Identification of Genetic Networks

Momiao Xiong,[1] Jun Li and Xiangzhong Fang

Human Genetics Center, University of Texas, Houston Health Science Center, Houston, Texas 77030

Manuscript received February 12, 2003
Accepted for publication October 15, 2003

ABSTRACT

In this report, we propose the use of structural equations as a tool for identifying and modeling genetic networks and genetic algorithms for searching the most likely genetic networks that best fit the data. After genetic networks are identified, it is fundamental to identify those networks influencing cell phenotypes. To accomplish this task we extend the concept of differential expression of the genes, widely used in gene expression data analysis, to genetic networks. We propose a definition for the differential expression of a genetic network and use the generalized T^2 statistic to measure the ability of genetic networks to distinguish different phenotypes. However, describing the differential expression of genetic networks is not enough for understanding biological systems because differences in the expression of genetic networks do not directly reflect regulatory strength between gene activities. Therefore, in this report we also introduce the concept of differentially regulated genetic networks, which has the potential to assess changes of gene regulation in response to perturbation in the environment and may provide new insights into the mechanism of diseases and biological processes. We propose five novel statistics to measure the differences in regulation of genetic networks. To illustrate the concepts and methods for reconstruction of genetic networks and identification of association of genetic networks with function, we applied the proposed models and algorithms to three data sets.

RECENT advances in genome sequencing and high-throughput technologies, such as DNA and protein chips, allow us to measure the spatio-temporal expression levels of thousands of genes or proteins (Brown and Botstein 1999; Lipschutz *et al.* 1999; Lockhart and Winzeler 2000; Young 2000; Figeys and Pinto 2001; Hughes and Shoemaker 2001; Houseman *et al.* 2002; Jong 2002). Mass spectrometry (Mann 1999; McLuckey and Wells 2001) provides further experimental tools to acquire knowledge of the genetic networks (Arnold *et al.* 2004). In the past several years, a number of statistical and computational methods for reconstructing genetic networks have been developed, such as Boolean networks (Liang *et al.* 1998; Akutsu *et al.* 2000; Ideker *et al.* 2001), probabilistic Boolean networks (Shmulevich *et al.* 2002), differential equations (Chen *et al.* 1999; D'Haeseleer *et al.* 1999; Von Dassow *et al.* 2000), neural networks (Wahde and Hertz 2000), fuzzy logic (Woolf and Wang 2000), and Bayesian networks (Friedman *et al.* 2000; Hartemink *et al.* 2001; Imoto *et al.* 2002).

Although a great advance in both experimental technology and computational methods for reconstructing genetic networks has been made, we still face significant challenges in understanding such networks. To accomplish the goal of identifying genetic networks and to explore their applications to biomedical research, several issues must be addressed. First, the development of dynamic models of genetic networks is severely compromised by the lack of experimental techniques to measure the dynamic quantities of such networks. Therefore, revealing information from steady states of genetic networks using gene expression profiles is of great interest.

Second, to identify physically connected genetic networks using gene expression profiles, which describe how genes directly activate or inhibit others, may be too ambitious to be accomplished at the current stage due to incomplete information on the structure of genetic networks. However, instead of reconstructing physically connected genetic networks, it may be feasible to model quasi-genetic networks, defined as a network that describes most likely functional relations between the genes in the network. The quasi-genetic network may not represent physical connection of the genes in the network, but represents the best fit of the network model to gene expression data.

Third, many current computational methods for the reconstruction of genetic networks have focused on the network structure. However, structure provides only partial information on genetic networks. To measure quantitatively the relationship between genes in the network is indispensable for studying regulatory properties of genetic networks (Ronen *et al.* 2002).

To model quantitatively genetic networks, we propose the use of structural equations (Bollen 1989) as a framework for modeling genetic networks. Structural

[1]*Corresponding author:* Human Genetics Center, University of Texas, 1200 Herman Pressler, Houston, TX 77225.
E-mail: mxiong@sph.uth.tmc.edu

equation models were first introduced into genetics (Wright 1921), econometrics (Haavelmo 1943), and social science (Duncan 1975). Since then, structural equations as a tool for causal inference have been widely explored in social science and engineering (Pearl 2000). However, to our knowledge, structural equation models have not been used for reconstruction of genetic networks. Although structural equations can be used to model both equilibrium and disequilibrium states of genetic networks, we focus on equilibrium states for the reasons described above. We provide (1) a mathematic representation of genetic networks based on structural equations, (2) statistical methods for estimating and testing model parameters, (3) probabilistic criteria for assessing how well models of the genetic networks explain the observed data, and (4) optimization procedures for searching the most likely structure of the genetic network.

Once a genetic network is identified, it is crucial to associate genetic networks with cell phenotypes. Differential expression of genes is a widely used concept for identifying genes that are able to discriminate cell phenotypes. To associate genetic networks with cell phenotypes, we generalize the notion of differentially expressed genetic networks and develop a statistic to test for the differential expression of such networks.

Coefficient parameters in the structural equations measure the regulatory effects of one gene on others or the strength of the gene-gene interactions. Functional mutations in the genes will often cause changes in regulatory effects. Thus, we expect that due to the accumulation of mutations in abnormal cells, the regulation of some genetic networks in abnormal cells will be significantly different from that in normal cells. Uncovering such differences may help us to identify the causes of disease. To accomplish this task, we provide five statistics to measure the differences in regulation between the genetic networks in normal and abnormal cells. We hope that by identifying differentially regulated genetic networks we are likely to discover a set of genes and genetic networks that influence the development of the diseases.

METHODS

Linear structural equation model: Linear structural equations can be used for construction of a first-order approximation model of a genetic network using steady-state gene expression measurements (Datta 2001). Rate equations expressing the rate of production of components in the system are often used to model the concentrations of mRNA, protein, and other molecules. Rate equations in a simplified form are given by Jong (2002),

$$\frac{dZ}{dt} = G(Z) - RZ, \qquad (1)$$

where R is a diagonal matrix and $G(Z)$ is a vector of nonlinear functions. The right-hand side of Equation 1 has two terms: the first one is the production of molecules, and the second is the degradation of existing molecules. The system of nonlinear differential equations (1) can be approximated to the first order by a linear system of equations near a steady state of the system

$$\frac{dY}{dt} = AY - RY,$$

where Y is a vector of the deviation of variables in Z from their means and A is a Jacobian matrix of $G(Z)$, *i.e.*, $A = \partial G(Z)/\partial Z$, measuring the strength of regulatory interactions between genes in the network. When the system reaches a steady state, which is equivalent to setting the time derivative of Y to zero, we have

$$RY = AY.$$

The above equations show that the Jacobian matrix involves feedback loops of a dynamic biological system and gene or protein expressions in cells or tissues are jointly or simultaneously determined. Gene expression data that are generated by biological systems must be described as a system of joint relations among the gene expression variables.

The naïve differential equation approach assumes that the genetic network is fully connected, ignoring the structural relations between genes in the network (D'Haeseleer *et al.* 1999). This assumption results in a large number of parameters in the differential equations. Due to a limited number of samples, it is difficult to develop any meaningful statistical methods for estimation of the parameters. However, most genetic networks are not fully connected (Gardner *et al.* 2003). The networks' relations contain structural or causal information on the gene expression variables. The matrix A is a sparse matrix and most elements of the matrix A are zero. Therefore, gene expression variables in genetic networks are modeled by structural equations, which consider both simultaneous and structural relations among the gene expression variables. Structural equations can simultaneously include all endogenous variables in one side of equations, which allows us to consider bidirectional causality. Unlike ordinary regression techniques that cannot deal with directed cyclic graphs, structural equation models allow bidirectional causality/feedback loops (which are referred to as nonrecursive models; Maruyama 1998). This remarkable feature makes structural equations a useful causal inference tool for reconstruction of genetic networks because many genetic networks contain feedback loops.

We begin to describe structural equations for modeling genetic networks by introducing a path diagram (Bollen 1989; Shipley 2000). A path diagram (or directed graph) is a graphical representation of a system of

structural equations and is used to describe graphically genetic networks as shown in Figure 1. The path diagram consists of nodes, represented by letters, and edges, represented by lines. The nodes of the path diagram correspond to variables. The directed edges between nodes denote the direction of the regulatory relationship between the nodes (variables) connected by the edges and indicate a directed regulatory influence of one gene on another. The directed edges can represent either activation (positive control) or inhibition (negative control).

Variables in path diagrams can be classified into two basic types of variables, observed variables that can be measured and residual error variables that cannot be measured and represent all other unmodeled causes of the variables. Most observed variables (*e.g.*, gene expression levels) are random. Some observed variables may be nonrandom or control variables (*e.g.*, drug doses) whose values remain the same in repeated random sampling or might be manipulated by the experimenter. The observed variables will be further classified into exogenous variables, which lie outside the model, and endogenous variables, whose values are determined through joint interaction with other variables within the system. All nonrandom variables and some of the gene (or protein) expression data (*e.g.*, initiators of pathway) can be viewed as exogenous variables. Most of the gene (or protein) expression data are viewed as endogenous variables. The terms exogenous and endogenous are model specific. It may be that an exogenous variable in one model is endogenous in another. The observed variables are enclosed in boxes and the error variables are not enclosed at all.

Let Y be a vector of the p endogenous variables and X be a vector of q exogenous variables. Occasionally, one or more of the X's are nonrandom. We denote the errors by e. We assume that $E[e] = 0$ and that e is uncorrelated with the exogenous variables in X. We also assume that e_i is homoscedastic and nonautocorrelated (Bollen 1989). Then, the structural equations for modeling gene expressions in the genetic network are given by

$$Y = BY + \Gamma X + e, \tag{2}$$

where B is a $p \times p$ matrix and Γ is a $p \times q$ matrix. The elements of the coefficient matrices B and Γ describe the regulatory effects of one gene on another or of a nonrandom variable on the gene, which is a direct regulatory influence of one variable on the other. Therefore, throughout the article, the matrices B and Γ are referred to as the *regulatory matrices*. Since the genetic networks are not fully connected, many elements in the matrices B and Γ will be zero. The matrices B and Γ are, in general, sparse. The matrix B can describe feedback relations in the path diagram. The structural equations can model directed cyclic graphs and hence genetic networks with feedback loops (Bollen 1989).

