Published in final edited form as:

Ann N Y Acad Sci. 2009 March ; 1158: 93–101. doi:10.1111/j.1749-6632.2008.03749.x.

DREAM Project: The five-gene-network data analysis with Local Causal Discovery Algorithm using Causal Bayesian networks

Changwon Yoo* and Erik M. Brilz

Computer Science, University of Montana, Missoula MT, USA

Abstract

Using microarray experiments we can model causal relationships of genes measured through mRNA expression levels. To this end, it is desirable to compare experiments of the system under complete interventions of some genes, e.g., knock-out of some genes, with experiments of the system under no interventions. However, it is expensive and difficult to conduct wet lab experiments of complete interventions of genes in a biological system. Thus, it will be helpful if we can discover promising causal relationships among genes with no interventions or incomplete interventions, e.g., applying a treatment that has unknown effects to modeled genes, in order to identify promising genes to perturb in the system that can later be verified in wet laboratories. In this paper we use Causal Bayesian networks to implement a causal discovery algorithm—the Equivalence Local Implicit latent variable scoring Method (EquLIM)—that identifies promising causal relationships even with a small dataset generated from no or incomplete interventions. We then apply EquLIM to analyze the five-gene-network data and compare EquLIM's predictions with true causal pairwise relationships between the genes.

Keywords

Causal Bayesian networks; causal discovery; gene networks; high throughput data analysis; microarray data

1. INTRODUCTION

Causal modeling and discovery are fundamental pursuits of science. Experimental studies, such as randomized controlled trials (RCTs), often provide the most trustworthy methods we have for establishing causal relationships from data. Observational data are passively observed from experiments with no or incomplete interventions. Such data are more readily available than are experimental data with complete interventions. As observational electronic databases become increasingly available, the opportunities for using them for causal discovery also increase. In an experimental study, one or more variables are manipulated (typically randomly) and the effects on other variables are measured. While potentially very informative, it may be expensive and difficult (if not impossible) to conduct wet lab experiments of complete interventions of genes in animal models, e.g., a mouse model. Thus, it will be helpful if we can discover promising causal relationships among

*corresponding author: cwyoo@cs.umt.edu.

genes with observational data alone in order to identify promising genes to perturb in the system that can later be verified in wet laboratories. Bayesian 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 [1]. In prior work on Bayesian discovery of causal networks, researchers have focused primarily on methods for discovering causal relationships from observational data [1–3]. A notable exception is a paper by Heckerman on learning influence diagrams as causal models. It contains key ideas for learning causal Bayesian networks from a combination of both experimental data under deterministic manipulation and observational data [4].

Graphical models hold great promise as representations of molecular biological processes, because they are both expressive and intuitive. The authors of five separate review articles on bioinformatics and related topics described graphical models as one of the most promising methods for representing cellular pathways [5–9]. Bayesian discovery of causal networks (with modeling possible latent variables) has been widely applied in systems biology research, especially in learning gene networks from various experiments [10–12].

We introduce a novel pairwise relationship scoring method—Equivalence Local Implicit latent variable scoring Method (EquLIM)—to learn causal networks from observational data alone. EquLIM extends earlier work [13] by improving the search methods to discover promising causal relationships on observational data alone. We describe Bayesian methods for learning Bayesian networks when variable manipulation is not necessarily deterministic, but rather stochastic. In the important special case in which manipulation is deterministic, there is a closed-form Bayesian scoring metric that is a simple variation on a previous scoring metric for Bayesian network learning [2, 14]. We then apply EquLIM to analyze the five-gene-network data analysis and output promising causal pairwise relationships between the genes. Additionally, in each causal relationship, we identify whether the relationship is excitatory or inhibitory connection.

2. MODELING METHOD

A Bayesian network is a directed acyclic graph in which each node represents a variable and each arc represents probabilistic influence. 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 [15]. Figure 1 illustrates the structure of a hypothetical causal Bayesian network structure that contains six nodes. Later in the modeling manipulation section, we will further discuss about manipulation node. The probabilities associated with this causal network structure are not shown.

The causal network structure in Figure 1 indicates, for example, that *Gene1* can regulate (causally influence) the expression level of *Gene3*, which in turn can regulate the expression level of *Gene5*.

The causal Markov condition gives the conditional independence relationships that are 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 distribution of the n variables in a causal Bayesian network to be factored as follows [15]:

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

where x_i denotes a state of variable X_i , π_i denotes a joint state of the parents of X_i , and K denotes background knowledge.

2.1. Structure Learning

We introduce six equivalence classes (E_1 through E_6) among the structures (Figure 2). The causal networks in an equivalence class are statistically indistinguishable for any observational and experimental data on X and Y . We denote an arbitrary pair of nodes in a given Bayesian network B 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 B . If X and Y share a common ancestor, we say that X and Y are *confounded* in B . As the first step toward Bayesian causal modeling of latent variables, which is a computationally challenging problem, we only look at pairwise relationships between two nodes (X and Y) and a latent variable H .

