10-610 KDD Lab
Spring 2001
Carnegie Mellon University

Causal Networks for Analysis of Gene Expression Data

Abstract

Our inherited characteristics and metabolism are controlled by the *expression* of our genes: we say that a gene is expressed when it produces proteins that interfere with the internal function of the cell and consequently of the whole organism. Many genes are never expressed. Those that are interact in complex relationships that have never been fully understood.

The discovery of gene regulation pathways is an extremely valuable alternative to solve several medical and biological problems, ranging from the growth of more resistant and nutritive crops to the cure of cancer. However, given the amount of data and genes on existence, it seems clear that this task turns out to be considerable difficult for these more ambitious goals.

This proposal relies on describing this data under the language of causal graphs due to its adequacy to represent the regulatory paths with hidden variables and to provide semantics for the experimental scenarios. The existing Bayesian network learning algorithms have to be specialized to the specific assumptions we have on this problem, and devising modified algorithms will be our major contribution. Simulated data will be used to evaluate the procedure, while real data for the yeast cell will be available, and the resulting model can be compared with existing publications on the literature.

Student: Ricardo Silva rbas@cs.cmu.edu
Advisor: Clark Glymour cg09@andrew.cmu.edu

1 Introduction

Understanding the role played by genes on the cell metabolism leads to important applications such as prevention, identification and treatment of diseases, as well as modifying the genetic structure of living beings in advantageous ways to humans. For instance, Golub et al. (1999) report that patterns of gene activation can be used to trace the evolution of a cancer tumor and aid class prediction (i.e., the cancer type) with better accuracy than traditional techniques that rely on the morphological appearance of the tumor. Advancements on DNA microarrays and other data collection techniques are increasing the feasibility of discovering regulatory connections among genes and proteins by statistical and computational approaches (Dutilh, 1999). Ideally, we would like to know what sets of genes should be controlled to achieve a desired effect.

This is not an easy problem. One expects to find a high number of genes present on a normal cell of a given specie and they may interact in long causal chains. This means that identifying the regulatory paths consists on searching on a very large combinatorial space (Edwards and Glass, 2000). Such a search must be conducted wisely, but scarce prior knowledge is often provided.

However, all prior work on automated procedures for studying gene regulation has ignored the intervening processes of protein synthesis and activation that mediate gene regulation and are of extraordinary biological interest. Forthcoming work in *Science* from the Institute for Systems Biology compares existing models of gene-protein metabolism with experimental data on gene expression from wild yeast strains and strains with specific genes removed, and finds the model and the data do not agree.

Devising modification on current models would require performing tedious and expensive cycles of theoretical formulation of hypothesis followed by experimental analysis. Evidently there is an important place for machine learning procedures for building data driven models. This project proposes the development and evaluation of algorithms for causal network induction from observational and experimental data measuring the activities of gene expression on yeast cells. Causal networks provide a suitable language for expression of important assumptions and interpretation of resulting models. They also provide mechanisms for dealing with intervening actions, i.e., situations where some variables are *controlled* instead of only observed (Glymour and Cooper, 1999; Spirtes, Glymour and Scheines, 2000b; Pearl, 2000).

The availability of data and assumptions contriving the model search, discussed at length on Section 3, are indications of the feasibility of this proposal. A rather simplified and incomplete algorithm is also provided as an illustration on Section 3.

The delivered algorithm intends to be general enough to be used for a large family of cells that satisfies the assumptions. Depending on the degree of success, the results that will follow are going to shed some light on useful aspects of gene expression on a variety of situations and allow biologists, physicians and genetic engineers to integrate this approach on their available analytical tools and further advance our comprehension of the mechanisms of life.

2 Related work

Clustering has been consistently applied in gene expression analysis (Eisen et al., 1998). For instance, the work described on (Spellman et al., 1998) is a well-know reference on analysis of yeast cells using a specific clustering method. However, even though clustering may be useful to detect genes that are correlated, one may want to find out how this correlation happens. Questions such as detecting the influence of common causes and direction of gene regulation cannot be answered by cluster analysis alone.

