A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays

Tianjiao Chu[1],*, Clark Glymour[2], Richard Scheines[1] and Peter Spirtes[2]

[1]Department of Philosophy, Carnegie Mellon University and [2]Department of Philosophy, Carnegie Mellon University and Institute for Human and Machine Cognition, University of West Florida

ABSTRACT

Motivation: One approach to inferring genetic regulatory structure from microarray measurements of mRNA transcript hybridization is to estimate the associations of gene expression levels measured in repeated samples. The associations may be estimated by correlation coefficients or by conditional frequencies (for discretized measurements) or by some other statistic. Although these procedures have been successfully applied to other areas, their validity when applied to microarray measurements has yet to be tested.

Results: This paper describes an elementary statistical difficulty for all such procedures, no matter whether based on Bayesian updating, conditional independence testing, or other machine learning procedures such as simulated annealing or neural net pruning. The difficulty obtains if a number of cells from a common population are aggregated in a measurement of expression levels. Although there are special cases where the conditional associations are preserved under aggregation, in general inference of genetic regulatory structure based on conditional association is unwarranted.

Contact: tchu@andrew.cmu.edu

INTRODUCTION

Two fundamentally different strategies have been proposed to determine networks of regulatory relationships among genes. One strategy (Yuh *et al.*, 1998; Ideker *et al.*, 2001; Davidson *et al.*, 2002) experimentally suppresses (or enhances) the expression of one or more genes, and measures the resulting increased or decreased expression of other genes. The method, while laborious, has proved fruitful in unraveling small pieces of the regulatory networks of several species. Its chief disadvantage is

that each experiment provides information only about the effects of the manipulated gene or genes. A single knockout of gene A resulting in changed expression of genes B and C, for example, does not of itself provide information as to whether A regulates both B and C directly, or whether A regulates B which in turn regulates C, etc. This implies that to identify a regulatory network, the number of experiments required will be super exponential in the number of distinct genes in the network. The requisite statistical procedures are essentially confined to the estimation of the expression level of each gene considered in each experiment, and of the uncertainties of those estimates.

A second strategy relies on the natural variation of expression levels of the same gene in different cells. The proposal is to measure—typically with microarrays—the expression levels in repeated samples from the same tissue source, or similar sources, and to infer the regulatory structure from the statistical dependencies and independencies among the measured expression levels (Akutsu *et al.*, 1998; D'haeseleer, 2000; D'haeseleer *et al.*, 2000; Friedman *et al.*, 2000; Hartemink, 2001; Liang *et al.*, 1998; Shrager *et al.*, 2002; Yoo *et al.*, 2002). The apparent advantage of the strategy is that it offers the possibility of determining multiple relationships without separate experimental interventions. If, for example, gene A regulates gene C only by regulating gene B which in turn regulates C, the expression level of A should be independent, or nearly independent, of the expression level of gene C conditional on the expression level of gene B. In principle, if adequate sample sizes were available, the method could also be used as a supplement to gain additional information from experiments in which the expression of particular genes are experimentally suppressed or enhanced. The requisite statistical procedures for this strategy are more elaborate, and require direct or indirect (e.g. implicit in the posterior probabilities)

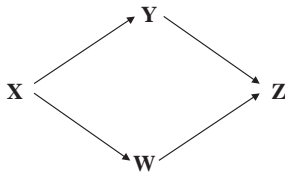*To whom correspondence should be addressed.

Fig. 1. A simple gene regulatory network.

estimates of conditional probability relationships among expression levels.

There are many statistical obstacles to the second strategy including: the joint influence of unmeasured factors (e.g. unmeasured gene expressions or extra-cellular factors), a variety of sources of measurement error, an unknown family of probability distributions governing the errors, and functional dependencies for the expression of any gene that may be Boolean for some regulating genes and continuous for other regulators. Some of these difficulties—in particular the presence of latent common causes—have, in principle, been overcome. (Spirtes *et al.*, 2001). We describe a more elementary statistical difficulty with the second strategy that calls its value into question and raises a set of important research problems.