In Figure 1 we assume that the expression levels of the genes *CDC28*, *CLB1*, and *CLB3*, denoted by x_1, x_2, and x_3, respectively, are exogenous variables and the expression levels of the genes *MCM1*, *MCM2*, *SWI4*, *CLN3*, *CDC47*, and *CDC6*, denoted by y_1, y_2, y_3, y_4, y_5, and y_6, respectively, are endogenous variables. The structural equations for the genetic network are written as

$$\begin{aligned}
y_1 &= 1.19x_1 + e_1\\
y_2 &= 0.16x_1 + 0.28x_2 - 0.34x_3 + e_2\\
y_3 &= 0.06y_1 + 0.19y_2 + e_3\\
y_4 &= 4y_3 + e_4\\
y_5 &= 0.19y_4 + e_5\\
y_6 &= 0.2y_5 + e_6.
\end{aligned}$$

We assume that the influence of the genes in the network is in one direction and that the errors in the equations are independent and uncorrelated with exogenous variables. Under these assumptions, if the genetic networks do not contain feedback loops, the B matrix can be made lower triangular by arranging the order of endogenous variables and the variance-covariance matrix of the errors is diagonal. Therefore, the structural equations for the genetic networks without feedback loops are recursive models, which ensure that parameters in the recursive model are identifiable (Bollen 1989).

Parameter estimation: To estimate the parameters of the structural equations, we assume that the structure of the network is known. How to identify network structure is discussed in the *Model selection* section. It is well documented that the ordinary least-squares estimator is biased and inconsistent for parameters in structural equations (Bollen 1989). To ensure that estimators are consistent and unbiased, we use the estimation procedures based on covariance analysis, which assumes that

$$\sum = \sum(\theta),$$

where Σ is the population covariance matrix of the variables Y and X, and $\Sigma(\theta)$ is the covariance matrix written as a function of the free model parameters in the models, which we denote by θ. Let Φ and ψ denote the covariance matrices of X and e, respectively. The matrix $\Sigma(\theta)$ consists of three parts: (1) the covariance matrix of Y, (2) the covariance matrix of X with Y, and (3) the covariance matrix of X. First we consider $\Sigma_{YY}(\theta)$, the implied covariance matrix of Y. From the Equation 2, we have $Y = (I - B)^{-1}(\Gamma X + e)$. Hence, $\Sigma_{YY}(\theta) = (I - B)^{-1}(\Gamma\Phi\Gamma' + \Psi)(I - B)^{-1'}$. The implied covariance matrix of Y and X is given by

$$\sum_{YX} = E\{(I - B)^{-1}(\Gamma X + e)\, X\} = (I - B)^{-1}\Gamma\Phi.$$

Therefore, we have

$$\Sigma(\theta) = \begin{bmatrix} (I - B)^{-1}(\Gamma\Phi\Gamma' + \Psi)(I - B)^{-1'} & (I - B)^{-1}\Gamma\Phi \\ \Phi\Gamma'(I - B)^{-1'} & \Phi \end{bmatrix}$$

(Bollen 1989). The above equation implies that each element of the covariance matrix is a function of model parameters. The unknown parameters in B, Γ, ψ, and Φ are estimated so that the implied covariance matrix $\Sigma(\theta)$ is as close to the sample covariance matrix S, the estimator of the matrix Σ, as possible. To know when our estimates are as "close" as possible, we must define close, that is, we require a fitting function that is minimized. The most widely used fitting function is based on the method of maximum likelihood (ML) defined by maximizing the likelihood function or its log,

$$F_{\mathrm{ML}} = \log|\Sigma(\theta)| + \mathrm{Tr}(S\Sigma^{-1}(\theta)) - \log|S| - (p + q),$$

where p and q are the number of endogenous and exogenous variables, and Tr denotes the trace of a matrix. The fitting function F_{ML} compares the difference between the observed and predicted covariance matrices. In general, F_{ML} is a complicated nonlinear function of the structural parameters, and explicit solutions are not always found. Instead, a Newton unconstrained optimization procedure is employed to find solutions (Bertsekas 1995).

It is well known that the ML estimators are consistent and asymptotically unbiased. Large sample theory ensures that $(N - 1)F_{\mathrm{ML}}$ is asymptotically distributed as χ^2 distribution with $\frac{1}{2}(p + q)(p + q + 1) - t$ d.f., where t is the number of free parameters, and the distribution of the estimator is asymptotically normal. Hence, the ratio of the estimated parameter to its standard error approximates a Z-distribution for large samples and can be used to test the parameters. The standard errors can be obtained from the following asymptotical covariance matrix for the ML estimators,

$$\left(\frac{2}{N-1}\right)\left\{E\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right\}^{-1},$$

where N is the number of samples.

Model selection: Learning about genetic networks consists of two parts: parameter learning and structure learning. For parameter learning, in the previous section we assume that the network structure is known. However, in most cases, the network structure is unknown and needs to be identified. To learn network structure from genome-wide gene expression profiles consists of two steps. The first step is to select the set of genes whose reconstructed network best fits the gene expression data. The second step is to learn the structure of the networks for a set of selected genes, which provides the best fit to the gene expression data.

To identify the structure of the network, an overall model fit measure is needed to assess how well a genetic network fits the data and to compare the merits of alternative network structure (Jordan 1999). The overall model fit measure is to calculate the difference between the covariance matrix predicted by the model and the sample covariance matrix from the observed data. Those differences measure how similar the hypothesized genetic network model is. The model fit measure allows us to rank genetic networks according to their ability to fit the observed data. A widely used model fit measure is the Akaike information criterion (AIC; Bollen 1989; Maruyama 1998), which is defined as

$$(N - 1)F_{\mathrm{ML}} - 2d,$$

where N is the number of samples, F_{ML} is the fitting function, $d = \frac{1}{2}(p + q)(p + q + 1) - t$ is degrees of freedom, and t is the number of free parameters in the model. The AIC value provides a relative ordering of different models fitting the data. The smaller the AIC value, the better the model fits the data.

However, AIC information cannot be employed to test whether the identified genetic network is valid. Fortunately, the statistic $(N-1)F_{\mathrm{ML}}$ is asymptotically distributed as a $\chi^2_{(d)}$ distribution under the null hypothesis H_0: $\Sigma = \Sigma(\theta)$. It should be noted that the null hypothesis means that the constraints on Σ imposed by the genetic network model are valid. In contrast to ordinary tests where the probability of obtaining a χ^2 value larger than a prespecified value is the probability of committing error for the rejection of the null hypothesis, in the model selection test here, the probability of obtaining a χ^2 value larger than a prespecified value is the probability of ensuring that the fitted model is correct and is referred to as the fitting probability. Therefore, the higher the probability of the χ^2, the closer is the fitted model for the genetic network to the true genetic network.

Genetic algorithms: Searching the genetic network is a very difficult problem because of the large number of possible networks. To exhaustively search all possible networks is infeasible, in practice, even with high-performance computers. Genetic algorithms (GAs) can be used for searching networks (Larranaga *et al.* 1996). Network search consists of two parts. First, we need to search a set of genes that are included in the network. Then, for the fixed set of genes we search the structures of the network that specify how the genes in the network are connected. We developed a new type of GA that accomplishes these two tasks simultaneously.

We use a $k \times k$ connective matrix C to represent the structure of a network with k genes. The elements of C are given by

$$c_{ij} = \begin{cases} 1 & \text{if node } j \text{ is directed to node } i \\ 0 & \text{otherwise.} \end{cases}$$

GAs begin with a population that consists of a large number of individuals. In our genetic algorithm, individuals of the population represent selected genes and network structures. This type of individual is denoted by a string,

$$g_1 g_2 \ldots g_k c_{11} c_{21} \ldots c_{k1} \ldots c_{1k} c_{2k} \ldots c_{kk},$$

which is usually referred to as a chromosome in the GA literature (as opposed to a real chromosome). The first part of the chromosome $g_1 g_2 \ldots g_k$ is a set of integer numbers representing genes selected in the network. The second part $c_{11} c_{21} \ldots c_{k1} \ldots c_{1k} c_{2k} \ldots c_{kk}$ is a binary string indicating the network structure. GAs attempt to find individuals from the search space with the best fitness (*e.g.*, smallest AIC value). The searching procedure of GAs can be briefly described as follows. First, the initial population is generated randomly, and the fitness of each individual is calculated. Second, individuals with good fitness are selected as parents. These parents produce children by the operations of crossover and mutation. A crossover operation in a GA algorithm produces two children by an exchange of chromosome segments between two parents. The mutation operation creates children by changing parents' chromosomes. All new produced children are added to the population. Some individuals with worse fitness (*e.g.*, higher AIC values) are removed from the extended population (including both parents and children) to generate a new population with its initial size, but with better fitness. Crossover and mutation play different roles in the genetic algorithm. Crossover increases the average fitness of the population. Mutation can help the algorithm to avoid local optima by exploring new states. After many iterations of GAs most likely or near most likely networks to fit the data can be found. When the difference between AIC values of two successive iterations is less than a prespecified threshold, the iteration of GAs is stopped.

The generalized T^2 statistic for testing the differential expression of genetic networks: Let $\overline{X}_1$ and $\overline{X}_2$ be the mean value of expression of all the genes in the network from normal and abnormal tissues, respectively. Let S_{pool} be the pooled estimate of common covariance matrix between gene expressions. It can be shown that

$$T^2 = \left[\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}\right]\left[\frac{n_1 n_2}{n_1 + n_2}\right]D^2$$

(Anderson 1984) follows an F-distribution with $v_1 = p$ and $v_2 = n_1 + n_2 - p - 1$ d.f., where

$$D^2 = (\overline{X}_1 - \overline{X}_2)^T S_{\text{pool}}^{-1}(\overline{X}_1 - \overline{X}_2),$$

n_1 and n_2 are the sample sizes of normal and abnormal tissues, respectively, and p is the number of genes selected in the test statistic. Consequently, T^2 can be used to test whether the population means, μ_1 and μ_2, differ significantly and to test for the significance of separation of two populations (normal and abnormal tissues). Formally, the null hypothesis H_0: $\mu_1 = \mu_2$ *vs.* the alternative hypothesis H_a: $\mu_1 \neq \mu_2$ is assumed. If H_0 is rejected on the basis of a T^2 test, we can conclude that the separation between normal and abnormal tissue populations is significant and the genetic network is differentially expressed.