Kaufmann (1993) exemplifies how Boolean networks can be used to describe many natural phenomena if one is willing to assume that expression levels are realistically represented by binary variables. The REVEAL algorithm by Liang, Fuhrman and Somogyi (1998) consists in constructing Boolean networks using information-theoretical measures, but without mechanisms to handle noisy data. These networks are intended not only to indicate the presence of causal relationship between genes, but also to infer the complete sequence of states that a cell would follow given the initial state. Since a Boolean network has a finite number of states, assuming that these networks are able to model the underlying process accurately, this implies that they also detect the existing cycles of states in a cell (the attractors of a dynamical system). Even though we do not propose modeling dynamical aspects, we still are able to unveil the cell behavior under stable states (i.e., when the concentration of proteins is fixed through time by equilibrium of the influence of each gene) without requiring discretization of local effects and non-noisy data.

D'haeseleer, Liang and Somogyi (2000) survey and compare clustering techniques and Boolean networks on gene expression domains[1]. It is not our intention to develop, on this proposal, algorithms to identify the temporal, dynamical aspect of gene networks. Our solutions are going to be designed to answer questions such as "Does gene X directly cause the expression of gene Y?" ignoring temporal aspects. Our data come from measures made on stabilized cells.

Friedman et al. (2000) provide one of the first accounts of applying Bayesian networks to gene expression data. Friedman et al. do not use any biological prior knowledge or constraints: they couple with the search space complexity by means of a domain-independent heuristic for parental selection. According to the authors, they were able to found many of the same conclusions devised by Spellman et al. (1998). However, their experiments demonstrated a considerable sensitivity to the kind of parameterization used to represent the local parent-child distributions: experiments using linear models and discretized multinomial distributions resulted in quite different relationships. We hope to achieve more stable models by introducing constraints. Friedman et al. also do not account for the effect of hidden variable such as proteins and using experimental data in conjunction with observed data. Cooper and Yoo (1999) formalize an approach to combine this data in a single Bayesian learning algorithm, but again these algorithms were designed for general problems and are arguably sub-optimal.

Spirtes et al. (2000a) provide a more extensive survey on possible applications of Bayesian networks to gene expression discovery, without proposing any specific algorithm. Gibson and Mjolness (2000) briefly give an introduction to biological aspects

---

[1] Boolean networks are also called "genetic networks" under this context.

of gene regulation and survey modeling tools, with a special emphasis on dynamical systems.

3 Approach

3.1 Bayesian and causal networks

Bayesian networks (Pearl, 1988) have been used on a variety of domains, and their success has sparked variations under names like *graphical models* (Jordan, 1998). These models associate probability distributions with graphs and certain characteristics of the distribution are reflected as properties of a graph. Random variables on a distribution are represented as nodes in the respective graphs and an edge between two nodes indicates a dependency between the variables. Much of the common terminology in graph theory is used, and concepts such as parents, descendants, directed paths, cycles appear on the definition of these models.

In general, graphical models are a rich language for knowledge representation, allowing the introduction of statements that cannot be reproduced only by probability distributions, such as causal relations. One of the important advantages of adding the graphical aspect to plain probability calculus is using the machinery of graph theory to describe or derive results that otherwise would be more difficult to obtain given only, for example, a joint distribution or a prior over parameters. If we have a joint distribution with many independence constraints, a Bayesian network is able to encode this same distribution using a much sparser representation (Pearl, 1988).

Usually, modeling with Bayesian networks relies on the following assumptions:

- the underlying data generating process can be represented as a directed acyclic graph (DAG);
- the Markov condition: a node is independent of its non-descendants (excluding its parents) given its parents;
- the faithfulness condition: a independence relation is entailed by the Markov condition applied to a graph if and only if it is also holds on the associated probability distribution.