Let $E = \{E_1, E_2, E_3, E_4, E_5, E_6\}$ and let E_i^{XY} denote the node pair X and Y with causal relationship E_i . Let us consider the posterior probability that variable X causes variable Y given data D on the measured variables. We can derive the posterior probability of E_i^{XY} as:

$$P(E_i^{XY} | D, K) = \frac{\sum_{S: E_i^{XY} \text{ is in } S} P(S | D, K)}{\sum_{S: E_i^{XY} \text{ is in } S} P(S | D, K)} \quad (2)$$

where the sum is taken over all admissible causal network structures S , such that S contains substructure E_i^{XY} . An *admissible causal network structure* is a structure that (1) contains all of the variables being modeled (of which X and Y are but two), and (2) has a non-zero prior probability. Based on the properties of probabilities, the term within the sum in Equation 2 may be rewritten as follows:

$$P(S | D, K) = \frac{P(S, D | K)}{P(D | K)} = \frac{P(S, D | K)}{\sum_S P(S, D | K)} \quad (3)$$

Since the probability $P(D | K)$ is a constant relative to the entire set of causal structures being considered, Equation 3 shows that the posterior probability of causal structure S is proportional to $P(S, D | K)$, which we can view as a *score* of S in the context of D . The probability terms on the right side of Equation 3 may be expanded as follows:

$$P(S, D|K) = P(S|K)P(D|S, K) = P(S|K) \int P(D|S, \theta_S, K) P(\theta_S|S, K) d\theta_S \quad (4)$$

where (1) $P(S|K)$ is a prior belief that network structure S correctly captures the qualitative causal relationships among all the modeled variables; (2) θ_S are the probabilities (parameters) that relate the nodes in S quantitatively to their respective parents; (3) $P(D|S, \theta_S, K)$ is the likelihood of data D being produced, given that the causal process generating the data is a causal Bayesian network given by S and θ_S ; and (4) $P(\theta_S|S, K)$ expresses a prior belief about the probability distributions that serve to model the underlying causal process.

With appropriate assumptions, we can evaluate $P(D|S, K)$ in Equation 4 with the following equation [2, 14]:

$$P(D|S, K) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \quad (5)$$

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 *passively observed* to have state k when its parents have states as given by j , Γ is the gamma function, α_{ijk} and α_{ij} express parameters of the Dirichlet prior distributions, and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. We

used the BDe metric [2] with $\alpha_{ijk} = \frac{1}{r_i q_i}$, which is a commonly used non-informative parameter prior for the BDe metric.

2.2. Modeling Manipulation

In the current section, we consider the situation in which manipulation of a variable may not be deterministic. A classic example from medicine is when a patient, who has volunteered to participate in a study, is randomized to receive a certain medication, but he or she decides not to take it. Let M_{X_i} be a variable that represents the value k (from 1 to r_i) to which the experimenter wishes to manipulate X_i . Let $M_{X_i} = o$ denote that the experimenter does not wish to manipulate X_i , but merely wants to observe its value. Augment the model variables to include M_{X_i} . Finally, carry out the analysis in section 2.1 assuming only observational data. The causal network hypotheses used in that analysis will include probabilities that specify prior beliefs about the causal influence of M_{X_i} on X_i . Those prior beliefs (on structure and parameters) will be updated by data on stated experimental intentions and observed variable value outcomes. For the special case of deterministic manipulation, we have that (1) with probability 1 variable M_{X_i} is a parent of X_i ; and (2)

$P(X_i = k | M_{X_i} = k, \pi'_i) = 1$, where π'_i are the parents of X_i other than M_{X_i} . When scoring X_i , deterministic manipulation is equivalent to ignoring the cases in which X_i was manipulated [13]. In particular, to incorporate experimental data, we evaluate Equation 5 by not adding the cases to N_{ijk} when X_i is manipulated [13]. In the remainder of this paper, we assume that manipulation is deterministic.

For example, Figure 1 shows a causal network structure that has an additional variable M_{X_2} . In Figure 1, variable X_2 (*Gene2*) is modeled as being manipulated in some cases, such as, knocking out *Gene2*.

3. EQUIVALENCE LOCAL IMPLICIT LATENT VARIABLE SCORING METHOD

In this section we introduce the implicit latent variable scoring (ILVS) method and then introduce a method called Local ILVS Method (LIM) that extends ILVS. At the end we introduce Equivalence LIM (EquLIM).

We denote $D(X, Y)$ as a set of cases (in dataset D) in which both node X and Y are observed. $D(mX, Y)$ denotes a set of cases in which node X is manipulated and Y is observed. Similarly, $D(X, mY)$ denotes a set in which node X is observed and Y is manipulated.