Index for measuring difference in regulation of genetic networks: Let $A = [B\Gamma]$ be a coefficient matrix of structural equations for modeling a genetic network. Let A_1 and A_2 be its corresponding coefficient matrices in the normal and abnormal tissue samples. Let $W = A_1 - A_2$ and w_{ij} be an element of the matrix W. Since w_{ij} is a parameter in the network, its asymptotic standard deviation can be calculated from the square root of the main diagonal of the asymptotic covariance matrix of the estimated parameters in the network and denoted by $S_{W_{ij}}$. We define the test statistic T_G as

$$T_G = \frac{W_{ij}}{S_{W_{ij}}}.$$

Although the exact distribution of T_G is unknown, its asymptotical distribution can be approximated by a t distribution with $N - 2$ d.f. This statistic can be used to test the difference of the regulatory effect of one gene on another between normal and abnormal tissues.

The difference of the regulatory effect of one gene on another cannot measure the difference in the global behavior of the genetic networks between normal and abnormal tissues. A simple quantity to measure the difference in global behavior of genetic networks between the normal and abnormal tissues is the largest absolute value of the difference of the regulatory effect of one gene on another in the network between the normal and abnormal tissues, *i.e.*, $w_0 = \max_{i,j}|w_{ij}| = |w_{i_0 j_0}|$. The statistic T_G for testing the difference of individual regulatory effect can be used to test the difference in global behavior of genetic networks. Specifically, the statistic for testing the differential regulation of the genetic networks is given by

$$T_{G_0} = \frac{w_{i_0 j_0}}{S_{w_{i_0 j_0}}}.$$

The P value is calculated by a permutation test. The gene expression profile matrix is randomly permuted, and the structural equation model and genetic algorithms are applied to randomly permutated gene expression data to reconstruct the genetic network hundreds or thousands of times. Then, we calculate T_{G_0} and obtain an empirical distribution of T_{G_0}. The P value of the test is then defined as the probability that T_{G_0} exceeds its observed value. The statistic T_{G_0} can be used to measure the difference in regulation of the genetic network.

The difference in global behavior of genetic networks between the normal and abnormal tissues depends on the whole regulatory coefficient matrix. A scalar associated with a matrix W is a norm of the matrix W that denotes a real valued function of W (of the elements w_{ij} of W). The norm is relevant with all elements of the

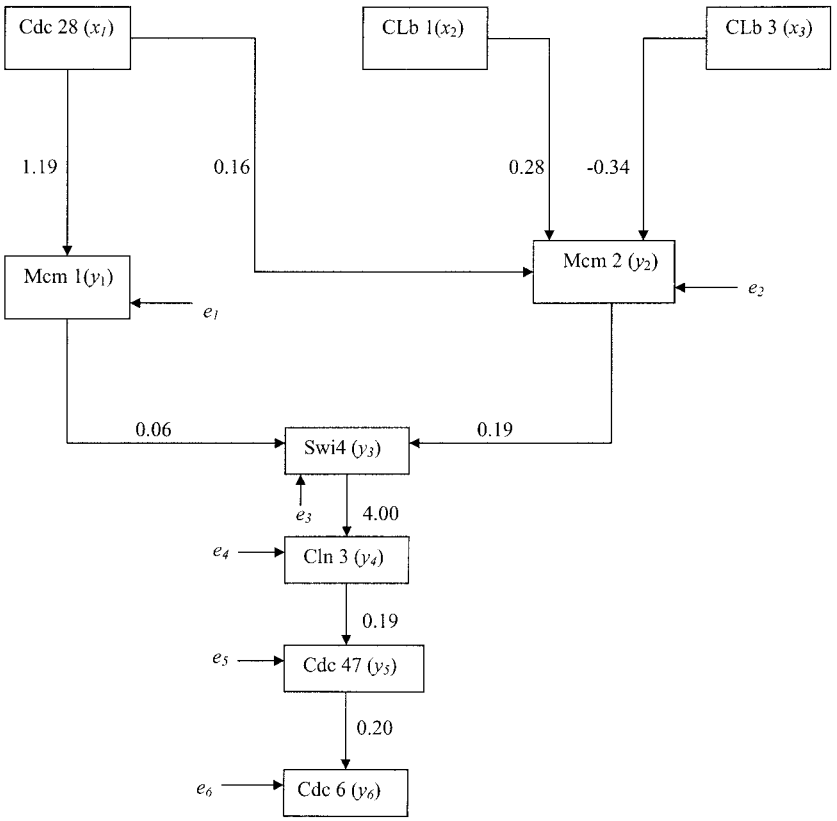

FIGURE 1.—Path diagram for a genetic network of nine genes reconstructed from yeast cell cycle data.

matrix and hence can be used to measure the difference in regulation of the whole genetic networks. Four metrics borrowed from the norms of the matrix for measuring the difference in regulation of the genetic networks are defined as follows (GRAYBILL 1976):

1. $||W||_1 = \max_j(\sum_{i=1}^{p+q}|w_{ij}|)$ = maximum of sums of absolute value of column elements of the matrix.
2. $||W||_\infty = \max_i(\sum_{j=1}^{p+q}|w_{ij}|)$ = maximum of sums of absolute values of row elements of the matrix.
3. $||W||_2$ = square root of the maximum eigenvalues of the matrix W^tW, a spectral norm.
4. $||W||_E = [\Sigma_i \Sigma_j (|w_{ij}|)^2]^{1/2}$, a Euclidean norm.

RESULTS

Illustration of structural equations for modeling genetic networks: To illustrate the use of structural equations for modeling genetic networks, we first analyze the expression profiles of 6220 genes using oligonucleotide arrays in synchronized yeast cells during the cell cycle (CHO *et al.* 1998). Data were collected at 17 points that included nearly two full cell cycles. Genetic expression profiles of the yeast cell cycle are time course data. The ideal method for reconstruction of genetic networks for time course data is dynamic (rather than ordinary) structural equations. Here, we use gene expression data of the yeast cell cycle as an example for illustration of genetic network modeling by structural equations. Although gene expression data from the yeast cell cycle are time course data, their dynamics are stable. When the time intervals at which gene expressions are measured are not small, the observed expression can be viewed as being sampled from near steady state of the yeast cell cycle dynamic system. Therefore, it is possible to use ordinary structural equations to model such systems. This example also serves to investigate how well ordinary structural equations are used to approximate stable dynamic systems.

The genetic network shown in Figure 1 was reconstructed by applying the proposed structural equation model to the expression profiles of the genes *CDC28*, *CLB1*, *CLB3*, *MCM1*, *MCM2*, *SWI4*, *CLN3*, *CDC47*, and *CDC6*, which play an important role in the M/G1 phase of the cell cycle. The regulatory relations between the genes in the network can be confirmed by the experiments (KOCH and NASMYTH 1994; MCINERNY *et al.* 1997). ZHANG (1999) investigated transcriptional regulation of the M/G1 phase in budding yeast using the same gene expression data and found that the genes *CLB1*, *CLB2*, *CLB3*, and *CDC28* regulated the expressions of the genes *CLN3*, *SWI4*, *CDC6*, and *CDC47*. However, he could not further infer the regulatory relations either among the genes *CLB1*, *CLB2*, *CLB3*, and *CDC28* or among the genes *CLN3*, *SWI4*, *Cdc6*, and *CDC47*. Therefore, our results gave a more detailed structure of the network than did those of ZHANG (1999).

A common approach for identifying networks is to use model selection to choose those networks with high-

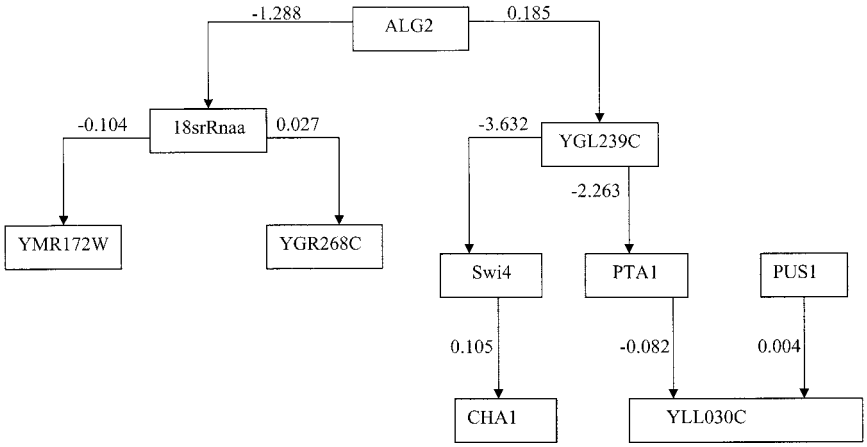

FIGURE 2.—The fully connected genetic network with the smallest AIC value (AIC = −64.02, fitting probability = 1.0000) among the 129 fully connected genetic networks with 10 genes for yeast cell cycle gene expression data.

scoring models and we use both AIC values and fitting probability to score the model. AIC values, which have a close relationship with the likelihood function, are widely used model selection criteria. However, AIC values measure only the relative goodness of fit. On the other hand, the fitting probability quantifies how well the model explains the observed data. Therefore, we use AIC values to select the model, but we also report the fitting probability of the selected models to indicate how reliable the selected models are.

The gene *SWI4* plays an important role in cell cycle progression (McInerny *et al.* 1997; Zhang 1999). As an example, we searched for genetic networks with 10 genes, including *SWI4*. We use genetic algorithms to search for optimal subsets of genetic networks with the smallest AIC values or the largest-fitting probability. Due to the high cost of microarray experiments, the number of tissue samples is often small relative to the number of genes in the data set. In this case, the number of genetic networks with a high score is usually large. Therefore, in searching for genetic networks with *SWI4* and the other 9 genes, we applied genetic algorithms to the data set 500 times, which yielded 500 genetic networks with AIC values ranging from −70.50 to −55.02 and fitting probabilities ranging from 1 to 0.997. It was interesting to observe that of these 500, 371 genetic networks were partitioned into more than two disconnected networks, while the remaining 129 genetic networks were fully connected. We ranked the fully connected genetic networks according to their AIC values. The highest-scoring fully connected genetic network with the smallest AIC value was plotted in Figure 2. Since the coefficients in the equations measure the magnitude of influence of one gene on the expression of another gene, they were referred to as the regulatory effects of the genes. Since the connected genes in a reconstructed network may or may not be physically connected in reality, the regulatory effect may be a direct effect that is unmediated by other genes or may be an indirect effect that is mediated by other genes that do not appear in the reconstructed networks. Unfortunately, the currently proposed method cannot untangle direct and indirect effects of the genes. The genetic network in Figure 2 has 10 genes: cell cycle gene *SWI4*; a glycosyltransferase gene, *ALG2*, involved in the dolichol pathway and regulated at two critical control points in the G1 phase of the cell cycle (G0/G1 and START; Lennon *et al.* 1995); an essential gene of *Saccharomyces cerevisiae* affecting pre-tRNA processing, *PTA1* (O'Connor and Peebles 1992); a pseudouridine synthetase gene, *PUS1*, which catalyzes the formation of pseudouridines in tRNAs (Arluison et al. 1999); a serine and threonine catabolism gene, *CHA1* (Bornaes *et al.* 1992); and five other unknown genes.

