

# DISCOVERY OF CAUSAL RELATIONSHIPS IN GENE-REGULATION PATHWAY FROM A MIXTURE OF EXPERIMENTAL AND OBSERVATIONAL DNA MICROARRAY DATA

C. YOO\*, V. THORSSON<sup>^</sup>, and G.F. COOPER\*

\*Center for Biomedical Informatics, University of Pittsburgh  
8084 Forbes Tower, 200 Lothrop St., Pittsburgh PA 15213

<sup>^</sup>The Institute for Systems Biology  
4225 Roosevelt Way NE, Suite 200, Seattle Washington 98105

This paper reports the methods and results of a computer-based search for causal relationships in gene-regulation pathway of galactose metabolism in the yeast *Saccharomyces cerevisiae*. The search uses recently published data from cDNA microarray experiments. A Bayesian method was applied to learn causal networks from a mixture of observational and experimental gene-expression data. The observational data were gene-expression levels obtained from unmanipulated “wild-type” cells. The experimental data were produced by deleting (“knocking out”) genes and observing the expression levels of other genes. Causal relations predicted from the analysis on 36 galactose gene pairs are reported and compared with known galactose pathway. Additional exploratory analyses are also reported.

## 1 Introduction

Causal knowledge makes up much of what we know and want to know in science. Thus, causal modeling and discovery are central to science. Experimental studies, such as biological interventions with corresponding controls, often provide the most trustworthy methods we have for establishing causal relationships from data. In such an experimental study, one or more variables are manipulated and the effects on other variables are measured. On the other hand, observational data result from passive (i.e., non-interventional) measurement of some system, such as a cell. In general, both observational and experimental data may exist on a set of variables of interest. For example, in biology, there is a growing abundance of observational gene expression data. In addition, for selected variables of high biological interest, there are data from experiments, such as the controlled alteration of the expression of a given gene.

Microarray technology has opened a new era in the study of gene regulation. It allows a relatively quick and easy way to assess the mRNA expression levels of many different genes. Large time-series datasets generated by microarray experiments can be informative about gene regulation. Microarray data have been analyzed using classification or clustering methods<sup>1,2</sup> and gene pathway (network) methods<sup>3,4,5,6,7</sup>. Dutilh<sup>8</sup> gives a short review of genetic networks. A more thorough review of genetic networks based on biological context was published by Smolen, et al.<sup>9</sup>. Wessels, et al.<sup>10</sup> conducted a limited comparison study of selected

continuous genetic network models<sup>3,11,12</sup>. Unlike these previous methods, we introduce a method that models experimental interventions explicitly when evaluating hypotheses about causal relationships. Independently, Pe'er, et al.<sup>13</sup> have similarly modeled interventions but they did not model latent variables.

This paper reports the results from the analysis of a gene-expression dataset that was gathered by experimentation on galactose genes in the yeast *Saccharomyces cerevisiae*<sup>14</sup>. Our analysis focuses on the discovery of pairs of genes ( $X, Y$ ) in which the expression of gene  $X$  has a causal influence on the expression of gene  $Y$ . As a representation of causation, we use causal Bayesian networks that include measured gene expression levels as well as possible latent causes that are not measured, such as the cellular level of proteins and small molecules<sup>15</sup>. The results of our causal analyses are compared with the known pathway. We also report novel causal relationships found in the analysis, which we believe deserve additional study.

## 2 Modeling Methods

A causal Bayesian network (or causal network for short) is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (variable) and a child node, relative to the other nodes in the network<sup>16</sup>. Figure 1 illustrates the structure of a hypothetical causal Bayesian network structure containing five nodes. The probabilities associated with this causal network structure are not shown.

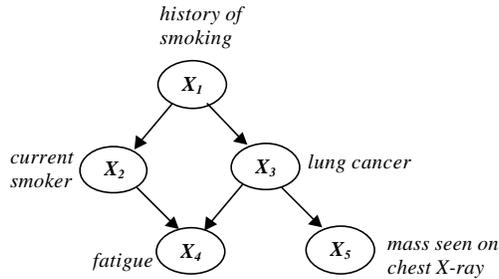


Figure 1. A hypothetical causal Bayesian network structure

The causal network structure in Figure 1 indicates, for example, that a *history of smoking* can causally influence whether *lung cancer* is present, which in turn can causally influence whether a patient experiences *fatigue*. The causal Markov condition gives the conditional independence relationships specified by a causal Bayesian network: *A node is independent of its non-descendants (i.e., non-effects) given its parents (i.e., its direct causes).*

