.296	0.401	0.197	0.296	0.401	0.148	36	10	25	5
$(\bar{B} \prec G)$	0.180	0.309	0.388	0.197	0.309	0.388	0.180	35	9	26	6
$(D \sim \bar{F})$	0.208	0.480	0.208	0.421	0.579	0.187	0.208	11	21	30	14
$(D \sim \bar{E})$	0.301	0.484	0.301	0.394	0.606	0.301	0.288	17	15	29	15
$(B \sim \bar{D})$	0.338	0.500	0.338	0.590	0.411	0.337	0.337	17	27	15	17
$(B \prec F)$	0.360	0.360	0.476	0.360	0.338	0.581	0.419	25	19	16	16
$(E \prec F)$	0.427	0.427	0.419	0.427	0.395	0.419	0.319	24	22	17	13

[a]The true and false relations (in parentheses) cross each other by only one case.

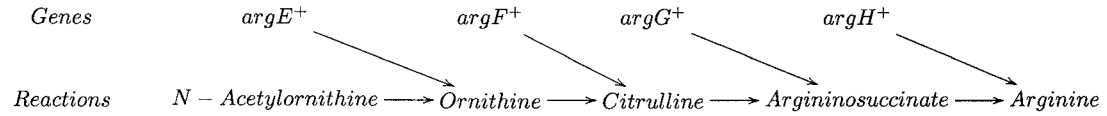

FIG. 6. Arginine biosynthetic pathway. The four genes code for the enzymes (not shown) that catalyze each reaction.

TABLE 9. THE STATES OF PRESENCE IN THE EXPERIMENTS OF GROWTH RESPONSE

Mutant strains	*Presence of elements*							
	E	*F*	*G*	*H*	*Ornithine*	*Citrulline*	*Arginino-succinate*	*Arginine*
Wildtype	1	1	1	1	1	1	1	1
argE	0	1	1	1	0	0	0	0
argF	1	0	1	1	?	0	0	0
argG	1	1	0	1	?	?	0	0
argH	1	1	1	0	?	?	?	0

Next, we have another look at this example from the perspective of the Boolean logic proposed in this paper. First, we rearrange the data from the experiments in an array; see Table 9. Please notice that this state table is different from the one shown in Chapter 9, page 275, in Russell (1995). The first four columns are definitions of the mutants. The next four columns show the presence state of the four arginine precursors when none of them is added externally. This can be deduced by the change of growth pattern after external controls. If we cannot determine the on–off status of an intermediate, we place a question mark.

The problem is to obtain the pathway in Fig. 6 from Table 9. By checking with Table 3, we can easily infer that (1) $E^+ \sim$ *Ornithine* or $E^+ \prec$ *Ornithine*, (2) $F^+ \prec$ *Citrulline*, (3) $F^+ \prec$ *Argininosuccinate*, (4) $F^+ \prec$ *Arginine*, (5) $G^+ \prec$ *Argininosuccinate*, (6) $G^+ \prec$ *Arginine*, and (7) $H^+ \prec$ *Arginine*. These pairwise relations are consistent with the sequence in Fig. 6. Even though the heuristic arguments of Russell (1995) can do the same job, the pairwise Boolean logic is more general. Also, we note that measurement error has not been considered in the example. When measurement error is unavoidable, we still can make inference by s-p-scoring. This is its advantage over no-measurement-error logic.

Yeast expression data

To study the signaling and circuitry of multiple mitogen-activated protein kinase MAPK pathways, Roberts *et al.* (2000) reported the expression data of yeast *Saccharomyces cerevisiae* for various knock-out cells under controlled experimental conditions. They particularly investigated four (MAPK) pathways: pheromone, PKC, HOG, and filamentous growth. We mentioned earlier that it is important to sample as much as possible from the state space of a network to avoid selection bias. This view highlights why various kinds of activation and perturbation, as done in this experiment, are valuable and necessary for the study of pathways. After activating relevant environmental factors (α-factor in this study), a cascade of biological activities occur sequentially. We want to use DAB networks to describe some aspects of these biological processes. We apply the s-p-scoring method to explore the expression profiles. Next, we show some exploratory result on the pheromone pathway.

During mating of *S. cerevisiae*, haploid MATa and MATα cells communicate with each other through secretion of pheromones α- and a-factor, respectively. Pheromone stimulates yeast cells to increase the expression of mating genes and arrest cell division in the G1 phase of the cell cycle. The responses to pheromone are initiated by a cell surface receptor that couples to a G protein and downstream MAPK kinase cascade; see (Fig. 1A) in Roberts *et al.* (2000). In some experiments, MATa cells are exposed to α-factor concentrations ranging from 0.15 to 500 nM. Cells with various knock-out genes are also tested. The genomewide expression levels are measured via the technique of cDNA microarrays. Namely, the abundance of each mRNA with respect to the reference is obtained in the form of expression ratios.

