Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data

Peter Spirtes[a], Clark Glymour[b], Richard Scheines[a], Stuart Kauffman[c], Valerio Aimale[c], Frank Wimberly[c]
[a]Department of Philosophy, Carnegie Mellon University
[b]Institute for Human and Machine Cognition [c]Bios Group

1. Introduction

Through their transcript products genes regulate the rates at which an immense variety of transcripts and subsequent proteins occur. Understanding the mechanisms that determine which genes are expressed, and when they are expressed, is one of the keys to genetic manipulation for many purposes, including the development of new treatments for disease.

Viewing each gene in a genome as a distinct variable that is either on (expresses) or off (does not express), or more realistically as a continuous variable (the rate of expression), the values of some of these variables influence the values of others through the regulatory proteins they express, including, of course, the possibility that the rate of expression of a gene at one time may, in various circumstances, influence the rate of expression of that same gene at a later time. If we imagine an arrow drawn from each gene expression variable at a given time to a gene variable whose expression it influences a short while after, the result is a network, technically a directed acyclic graph (DAG). For example, the DAG in Figure 1 is a representation of a system in which the expression level of gene G_1 at time 1 (denoted as $G_1(1)$) causes the expression level of $G_2(2)$, which in turn causes the expression level of $G_3(3)$. The arrows in Figure 1 which do not have a variable at their tails are "error terms" which represent all of the causes of a variable other than the ones explicitly represented in the DAG. The DAG describes more than associations—it describes causal connections among gene expression rates. A shock to a cell—by mutation, heating, chemical treatment, etc. may alter the DAG describing the relations among gene expressions, for example by activating a gene that was otherwise not expressed, producing a cascade of new expression effects.

Although "knockout" experiments (which lower a gene's expression level) can reveal some of the underlying causal network of gene expression levels, unless guided by information from other sources, such experiments are limited in how much of the network structure they can reveal, due to the sheer number of possible combinations of experimental manipulations of genes necessary to reveal the complete causal network.

Recent developments have made it possible to compare quantitatively the expression of tens of thousands of genes in cells from different sources in a single experiment, and to trace gene expression over time in thousands of genes simultaneously. cDNA microarrays are already producing extensive data, much of it available on the web. Thus there are calls for analytic software that can be applied to microarray and other data to help infer regulatory networks (Weinzierl, 1999). In this paper we will review current techniques that are available for searching for the causal relations between variables, describe algorithmic and data gathering obstacles to applying these techniques to gene expression levels, and describe the prospects for overcoming these obstacles.

2. Bayesian Networks

A number of different models have been suggested for gene expression networks. These include linear models (D'haeseleer et al 1999), nonlinear models (Weaver et al. 1999), and Boolean networks (Kauffman 1993, Somogyi and Sniegoski, 1996). In all of these models all variables are assumed to be observed and the relationships among them are deterministic. Liang, et al, (1998) describe a search, the REVEAL program, for Boolean dependencies in systems free of noise and of unmeasured common causes using mutual information measures. However, the system as described, is not robust over aggregation, omitted common causes, measurement error, non-synchronized cells, or feedback.

Murphy and Mian (1998) and Friedman et al. (1999) have suggested using Bayesian network models of gene expression networks. Among the advantages of Bayesian networks models are that 1) they explicitly relate the directed acyclic graph model of the causal relations among the gene expression levels to a statistical hypothesis; 2) they include all of the aforementioned models, and Hidden Markov Models, as special cases; 3) there are already well developed algorithms for searching for Bayesian networks from observational data (see reviews in Spirtes et al. 2000, and Cooper 1999); 4) they allow for the introduction of a stochastic element and hidden variables; 4) they allow explicit modeling of the process by which the data are gathered.

A Bayesian network consists of two distinct parts: a directed acyclic graph (DAG or belief-network structure) and a set of parameters for the DAG. The DAG in a Bayesian network can be used to represent causal relationships among a set of random variables (such as gene expression levels). A DAG represents the causal relations in a given population with a set of vertices **V** when there is an edge from A to B if and only if A is a direct cause of B relative to **V**. (We adopt the convention that sets of variables are boldfaced.)