The causal Markov condition permits the joint probability distribution of the  $n$  variables in a causal Bayesian network to be factored as follows<sup>16</sup>:

$$P(x_1, x_2, \dots, x_n | K) = \prod_{i=1}^n P(x_i | \mathbf{p}_i, K) \quad (1)$$

where  $x_i$  denotes a state of variable  $X_i$ ,  $\mathbf{p}_i$  denotes a joint state of the parents of  $X_i$ , and  $K$  denotes background knowledge.

Discovery of causal networks is an active field of research in which numerous advances have been—and continue to be—made in areas that include causal representation, model assessment and scoring, and model search<sup>17,18,19,20</sup>.

### 2.1 The Modeling of Experimental Interventions

In this section, we briefly describe how to represent experimental intervention in a causal Bayesian network. First, consider that we have a Bayesian network  $S$  that represents the causal relationships among a set of genes (in terms of the regulation of expression). We need to augment this network to represent the experimental interventions (manipulations) that were performed to obtain the microarray data at hand. To do so, let  $M_{X_i}$  be a variable that represents the value  $k$  (from 1 to  $r_i$ ) to which the experimenter deterministically manipulated gene  $X_i$  (e.g., a “knock out” of  $X_i$ ). To represent this deterministic manipulation, we augment  $S$  so that (1) variable  $M_{X_i}$  is a parent of  $X_i$  in  $S$ , and (2) for all the joint states of  $\mathbf{p}'_i$  we have that  $P(X_i = k | M_{X_i} = k, \mathbf{p}'_i) = 1$ , where  $\mathbf{p}'_i$  are the parents of  $X_i$  other than  $M_{X_i}$ . Details about this representation are discussed by Cooper and Yoo<sup>21</sup>.

For given microarray data  $D$ , we are interested in deriving the posterior probability of a causal network hypothesis  $S$  given data  $D$  and background knowledge (priors)  $K$ ; that is, we want to know  $P(S | D, K)$ . In particular, we would like to find causal networks with posterior probabilities that are relatively high. A key step in the Bayesian derivation of  $P(S | D, K)$  is to derive the marginal likelihood, namely  $P(D | S, K)$ . Specifically,  $P(S | D, K)$  is proportional to  $P(D | S, K) \times P(S | K)$ , where  $P(S | K)$  denotes prior belief (perhaps from background biological knowledge) that  $S$  is a valid causal hypothesis.

If  $D$  contains only passively observed data (no interventions), then Equation 2 provides a method for deriving the marginal likelihood, where  $r_i$  is the number of states that  $X_i$  can have,  $q_i$  denotes the number of joint states that the parents of  $X_i$  can have,  $N_{ijk}$  is the number of cases in  $D$  in which node  $X_i$  is observed to have state  $k$  when its parents have the states that are indexed by  $j$ ,  $\Gamma$  is the gamma function,  $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ ,  $\alpha_{ijk}$  and  $\alpha_{ij}$  express parameters of the Dirichlet prior distributions,

and  $\mathbf{a}_{ij} = \sum_{k=1}^{r_i} \mathbf{a}_{ijk}$ . The derivation and detailed explanation of Equation 2 are given in Cooper and Herskovits<sup>22</sup> and Heckerman, et al.<sup>23</sup>. Briefly, the equation assumes (1) discrete variables, (2) causal mechanisms that are local and independent (e.g., belief about the causes of gene  $X_i$  do not influence belief about the causes of gene  $X_j$ ), (3) data exchangeability (i.e., the order in which the experiments were performed is irrelevant), (4) a particular representation of parameter prior probabilities that is based on Dirichlet probability density functions, and (5) no missing data or latent variables. The marginal likelihood given by Equation 2 is sometimes called the BDe metric<sup>23</sup>.

$$P(D | S, K) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\mathbf{a}_{ij})}{\Gamma(\mathbf{a}_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\mathbf{a}_{ijk} + N_{ijk})}{\Gamma(\mathbf{a}_{ijk})} \quad (2)$$