DIRECTED ACYCLIC GRAPHS AND MARKOV FACTORIZATION

Qualitative regulatory relationships among genes are often represented by directed graphs. Each vertex is a random variable whose values represent levels of expression of a particular gene. Each directed edge from a variable X to a variable Y in such a graph indicates that X produces a protein that regulates Y. In principle, the graph may be cyclic or acyclic, and may even have self-loops—a directed edge from a variable to itself—but most proposed search methods have been confined to acyclic graphs. In the simplest case, one assumes an acyclic graph with noises and random measurement errors for each measurement of each gene that are independent of those for any other gene.

We consider a simplest case: the true, but unknown regulatory structure can be represented by a directed acyclic graph, with independent errors. Consider, for example, four genes, X, Y, Z, W whose regulatory connections can be represented by Figure 1.

Suppose that the measured values of X, Y, Z, W satisfy:

$$Z = f(Y, W) + \epsilon_z$$
$$Y = g(X) + \epsilon_y \qquad (1)$$
$$W = h(X) + \epsilon_w$$

Where f, g, h are any functions and ϵ_z, ϵ_y, ϵ_w are independently distributed noises. It follows that the joint probability density of Z, Y, W, X admits a Markov factorization

$$d(X, Y, Z, W) = d(Z|Y, W)d(Y|X)d(W|X)d(X) \quad (2)$$

The Markov factorization implies that Y, W are independent conditional on X, and that X, Z are independent conditional on Y, W, and is in fact equivalent to specifying that these two relationships hold. More generally, assuming each random variable has an independent noise source but is otherwise a deterministic function of its parents in the graph, the system described by any directed acyclic graph has a density that admits a Markov factorization that can be written of as the product, over all variables, of the density of each variable conditional on its parent variables in the graph. Markov equivalent graphs imply the same independencies and conditional independencies. In the example of Figure 1, the Markov equivalence class consists of the graph shown and the graphs obtained by reorienting exactly one of the edges from X to Y or X to W. Absent extra knowledge from other sources, the Markov equivalence class represents the most information that could be obtained from conditional independencies among the variables.

Where data are obtained in a time series, regulatory relationships can still be represented by a directed acyclic graph and probabilities admitting a Markov factorization, but with vertices appropriately labeled by gene and time.

SUMS OF VARIABLES AND PRESERVATION OF CONDITIONAL INDEPENDENCE

The aim is to discover the regulatory structure in individual cells, but measurements are typically of relative concentrations of mRNA transcripts obtained from thousands, or even millions, of cells. Such measurements are not of variables such as X in the graph above, but are instead, ideally, of the sum of the X values over many cells. We will denote such measured sums over n cells by $\sum_{i=1}^{n} X_i$.

The difficulty with the second strategy for regulatory structure inference, which relies on the statistical dependencies among the gene expression levels, is that the conditional dependencies/independencies among the gene expression levels of a single cell in general are not the same as those among the sums of gene expression levels over a number of cells. For example, if the variables in Figure 1 are binary, and each measurement is of the aggregate of transcript concentrations from two or more cells, $\sum_{i=1}^{n} X_i$, $\sum_{i=1}^{n} Z_i$ are not independent conditional on $\sum_{i=1}^{n} Y_i$, $\sum_{i=1}^{n} W_i$, and the associations obtained from repeated samples will not therefore satisfy the Markov factorization (Danks and Glymour, 2002).

Interestingly, there are some special cases where the conditional independencies are invariant under aggregation. For example, if binary regulatory relations among

genes X, Y and Z are described by a singly connected graph, i.e. $X \longrightarrow Y \longrightarrow Z$ or $X \longleftarrow Y \longleftarrow Z$ or $X \longleftarrow Y \longrightarrow Z$, then the implied conditional independence of X, Z given Y holds as well for sums of independent measurements of X, Y and Z respectively (Danks and Glymour, 2002).