To investigate the effect of removing a gene from the genetic network, we plotted Figure 3, in which the gene *SWI4* was removed from the genetic network shown in Figure 2. It was interesting to note that most of the regulatory effects in the genetic network were not changed except for the regulatory effect of YGL239C on *CHA1*. This had an important implication: removing a gene will influence only the effects of the genes that were directly connected with the removed gene, but it did not have a significant impact on other parts of the genetic network.

As the number of genes in genome-wide gene expression profiles increases, the total number of all possible genetic networks exponentially increases. This number of possible genetic networks is too large to be exhaustively searched. There are two approaches to treat this problem. One approach is an ensemble method for identifying genetic networks that are consistent with existing gene expression profiling data (Battogtokh *et al.* 2002). The second, which we proposed, is to use genetic algorithms for searching genetic networks with smallest AIC values. We hope that when we run a large number of iterations we can search and identify the most likely genetic network model with the smallest AIC value and the largest fitting probability. The AIC value and fitting probability are referred to as the score of the genetic networks. To investigate whether genetic algorithms can identify the networks with the highest

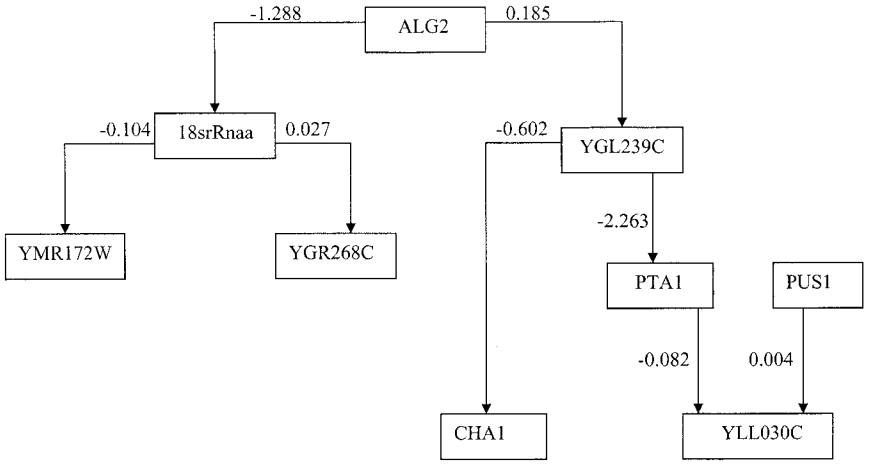

FIGURE 3.—The genetic network in which the gene *SWI4* was removed from the genetic network shown in Figure 2.

score and how many iterations are required to achieve the highest score, genetic algorithms were applied to yeast cell cycle data (CHO *et al.* 1998) to search networks with 12 genes. Figures 4 and 5 plot the AIC value and the fitting probability against the number of iterations, respectively. From Figure 4 we can see fluctuations in AIC values, but we still can observe the decreasing trend of AIC values. From Figure 5 we can see that, after 80 iterations, these runs reach a fitting probability of 1.

The largest fitting probability that we can reach after a number of iterations is a function of the number of genes in the network. The fitting probability will decrease when the number of genes in the network increases. To demonstrate this, we first fix the number of genes in the network and run 100 genetic algorithms to search for networks with the fixed number of genes. In this way, for each fixed number of genes in the network we can obtain the largest fitting probability. We can see from Figure 6 that when the number of genes in the network was >14 the fitting probability became small, which implied that the genetic network did not fit the data well. The size of the genetic network (*i.e.*, the number of genes in the network) is limited by the number of tissue samples.

To further evaluate the performance of the proposed model for reconstructing genetic networks, we take 85 regulators of yeast listed in LEE *et al.* (2002), where the remaining 21 of 106 regulators in Lee *et al.* cannot be found in the CHO *et al.* (1998) data set, as the primary genes for reconstruction of genetic networks with six

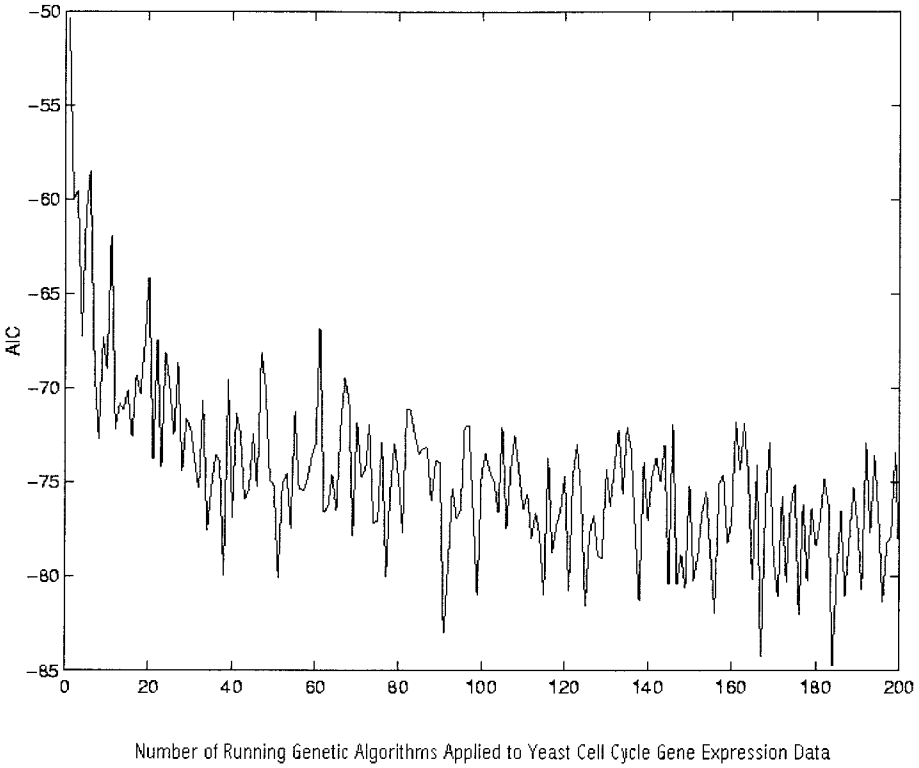

FIGURE 4.—The AIC values of 200 genetic networks with 12 genes as a function of the number of running genetic algorithms that were applied to yeast cell cycle gene expression data.

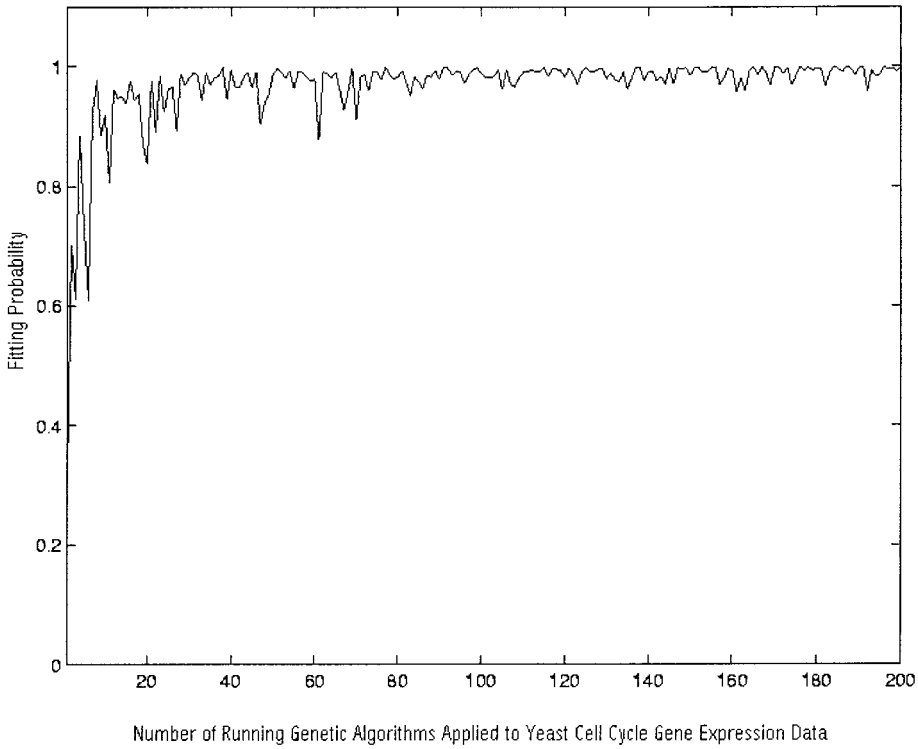

FIGURE 5.—The fitting probability of 200 genetic networks with 12 genes as a function of the number of running genetic algorithms that were applied to yeast cell cycle gene expression data.

genes. For each regulator that was used as an exogenous variable in the structural equations or a primary gene of the genetic network being reconstructed, genetic algorithms were applied 300 times to the yeast cell cycle data for searching genetic networks with six genes. From the reconstructed genetic networks with the regulators as primary genes of the genetic networks we can find pairs of the regulator-regulated target gene. The identified regulator-regulated gene interactions were compiled in Table S1 (http://www.genetics.org/supplemental/), where *P* values were given by location analysis in LEE *et al.* (2002). From Table S1 we can see that those regulator-regulated gene interactions predicted by the proposed structural equation model had small *P* values in genome-wide location analysis (LEE *et al.* 2002), which indicated that those predicted regulator-regulated gene interactions were

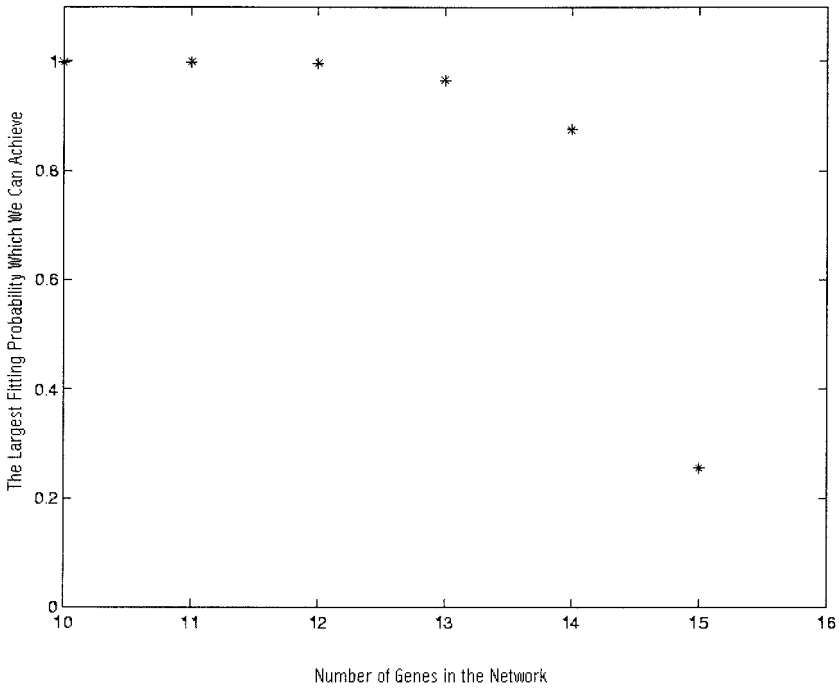

FIGURE 6.—The largest fitting probability that we can achieve after 100 iterations of genetic algorithms as a function of the number of genes in the genetic networks.