Consider microarray data  $D$ , some of which was obtained under intervention and some of which was obtained by passive observation (i.e., data on “wild types”). As proven by Cooper and Yoo<sup>21</sup>, augmenting  $S$  to contain manipulation variables of the type  $M_{X_i}$ , as described above, is equivalent to having the terms  $N_{ijk}$  in Equation 2 denote just those cases in which  $X_i$  was *passively observed* (e.g., not manipulated) to have value  $k$  when its parents were in state  $j$ . We used Equation 2 under this modification to derive  $P(D | S, K)$  when  $D$  contains a mixture of data obtained under manipulation and under passive observation. For parameter priors, we assumed that  $\mathbf{a}_{ijk} = \frac{1}{r_i q_i}$ , which for the BDe metric is a commonly used non-informative parameter prior.

### 3 Scoring Methods of Structures with a Latent Variable

As mentioned in Section 2, Equation 2 assumes no latent variables. If we wish to model the possibility of a latent causal influence (which indeed we do), we need to extend Equation 2. Exact extensions involve applying Equation 2 an exponential number of times, one application for each possible state of the latent variables<sup>24</sup>. Such an exact method is not computationally feasible. Thus, in practice, we need to use approximation methods to evaluate  $P(S | D, K)$  when  $S$  contains latent variables. In the remainder of this section, we explain the approximation method that we applied in this paper.

#### 3.1 Causal Hypotheses Being Modeled

Figure 2 displays six local causal hypotheses ( $E_1$  through  $E_6$ ) that we model, shown as causal network structures. Variable  $X$  is the expression level of a given gene. Variable  $Y$  is the expression level of another gene.  $H$  is a latent (hidden) variable.

We denote an arbitrary pair of nodes in a given causal network  $S$  as  $(X, Y)$ . If there is at least one directed causal path from  $X$  to  $Y$  or from  $Y$  to  $X$ , we say that  $X$  and  $Y$  are *causally related* in  $S$ . If  $X$  and  $Y$  share a common ancestor, we say that  $X$  and  $Y$  are *confounded* in  $S$ .

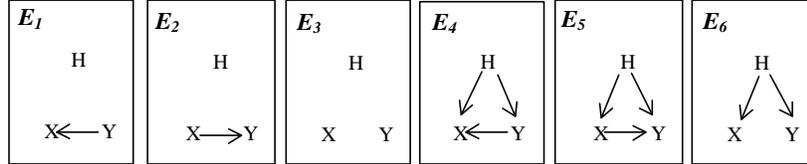


Figure 2. Six causal hypotheses on a pair  $(X, Y)$  of measured variables.

To derive the marginal likelihood (i.e., *score*) of causal structures  $E_1$ ,  $E_2$ , and  $E_3$ , we can use Equation 2, since the hidden variable  $H$  does not influence either  $X$  or  $Y$ , and thus, can be ignored<sup>21</sup>. For structures  $E_4$ ,  $E_5$ , and  $E_6$ , for which  $H$  is a confounding influence of  $X$  and  $Y$ , we use the scoring method discussed in the next section.

### 3.2 Implicit Latent Variable Scoring (ILVS) Method

Since explicit scoring of latent-variable models needs exponential time (in the number of database samples), approximation methods have been introduced in the literature, including methods based on stochastic simulation and on expectation maximization<sup>23</sup>. Unfortunately, these methods often require long computation times before producing acceptable approximations. Therefore, we developed a new method called the Implicit Latent Variable Scoring (ILVS) method<sup>15</sup>.

The basic idea underlying ILVS is to (1) transform the scoring of a latent model  $S$  (e.g., model  $E_5$  in Figure 2) into the scoring of multiple non-latent variable models, (2) score those non-latent models efficiently using Equation 2, and then (3) combine the results of those scores to derive an overall score (i.e., marginal likelihood). For instance, consider scoring  $E_5$  with two types of samples. One type is data for which  $X$  and  $Y$  were passively observed. We can derive the marginal likelihood of this data using the causal network in Figure 3(a), which contains no latent variable. Let  $P(D_o | E_5, K)$  denote this marginal likelihood. The other type of samples is data for which  $X$  was manipulated and  $Y$  was observed. We use the causal network in Figure 3(b) to derive the marginal likelihood of this data, namely  $P(D_m | E_5, K)$ . The different appearance of the arcs in Figures 3(a) and 3(b) signifies that these arcs are representing different distributions of  $X$  and  $Y$ . Figure 3(a) represents a situation in which  $X$  and  $Y$  are dependent based on a combination of direct causal influence of  $X$  on  $Y$  and on the confounding of  $X$  and  $Y$  by hidden process  $H$ . For the situation modeled by Figure 3(b), the experimental manipulation of  $X$  removes all causal influence of  $H$  on  $X$ . Therefore, Figure 3(b)