Linear, normal distributions have special virtues for invariance. Whatever the directed acyclic graph of cellular regulation may be, if the noise terms, as in Equation (1), are normally distributed and each variable is a linear function of its parents and an independent Gaussian noise, then the Markov factorization holds for the summed variables. For in that case, conditional independence is equivalent to vanishing partial correlation, and the partial correlation of the two variables, each respectively composed of the sum of n-like variables, will be the same as the partial correlation of the unsummed variables.

Two less restrictive sufficient conditions for conditional independence of variables to be the same as the conditional independence of their sums, are given in the following two theorems. The general setting is an acyclic graph such that each node is a function—not necessarily additive—of its parents and an independent noise term.

THEOREM 1 (LOCAL MARKOV THEOREM). *Given an acyclic graph G representing the causal relations among a set* $\mathbf{V}$ *of random variables. Let* $Y, X^1, \ldots, X^k \in \mathbf{V}$, *and* $\mathbf{X} = \{X^1, \ldots, X^k\}$ *be the set of parents of Y in G. If* $Y = \mathbf{c}^{\mathrm{T}}\mathbf{X} + \epsilon,$[†] *where* $\mathbf{c}^{\mathrm{T}} = (c^1, \ldots, c^k)$, *and* ϵ *is a noise term independent of all non-descendents of Y, then Y is independent of all its non-parents, non-descendents conditional on its parents* $\mathbf{X}$, *and this relation holds under aggregation.*

PROOF. Let $\mathbf{U}$ be the set of the variables in $\mathbf{V}$ that are neither parents nor descendents of Y. That Y is independent of $\mathbf{U}$ conditional on its parents $\mathbf{X}$ is a direct consequence of the local Markov condition for acyclic graphs (Spirtes *et al.*, 2001).

Let Y_i, ϵ_i, $\mathbf{X}_i$, and $\mathbf{U}_i$ be the ith i.i.d. copy of Y, ϵ, $\mathbf{X}$, and $\mathbf{U}$ respectively, we have,

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}(\mathbf{c}^{\mathrm{T}}\mathbf{X}_i + \epsilon_i) = \mathbf{c}^{\mathrm{T}}\sum_{i=1}^{n}\mathbf{X}_i + \sum_{i=1}^{n}\epsilon_i.$$

Clearly, $(\epsilon_1, \ldots, \epsilon_n)$ is independent of $(\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{U}_1, \ldots, \mathbf{U}_n)$. This means that $\sum_{i=1}^{n}\epsilon_i$ is independent of $(\sum_{i=1}^{n}\mathbf{U}_i, \sum_{i=1}^{n}\mathbf{X}_i)$, which again implies that $\sum_{i=1}^{n}\epsilon_i$ is independent of $\sum_{i=1}^{n}\mathbf{U}_i$ conditional on $\sum_{i=1}^{n}\mathbf{X}_i$. Consequently, $\mathbf{c}^{\mathrm{T}}\sum_{i=1}^{n}\mathbf{X}_i + \sum_{i=1}^{n}\epsilon_i$ is independent of $\sum_{i=1}^{n}\mathbf{U}_i$ given $\sum_{i=1}^{n}\mathbf{X}_i$. (Note that $\mathbf{c}^{\mathrm{T}}\sum_{i=1}^{n}\mathbf{X}_i$ is a

constant conditional on $\sum_{i=1}^{n}\mathbf{X}_i = \mathbf{x}$, where $\mathbf{x}$ is an arbitrary constant vector.) □

The above theorem states that, under the local linearity condition, the conditional independence relation between a random variable and its non-descendent and non-parent is invariant under aggregation. In the next theorem, we give another sufficient condition for the conditional independence relation to be invariant under aggregation.