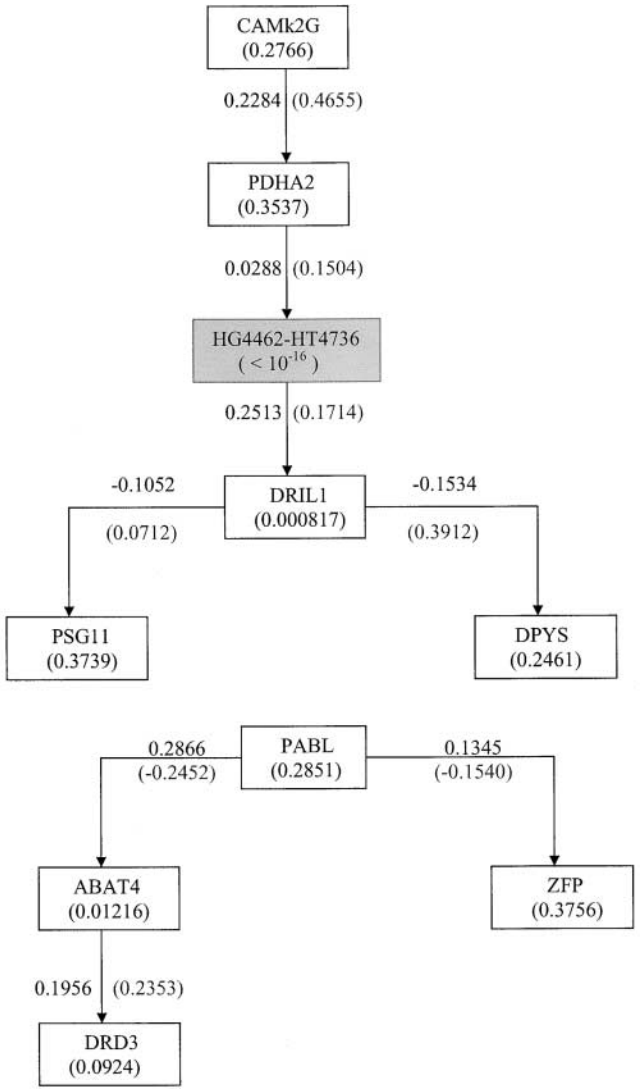

FIGURE 7.—The most significantly differentially expressed genetic network among the 200 genetic networks with 10 genes for the MM data set where AIC = −65.70 and the fitting probability = 1.000. The numbers in parentheses below the name of the gene denote the *P* value of evidence of showing differential expression of the individual gene, the numbers along the edge denote the gene regulatory effect in the tumor tissues, and the numbers in parentheses along the edge denote the gene regulatory effect in the normal tissues.

confirmed by results of genome-wide location analysis experiments.

Differentially expressed genetic networks: Differentially expressed genetic networks are a property of the network as a whole. The differential expression of the genetic network may be due to the differential expression of some individual genes in the network or other factors such as gene-gene interaction. To show that highly differentially expressed genetic networks may contain highly differentially expressed genes, we analyzed the expression profiles of 5483 genes using oligonucleotide arrays in 74 multiple-myeloma (MM) tissue samples and 31 normal tissue samples (ZHAN *et al.* 2002). These data were transformed by a natural logarithm and normalized by subtracting the mean of each gene and then dividing by its standard deviation. Genetic algorithms were applied to the data set 200 times to search for the most likely genetic networks with 10 genes that best fit the data. The AIC values for the resulting 200 genetic networks ranged from −56.31 to −67.55 and the fitting probability ranged from 0.9982 to 1. For each resulting network we calculated the test statistic T^2 and *P* values for testing the difference in expression of the genetic network between normal and abnormal samples. A specific AIC value was taken as a threshold and all genetic networks whose AIC values were larger than the threshold were discarded. We then ranked the genetic networks according to their T^2 values. The genetic network representing the most significant difference in the T^2 test had a *P* value $<10^{-16}$ and is shown in Figure 7. This network consisted of two subnetworks. The *P* value of one subnetwork was $<10^{-16}$ and another subnetwork had a *P* value of 0.043. Several features emerged from Figure 6. First, the gene *HG4462-HT4736* (immunoglobulin heavy chain) that was most significantly differentially expressed in the set of total genes (*P* value $<10^{-16}$) was included in the subnetwork with a significantly differentially expressed subnetwork. Second, in the network we observed another differentially expressed gene, *DRIL1* (*P* value = 0.000817), which was directly linked to the gene *HG4462-HT4736*. The gene *DRIL1* is a dead-ringer transcription regulator and was recently identified as an oncogene (PEEPER *et al.* 2002). Third, the remaining genes in the network were not significantly differentially expressed. This demonstrated that the differential expression of the genetic network is a systems property, which does not imply that all genes in the network are differentially expressed.

The differential expression of the genetic network may be largely due to the differential expression of some genes in the network. However, this need not always be the case. For example, it is possible that all genes in a genetic network are not highly differentially expressed, but the network as a whole is highly differentially expressed. To show this, we analyzed the expression profiles for 12,531 genes using an Affymatrix oligonucleotide array in 50 normal and 52 tumor prostate tissues (SINGH *et al.* 2002). Again, genetic algorithms were applied to this data set 200 times to search for the most likely genetic networks with 10 genes. The AIC values for the resulting 200 genetic networks ranged from −36.31 to −61.84 and the fitting probabilities ranged from 0.53132 to 0.99999. The second most significantly differentially expressed genetic network for the prostate data set, which is shown in Figure 8, had an AIC value of −52.93, a fitting probability of 0.9953, and a *P* value for testing significance of differential expression of the genetic network of 2.93×10^{-11}. However, the *P* value of the most significantly differentially expressed gene

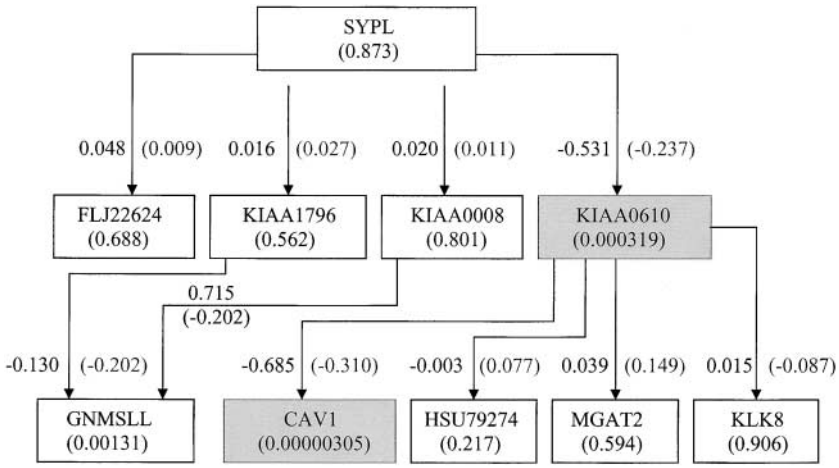

FIGURE 8.—The second-most significantly differentially expressed genetic network among the 200 genetic networks with 10 genes for the prostate data set where AIC = −52.93 and the fitting probability = 0.9953. The numbers in parentheses below the name of the gene denote the *P* value of evidence of showing differential expression of the individual gene, the numbers along the edge denote the gene regulatory effect in the normal tissues, and the numbers in parentheses along the edge denote the gene regulatory effect in the tumor tissues.

in the network (*CAV1*) was equal to 3.047×10^{-6} and was much larger than that of the genetic network as a whole. In addition, although 2 genes, *GNMSLL* (*P* value = 0.001308) and *KIAA0610* (*P* value = 0.0003191), showed mild evidence of significance of difference expressions, the remaining 7 genes in the network did not show any evidence of differential expression. This example demonstrated that the genetic network as a whole might be more significantly differentially expressed than the individual genes in the network. It was reported that the gene *CAV1* was involved in breast cancer (FIUCCI *et al.* 2002; LEE *et al.* 2002) and ovarian carcinoma (WIECHEN *et al.* 2001).

Differentially regulated genetic networks: Identification of differentially regulated genetic networks consists of three steps. First, we reconstruct genetic networks using structural equations and gene expression data in all available samples. Second, we fix the structure of the genetic networks and then estimate network parameters by using gene expression data of normal and abnormal samples. Third, we rank the genetic networks according to some statistics, which measure the extent of the difference in regulatory effects of the genetic networks between normal and abnormal tissue samples.