represents a situation in which  $X$  and  $Y$  are dependent based only on the direct causal influence of  $X$  on  $Y$  – there is no additional confounding influence. Continuing the Bayesian analysis, if (as in ILVS) we assume our beliefs about the distribution of  $X$  and  $Y$  in the Figure 3(a) situation are independent of the beliefs about their distribution in the Figure 3(b) situation, then the overall marginal likelihood of all the data (the passively observed data and the experimentally manipulated data) is  $P(D | E_5, K) = P(D_o | E_5, K) \times P(D_m | E_5, K)$ . It is straightforward to extend the analysis to also include data in which  $Y$  was manipulated and  $X$  was passively observed.

In deriving the marginal likelihood of  $E_4$  and  $E_6$ , ILVS uses a technique similar to the one just described for  $E_5$ . Yoo and Cooper<sup>15</sup> provide algorithmic details of ILVS and a proof of its convergence to the correct generating structure in the large sample limit.



Figure 3. Two non-latent variable structures used to score a latent-variable structure.

## 4 Gene Expression Dataset Analyses

We applied the ILVS algorithm to a gene-expression dataset to produce putative causal relationships among the genes. This section briefly describes the dataset and summarizes the steps we followed in preparing the data for analysis.

### 4.1 Dataset Description

The cDNA microarray data we analyzed were obtained from experiments that focused on the galactose utilization pathway in the yeast *Saccharomyces cerevisiae* as reported by Ideker, et al.<sup>14</sup>. The experiments included single gene deletion involving nine of the key genes<sup>1)</sup> that participate in yeast galactose metabolism. All microarray experiments were repeated four times. For each experiment, one of the nine genes was deleted, or alternatively, the experiment used a *wild-type* cell wherein no genes were deleted. For each of those 10 experimental conditions, galactose was available extracellularly in one set of experiments and absent in another set. Thus, there were a total of 20 different experimental conditions. Since each of those 20 experiments was repeated four times, the overall dataset contains results from 80 experiments. In each

---

<sup>1)</sup> Nine galactose genes are *Gal1*, *Gal2*, *Gal3*, *Gal4*, *Gal5(PGM2)*, *Gal6(LAP3)*, *Gal7*, *Gal10*, and *Gal80*.

experiment, 5,936 gene expression levels were measured, corresponding to almost the entire *Saccharomyces cerevisiae* genome.

#### 4.2 Dataset Preparation

This section describes the five data preparation steps that we applied to the data. We tried different methods of discretization (i.e., forcing all knock out genes to have expression level of 0; discretization with clustering) but found no significant difference in the overall predicted performances.

- 1) Genes that had expression levels missing in all four repetitions of a given experiment were excluded from the analysis ( $n = 195$  genes were excluded).
- 2) If the expression level of a gene was missing in some experiment, its value was estimated as the mean value of the available measurements for that gene.
- 3) Negative intensities were assumed to be 0.
- 4) Let  $X_i^*$  denote the intensity for gene  $X_i$ , which serves as an indicator of the expression level of  $X_i$  in an experiment in which some gene (not necessarily  $X_i$ ) was knocked out. Similarly, let  $X_i^r$  denote the intensity, which is an indicator of the expression level of  $X_i$  when no genes were manipulated (wild type). The relative intensity for gene  $X_i$  was calculated as  $\log(X_i^*/X_i^r)$ . If  $X_i^*$  was 0, we used the minimum  $\log(X_i^*/X_i^r)$  value for gene  $X_i$  over all 80 experiments for gene  $X_i$ . If  $X_i^r$  was 0, we used maximum  $\log(X_i^*/X_i^r)$  value for gene  $X_i$  over all 80 experiments. If both  $X_i^*$  and  $X_i^r$  were 0,  $\log(X_i^*/X_i^r)$  was set to 0.
- 5) Discretization was performed based on each gene's expression level mean  $m$  and standard deviation  $d$  over all 80 samples. All genes were assigned three states: 0 was assigned to any value less than  $m-d$ , 1 was assigned to any value greater than or equal to  $m-d$  and less than  $m+d$ , and 2 was assigned to any value greater than or equal to  $m+d$ .