An edge between two nodes on a Bayesian network does not need to represent a causal relationship. In fact, without the faithfulness condition, one can construct a network out of an arbitrary ordering of variables: when adding a new node Y and considering a previously added node X, one would add an edge $X \rightarrow Y$ if there was no set **S** containing previous nodes such that X would be independent of Y given S (i.e., $X \perp Y \mid \mathbf{S}$). With a proper parameterization, this network could represent various probability families. However, constructing a network following the causal order is expected to result in much simpler models (Russell and Norvig, 1995) and allows the measurement of the effect of interventions in a variety of situations (Spirtes, Glymour and Scheines, 2000b; Pearl, 2000). On causal networks, a directed edge $X \rightarrow Y$ is interpreted as “X is the direct cause of Y”, where a direct cause X is a factor that influences the value of its effect Y regardless of the state of the other factors on the system.

Our samples for yeast cell gene expression are small, consisting of a few dozens of cases for each scenario. Friedman et al. (2000) emphasize that even with the technological advancements in obtaining relevant data, the samples are still very small compared with the amount of variables. However, according to current knowledge, each gene is influenced by at most a few others. In our case, the original space of 6,000 genes will be reduced to a set of only 9 ones that are considered a good approximation of a closed system representing the influence of the galactose nutrient on the metabolism of the cell.

On the next two subsections, we present our approach. Section 3.2 discusses how we are going to represent the mechanisms of gene expression by specialized Bayesian networks. Section 3.3 makes some points about how to generate these models by learning algorithms that search through our constrained hypothesis space. Section 3.4 introduces our measures of evaluation of results.

3.2 Modeling

The basic mechanisms of gene regulation metabolism have a well-known general structure (Weinzierl, 1999): sequences of DNA are copied (transcribed) into messenger RNA, mRNA, which is then spliced to create new mRNA sequences. The mRNA sequences are then translated into chains of amino acids, or proteins. The proteins in turn interact with one another — bind, fold, intertwine, etc. — to form “activated” proteins. Some of the activated proteins in turn control which sequences of DNA are transcribed. Other factors on the cell may exert influence over this process. For examples, nutrients such as galactose may be essential for production of some proteins. Figure 1 illustrates a model of this process using graphical representations that can be interpreted as Bayesian networks.

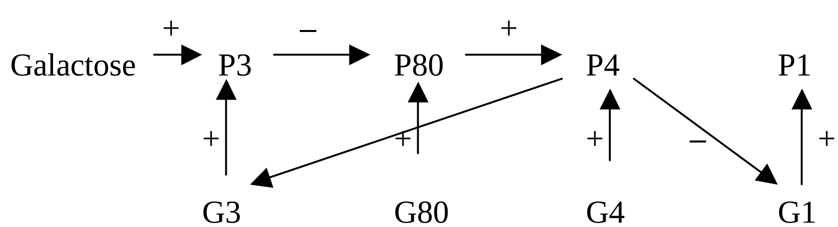

Figure 1: in this simplified model of cell metabolism, *G* nodes represent genes that are either activated or deactivated. *P* nodes represent concentrations of proteins. On this figure, galactose is a nutrient that is required to activate *P3*. Signs attached to the arrows indicate whether increases in concentrations of variables at the tail of the arrow increase or decrease concentrations at the head of the arrow.

In this example, data consist of correlations among *Galactose* and the *G* variables, including experiments in which single *G*s are knocked out experimentally. The proteins are *not* measured. Note some assumptions of the representation. Knocking out a gene is essentially setting it to zero, which sets its respective protein to zero, creating independence relations[2]. Thus *Galactose* is independent of G_1 given that G_3 is knocked out in the model above. *Galactose* is not independent of G_1 if G_3 is activated. The reason is that a nonzero value of G_3 generates a concentration of P_3, and for that concentration of P_3, variations in other inputs to P_3 (in this case, *Galactose*) will vary the concentration of this protein.