There are three important cases: (i) the genetic network is differentially regulated but not differentially expressed; (ii) the genetic network is differentially expressed but not differentially regulated; or (iii) the genetic network is both differentially regulated and expressed. We first use the largest difference of the gene regulatory effect in the network between normal and abnormal samples as a measure to quantify the difference in regulation of the network. Then we compare all five measures. The most differentially regulated genetic network for the MM data set that had an AIC value of −65.45 and a fitting probability of 1 is plotted in Figure 9. The network with 10 genes was partitioned into two subnetworks: one subnetwork with 8 genes and one subnetwork with 2 genes. The largest difference of the gene regulatory effect was 2.7953 ($T_{G_0} = 25.68$, *P* value = 0.00062), which was associated with the regulation of the gene *DF* (D component of complement adipsin) on *AX1*, where the *P* value was obtained by a permutation test. From Figure 9, we could also observe that other regulatory effects in the network for the tumor and normal samples were not significantly different. It was interesting to note that the gene *DF* (*P* value = 6.77×10^{-9}) and the receptor *AX1* (*P* value= 4.78×10^{-10}) as well as the network (*P* value = 3.87×10^{-14}) were differentially expressed. The expression of the genes *AX1* and *DF* and the fitted structural equation line of the expression of gene *DF* as a function of the expression of the gene *AX1* in tumor and normal samples are shown in Figure 10. The slope of the line represented the regulatory effect of the gene *DF* on the gene *AX1*. We could clearly see the different regulatory effects in the tumor and normal samples from Figure 10. It was reported that the gene *DF* was a novel serine protease (VOLANAKIS and NARAYANA 1996) and was involved in myeloid cell differentiation (WONG *et al.* 1999). The gene *AX1* was a tyrosine kinase receptor and was recently found downregulated in mature bone marrow-derived dendritic cells (CHEN *et al.* 2002).

The secondmost differentially regulated genetic network for the MM data set with 10 genes that had an AIC value of −63.17 and a fitting probability of 0.9999 is shown in Figure 11. Again, the network was partitioned into two subnetworks: one subnetwork with 4 genes and one subnetwork with 6 genes. The largest difference in the regulatory effect was 2.523 ($T_{G_0} = 21.99$, *P* value = 0.001), which was associated with the regulation of the gene *ABCA2* on the gene *GABA-A*. It was interesting that the *P* value for testing the differential expression of this subnetwork (with 6 genes) was equal to $P = 0.2672$. Also we can see from Figure 11 that neither *ABCA2* nor *GABA-A* was differentially expressed. This demonstrated that differentially regulated genetic networks may not be differentially expressed. Expression of *ABCA2* and *GABA-A* and the fitted structural equation line of *GABA-A* expression as a function of

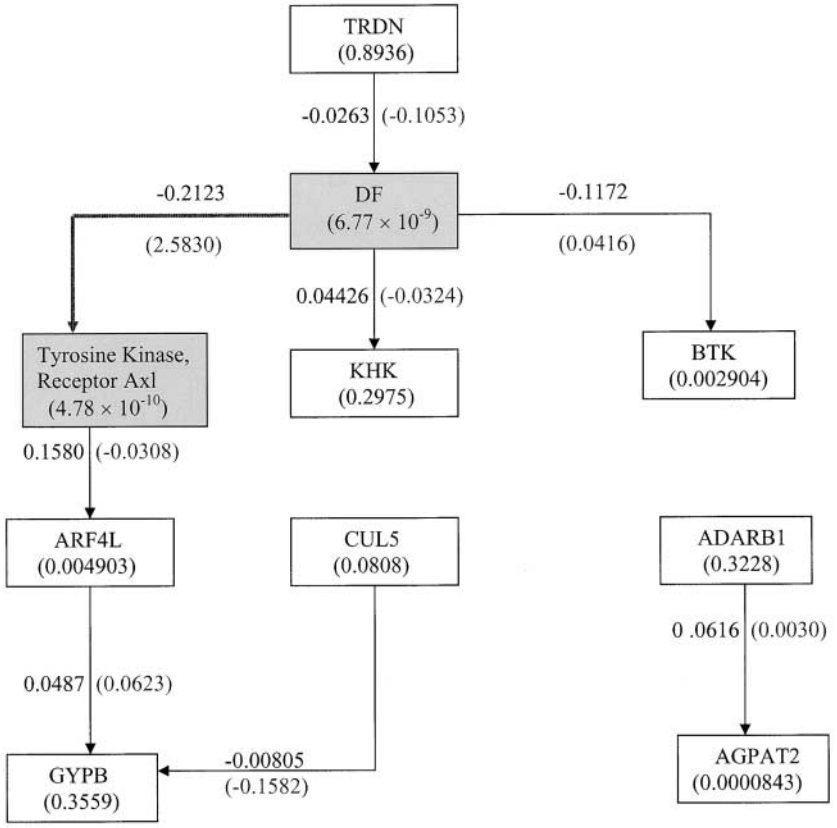

FIGURE 9.—The most significantly differentially regulated genetic network among the 200 genetic networks with 10 genes for the MM data set where AIC = −65.45 and the fitting probability = 1.0000. The numbers in parentheses below the name of the gene denote the *P* value of evidence of showing differential expression of the individual gene, the numbers along the edge denote the gene regulatory effect in the tumor tissues, and the numbers in parentheses along the edge denote the gene regulatory effect in the normal tissues.

the expression of *ABCA2* in tumor and normal samples are shown in Figure 12. We can see from Figure 12 that tumor and normal samples cannot be separated by the expression of *GABA-A* and *ABCA2*; however, the slope of the structural equation lines of *GABA-A* on *ABCA2* in tumor and normal samples can be significantly different. It was reported that *ABCA2* was a regulator of neural transmembrane lipid transport (SCHMITZ and KAMINSKI

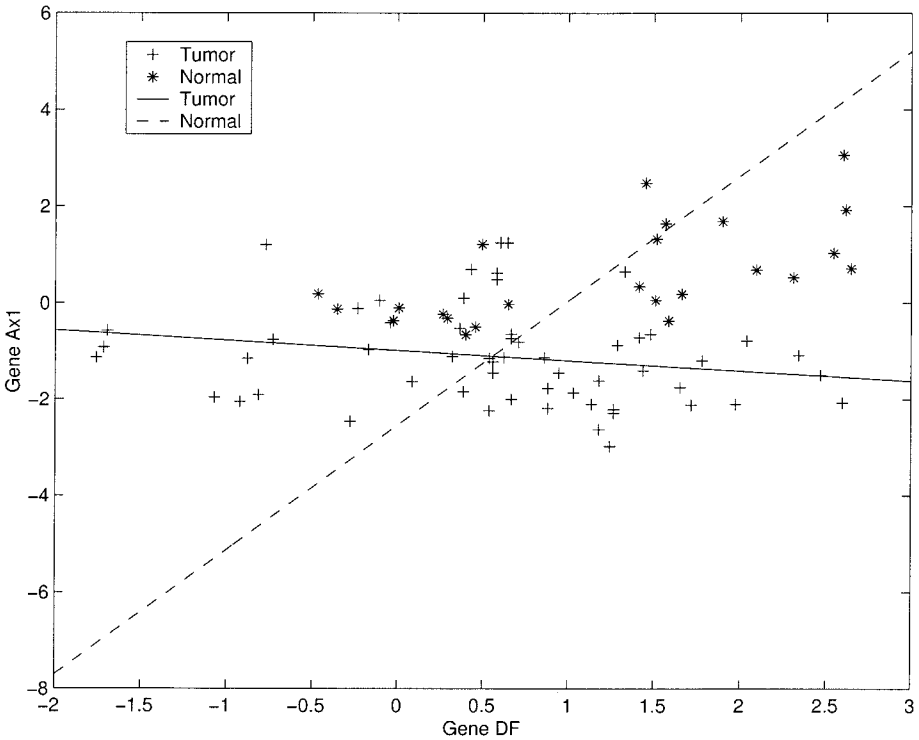

FIGURE 10.—Fitted structural equation line of the expression of the gene *DF* as a function of the expression of the gene *AX1*.

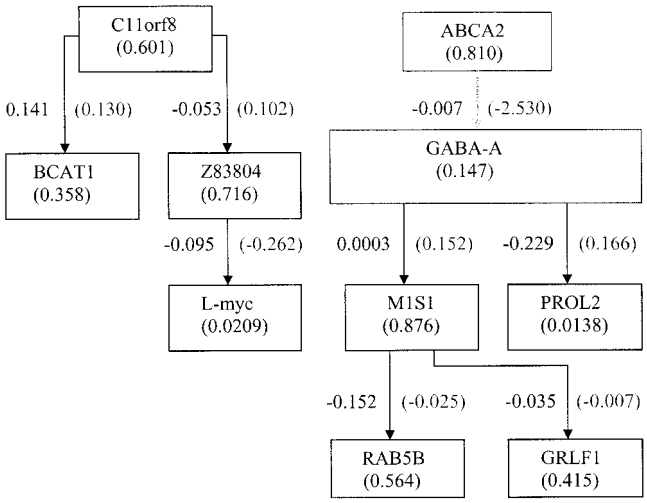

FIGURE 11.—The second-most significantly differentially regulated genetic network among the 200 genetic networks with 10 genes for the MM data set where AIC = −63.17 and the fitting probability = 0.9999. The numbers in parentheses below the name of the gene denote the *P* value of evidence of showing differential expression of the individual gene, numbers along the edge denote the gene regulatory effect in the tumor tissues, and the numbers in parentheses along the edge denote the gene regulatory effect in the normal tissues.

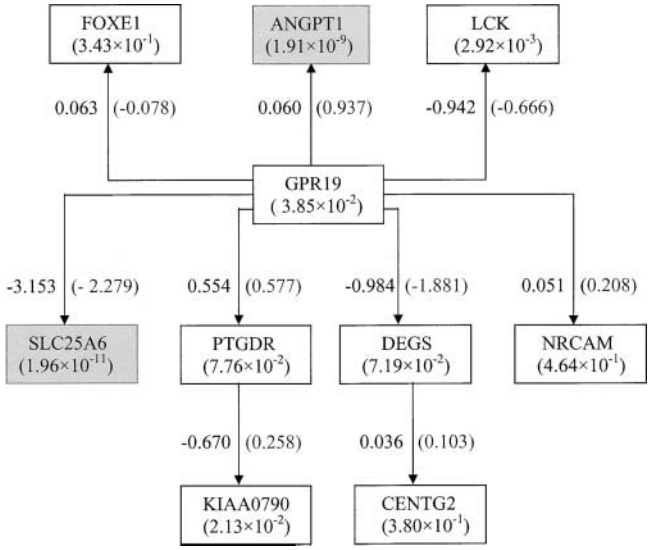

FIGURE 13.—The most significantly differentially expressed genetic network among the 200 genetic networks with 10 genes for the prostate data set where AIC = −51.56 and the fitting probability = 0.9827. The numbers in parentheses below the name of the gene denote the *P* value of evidence of showing differential expression of the individual gene, the numbers along the edge denote the gene regulatory effect in the normal tissues, and the numbers in parentheses along the edge denote the gene regulatory effect in the tumor tissues.