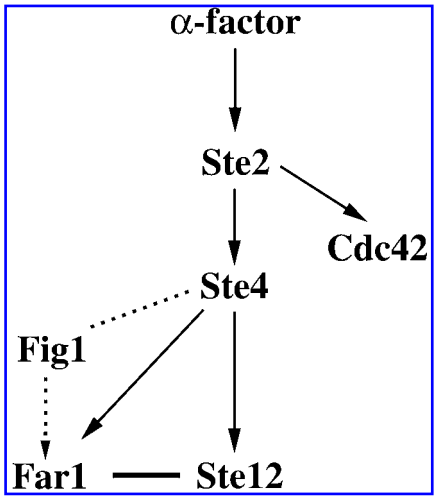

FIG. 7. Some pairwise relations identified by s-p-scoring method from the expression data of yeast *Saccharomyces cerevisiae* with knock-out and activation; see Roberts *et al.* (2000).

In our analysis, we exclude those experiments carried out under a different condition of 2% galactose for 3 hours, and two experiments measured at 0 and 15 minutes after the α-factor exposure. In total, we consider expression profiles from 45 experiments. We include the α-factor as an element and discretize it by setting it on if the concentration is larger than 0.50 nM and off otherwise. Figure 7 shows a DAB network obtained from our analysis. The part of network close to the α-factor is well reconstructed. That is, the pheromone α-factor activates the receptor Ste2p. Then, receptor stimulation releases free Gbg (Ste4p/Ste18p). The transcription factor Ste12p, which activates the promoters of mating, is also identified as one element downstream of the MAPK cascade. The positions of those genes in the middle of the pathway, such as Ste20p, Ste11p, Ste7, are missed. FIG1 is a transcriptional reporter gene for activation of the MAPK. Our analysis indicates its position in the pathway as shown in Fig. 7. We found that those genes whose expressions stay steady after some exposure to a concentration of α-factor are more easily identified.

5. DISCUSSION

Discretization

The data types in the DAB networks are binary. If elements such as expression levels are observed on a continuous scale, then we need to discretize them. In cDNA microarrays, a reference sample is also hybridized to probe. The ratios of expression levels (or differences in the logarithm scale) lead to a natural way of discretization. That is, an element is on if the log-ratio is larger than zero, and is off otherwise. If other information is available for some elements, we can exploit it to achieve better discretization. Consider expression levels of a gene A. Suppose the log-ratio of its expression is l_{-A} in a knock-out experiment $\triangle$A, and is l_{+A} in an experiment in which we know it is overexpressed. Then the threshold L must satisfy $l_{-A} \leq L \leq l_{+A}$. Histograms of the expression levels are also helpful for discretization. In the case that discretization is not perfect, the error mechanism introduced in the model still allows us to run the s-p-scoring analysis. In Xing and Karp (2001), a mixture model is used as a quantizer for their clustering method, and the result is quite good.

Coding issues

Each element in a DAB network is a dichotomous variable. In practice, an element may have more than two levels. In this case, we introduce multiple pseudo elements to code for its values. For example, if an element A has four levels, then we code it by two pseudo elements as shown in Table 10. In general, the information in a binary element is equivalent to a bit, and n bits can encode up to 2^n values.

Table 10. Coding an Element with Four Expression Levels by Two Pseudo Elements

Level	*0*	*1*	*2*	*3*
Element A_1	0		1	
Element A_2	0	1	0	1

If samples are obtained from a time course, then it is possible to consider differences of expressions between two consecutive time points. In this way, the dynamics of the networks are included in the analysis. For networks with feedback, caution is necessary to apply the s-p-scoring analysis. One strategy is to consider data in a time window and then examine how the pairwise relations evolve as the time window moves.

Computational complexity

The key step of the procedure is the computation of s- and p-scores for each of the $\frac{m(m-1)}{2}$ pairs of elements, where m is the number of elements. The E-M procedure used to compute the MLE is an iterative algorithm. It converges at a linear rate that depends on the fraction of missing data; see McLachlan and Krishnan (1997). The number of iterations required for convergence varies depending on initial values of parameters. A common practice in numerical implementation is setting an upper bound for iterations. Consequently, this keeps the $O(m^2)$ complexity for the computation of MLE. According to our numerical experience, the convergence is quite fast for the 2 by 2 count data. The sorting algorithm, such as heapsorting, can rank the $\frac{m(m-1)}{2}$ s-p-scores in $O(m^2 \log m)$ time and in place. Thus, the overall complexity is $O(m^2 \log m)$ in time and $O(m^2)$ in memory.