The assumptions introduced by this example can be formalized as follows. Figure 2 illustrates some of them. Assumptions 3 and 6 might well be implausible - for now we allow them.

1. The expression level of N genes G_i can be measured, and each directly causes the level of a protein P_i which *cannot* be measured. They are treated as hidden variables, and we specify in advance at least the ones that are directed affected by the genes.
2. No G_i directly causes any other gene G_j.
3. No G_i directly causes any other protein $P_{j \neq i}$.
4. Proteins may cause each other and genes.
5. No cycles among only protein variables.
6. No P_i directly causes its own gene G_i.
7. Relationships are approximately linear in the logs of concentrations, appropriately scaled
8. The Markov and faithfulness conditions are valid;

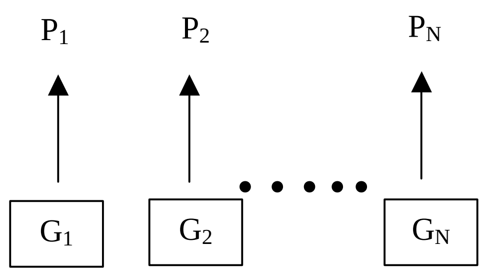

Figure 2: graphical representation of assumptions regarding the interaction between single genes and single proteins.

We are not assuming here that the true graph is acyclic. Cycles among proteins are not allowed, but it is possible to have a path H = ($G_i \rightarrow P_i \rightarrow P_{i+1} \rightarrow \ldots \rightarrow G_i$) starting and finishing on the same gene, but no node on H appear more than once with the exception of G_i. On this case, it is expected that the knockout experiments will be able to break the cycles.

Within these restrictions, we propose algorithms to search among causal graphs motivated by developments on data acquisition techniques performed in the last five years. These advancements have made it possible to measure the concentration of mRNA transcripts produced by cells under various treatments, and also to develop cell lines in which specific genes have been "knocked out" or rendered incapable of transcription. Data will be available to us through the Institute for Systems Biology, Seattle Washington. This data contains samples in which selected genes of yeast cells are active, and also for experiments in which each G_i alone is knocked-out, i.e., set at 0. This

[2] Easily identified using criteria such as d-separation (Pearl, 1988).

intervention also sets P_i at 0. There are no experiments intervening on more than one gene at the same time due to experimental costs. However, if our research finds out that double or higher order knock-outs would greatly increase the amount of structure that could be discovered, it can suggest new experiments.

3.3 Searching

Several approaches for learning the structure of Bayesian networks from data haven been published over the last few years, which are usually classified in two families of strategies (Cooper, 1999):

- score-based search, which assigns a measure of fitness to a Bayesian network as a whole. The score is usually obtained by Bayesian learning, a combination of a prior over parameters, network structures and the likelihood of the model (Cooper and Herskovits, 1992; Heckerman, 1995). The score can be the probability of the data given the model, as usual on model selection.

- constraint-based search, which evaluates each edge individually by detecting independencies using the Markov condition and statistical measures such as partial correlations (Glymour and Cooper, 1999; Pearl, 2000; Spirtes, Glymour and Scheines, 2000b).

Both have advantages and disadvantages[3]. An example of a simplified and incomplete constraint-based algorithm for reconstructing an idealized gene network was recently proposed by Scheines (personal communication). The general idea is starting with a complete graph, where all proteins are connected to each other and to every gene. From this dense graph, we perform tests for each pair of nodes in order to decide if there is a way to make these nodes statistically independent by conditioning on other ones. If this is possible, then the respective have to be pruned, since its presence would indicate that there is no way to block the effect of one node in the other. Then, use these same independencies to determine the directions of the edges, as discussed at length by Spirtes, Glymour and Scheines (2000) for a general class of Bayesian networks.