2.1. The Causal Markov Assumption

We say that a set of variables **V** is *causally sufficient* when no two members of **V** are caused by a third variable not in **V**. According to the Causal Markov Assumption, each vertex is independent of its non-descendants in the graph conditional on its parents in the graph. For example, in Figure 1, the Causal Markov Assumption entails that $G_3(3)$ is independent of $G_1(1)$ (which is neither a parent nor a descendant of $G_3(3)$), conditional on $G_2(2)$ (which is a parent of $G_3(3)$). The Causal Markov Assumption entails, for example, that if there is no edge between two variable X and Y in a DAG G, then X and Y are independent conditional on some subset of the other variables

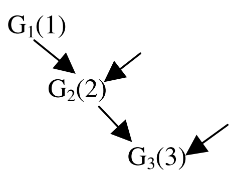

Figure 1: Example 1

2.2. The Causal Faithfulness Assumption

In order to draw any conclusions about the structure of the DAG from an observed sample, one must make some kind of simplicity assumption. One such assumption is the Causal Faithfulness Assumption, which states that any conditional independence relations in the population are entailed by the Causal Markov Assumption. For a number of different parametric families, the set of parameters that lead to violations of the Causal Faithfulness Assumption are Lebesgue measure 0.

Under the Causal Faithfulness Assumption, conditional independence relations give direct (but partial) information about the structure of the graph. For example, in Example 1 of Figure 1 we can conclude that there is no direct edge between $G_1(1)$ and $G_3(3)$ if a statistical test indicates that $G_1(1)$ is independent of $G_3(3)$ conditional on $G_2(2)$.

A number of methods of assigning scores to Bayesian networks based on observed data do not explicitly make the Causal Faithfulness Assumption, but do so implicitly. See Heckerman et al., (1999).

3. Search

3.1. Assuming Causal Sufficiency

We first consider search algorithms when it is assumed that the measured set of variables **V** is causally sufficient, or equivalently, there are no hidden common causes of members of **V**.

The problem of finding the best DAGs from a given sample is difficult because the number of DAGs is super-exponential in the number of observed variables. While background information, such as the time order in which events occur greatly reduces the complexity of the problem, it still remains a large search space.

There are two main approaches to searching for Bayesian network models. The first approach (as exemplified in the PC algorithm, Spirtes, et al., 2000) performs a series of tests of conditional independence on the sample, and uses the results to construct the set of DAGs that most closely implies the results of the tests If the time order of the variables is known (as would be the case in a time series of measurement of gene expression levels) is known, the output is a single DAG. For example, in Figure 1, if a statistical test indicates that the $G_3(3)$ is independent of $G_1(1)$ conditional on $G_2(2)$, the PC algorithm concludes that there is no direct edge from $G_1(1)$ to $G_3(3)$. For either discrete or normally distributed variables, under the Causal Markov and Faithfulness Assumptions the algorithm pointwise (but not uniformly) converges to the correct answer in the large sample limit, and, as long as the maximum number of parents of any given variable is held fixed, the algorithm's complexity is polynomial in the number of measured variables.

The second approach to searching for Bayesian networks assigns a score to each DAG based on the sample data, and searches for the DAG with the highest score. The scores that have been assigned to DAGs for variables that are discrete or distributed normally include posterior probabilities, the Minimum Description Length, and the Bayesian Information Criterion. A variety of methods of search for DAGs with the highest score have been proposed, including hill-climbing, genetic algorithms, and simulated annealing. (Heckerman et al.,1999; Spirtes, et al., 2000). If the time order of the variables is known, then there is in general a single DAG with the highest score. The scores have been shown to be asymptotically correct in the sense that in the large sample limit no DAG receives a higher score than the true DAG. Generally, however, scoring searches are heuristic.

3.2. Not Assuming Causal Sufficiency

When causal sufficiency is not assumed, search for DAGs becomes much more difficult. The FCI algorithm is an extension of the PC algorithm to DAGs with latent variables and (pointwise, but not uniformly) converges to the correct output under the Causal Markov and Faithfulness Assumptions for normal or discrete variables. However, even if the time order of the measured variables is known, the output of the algorithm is not a unique DAG. The informativeness of the output depends whether the set of DAGs output have any interesting features in common. This in turn depends heavily upon the true number or hidden common causes, and their precise causal relationship to the measured variables. (Spirtes, et al., 2000) In the worst case, if there is a hidden common cause of every pair of measured variables, there is essentially no useful information about the true DAG.