Implicit Latent Variable Scoring (ILVS) Method

Explicit scoring of latent-variable models requires exponential time in the number of database samples. Therefore, approximation methods have been introduced, including methods based on stochastic simulation and on expectation maximization [2]. 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 [16].

The basic idea underlying ILVS is to (1) transform the scoring of a latent model E_i (e.g., 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 sample 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 Figure 3 (a) and Figure 3(b) signifies that these arcs are representing different distributions of X and Y . 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 data generated by experimental manipulation) 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 [16] provide algorithmic details of ILVS and a proof of its convergence to the correct generating structure in the large sample limit.

ILVS scores each E_i in Figure 2 by only considering pairwise measured nodes. Thus, ILVS evaluates Equation 5 for only two measured nodes at a time. In earlier studies, we applied

ILVS to simulated data [16] and to yeast DNA microarray data [17]. We have also extended ILVS to create a system called extILVS that scores more than pairwise relationships.

Local ILVS Method (LIM)

Let L_i^{XY} denote a set of local structures that includes E_i^{XY} and let $L^{XY} = \cup_i L_i^{XY}$. For example, in Figure 1 let $X=Gene2$ and $Y=Gene1$. Then L_1^{XY} could be the causal structure shown in Figure 1. LIM (Local ILVS method) calculates $P(E_i^{XY} | D, K)$ by first, searching for the best L_i^{XY} that fits the data; and second, using all unique L_i^{XY} that were visited so far. Scores of the node pairs, calculated by extILVS, are used to guide the search for the best L_i^{XY} . Finally, we estimate Equation 2 by the following equation:

$$P(E_i^{XY} | D, K) \approx \frac{\sum_{S: E_i^{XY} \text{ is in } T} P(D, S | K)}{\sum_T P(D, T | K)} \quad (6)$$

where T denotes all the structures generated in the search. Many heuristic methods have been used to search for the best structure that fits the data [2]. Note that unlike the previous methods, we concentrate on L_i^{XY} , i.e., the local structure of E_i^{XY} . In this paper we use structure search as defined in the following steps: (Step 1) Construct a set V that represents strongly related variables with X and Y . Let W equal $V \cup \{X, Y\}$. We limit the number of variables in W to be less than k and use those variables to define the structures in L^{XY} . Now any structure $S \in L^{XY}$ can be denoted as $S = \{E_i^P | P \in \{\text{all pairs in } W\}\}$. (Step 2) We initialize S to a random structure by randomly choosing E_i^P for all P . (Step 3) For a given structure S , we score with extILVS six different structures by substituting E_i^P with one of the six hypotheses (from Figure 2) for all node pairs P in W ; (Step 4) Select the E_j^{P*} that in Step 3 generated the structure with the highest score; update S by substituting E_j^{P*} for E_i^P in S and repeat Step 3 with the new S . Stop the search if either there is no improvement in the structure score (it has reached a local maximum) repeat from Step 2; otherwise, repeat from Step 3 with the original node pair P . We repeat the search from Step 2 for a user defined number of times. We note that the Local Structure Size (LSS) is the number of nodes included in L^{XY} .

Equivalence Local ILVS Method (EquLIM)

EquLIM extends LIM by scoring additional structures in (Step 3) of LIM: when we score with extILVS six different structures by substituting E_i^P with one of the six hypotheses (from Figure 2) for a node pair P in W , we additionally search for the reverse arc structure of each structure and score it. More specifically, assume we are currently substituting six hypotheses for a certain pair $q \in P$. Further assume we are substituting E_i^q with E_j^q . LIM will simply score a structure S that includes E_j^q and in the next step, substitute E_j^q with the other five hypotheses. EquLIM will not only score a structure S that includes E_j^q but also score the structure with all arcs in S reversed. After this additional process, EquLIM will move on and substitute E_j^q with the other five hypotheses, each time also scoring the reverse structure. Note that the structure S' that is achieved by reversing all the arcs in S is guaranteed to be the structure that has never been scored before by EquLIM. This is because EquLIM records all the structures that has been scored to make sure it does not score the same structure

twice, thus if S' was scored before that means either S' was scored by reversing the arc of S or S was scored by reversing the arc of S' (either way, it contradicts the fact that S is being scored, which means S' was never scored before). EquLIM has shown stable predictions with datasets of small size (< 500 cases) of simulated microarray experiments compared to LIM [18].