2002) and played a role during myelination (ZHOU *et al.* 2002). The GABA-A receptor gene was a neurotransmitter receptor gene (IWAMA and GOJOBORI 2002) associated with epilepsy (BOWSER *et al.* 2002) bipolar disorder (PAPADIMITRIOU *et al.* 2001). It was also reported that GABA-A was an inhibitory regulator for the migration of SW 480 colon carcinoma cells (JOSEPH *et al.* 2002) and played a role in breast cancer (GARIB *et al.* 2002), renal carcinoma (DOTAN *et al.* 2000), and hepatocellular carcinoma (ZHANG *et al.* 2000).

The most significantly differentially expressed genetic

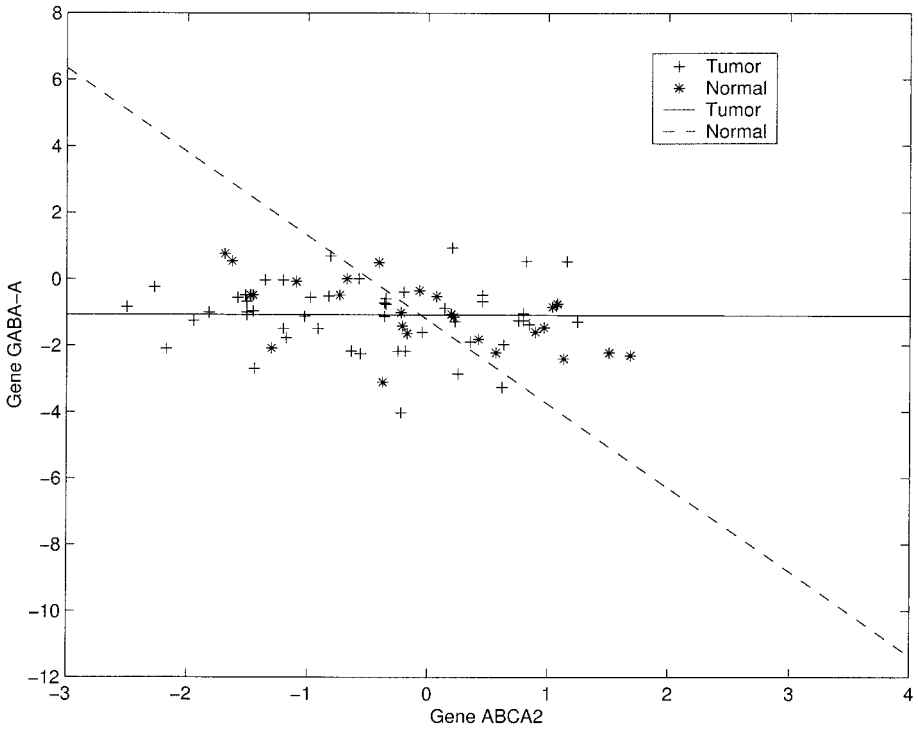

FIGURE 12.—Fitted structural equation line of the expression of the gene *GABA-A* as a function of the expression of the gene *ABCA2*.

TABLE 1

Correlation coefficients between rankings of the genetic networks for prostate normal and tumor tissue sample gene expression profiles made by the five metrics

	$\max_{i,j}\|w_{ij}\|$	$\|\|W\|\|_1$	$\|\|W\|\|_\infty$	$\|\|W\|\|_2$	$\|\|W\|\|_E$
$\max_{i,j}\|w_{ij}\|$	1	0.9418	0.9864	0.9921	0.9839
$\|\|W\|\|_1$		1	0.9234	0.9676	0.9509
$\|\|W\|\|_\infty$			1	0.9841	0.9773
$\|\|W\|\|_2$				1	0.9885
$\|\|W\|\|_E$					1

network for the prostate data set, which is shown in Figure 13, had an AIC value of −51.56, a fitting probability of 0.9827, and a *P* value for testing the significance of differential expression of 2.47×10^{-12}. The gene *SLC25A6* (*P* value = 1.96×10^{-11}) and the gene *ANGPT1* (*P* value = 1.91×10^{-9}) in the network showed significant evidence of differential expressions. However, the rank of the genetic network in the differentially regulated genetic network for the prostate data was 122. The largest difference in the gene regulatory effects of the network was 0.877 ($T_G = 8.5253$, *P* value = 0.15). This demonstrated that although this genetic network was highly differentially expressed, it was not differentially regulated.

To compare the five metrics for characterizing the difference in regulation of the genetic networks under different conditions, we present Table 1, which shows the correlation coefficients between rankings of the genetic networks made by the five metrics. We can see that the correlation coefficients between rankings of the five metrics were very high. This suggested that the five metrics can provide similar evidence showing differential regulation of the genetic networks in normal and abnormal tissues.

DISCUSSION

Genetic networks have two aspects: structure of the networks and strength of the interaction between the genes in the networks. To understand comprehensively genetic networks, in addition to studying the nature of structure, we also need to quantify the strength of the interaction between the genes. Due to the large variation in observed gene expression profiles, quantitative models for genetic networks may not be accurate, but they will still be a useful tool for guiding experiments and understanding complex biological systems, particularly when advances in experimental technologies are made and the precision of experimental data is improved.

Regulation of genetic networks has a cause-effect feature. Causal inference may provide an ideal conceptual framework for reconstruction of genetic networks. In the past decades, several causal inference tools have been developed. Pearl (2000), who coined the term "Bayesian networks in 1985, expressed his preference for functional causal models in his book *Causality: Models, Reasoning, and Inference*. His reasoning is as follows. First, the functional causal models are more general than Bayesian networks. Second, the functional causal models share more features with human intuition. Finally, some concepts that are ubiquitous in human language can be described only in functional causal models. The general form of functional causal models is structural equations. In this report, we proposed structural equations as a useful tool for quantitatively studying genetic networks.

Identification of genetic networks consists of two steps: parameter estimation and structure discovery. In the first step, we assume that the structure of the network is known. A remarkable feature of the regulatory relation among genes in the network is that the expression levels of the genes are determined by the simultaneous interaction of the regulatory relations in the network. Using ordinary regression and the least-squares method for estimation of the parameters will result in inconsistent estimates of the parameters in the network. The proposed structural equation models and estimation procedures based on covariance analysis can avoid this problem and lead to consistent estimates of the parameters in the networks.

The second step is to identify the structure of the networks when it is unknown. The genetic networks that best fit the data may not be truly physically connected, but can reveal causal relations between variables in the network and predict the behavior of biological systems. We used model selection to accomplish this task. Since searching optimal models from an extremely large number of potential networks is computationally expensive, we proposed using genetic algorithms to search the most likely genetic networks fitting the data.

Structure discovery is, essentially, to identify causal relations between variables in the model. The definition of cause has three crucial components: isolation, association, and direction of influence (Bollen 1989). Much of the debate about causal relations comes from inability to completely isolate the variables. Although structural equation models make various assumptions to approximate isolation, it is impossible to achieve perfect isolation in practice. Therefore, the limitation of structural equation models for genetic networks is that they may not reveal true causal relations of the variables in the models, which will affect their precision in predicting the behavior of biological systems.

When genetic networks are reconstructed, either from experiments or from computational modeling, it is essential to link genetic networks with cell function. It has been noted that the function of complex systems is accomplished through networks (Hasty *et al.* 2002). One step toward linking genetic networks with cell phe-

notypes is to characterize genetic networks in terms of cell phenotypes. It is known that the generalized T^2 statistic has a close relationship with discriminant analysis. The highly differentially expressed genetic networks can discriminate the phenotypes of the cell. Therefore, differentially expressed genetic networks can provide information on linking genetic networks to function of cells. Here, we also need to point out that the generalized T^2 test statistic uses only information from the expression levels of the genes in the networks, and does not explicitly use information on structure of the networks although it includes the covariance between expression levels of the genes in the network. It will be useful to develop test statistics to explore both structural information and expression profiles of the genes in the network in the future.

Knowing differential expressions of the genetic networks is not enough for understanding complex biological systems. Differences in gene expressions do not directly reflect strength between gene activities. The differentially expressed genes may not be the cause of diseases. To overcome this problem, we proposed the concept of differentially regulated genetic networks. Therefore, to investigate how the activities of the genetic networks influence the phenotypes of cells, we need not only to discover the structure of genetic networks, which specify the interaction of genes in the networks, but also to quantify the strength of the interaction and to measure the effect of induction and repression of target genes on other genes. Identification of differentially regulated genetic networks may help us to discover the cause of diseases. We proposed five statistics to measure the differences in regulation of the genetic networks, some of which quantify the differences in regulatory effects of the individual pair of the regulator and regulated genes, while others take differences in regulatory effects of all genes in the network into account. We showed that in one important application these five measures will identify a common set of differentially regulated genetic networks.

The expression level of each gene is a nonlinear function of its regulatory input. Linear tools for reconstruction of genetic networks can only approximately model the response of the gene to the regulatory input. Therefore, in some cases, the identified genetic networks based on the linear structural equations may not represent the true biological systems. To overcome this problem, we need to develop nonlinear structural equation models for genetic networks in the future.

The authors thank Joshua M. Akey for his helpful comments on this article, which helped to improve its presentation. We also thank the associate editor, Bruce Walsh, and two anonymous reviewers for helpful comments on the manuscript, which led to much improvement of the article. M. M. Xiong and L. Jun are supported by National Institutes of Health-National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIH-NIAMS) grant IP50AR44888 and NIH grant ES09912.

LITERATURE CITED

Akutsu, T., S. Miyano and S. Kuhara, 2000 Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. J. Comput. Biol. **7:** 331–343.

Anderson, T. W., 1984 *An Introduction to Multivariate Statistical Analysis*, Ed. 2. John Wiley & Sons, New York.

Arluison, V., G. Batelier, M. Ries-Kautt and H. Grosjean, 1999 RNA:pseudouridine synthetase Pus1 from Saccharomyces cerevisiae: oligomerization property and stoichiometry of the complex with yeast tRNA(Phe). Biochimie **81:** 751–756.

Arnold, J., H.-B. Schuttler, D. Logan, J. Griffith, B. Arpinar *et al.*, 2004 Metabolomics, in *Handbook of Industrial Mycology*. Marcel Dekker, New York (in press).

Battogtokh, D., D. K. Asch, M. E. Case, J. Arnold and H.-B. Schuttler, 2002 An ensemble method for identifying regulatory circuits with special reference to the *qa* gene cluster of Neurospora crassa. Proc. Natl. Acad. Sci. USA **99:** 16904–16909.

Bertsekas, D. P., 1995 *Nonlinear Programming*. Athena Scientific, Belmont, MA.