It has proved especially difficult to extend score-based searches to hidden variable models. (One heuristic approach is described in Friedman, 1998.) When there is no bound on the number of hidden variables, the search space for scoring algorithms is infinite. Perhaps more important, there are difficult unsolved computational and conceptual problems in calculating scores for hidden variable models (e.g. it is not known whether such scores are asymptotically correct). Moreover, extending scores beyond the discrete and normal cases faces serious difficulties because many other families of distributions are not closed under marginalization.

4. Applying Search to Microarrays

4.1. Sample Size Issues

Simulation experiments on inferring DAG structure from sample data indicate that even for relatively sparse graphs sample sizes of several hundred are required for high accuracy. A typical single microarray chip produces a sample of one for each gene. Thus in order to gather sample sizes of several hundred, data will need to be gathered from hundreds of microarray chips. For example, the sample size in the well known cell cycle experiments of Spellman, et al 1998 is 76. If current trends in decreases in the price of microarray chips continue, then it is may be possible to gather the sample sizes required for high accuracy in the near future.

The price of a microarray chip is related to how many genes it measures. Only a fraction of genes appear to vary in their expression in response to endogenous or exogenous changes (Kauffman, 1993). Selecting a subset of genes that are causally interacting with each other or are reacting to an external stimulus would decrease the cost of gathering a large number of samples. In addition, since the uncertainties of estimation depend on, among other things, the ratio of the number of variables to the number of independent measurements of those variables, the ability to restrict attention to a fraction of the genome is crucial. Spellman, et al. (1998) describes one method of selecting a subset of genes whose expression varied as a function of the cell cycle phase.

Another possible method for selecting a subset of genes is to use clustering algorithms on a relatively small sample, and then select a subset of genes that occur in a single cluster for further analysis. There are a number of clustering techniques that have been proposed for gene expression data, including Eisen et al. (1999), Hastie et al. (2000), and Lazzeroni and Owen (1999). Tibshirani et al. (1999) provides an overview of clustering techniques for gene expression data. However, it is an open question whether any of the clusters created by these different algorithms contain genes that are strongly *causally* interacting with each other.

4.2. Measurement Error

If $G_1(1)$ influences the expression of $G_3(3)$ only indirectly through the influence of $G_2(2)$, the fact that $G_2(2)$ is an intermediate can be recovered if the joint probabilities are known, because the expression level of $G_3(3)$ will be independent of the expression level of $G_1(1)$ conditional on $G_2(2)$. But if the gene expressions are measured with error, the corresponding conditional independence among the measured variables will not hold. Suppose, for example, in Figure 2, G_{M1} measures the true expression level G_1 but the value of G_{M1} is also affected by noise (represented by the error term into G_{M1}), and similarly for G_{M2} and G_{M3}. It will not in general be the case that G_{M1} is independent of G_{M3} conditional on G_{M2}. (This may not be obvious, but like all of our claims about conditional independence relations in particular examples, they can be proved by applying Pearl's d-separation relation to the graph in question. See Pearl (1988)). However, the independence of G_{M1} and G_{M3} conditional on G_{M2} will hold approximately if the measured values are strongly correlated with the true values (i.e. the noise is small), in which case it may be possible to recover useful information about the underlying causal relations between $G_1(1)$ through $G_3(3)$. Information about the size of random measurement error could be gathered by putting multiple copies of the same gene on a single microarray chip, and estimating the variance.

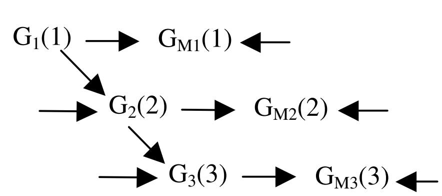

Figure 2: Measurement Error

4.3. Averaging

While the aim is to describe the regulatory network at the cellular level, typical microarray experiments do not measure the concentration of transcripts in a single cell (although with film and laser technology, that is increasingly feasible—see Weinzierl, 1999) but instead of a large collection of cells. That means that the recording from each spot on a microarray is not a measurement of the transcripts from any one cell, but is instead a measurement of the *sum* or *average* of the transcripts (or their concentrations) from a collection of cells.