THEOREM 2 (MARKOV WALL THEOREM). *Given an acyclic graph G representing the causal relations among a set* $\mathbf{V}$ *of random variables. Let* $\mathbf{X} = \{X^1, \ldots, X^h\}$, $\mathbf{Y} = \{Y^1, \ldots, Y^k\}$, $\mathbf{W} = \{W^1, \ldots, W^m\}$, *and* $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{W} = \mathbf{V}$. *Suppose that the following three conditions hold:*

(1) *The joint distribution of* $X^1, \ldots, X^h, Y^1, \ldots, Y^k$ *is multivariate normal with non-singular covariance matrix.*

(2) *For* $i = 1, \ldots, k$, Y^i *is neither a parent, nor a child, of any variable* $W^j \in \mathbf{W}$. *That is, there is no direct edge between a variable in* $\mathbf{Y}$ *and a variable in* $\mathbf{W}$.

(3) *For* $i = 1, \ldots, h$, X^i *is not a child of any variable* $W^j \in \mathbf{W}$. *That is, if there is an edge between a variable in* $\mathbf{X}$ *and a variable in* $\mathbf{W}$, *the direction of the edge must be from the variable in* $\mathbf{X}$ *to the variable in* $\mathbf{W}$.

Then conditional on $\mathbf{X}$, $\mathbf{Y}$ *is independent of* $\mathbf{W}$, *and this relation holds under aggregation.*

PROOF. The conditional independence of $\mathbf{Y}$ and $\mathbf{W}$ given $\mathbf{X}$ is obvious, because $\mathbf{W}$ can be represented as a function of $\mathbf{X}$ and some other random variables independent of $(\mathbf{X} \cup \mathbf{Y})$. [‡]

Now let $\mathbf{Z} = (X^2, \ldots, X^h, Y^1, \ldots, Y^k)^{\mathrm{T}}$, suppose the joint distribution of X^1 and $\mathbf{Z}$ is:

$$\begin{bmatrix} X^1 \\ \mathbf{Z} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \vec{\nu} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \vec{\alpha}^{\mathrm{T}} \\ \vec{\alpha} & \Sigma_Z \end{bmatrix}\right).$$

Let $\mathbf{Z}_i = (X_i^2, \ldots, X_i^h, Y_i^1, \ldots, Y_i^k)^{\mathrm{T}}$, which is the ith i.i.d. copy of $\mathbf{Z}$, we are going to show that X_1^1 is independent of $\sum_{i=1}^{n}\mathbf{Z}_i$ given $\sum_{i=1}^{n}X_i^1$. First, let us see the joint distribution of X_1^1, $\sum_{i=1}^{n}X_i^1$, and $\sum_{i=1}^{n}\mathbf{Z}_i$:

$$\begin{bmatrix} X_1^1 \\ \sum_{i=1}^{n}X_i^1 \\ \sum_{i=1}^{n}\mathbf{Z}_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ n\mu \\ n\vec{\nu} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_1^2 & \vec{\alpha}^{\mathrm{T}} \\ \sigma_1^2 & n\sigma_1^2 & n\vec{\alpha}^{\mathrm{T}} \\ \vec{\alpha} & n\vec{\alpha} & n\Sigma_Z \end{bmatrix}\right).$$

We claim that conditional on $\sum_{i=1}^{n}X_i^1 = nx$ and $\sum_{i=1}^{n}\mathbf{Z}_i = n\vec{z}$, the mean of X_1^1 is x.

[†] In this and the next theorems, we shall use the same bold face symbol to represent both a set of variables, and a vector of that set of variables.

[‡] More precisely, these variables are the exogenous variables in $\mathbf{W}$ and the independent noise terms associated with the endogenous variables in $\mathbf{W}$.

Note that:

$$E[X_1^1 \Big| \sum_{i=1}^n X_i^1 = nx, \sum_{i=1}^n \mathbf{Z}_i = n\vec{z}]$$

$$= \mu + \begin{bmatrix} \sigma_1^2 & \vec{\alpha}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} n\sigma_1^2 & n\vec{\alpha}^{\mathrm{T}} \\ n\vec{\alpha} & n\Sigma_Z \end{bmatrix}^{-1} \begin{bmatrix} nx - n\mu \\ n\vec{z} - n\vec{\nu} \end{bmatrix}.$$