Bollen, K. A., 1989 *Structural Equations With Latent Variables*. John Wiley & Sons, New York.

Bornaes, C., J. G. Petersen and S. Holmberg, 1992 Serine and threonine catabolism in *Saccharomyces cerevisiae*: the CHA1 polypeptide is homologous with other serine and threonine dehydratases. Genetics **131:** 531–539.

Bowser, D. N., D. A. Wagner, C. Czajkowski, B. A. Cromer, M. W. Parker *et al.*, 2002 Altered kinetics and benzodiazepine sensitivity of a GABAA receptor subunit mutation [gamma 2(R43Q)] found in human epilepsy. Proc. Natl. Acad. Sci. USA **99:** 15170–15175.

Brown, P. O., and D. Botstein, 1999 Exploring the new world of the genome with DNA microarrays. Nat. Genet. **21:** 33–37.

Chen, T., H. L. He and G. M. Church, 1999 Modeling gene expression with differential equations. Pac. Symp. Biocomput. **4:** 29–40.

Chen, Z., J. R. Gordon, X. Zhang and J. Xiang, 2002 Analysis of the gene expression profiles of immature versus mature bone marrow-derived dendritic cells using DNA arrays. Biochem. Biophys. Res. Commun. **290:** 66–72.

Cho, R. J., M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway *et al.*, 1998 A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell **2:** 65–73.

Datta, S., 2001 Exploring relationships: a partial least square approach. Gene Exp. **9:** 257–264.

D'Haeseleer, P., X. Wen, S. Fuhrman and R. Somogyi, 1999 Linear modeling of mRNA expression levels during CNS development and injury. Pac. Symp. Biocomput. **4:** 41–52.

Dotan, Z. A., T. Litmanovitch, Y. Ravia, N. Oniashvili, L. Leibovitch *et al.*, 2000 Modification in the inherent mode of allelic replication in lymphocytes of patients suffering from renal cell carcinoma: a novel genetic alteration associated with malignancy. Genes Chromosomes Cancer **27:** 270–277.

Duncan, O. D., 1975 *Introduction to Structural Equation Models*. Academic Press, New York.

Figeys, D., and D. Pinto, 2001 Proteomics on a chip: promising developments. Electrophoresis **22:** 208–216.

Fiucci, G., D. Ravid, R. Reich and M. Liscovotch, 2002 Caveolin-1 inhibits anchorage-independent growth, anoikis and invasiveness in MCF-7 human breast cancer cells. Oncogene **21:** 2365–2375.

Friedman, N., M. Linial, L. Nachman and D. Pe'eré, 2000 Using Bayesian networks to analyze expression data. J. Comput. Biol. **7:** 601–620.

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

Garib, V., B. Niggemann, K. S. Zanker, L. Brandt and B. S. Kubens, 2002 Influence of non-volatile anesthetics on the migration behavior of the human breast cancer cell line MDA-MB-468. Acta Anaesthesiol. Scand. **46:** 836–844.

Graybill, A. A., 1976 *Matrices With Applications in Statistics*, Ed. 2. Wadsworth International Group, Belmont, CA.

Haavelmo, T., 1943 The statistical implications of a system of simultaneous equations. Econometrica **11:** 1–12.

Hartemink, A. J., D. K. Gifford, T. S. Jaakkola and R. A. Young, 2001 Using graphical models and genomic expression data to

statistically validate models of genetic regulatory networks. Pac. Symp. Biocomput., 422–433.

Hasty, J., D. Mcmillen and J. J. Collins, 2002 Engineered gene circuits. Nature **420:** 224–230.

Houseman, B. T., J. H. Huh, S. J. Kron and M. Mrksich, 2002 Peptide chips for the quantitative evaluation of protein kinase activity. Nat. Biotechnol. **20:** 270–274.

Hughes, T. R., and D. D. Shoemaker, 2001 DNA microarrays for expression profiling. Curr. Opin. Chem. Biol. **5:** 21–25.

Ideker, T., V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler *et al.*, 2001 Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science **292**: 929–934.

Imoto, S., T. Goto and S. Miyano, 2002 Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. Pac. Symp. Biocomput., 175–186.

Iwama, H., and T. Gojobori, 2002 Identification of neurotransmitter receptor genes under significantly relaxed selective constraint by orthologous gene comparisons between humans and rodents. Mol. Biol. Evol. **19:** 1891–1901.

Jong, H. D., 2002 Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. **9:** 67–103.

Jordan, M. I., 1999 *Learning in Graphical Models*. MIT Press, Cambridge, MA.

Joseph, J., B. Niggemann, K. S. Zaenker and F. Entschladen, 2002 The neurotransmitter gamma-aminobutyric acid is an inhibitory regulator for the migration of SW 480 colon carcinoma cells. Cancer Res. **62:** 6467–6469.

Koch, C., and K. Nasmyth, 1994 Cell cycle regulated transcription in yeast. Curr. Opin. Cell Biol. **6:** 451–459.

Larranaga, P., M. Poza, Y. Yurramendi, R. H. Marga and C. M. H. Kuijpers, 1996 Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters. IEEE Trans. Patt. Anal. Mach. Intell. **18:** 912–926.

Lee, H., D. S. Park, B. Razani, R. G. Russell, R. G. Pestell *et al.*, 2002 Caveolin-1 mutations (P132L and null) and the pathogenesis of breast cancer: caveolin-1 (P132L) behaves in a dominant-negative manner and caveolin-1 (−/−) null mice show mammary epithelial cell hyperplasia. Am. J. Pathol. **161:** 1357–1369.

Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Barjoseph *et al.*, 2002 Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science **298:** 799–804.

Lennon, K., R. Pretel, J. Kesselheim, S. Heesen and M. A. Kukuruzinska, 1995 Proliferation-dependent differential regulation of the dolichol pathway genes in Saccharomyces cerevisiae. Glycobiology **5:** 633–642.

Liang, S., S. Fuhrman and R. Somogyi, 1998 Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac. Symp. Biocomput. **3:** 18–29.

Lipschutz, R. J., S. P. A. Fodor, T. R. Gingeras and D. J. Lockhardt, 1999 High density synthetic oligonucleotide arrays. Nat. Genet. **21** (Suppl.): 20–24.

Lockhart, D. J., and E. A. Winzeler, 2000 Genomics, gene expression and DNA arrays. Nature **405:** 827–836.

Mann, M., 1999 Quantitative proteomics. Nat. Biotechnol. **17:** 954–955.

Maruyama, G. M., 1998 *Basics of Structural Equation Modeling*. SAGE Publications, Thousand Oaks, CA.

McInerny, C. J., J. F. Partridge, G. E. Mikesell, D. P. Creemer and L. L. Breeden, 1997 A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G1-specific transcription. Genes Dev. **11:** 1277–1288.

McLuckey, S. A., and J. M. Wells, 2001 Mass analysis at the advent of the 21st century. Chem. Rev. **101:** 571–606.

O'Connor, J. P., and C. L. Peebles, 1992 PTA1, an essential gene of Saccharomyces cerevisiae affecting pre-tRNA processing. Mol. Cell. Biol. **12:** 3843–3856.

Papadimitriou, G. N., D. G. Dikeos, G. Karadima, D. Aaramopoulos, E. G. Daskalopoulou *et al.*, 2001 GABA-A receptor beta3 and alpha5 subunit gene cluster on chromosome 15q11-q13 and bipolar disorder: a genetic association study. Am. J. Med. Genet. **105:** 317–320.

Pearl, J., 2000 *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK/London/New York.

Peeper, D. S., A. Shvarts, T. Brummelkamp, S. Douma, E. Y. Koh *et al.*, 2002 A functional screen identifies hDRIL1 as an oncogene that rescues RAS-induced senescence. Nat. Cell Biol. **4:** 148–153.

Ronen, M., R. Rosenberg, B. L. Shraiman and U. Alon, 2002 Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. Proc. Natl. Acad. Sci. USA **99:** 10555–10560.

Schmitz, G., and W. E. Kaminski, 2002 ABCA2: a candidate regulator of neural transmembrane lipid transport. Cell. Mol. Life Sci. **59:** 1285–1295.

Shipley, B., 2000 *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, Cambridge, UK/London/New York.

Shmulevich, I., E. R. Dougherty and W. Zhang, 2002 Gene perturbation and intervention in probabilistic Boolean networks. Bioinformatics **18:** 1319–1331.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola *et al.*, 2002 Gene expression correlates of clinical prostate cancer behavior. Cancer Cell **1:** 203–209.

Volanakis, J. E., and S. V. Narayana, 1996 Complement factor D, a novel serine protease. Protein Sci. **5:** 553–564.

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

Wahde, M., and J. Hertz, 2000 Coarse-grained reverse engineering of genetic regulatory networks. Biosystems **55:** 129–136.

Wiechen, K., L. Diatchenko, A. Agoulnik, K. M. Scharff, H. Schober *et al.*, 2001 Caveolin-1 is down-regulated in human ovarian carcinoma and acts as a candidate tumor suppressor gene. Am. J. Pathol. **159:** 1635–1643.

Wong, E. T., D. E. Jenne, M. Zimmer, S. D. Porter and C. B. Gilks, 1999 Changes in chromatin organization at the neutrophil elastase locus associated with myeloid cell differentiation. Blood **94:** 3730–3736.

Woolf, P. J., and Y. Wang, 2000 A fuzzy logic approach to analyzing gene expression data. Physiol. Genomics **3:** 9–15.

Wright, S., 1921 Correlation and causation. J. Agric. Res. **10:** 557–585.

Young, R. A., 2000 Biomedical discovery with DNA arrays. Cell **102:** 9–15.

Zhan, F., J. Hardin, B. Kordsmeier, K. Bumm, M. Zheng *et al.*, 2002 Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. Blood **99:** 1745–1757.

Zhang, M. Q., 1999 Large-scale gene expression data analysis: a new challenge to computational biologists. Genome Res. **9:** 681–688.

Zhang, M., Y. Gong, N. Assy and G. Y. Minuk, 2000 Increased GABAergic activity inhibits alpha-fetoprotein mRNA expression and the proliferative activity of the HepG2 human hepatocellular carcinoma cell line. J. Hepatol. **32:** 85–91.

Zhou, C. J., N. Inagaki, S. J. Pleasure, L. X. Zhao, S. Kikuyama *et al.*, 2002 ATP-binding cassette transporter ABCA2 (ABC2) expression in the developing spinal cord and PNS during myelination. J. Comp. Neurol. **451:** 334–345.

Communicating editor: J. B. Walsh