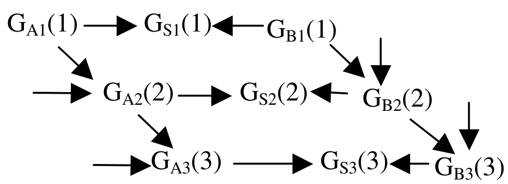

Figure 3: Averaging

Figure 3, shows a hypothetical causal structure for gene expression levels in two cells, A and B. $G_{A1}(1)$ represents the gene expression level of G_1 in cell A at time 1, $G_{B1}(1)$ represents the gene expression level of G_1 in cell B at time 1, and $G_{S1}(1)$ represents the average of $G_{A1}(1)$ and $G_{B1}(1)$. Although $G_{A1}(1)$ and $G_{A3}(3)$ are independent conditional on $G_{A2}(2)$, and $G_{B1}(1)$ and $G_{B3}(3)$ are independent conditional on $G_{B2}(2)$, in general $G_{S1}(1)$ and $G_{S3}(3)$ are *not* independent conditional on $G_{S2}(2)$. However, if the variance of the gene expression levels across different cells is small, $G_{S1}(1)$ and $G_{S3}(3)$ will be approximately independent conditional on $G_{S2}(2)$.

We have also shown that if there are experimentally realizable conditions in which the underlying influences of the genes on one another are approximately linear, or piecewise linear, then the PC and FCI algorithms, for sufficiently large samples and under reasonable further assumptions, recover features of the network structure even from data that consists of averages of gene expression levels from many cells, rather than gene expression levels from individual cells. Linearity is sufficient, we do not know that it is necessary. For example, there are parameterizations of networks of binary variables, so-called noisy "or" gates, that have many of the properties of linear systems (Pearl, 1988, Cheng, 1997, Glymour, in press) and we have not investigated whether they have requisite invariance properties, although we plan to.

4.4. Families of Distributions

Another solution to the problem of averaging is to measure gene expressions in a single cell. Most Bayesian network discovery algorithms have assumed that the data is either normal, or discrete. Gene expression level data, even for single cells, may satisfy neither of these assumptions. One way of applying existing Bayesian network discovery algorithms to non-normal continuous data is to discretize it. However, if continuous variables are collapsed into discrete variables using cutoffs, for example dividing expression levels above and below a certain value into "on" and "off," or twice dividing into "high," "medium" and "low, " the conditional independence relations among the original continuous variables are not in general retained. Hence, further research in this area is needed.

4.5. Other Hidden Common Causes

One or more varying non-genetic factors may influence the expression of multiple genes within the cell. These factors will not be recorded in microarray data, so that the correct graphical architecture would not only include the genes expressed, but also some representation of associations among gene expression produced by unrecorded common causes. As long as there are not too many hidden common causes, search algorithms such as FCI can in principle find useful information about the true structure.

4.6. Feedback

If measurements are taken from synchronized systems at intervals close enough in time so that there is no, or little, feedback the issue does not arise, but if data taken at different times are aggregated, or data are taken from a mixture of non-synchronized cells, the measured expression levels may have been produced by feedback. A finite graphical representation of the conditional independence and causal relations among such variables will be a directed *cyclic* graph. An algorithm for extracting such a graph from suitable data is known for linearly related variables (Richardson, 1994), but no algorithms have been developed for cases in which there are feedback and unmeasured common causes.

4.7. Lack of Synchronization

Even if cells start off in a synchronized state (see Spellman et al. 1998) further complications in analyzing data will occur if the time it takes for one gene expression levels to affect subsequent gene expression levels differs from sample to sample. This is illustrated in Figure 4, where depending on the value of r(1), $G_1(1)$ either affects $G_2(2)$ directly, or $G_2(3)$ directly. This complicates search because it implies that parents in a graph may occur not just at one time step earlier (as assumed in the REVEAL algorithm for example) but parents may occur at multiple time steps earlier. It also implies that a pair of variables may be independent conditional only on subsets containing variables from multiple time segments, complicating the search for conditioning sets which make pairs of variables independent. For example, in Figure 4, $G_1(1)$ and $G_3(4)$ are independent conditional only on subsets of measured variables containing both $G_2(2)$ and $G_2(3)$.