Let $\vec{\beta}^{\mathrm{T}} = n\vec{\alpha}^{\mathrm{T}}(n\Sigma_Z)^{-1}, \gamma = 1/(n\sigma_1^2 - \vec{\beta}^{\mathrm{T}}n\vec{\alpha})$, inverting by partition, we have:

$$\begin{bmatrix} n\sigma_1^2 & n\vec{\alpha}^{\mathrm{T}} \\ n\vec{\alpha} & n\Sigma_Z \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \gamma & -\gamma\vec{\beta}^{\mathrm{T}} \\ -\gamma\vec{\beta} & (n\Sigma_Z)^{-1}[I + (n\vec{\alpha})\gamma\vec{\beta}^{\mathrm{T}}] \end{bmatrix}.$$

It then can be shown that:

$$\begin{bmatrix} \sigma_1^2 & \vec{\alpha}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} n\sigma_1^2 & n\vec{\alpha}^{\mathrm{T}} \\ n\vec{\alpha} & n\Sigma_Z \end{bmatrix}^{-1} = \begin{bmatrix} 1/n & \vec{0}^{\mathrm{T}} \end{bmatrix}.$$

It then follows:

$$E[X_1^1 \Big| \sum_{i=1}^n X_i^1 = nx, \sum_{i=1}^n \mathbf{Z}_i = n\vec{z}]$$

$$= \mu + \begin{bmatrix} 1/n & \vec{0}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} nx - n\mu \\ n\vec{z} - n\vec{\nu} \end{bmatrix} = x.$$

The conditional variance of X_1^1 given $\sum_{i=1}^n X_i^1 = nx$ and $\sum_{i=1}^n \mathbf{Z}_i = n\vec{z}$ is:

$$\mathrm{Var}\left(X_1^1 \Big| \sum_{i=1}^n X_i^1 = nx, \sum_{i=1}^n \mathbf{Z}_i = n\vec{z} \right)$$

$$= \sigma_1^2 - \begin{bmatrix} \sigma_1^2 & \vec{\alpha}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} n\sigma_1^2 & n\vec{\alpha}^{\mathrm{T}} \\ n\vec{\alpha} & n\Sigma_Z \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1^2 \\ \vec{\alpha} \end{bmatrix}$$

$$= \frac{n-1}{n}\sigma_1^2.$$

Thus, we have shown that both the conditional mean and the conditional variance of X_1^1 is constant in $n\vec{z}$. Given that the conditional distribution of X_1^1 is normal, this implies that X_1^1 is independent of $\sum_{i=1}^n \mathbf{Z}_i$ given $\sum_{i=1}^n X_i^1$. Note that by the same argument, we could show that, conditional on $\sum_{i=1}^n X_i^1$, X_1^1 is independent of $\sum_{i=1}^n X_i^2, \ldots, \sum_{i=1}^n X_i^h$. Let $\mathbf{X}_i$ be the ith copy of $\mathbf{X}$, it follows that, conditional on $\sum_{i=1}^n \mathbf{X}_i$, X_1^1 is independent of $\sum_{i=1}^n \mathbf{Y}_i$. Because the choice of X_1^1 is arbitrary, we actually have shown that, conditional on $\sum_{i=1}^n \mathbf{X}_i$, X_i^j is independent of $\sum_{i=1}^n \mathbf{Y}_i$ for any $1 \leq i \leq n$ and $1 \leq j \leq h$. Moreover, the joint distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n$

and $\sum_{i=1}^n \mathbf{Y}_i$ conditional on $\sum_{i=1}^n \mathbf{X}_i$ is multivariate normal, and for multivariate normal, marginal independence relations imply the joint independence relation.[§] It then follows that $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is independent of $\sum_{i=1}^n \mathbf{Y}_i$ given $\sum_{i=1}^n \mathbf{X}_i$.