Software

We have developed MATLAB code for the s-p-scoring method.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Michael Waterman, Wing Wong, and Simon Tavaré, for their help. The research is partially supported by the Functional Genomics Program, Institute for Pure and Applied Mathematics, UCLA. Lu's research is partially supported by the National Science Council in Taiwan. Li's research is partially supported by the CEGS grant from NIH.

REFERENCES

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96, 6745–6750.

Ben-Dor, A., Shamir, R., and Yakhini, Z. 1999. Clustering gene expression patterns. *J. Comp. Biol*. 6, 281–297.

Bickel, P.J., and Doksum, K.A. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco.

Brightwell, G. 1997. Partial orders. *In* Beineke, L.W., and Wilson, R.J., eds., *Graph Connections: Relationships between Graph Theory and Other Areas of Mathematics*, Clarendon Press, Oxford.

Cover, T.M., and Thomas, J.A. 1991. *Elements of Information Theory*, Wiley, New York

De Jong, H. 2002. Modeling and simulation of genetic regulatory systems: A literature review. *J. Comp. Biol*. 9, 67–103.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Series B* 39, 1–22.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. 1998. Clustering analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863–14868.

Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T. 2000. Protein function in the post-genomic era. *Nature* 405, 823–826.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. 2000. Using Bayesian networks to analyze expression data. *J. Comp. Biol*. 7, 601–620.

Heckerman, D., Geiger, D., and Chickering, D.M. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20, 197–243.

Jensen, F.V. 1996. *An Introduction to Bayesian Networks*, University College London Press, London.

Kauffman, S. 1977. Gene regulation networks: A theory for their global structure and behaviors, in *Current Topics in Developmental Biology*, vol. 6, 145–182, Academic Press, New York.

Kauffman, S. 1979. Assessing the probable regulatory structures and dynamics of the metazoan genome, *in* Thomas, R., ed., *Kinetic Logic: A Boolean Approach to the Analysis of Complex Regulatory Systems*, vol. 29 of *Lecture Notes in Biomathematics*, 30–60, Springer-Verlag, Berlin.

Lehmann, E.L. 1986. *Testing Statistical Hypotheses*, Wiley, New York.

Li, L., and Lu, H.S. 2001. "Span" directed acyclic Boolean networks from array data. Technical report, Florida State University and University of Southern California.

McGuire, A.M., and Church, G.M. 2000. Predicting regulons and their *cis*-regulatory motifs by comparative genomics. *Nucl. Acids Res*. 28, 4523–4530.

McLachlan, G.J., and Krishnan, T. 1997. *The EM Algorithm and Extensions*, John Wiley, New York.

Murphy, K., and Mian, S. 1999. Modeling gene expression data using dynamic Bayesian networks. Technical report, University of California at Berkeley, Department of Computer Science.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Francisco.

Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A., He, Y., Dai, H., Walker, W.L., Hughes, T.R., Tyers, M., Boone, C., Friend, S.H. 2000. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287, 873–880.

Russell, P.J. 1996. *Genetics*, 4th edition. HarperCollins, New York.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

Spirtes, P., Glymour, G., Kauffman, S., Scheines, R., Aimalie, V., and Wimberly, F. 2000a. Constructing Bayesian network models of gene expression networks from microarray data. *Proc. Atlantic Symposium on Computational Biology, Genome Information Systems and Technology*.

Spirtes, P., Glymour, C., and Scheines, R. 2000b. *Causation, Prediction, and Search*, 2nd ed., MIT Press, Cambridge, MA.

Xing, E.P., and Karp, R.M. 2001. CLIFF: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics* 17, 306–315.

Address correspondence to:
Lei M. Li
Molecular and Computational Biology Program
Department of Biological Sciences
University of Southern California
1042 West 36th Place, DRB 289
Los Angeles, CA 90089

E-mail: lilei@hto.usc.edu

This article has been cited by:

1. Fabrício M. Lopes , Roberto M. Cesar , Jr. , Luciano Da F. Costa . 2011. Gene Expression Complex Networks: Synthesis, Identification, and Analysis. *Journal of Computational Biology* **18**:10, 1353-1367. [Abstract] [Full Text] [PDF] [PDF Plus]
2. S. Mukherjee, S. Pelech, R. M. Neve, W.-L. Kuo, S. Ziyad, P. T. Spellman, J. W. Gray, T. P. Speed. 2009. Sparse combinatorial inference with an application in cancer biology. *Bioinformatics* **25**:2, 265-271. [CrossRef]
3. T CHEN, H LU, Y LEE, H LAN. 2008. Segmentation of cDNA microarray images by kernel density estimation. *Journal of Biomedical Informatics* **41**:6, 1021-1027. [CrossRef]