#### 4.3 Analyses

We applied ILVS to every pair of the 5,936 yeast genes that includes one or both genes from the nine galactose genes. For each gene pair, we used the method in Section 3.2 to derive a posterior probability for each of the six causal hypotheses in Figure 2. We analyzed our results in two main parts.

The first part focused on just the nine galactose genes that were manipulated. Since the causal relationships among these nine genes are understood relatively well, we assume that these generally accepted relationships are correct and can serve as a standard against which to compare the output of ILVS. The ILVS method was considered to label a given gene pair as having causal relationship  $R$  (among the six possibilities in Figure 2) if  $P(R | D, K)$  was greater than 0.5. For each gene

pair  $(X, Y)$ , there were 8 cases in each  $X$  and  $Y$  (and all other genes for that matter) were passively observed, 8 cases in which  $X$  was knocked out and  $Y$  observed, and 8 cases in which  $Y$  was knocked out and  $X$  observed.

The second part of our analysis was more exploratory than evaluative. In particular, we examined all  $9 \times 5,732 = 51,588$  gene pairs consisting of one gene from the nine galactose genes and one gene from outside that set. Let  $X$  denote one of the nine galactose genes and let  $Y$  denote one of the other 5,732 genes studied. For each gene pair  $(X, Y)$ , there were 8 cases in which  $X$  and  $Y$  were observed, 8 cases in which  $X$  was knocked out and  $Y$  observed, and *no cases* in which  $Y$  was knocked out and  $X$  observed. We identified every gene pair for which one of the six causal hypotheses in Figure 2 was greater than probability 0.9. Each pair represents a hypothesis about the nature of a causal relationship that has yet to be characterized.

## 5 An Investigation of ILVS: Results

This section first presents the results of exploring relationships just among the nine galactose genes, and then between those nine genes and all other genes in the dataset.

### 5.1 Galactose Genes

The first column in Table 1 shows pairwise causal relationships that represent generally accepted biological knowledge about galactose gene-regulation pathway, as summarized in Ideker, et al.<sup>14</sup>. The table also shows the results of applying ILVS to the 36 pairs of galactose genes. Only the causal relationships that had a posterior probability greater than 0.5 are shaded. No relationship had a causal hypothesis with a probability higher than 0.9, possibly due to the small number of samples in the dataset.

Upon comparing the ILVS predictions with the known galactose metabolic pathway, we show the results in Table 1. Each shaded row in reversed font represents ILVS predictions that agree with accepted biological knowledge. The shaded rows in bold font denote that for reference structure T, there exists uncertainty about its validity. Other shaded rows represent predictions that are inconsistent with accepted biological knowledge.

The errors in Table 1(a) are due to ILVS assuming genes are independent, when biological knowledge indicates an expected dependence. We label these as false negatives. There are at least two plausible reasons for these errors. First, the sample size (24 samples per pair) is small, and thus, unless the dependence is strong, that dependence may not be apparent from the data. Second, the biological knowledge we used as a reference standard expresses general patterns of causal

dependency among the galactose genes; not all of those patterns were necessarily revealed by the experiments performed in creating the dataset that was given as input to ILVS.

Table 1. The most probable causal hypotheses predicted by ILVS as representing relationships among the nine manipulated galactose genes under study

Reference Structure T	ILVS Predicted Structure		P(T/D,K) <sup>§</sup>	
	Structure S	P(S/D,K)		
Gal6---Gal7	Gal6 Gal7	0.80	0.02	
<b>Gal2▷Gal7</b>	<b>Gal2 Gal7</b>	<b>0.80</b>	<b>0.01</b>	
Gal1---Gal5	Gal1 Gal5	0.80	0.03	
<b>Gal2▷Gal5</b>	<b>Gal2 Gal5</b>	<b>0.75</b>	<b>0.01</b>	
<b>Gal2▷Gal10</b>	<b>Gal2 Gal10</b>	<b>0.74</b>	<b>0.01</b>	
Gal1---Gal6	Gal1 Gal6	0.73	0.03	
Gal6---Gal10	Gal6 Gal10	0.73	0.02	
Gal80⇒Gal5	Gal5 Gal80	0.70	0.04	
<b>Gal2▷Gal1</b>	<b>Gal1 Gal2</b>	<b>0.57</b>	<b>0.16</b>	
<b>Gal2▷Gal6</b>	<b>Gal2 Gal6</b>	<b>0.56</b>	<b>0.01</b>	
Gal5---Gal7	Gal5 Gal7	0.55	0.02	
Gal5---Gal10	Gal5 Gal10	0.54	0.07	
Gal3⇒Gal6	Gal3 Gal6	0.51	0.10	
Gal3▷Gal5	Gal3 Gal5	0.45	0.08	
Gal3▷Gal10	Gal3 Gal10	0.43	0.25	
Gal4@Gal6	Gal4 Gal6	0.41	0.08	
Gal3---Gal80	Gal3 Gal80	0.33	0.11	
Gal4@Gal5	Gal4 Gal5	0.28	0.08	
Gal5---Gal6	Gal5 Gal6	0.28	0.09	