We note that $\mathbf{W}_i$, the ith copy of $\mathbf{W}$, can be represented as a function of $\mathbf{X}_i$ and some other random variables independent of $(\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. Thus, as a function of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and other random variables independent of $(\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, $\sum_{i=1}^n \mathbf{W}_i$ is independent of $\sum_{i=1}^n \mathbf{Y}_i$ given $\sum_{i=1}^n \mathbf{X}_i$. □

Although there are established regulatory mechanisms in which some regulators of a gene act linearly in the presence of a suitable combination of other regulators of the same gene (Yuh *et al.*, 1998), there does not appear to be any known regulatory system that is simply linear. One of the best-established regulatory functional relations seems to be the expression of the Endo16 gene of the sea urchin (Yuh *et al.*, 1998). The expression level of the gene is controlled by a Boolean regulatory switch between two functions, each of which is a product of a Boolean function of regulator inputs multiplied by a linear function of other regulator inputs. Even much simplified versions of such transmission functions do not preserve conditional independence over sums of variables.

Suppose in each of n cells genes X, Y, Z and W have the regulatory structure $X \longrightarrow Y \longrightarrow Z \longleftarrow W$ with $Y = g(X)$; $Z = aYW$, where a is a positive real number, W is Boolean such that $P(W = 1) = p$, and $g(X) = X^2$. Assume without loss of generality that $a = 1$. Assume X takes values in $\{0, 1, 2, 3, 4\}$ with uniform probability. Let $\sum_{i=1}^n X_i$, $\sum_{i=1}^n Y_i$, $\sum_{i=1}^n Z_i$ and $\sum_{i=1}^n W$ denote the sums of values of X, Y, Z and W respectively over $n = 4$ cells. Z is independent of X given Y; however, we will show that $\sum_{i=1}^n Z_i$ is not independent of $\sum_{i=1}^n X_i$ given $\sum_{i=1}^n Y_i$.

For each cell i, Z_i is Y_i if the value of W_i is 1, and zero otherwise. Hence the probability that $Z_i = y_i$ given that $Y_i = y_i$ is p. Let $\sum_{i=1}^n Y_i = \sum_{i=1}^n X_i^2 = 16$. There are just five possible vector values for $\mathbf{X} = \langle X_1, X_2, X_3, X_4 \rangle$ consistent with $\sum_{i=1}^n X_i^2 = 16$: $\langle 4, 0, 0, 0 \rangle$; $\langle 0, 4, 0, 0 \rangle$; $\langle 0, 0, 4, 0 \rangle$; $\langle 0, 0, 0, 4 \rangle$ and $\langle 2, 2, 2, 2 \rangle$. The first four vectors in the list have $\sum_{i=1}^n X_i = 4$ and the last has $\sum_{i=1}^n X_i = 8$. We show that the probability $\sum_{i=1}^n Z_i = 16$ given that $\sum_{i=1}^n Y_i = 16$ and $\sum_{i=1}^n X_i = 4$ is not in general equal to the probability that $\sum_{i=1}^n Z_i = 16$ given that $\sum_{i=1}^n Y_i = 16$ and $\sum_{i=1}^n X_i = 8$.

For example, if $\mathbf{X} = \langle 4, 0, 0, 0 \rangle$, then $\sum_{i=1}^n Z_i$ equals 16 if and only if $W_1 = 1$. The probability that $W_1 = 1$ is p. Similarly for the vectors $\langle 0, 4, 0, 0 \rangle$, $\langle 0, 0, 4, 0 \rangle$ and $\langle 0, 0, 0, 4 \rangle$. Given that $\sum_{i=1}^n X_i = 4$ and $\sum_{i=1}^n Y_i =$

[§] Suppose X, Y, Z are multivariate normal. If X is independent of Y, and X is also independent of Z, then X is independent of (Y, Z).
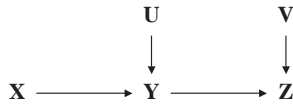
Fig. 2. A Sea Urchin type regulatory network.