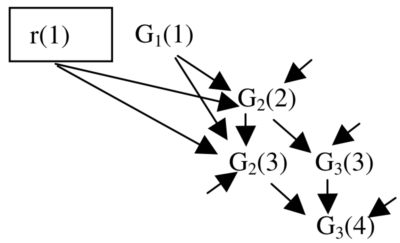

Figure 4: Lack of Synchronization

4.8. Conclusion

Existing proposals for obtaining genetic regulatory networks from microarray data have ignored many of the difficulties of reliable data analysis. The prospects for success depend both upon generalizing the current algorithms (e.g. by extending them to larger classes of distribution families) and upon being able to gather data in a way that simplifies the task of the data analyst. It would greatly improve the prospects for successful application of current techniques if sample sizes of several hundred to several thousand could be gathered, with low measurement error, with each sample either gathered from a single cell or from a collection of cells with very low variance. There do not seem to be any fundamental obstacles to being able, within the next few years, to gather data of the kind that would greatly improve the performance of current Bayesian network discovery algorithms

The research presented here was supported in part by grant DMS-9873442 from the National Science Foundation

5. Bibliography

Cheng, P. (1997) From correlation to causation: causal power theory. *Psychological Review*, 107.

D'haeseleer, P., Wen, X., Furhman, S. and Somogyi. (1999) R. Linear modeling of mRNA expression levels during CNS development and injury. In *Proc. of the Pacific Symp. on Biocomputing*,. 4:41-52

Eisen, M., Spellman, P., Brown, P, and Botstein. D. (1999) Cluster analysis and display of genome-wide expresion patterns. Proce. Natl. Acad. Sci. (in press).

Friedman, N., Lineal, M., Nachman, I., and Pe'er, D. (2000) Using Bayesian Networks to Analyze Expression Data. Accepted to *Journal of Computational Biology*.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. *Proceedings of the 14th International Conference on Machine Learning*.

Glymour, C. and Cooper, G. (1999) *Computation, Causation and Discovery*. Cambridge, MA, MIT Press.

Glymour, C. (in press) Bayes nets and graphical causal models in psychology. MIT press.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizaeh, A. Staudt, L., and Botstein, D. (2000) Gene Shaving: a New Class of Clustering Methods for Expression Arrays. Stanford University Department of Statistics Technical Report.

Heckerman, D., Meek, C., and Cooper, G. (1999) A Bayesian approach to causal discovery. in *Computation, Causation and Discovery*. C. Glymour and G. Cooper. Cambridge, MA, MIT Press.

Kauffman, S. (1993) *The Origins of Order. Self-organization and Selection in Evolution.* Oxford University Press.

Lazzeroni, L. and Owen, A. (1999) Plaid Models for Gene Expression Data. Stanford University Department of Statistics Technical Report.

Liang, S. Fuhrman, S., and Somogyi, R. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing,* 3, 18-29.

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

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, Morgan Kaufmann.

Richardson, T. (1996b) A discovery algorithm for directed cyclic graphs. *Proceedings of the 12th Conference of Uncertainty in AI*, Portland, OR, Morgan Kaufmann: 454-461.

Somogyi, R. and Sniegoski, C. (1996) Modeling the complexity of genetic networks: understanding multigenetic and pleiotropic regulation. *Complexity*, 1, 45-63.

Spellman, P, Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9, 3273-3298.

Spirtes, P. Glymour, C. and Scheines, R. (2000) *Causation, Prediction, and Search*, 2nd edition, MIT Press.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., and Brown, P. Clustering methods for the analysis of DNA microarray data. (1999) Stanford University Department of Statistics Technical Report.

Weaver, C., Workman, C., and Stormo, G. (1999) Modeling regulatory networks with weight matrices. In *Proc. of the Pacific Symp. on Biocomputing*, 4:112-123.

Weinzierl, R. (1999) *Mechanisms of Gene Expression: Structure, Function and Evolution of the Basal Transcriptional Machinery*. World Scientific Publishing Company.