(a)  $E_3$

Reference Structure T	ILVS Predicted Structure		P(T/D,K) <sup>§</sup>	
	Structure S	P(S/D,K)		
Gal80⇒Gal1	Gal80→Gal1	0.56	0.16	
Gal3▷Gal1	Gal3@Gal1	0.33	0.33	
Gal3▷Gal7	Gal3@Gal7	0.31	0.31	
Gal80▷Gal10	Gal80@Gal10	0.30	0.17	
Gal80▷Gal7	Gal7@Gal80	0.25	0.11	

(b)  $E_1$  and  $E_2$

Reference Structure T	ILVS Predicted Structure		P(T/D,K) <sup>§</sup>	
	Structure S	P(S/D,K)		
<b>Gal4@Gal7</b>	<b>Gal4▷Gal7</b>	<b>0.83</b>	<b>0.01</b>	
<b>Gal10---Gal7</b>	<b>Gal10▷Gal7</b>	<b>0.81</b>	<b>0.004</b>	
<b>Gal4@Gal80</b>	<b>Gal4▷Gal80</b>	<b>0.58</b>	<b>0.10</b>	
Gal4@Gal1	Gal4▷Gal1	0.47	0.04	
Gal2∪Gal3	Gal2▷Gal3	0.42	0.45 <sup>*</sup>	
Gal4@Gal10	Gal4▷Gal10	0.41	0.003	
Gal2«Gal4	Gal2▷Gal4	0.40	0.33 <sup>†</sup>	
Gal3▷Gal1	Gal3▷Gal1	0.33	0.33	
Gal3▷Gal7	Gal3▷Gal7	0.31	0.31	

(c)  $E_4$  and  $E_5$

Reference Structure T	ILVS Predicted Structure		P(T/D,K) <sup>§</sup>	
	Structure S	P(S/D,K)		
Gal80⇒Gal6	Gal6---Gal80	0.63	0.06	
<b>Gal1---Gal7</b>	<b>Gal1---Gal7</b>	<b>0.60</b>	<b>0.60</b>	
<b>Gal1---Gal10</b>	<b>Gal1---Gal10</b>	<b>0.57</b>	<b>0.57</b>	
<b>Gal2∪Gal80</b>	<b>Gal2---Gal80</b>	<b>0.57</b>	<b>0.35<sup>†</sup></b>	
Gal4@Gal3	Gal3---Gal4	0.49	0.29	

(d)  $E_6$

Notation: The symbol  $\rightarrow$  represents the relationship given by causal structure  $E_1$  and  $E_2$  (from Figure 2). Likewise, a blank space is used for  $E_3$ , a  $\Rightarrow$  for  $E_4$  and  $E_5$ , and a  $---$  for  $E_6$ . In column 1, a bidirectional arc indicates that there is a known feedback pathway between the two genes (e.g.,  $Gal2\cup Gal3$ ). The symbol  $\dagger$  indicates summing the posterior probabilities for  $E_1$  and  $E_2$ . A  $*$  indicates summing the probabilities for  $E_4$  and  $E_5$ . The column labeled  $P(T/D,K)$ <sup>§</sup> gives the probability that ILVS assigned to the reference structure in column 1. See the text for an explanation of the shaded results.