For the first step, initialize the graph as depicted on Figure 2 and add the following edges:

- Type 1: $P_i \rightarrow G_j$, for all pairs of P_i and G_j nodes where $i \neq j$;
- Type 2: $P_i — P_j$, for all pair of nodes P_i and P_j where $i \neq j$;

The original $G_i \rightarrow P_i$ edges are fixed by assumption and will not be removed at any point of the search algorithm. The task will be identifying edges of type 1 and 2 that are consistent with the data. The algorithm should eliminate edges when there are no direct dependencies between the correspondent proteins and genes. Ideally, it should also indicate the direction for all the remaining edges of type 2. We can partially identify

[3] For example, constraint-based approaches are more computationally efficient, but less robust to sampling variation. Score-based approaches are specially expensive when searching over hidden variables, which are treated more naturally within constraint-based search.

which of the first type do not exist with the two following rules. The symbol $\perp$ represents probabilistic independence.

R1: If $G_i \perp G_j$, then
i) eliminate $P_i \rightarrow G_j$ and ii) eliminate $P_j \rightarrow G_i$

R2: If $G_i \not\perp G_j$ but $G_i \perp G_j \mid G_k$ is knocked-out, then
i) eliminate $P_i \rightarrow G_j$ and ii) eliminate $P_j \rightarrow G_i$

The intuition behind the first rule is that is impossible to have a directed path from G_i to G_j (and vice-versa) if these genes are marginally independent. The second rule is similar, but on this case we are evaluating if the independence holds when another gene G_k is knocked-out. One example where R1 does not hold but R2 is valid is when P_k lies between G_i and G_j on the only directed path that exists between these nodes.

These rules are clearly sound according to the representational assumptions of Bayesian networks, but they are not complete. That is, edges eliminated by them are surely not in the generating graph, but edges left in might not be in the graph. For example, consider the following model:

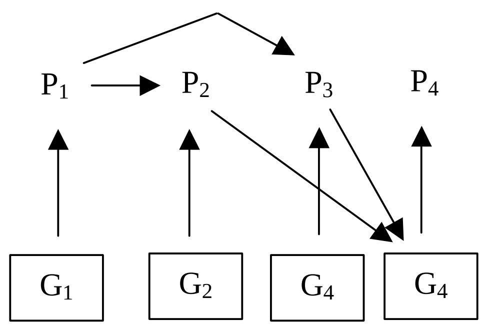

Figure 3: example of a simplified gene network.

The two rules above could not eliminate the edge from $P_1 \rightarrow G_4$ added on the first step, even though it is not there. This is because we need to condition on both P_2 and P_3 at the same time to block all the paths from P_1 to G_4, and we cannot because of the limitation on the knockout experiments[4].

The last step is directing the edges that are not yet directed. However, one cannot expect to direct all edges. For a given probability distribution, there are families of graphs able to hold the same independencies. *This is the main point of our proposal*: the simplified algorithm given in the previous paragraphs is too simple, and general algorithms available on the literature do not add much. *We propose the development of more complete algorithms that unveil as much of the structure as possible while ideally leaving only patterns that could not be distinguished by any other algorithm*[5]. The experimental data alone is sufficient to detect if a gene causes another (G_i is a cause of G_j if and only if $E(G_j)$ when G_i is not manipulated is different from $E(G_j)$ when G_i is

[4] Remember that the each experiment manipulate only one gene.

[5] Proving completeness of this family of algorithms is usually too hard, and in practice we just make comparisons with other existing algorithms.

knocked-out). But this causation can be indirect. We want to reveal a bit more of how this causation happens through the relationship between proteins and genes.

As a final observation, constraint-based search for learning Bayesian networks is usually a problem of exponential complexity on the size of the maximum number of parents of a node. It may turn out that the computational costs for this specific domain are exceedingly high, and definition of heuristics based on them would be necessary.