$\sum_{i=1}^{n} X_i^2 = 16$, the set of the first four vectors has probability 1, and each individual vector of the first four has probability 0.25. Therefore the probability that $\sum_{i=1}^{n} Z_i = 16$ given that $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} X_i^2 = 16$ and that $\sum_{i=1}^{n} X_i = 4$ is p. On the other hand, the probability that $\mathbf{X} = \langle 2, 2, 2, 2 \rangle$ is 1 given that $\sum_{i=1}^{n} X_i = 8$ and $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} X_i^2 = 16$. The probability that $\sum_{i=1}^{n} Z_i = 16$ given $\sum_{i=1}^{n} Y_i = 16$ and $\sum_{i=1}^{n} X_i = 8$ is therefore just the probability that $W_i = 1$ for $i = 1, 2, 3, 4$, which is p^4.

Although we have no general, interesting sufficient condition for invariance to fail, many of the assumptions in the preceding example, e.g. that $n = 4$, that X is uniformly distributed, that X has 5 distinct values, that $Y = X^2$, are obviously inessential, and $Y = X^2$ was used only because it is the simplest non-linear, non-Boolean function proposed for a regulator (Schilstra, 2002). (Note that by the previous results if the dependency of Z were linear in Y and additive in a function of W, the conditional independence would hold for the sums of variables.) Similar arguments would apply to a variety of non-linear dependencies of Y on X. For example, consider the Sea Urchin type causal structure shown in Figure 2, where $Y = UX$ and $Z = VY$. Suppose X has a Poisson distribution with parameter λ, U and V are Bernoulli random variables with parameters p_1 and p_2 respectively.

It is obvious that X and Z are independent conditional on Y. However, it can be shown that this relation does not hold under aggregation. For example, let X_1, U_1, Y_1, V_1, Z_1 and X_2, U_2, Y_2, V_2, Z_2 be two independent samples generated from the same causal structure. Assuming that U and V are not degenerate, that is, $p_1, p_2 \neq 1$ and $p_1, p_2 \neq 0$, through straightforward calculation, we can show that:

$$P(Z_1 + Z_2 = 2 | Y_1 + Y_2 = 2)$$

$$= p_2 - p_2(1 - p_2) \frac{p_1 \lambda e^{-\lambda}}{1 - p_1 + p_1 e^{-\lambda} + p_1 \lambda e^{-\lambda}}$$

$$P(Z_1 + Z_2 = 2 | Y_1 + Y_2 = 2, X_1 + X_2 = 4) = p_2.$$

Clearly, conditional independence relation is not preserved under aggregation for the causal structure shown in Figure 2, because as long as $p_1, p_2 \neq 1$ and $p_1, p_2 \neq 0$,

$$P(Z_1 + Z_2 = 2 | Y_1 + Y_2 = 2) \neq$$
$$P(Z_1 + Z_2 = 2 | Y_1 + Y_2 = 2, X_1 + X_2 = 4).$$

In the above examples, we treat the number n of cells in an aggregated sample as a constant. In practice, however, as pointed by a referee, when several samples are obtained, the number of cells in each sample is a random variable. This could make the inference of conditional association even more problematic. When n is held constant, we know that there is a fixed set of conditional associations among the aggregated genes, though they are not the same as the genes within each individual cell. If n is a random variable, we are not sure if the aggregated genes in different samples share the same set of conditional associations.

CONCLUSION

The considerations we have advanced argue that, other than by chance, inference to genetic regulatory networks from associations among measured expression levels is possible only if the graphical structure and transmission functions from regulator concentrations to expression concentrations of regulated genes preserve conditional independence relations over sums of i.i.d. units, or if the aggregated variations from unit level conditional independence are small. The few sufficient conditions we have provided are not biologically relevant, but, unfortunately, the negative example based on a simplification of Endo 16 regulation is relevant. We have not as yet found interesting, general sufficient conditions for conditional independence *not* to be invariant.