The errors in Tables 1(b), 1(c), and 1(d) result from the most probable hypothesis (according to ILVS) being inconsistent with assumed biological knowledge. There are only two pairs (the shaded ones in Table 1(d)) for which

ILVS obtains exactly the correct hypothesis. Consider, however, the following relaxation: A hypothesis is correct if it indicates that there is a causal pathway from  $X$  to  $Y$  (with or without confounding) and according to existing biological knowledge there is indeed a causal pathway from  $X$  to  $Y$  (with or without confounding). For example, under this interpretation  $Gal80 \rightarrow Gal1$  would be correct, since the reference structure is  $Gal80 \Rightarrow Gal1$ , which includes a causal path from  $Gal80$  to  $Gal1$ . Under this relaxation of correctness, 12 of the 17 unique relationships (71%) in Tables 1(b), 1(c), and 1(d) are correct.

### 5.2 Galactose Genes and Other Genes

In this section, we report the results of an exploratory analysis. The purpose of this section is to illustrate an initial step in using computer-based, data-intensive methods to hypothesize causal relationships.

Table 2. Types of highly probable ( $>0.9$ ) gene pairs predicted by ILVS from 51,588 considered pairs.

$E_1, E_4,$ and $E_6$	$E_2$	$E_3$	$E_5$
1,329 (2.58%)	4 (0.008%)	586 (1.14%)	1,113 (2.16%)

Table 3. Conditional distributions of four genes that are reported by ILVS in Table 2 to be highly probable ( $>0.9$ ) effects of  $Gal1$  or  $Gal2$ . For example, Table 3(a) presents  $P(YBR096W | Gal2)$ .

		<i>Gal2</i>					<i>Gal2</i>		
		low	no ch	high			low	no ch	high
<i>YBR096W</i>	low	0.021	0.301	0.083	<i>YMR086W</i>	low	0.021	0.301	0.083
	no change	0.021	0.688	0.833		no change	0.021	0.688	0.833
	high	0.958	0.011	0.083		high	0.958	0.011	0.083
		(a)					(b)		
		<i>Gal2</i>					<i>Gal1</i>		
		low	no ch	high			low	no ch	high
<i>SSU1</i>	low	0.958	0.011	0.083	<i>SER3</i>	low	0.026	0.493	0.026
	no change	0.021	0.688	0.083		no change	0.949	0.493	0.026
	high	0.021	0.301	0.833		high	0.026	0.013	0.949
		(c)					(d)		

In an analysis of 51,588 gene pairs, ILVS scored 5.9% of the node pairs as having a high probability ( $>0.9$ ) causal hypothesis (see Table 2). Since the 5,732 genes were not experimentally deleted, ILVS could not distinguish among  $E_1$ ,  $E_4$ , and  $E_6$  when one of the nine galactose genes were treated as variable  $X$  in Figure 2; therefore,  $E_1$ ,  $E_4$ , and  $E_6$  are grouped in Table 2. The four unconfounded causal relationships are  $Gal2 @ YBR096W$ ,  $Gal2 @ YMR086W$ ,  $Gal2 @ SSU1$ , and  $Gal1 @$

*SER3*. Conditional distributions in Table 3 suggest that *Gal2* is acting as a relative inhibitor for *YBR096W* and *YMR086W*, and as an activator of *SSU1*. Table 3(d) suggests that *Gal1* is acting as an activator for *SER3*. Interestingly *SER3* is one of only two proteins that are known to bind with *Gal1* and its regulatory role is unknown<sup>25</sup>.

To evaluate all 51,588 gene pairs, ILVS required about four and a half hours of CPU time on a Pentium 500MHz Linux machine with 384M RAM.

## 6 Summary and Discussion

ILVS is a novel, efficient causal discovery algorithm that can model causal hypotheses with (and without) latent variables. The method attains its efficiency by modeling one pair of variables at a time and by evaluating latent-variable models implicitly, rather than explicitly. In previous work, the ILVS algorithm has been shown to be asymptotically correct in the large sample limit<sup>15</sup>. Thus, with enough valid data, it is guaranteed to find the correct causal relationship between each pair of variables in a dataset. The ILVS method can use data obtained from passive observation and from active experimental manipulation. Since much gene-expression data is of both types, the ILVS method is of particular relevance to work on discovery of gene-regulation pathways from gene-expression data.