3.4 Evaluating

We will divide our evaluation procedure in two subtasks:

- testing the effectiveness of the algorithm assuming correctness of the model. In this case, simulated graphs respecting the assumptions delineated on Section 3.2 will be randomly generated, and samples obtained from these idealized models. We will run our algorithm with these samples and compare with the original graphs. With the artificially generated models we will be able to evaluate the structure of the subgraph among the unobserved proteins: measures of success include the number of edges that are correctly pruned and oriented.

- testing the algorithm with real-world experiments with yeast cells. The yeast cell is specially interesting due to the existence of previous published work unveiling part of its regulatory mechanisms (Spellman et al., 1998). This evaluation is not as well defined as the previous one since our current understanding of yeast cells metabolism is incomplete, but we have partial knowledge of how some genes interact on the production of given proteins.

4 Work plan

February, 02 – 23

1. develop algorithms for finding the indistinguishable classes of transcript-protein models of gene expression data with and without single gene knockouts. Both constraint-based and scoring algorithms will be considered, and if strict algorithms cannot be found, or are infeasible, heuristics will be developed.

February, 24 – March, 24

2. implement the algorithms/heuristics, probably in Java.
3. test the algorithms/heuristics on simulated data.

March, 25 – April, 02

4. apply the implementation to data on yeast metabolism to be obtained from the Institute for Systems Biology, Seattle Washington.
5. compare the results with received accounts of cell metabolism (which are known not to be entirely accurate.) and report the final conclusions.

5 Summary

The discovery of the mechanisms of gene expression of several species can unleash new technologies to improve our quality of life by treatment of diseases and new techniques for agriculture, among other applications. However, the complexity of nature mechanisms requires the aid of computer-based data analysis tools. These tools are only now becoming feasible due to new data acquisition methodologies. Our proposal introduces a new category of causal networks that express these relationships in human-understandable format and allow combination of observational and experimental data. It We hope that this approach will be able point out parts of the interactions between genes and proteins and ultimately aid the most ambitious reverse engineering project ever conceived: understand the ways by which life was possible.

References

Cooper, G. (1999). An overview of representation and discovery on causal Bayesian networks. In Glymour, C., and Cooper, G., *Computation, Causation and Discovery,* 3-62. MIT Press.

Cooper, G., and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning,* 9:309-347.

Cooper, G., and Yoo, C. (1999). Causal discovery from a mixture of experimental and observational data. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (UAI 99), 116-125.

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

Dutilh, B. (1999). Analysis of data from microarray experiments, the state of the art in gene network reconstruction. Literature thesis, Theoretical biology and Bioinformatics groups, Utrecht University.

Edwards, R., and Glass, L. (2000). Combinatorial explosion in model gene networks. *Chaos* 10:691-704.

Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95:14683-14868.

Friedman, N., Linial, M., Nachman, I., and Pe'er D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* (to appear).

Gibson, M., and Mjolsness, E. (2000). Modeling the Activity of Single Genes. In Bolouri and Bower, eds.: *Computational Methods in Molecular Biology.* MIT Press (in press).

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

Heckerman, D. (1995). A tutorial on learning Bayesian networks. Microsoft Research technical report MSR-TR-95-06.

Jordan, M. (ed.) (1998). *Learning in Graphical Models*. MIT Press.

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

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

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Russell, S., and Norvig, P. (1995). *Artificial Intelligence: a Modern Approach*. Prentice Hall.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botsein, D. and Futcher, B. (1998). Comprehensive indentification of cell cycle-regulated genes of the yeast *Saccharomyces_cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*:9, 3273-3297.

Spirtes, P., Glymour, C., Scheines, R., Kauffman, S, Aimalte, V. and Winberly, F. (2000a). Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data. *Proceedings of the Atlantic Symposium on Computational Biology*, Genome Information Systems & Technology (to appear).

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

Weinzierl, R. (1999). *Mechanisms of Gene Expression*. Imperial College Press.