These results appear to conflict with many reports of successful machine learning searches for regulatory structure. In many cases, however, the successes are with simulated data in which the simulated values for individual cell representatives are not summed in forming the simulated measured values, and are therefore unfaithful to the actual measurement processes. In several other cases results with real data are not independently confirmed, but merely judged plausible. Rarely, results are obtained that agree with independent biological knowledge; in these cases the actual regulatory structure among the genes considered may approximately satisfy invariance of conditional independence for summed variables, or the procedures may simply have been lucky. Feasible, economical techniques for measuring concentrations of transcripts in single cells could make machine learning techniques based on associations of expressions valuable in identifying regulatory structure, but such techniques are not yet available. In the meanwhile, absent biological evidence that regulatory dependencies have the requisite invariance over sums of variables, there seems little warrant for thinking accurate methods are possible for inferring regulatory structures that depend on conditional associations.

We know that two important features of the joint distribution of the gene expression levels—the mean

vector and the covariance matrix—are invariant under aggregation up to a simple linear transformation. More precisely, let $\mathbf{G} = (G^1, \ldots, G^k)^{\mathrm{T}}$ be a random vector representing the expression levels of k genes in a single cell, and $\mathbf{G}_i = (G_i^1, \ldots, G_i^k)^{\mathrm{T}}$ be the ith i.i.d. copy of $\mathbf{G}$ for $i = 1, \ldots, n$, then it is trivial to see that the following two equations hold:

$$n\mathrm{E}[\mathbf{G}] = \mathrm{E}\left[\sum_{i=1}^n \mathbf{G}_i\right]$$

$$n\mathrm{Cov}(\mathbf{G}) = \mathrm{Cov}\left(\sum_{i=1}^n \mathbf{G}_i\right).$$

It is also easy to see that the independence relations between the random variables are invariant under aggregation, for if G^1 and G^2 are independent, then $(G_1^1, \ldots, G_n^1)$ and $(G_1^2, \ldots, G_n^2)$ are also independent, hence $\sum_{i=1}^n G_i^1$ and $\sum_{i=1}^n G_i^2$ are independent. Thus, while waiting for the technologies capable of measuring efficiently the expression levels in single cells, in experimental studies, we can still make valid—although probably more limited—inferences about the regulatory networks based only on the first two moments of the joint distribution and the independence relations.

ACKNOWLEDGMENTS

REFERENCES

Akutsu,T., Miyano,S. and Kuhara,S. (1998) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In *Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms*. pp. 695–702.

Danks,D. and Glymour,C. (2002) Linearity properties of Bayes nets with binary variables. *Proceedings of the Conference on Uncertainty in Artificial Intelligence 2001*. Seattle.

Davidson,E., Rast,J., Oliveri,P., Ransick,A., Calestani,C., Yuh,C., Minokawa,T., Amore,G., Hinman,V., Arenas-Mena,C. *et al.* (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

D'haeseleer,P. (2000) *Reconstructing gene networks from large scale gene expression data*, PhD thesis, University of New Mexico.

D'haeseleer,P., Liang,S. and Somogyi,R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.

Ideker,T., Thorsson,V., Ranish,J., Christmas,R., Buhler,J., Eng,J., Bumgarner,R., Goodlett,D., Aebersold,R. and Hood,L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Friedman,N., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *Recomb 2000*. Tokyo.

Hartemink,A. (2001) *Principle computational methods for the validation and discovery of genetic regulatory networks*, PhD thesis, MIT.

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

Schilstra,M. (2002) NetBuilder http://strc.herts.ac.uk/bio/maria/NetBuilder.

Shrager,J., Langley,P. and Pohorille,A. (2002) Guiding revision of regulatory models with expression data. *Pac. Symp. Biocomput.*, **7**, 486–497.

Spirtes,P., Glymour,C. and Scheines,R. (2001) *Causation, Prediction and Search*. MIT Press, Cambridge.

Yoo,C., Thorsson,V. and Cooper,G.F. (2002) Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. *Pac. Symp. Biocomput.*, **7**, 498–509.

Yuh,C., Bolouri,H. and Davidson,E. (1998) Genomic cis-regulatory logic: experimental and computational analysis of a Sea Urchin gene. *Science*, **279**, 1896–1902.