We applied the ILVS method to an available dataset containing gene expression levels from experiments that focused on galactose metabolism. These early results are promising, but in need of improvement. The error rates in re-discovering known galactose gene-regulation pathway were high. Possible reasons include a small set of samples and limited experimental conditions and variation; the influence of these two limitations is supported by the fact that ILVS did not give a high probability ( $> 0.9$ ) to any of the galactose causal relationships that it hypothesized. For the false positives output by ILVS, some may simply be wrong, while others may represent unknown causal relationships within galactose gene regulation.

An exploratory analysis of the galactose dataset yielded approximately 3,000 causal relationships (out of 51,588) that appear highly probably (according to ILVS). Those hypothesized relationships have not yet been investigated in the laboratory. In future work, we intend to work with yeast biologists to begin to explore some of the most interesting and promising of these hypotheses.

Algorithmic improvements that we are pursuing include performing Bayesian model averaging (rather than model selection), modeling continuous data directly (rather than discretizing the data), developing and using informative models of the types of measurement noise that exists in microarray experiments, modeling covariates (e.g., galactose levels) in evaluating the causal relationships among a

pair of genes, modeling gene-regulation feedback, and incorporating into our Bayesian analyses informative structure and parameter prior probabilities.

### Acknowledgments

We thank the Computational Systems Biology Group at the University of Pittsburgh and Carnegie Mellon University for helpful discussions. This research was supported by NSF grant IIS-9812021 and NASA grant NRA2-37143.

### References

1. Spellman, P.T., G. Sherlock, M.Zhang, V.Iyer, K.Anders, M.Eisen, P. O. Brown, D. Botstein, and B. Futcher (1998). *Mol. Bio. Cell* 9, p3273–3297.
2. Getz, G., E. Levine, and E. Domany (2000). *Proc. Na. Ac. Sc.*, 97(22).
3. Chen, T., V. Filkov, and S. S. Skiena (1999). *RECOMB*.
4. D'haeseleer, P., S.Liang, and R.Somogyi (2000). *Bioinfo.*, 16(8):707-726.
5. Friedman, N., M. Linial, I. Nachman, and D Pe'er (2000). *J. Comp. Bio.*
6. Hartemink, A.J., D.K. Gifford, T.S. Jaakkola, and R.A. Young (2001). *PSB*.
7. Maki, Y., D. Tominaga, M. Okamoto, S. Watanabe, Y. Eguchi (2001) *PSB*.
8. Dutilh, B. (1999). Gene Networks from Microarray Data. Unpublished paper.
9. Smolen, P., D.Baxter, and J.Byrne (2000). *Bul. Math. Bio.* 62, p247–292.
10. Wessels, L.F.A., E.P. Van Someren, and M.J.T. Reinders (2001). *PSB*.
11. Arkin, A., P. Shen, and J. Ross (1997). *Science* 277, p1275–1279.
12. Weaver, D., C.Workman, and G. Stormo (1999). *PSB* 4, p112–123.
13. Pe'er, D., A.Regev, G.Elidan, and N.Friedman (2001). *ISMB*.
14. Ideker, T., V.Thorsson, J. Ranish, C.Rowan, J.Buhler, J.Eng, R.Bumgarner, D.Goodlett, R.Aebersold, L. Hood (2001). *Science* 292: 929–934. Dataset is available at: <http://www.sciencemag.org/cgi/content/full/292/5518/929/DC1>
15. Yoo, C. and G. F. Cooper (2001). *CBMI research report* CBMI-173.
16. Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems. *Morgan Kaufmann*, San Mateo, CA.
17. Glymour, C., G.Cooper (1999). Computation, Causation, Discovery. *MIT Pr.*
18. Spirtes, P., C.Glymour, R.Scheines (2000). Causation, Prediction, and Search. *MIT Press*.
19. Pearl, J. (2000) Causality. *Cambridge University Press*.
20. UAI 2001: <http://robotics.stanford.edu/~uai01/pastconferences.htm>.
21. Cooper, G.F., and C. Yoo (1999). *UAI '99*, Stockholm, Sweden.
22. Cooper, G.F., and E. Herskovits (1992). *Mach. Learning*, 9, 309-347.
23. Heckerman, D., D.Geiger, D.Chickering (1995). *Mach. Learning* 20 197-243.
24. Cooper, G.F. (1995). *Journal of Intelligent Information Systems*, 4, p71-88.
25. Ito, T., Chiba, Ozawa, Yoshida, Hattori, Sakaki (2001) *Proc. Natl. Acad. Sci.* 10.