Example Run of EquLIM

For example, let us assume there are only five modeled nodes: U , V , X , Y , and Z . Note that there are 10 possible pairs in P . Further assume we are limiting $k = |W| < 4$, or LSS of 3, and in Step 1 we choose $W = \{X, Y, Z\}$. Among the 10 pairs in P , assume we are now focusing in L^{XY} . In Step 2 we randomly initiate a structure, e.g., $S = \{E_I^{XY}, E_2^{XZ}, E_6^{YZ}\}$. In Step 3, (1) we consider the six different structures derived from S by substituting E_I^{XY} with any of $\{E_I^{XY}, E_2^{XY}, E_3^{XY}, E_4^{XY}, E_5^{XY}, E_6^{XY}\}$. (2) When we score $S = \{E_2^{XY}, E_2^{XZ}, E_6^{YZ}\}$ in (1), we also score the additional structure $S_I = \{E_I^{XY}, E_I^{XZ}, E_6^{YZ}\}$. Here note that S_I is the structure achieved by reversing all the arcs in S . We do the same for E_3^{XY} , E_4^{XY} , E_5^{XY} , and E_6^{XY} . In Step 4 we choose the highest scored structure and then repeat from Step 3 for E_2^{XZ} and E_6^{YZ} . Upon reaching a stopping condition, to score $P(E_I^{XY}/S, K)$, for example, we sum all scores of visited structures that include E_I^{XY} and divide by the sum of the scores of all visited structures. Note that indirect causal relationships, e.g., $X \leftarrow Z \leftarrow Y$, are also used in scoring E_I^{XY} .

4. EXPERIMENTAL METHODS

In evaluating causal learning, ideally we would know the real-world causal relationships (both the structure and parameters) among a set of variables of interest. With such knowledge we could generate experimental and observational data. Using these datasets as input, a learning method could predict the causal structure and estimate the causal parameters that exist among the modeled variables. These predictions and estimates would then be compared to the true causal relationships. Since confident knowledge of underlying causal processes is relatively rare, to evaluate the results of causal discovery from mixed high throughput data, we typically use a causal model that is constructed by an expert biologist as a gold standard. This process is especially useful for understanding what the expert biologist already knows. Competitions such as DREAM are also useful in testing and evaluating such causal discovery algorithms.

EquLIM currently models gene expression levels using discrete variables only, although it could be extended to model with continuous variables as well. Thus, we discretized each gene's expression level into three states (i.e., low, no change, and high) based on whether each gene's expression levels were above, below, or between one standard deviation from their mean [19].

The DREAM five-gene-network dataset contains two time series corresponding to two different treatments. 588 genes from the original Affymetrix microarray data were selected, which include the five genes in the synthetic network plus genes known in the literature to be regulated by some of these five genes. The five-gene network, which is a subnet of the bigger network, is oscillating with the cell cycle.

We have used LSS 6 for this analysis because in simulation studies, LSS 6 ran with a reasonable time (e.g., < four hours) to produce stable results. We have selected the six variables in LSS 6 using the following method (A is a set of all 588 genes plus the random variable *Treatment* and G is a set of variables that are in LSS 6):

First, G includes a random variable called *Treatment* which can have two possible values, i.e., Treatment 1 and Treatment 2. *Treatment* was used as a possible manipulation (as M_{X_i} shown in Figure 1) of any genes in LSS 6.

Next we add the following genes into G : *gene_1*, *gene_2*, *gene_3*, *gene_4*, and *gene_5*. We report the EquLIM analyses results and compare them with the generating structure (call it GS ; it is shown in Figure 4) of the five-gene-network data of the DREAM competition. More precisely, (1) we apply EquLIM using random variables in G and the five-gene-network data; and (2) compare the causal pairwise relationships among the genes in G that are from GS and are predicted by EquLIM. Additionally, in each causal relationship, we identify whether it is an excitatory or inhibitory connection.

We have implemented EquLIM with C++ and used two linux machines to run EquLIM. Since EquLIM and LIM are anytime algorithms, we have let EquLIM run for about two hours for the DREAM five-gene-network dataset.

5. RESULTS

Results of EquLIM pairwise causal predictions on genes in G are shown in Table 1. Since we assume $P(X \rightarrow Y) = P(Y \rightarrow X) = P(X \perp Y) = 1/3$ for all genes X and Y , we only report $P(X \rightarrow Y)$ (or $P(Y \rightarrow X)$) $> 1/3$ in Table 1(a). There was only one relationship that was reported by EquLIM under this condition. We believe this is because of the feedback mechanisms that are presented in the generating structure (GS ; see Figure 4.) Note that EquLIM scores hypothesis E_6 (from Figure 2) that might be promising to identify feedback mechanism between genes. In Table 1(b) we report pairwise relationships that showed $P(E_3) < 1/3$. Except relationships between *gene_1* and *gene_3*, all other relationships received causally independent relationship, $X \perp Y$, to be the most promising relationship. However, note that relationships between (*gene_3* and *gene_4*), (*gene_3* and *gene_5*), and (*gene_2* and *gene_3*) all show $P(E_6) > 2 \cdot P(E_3)$. Although it needs further investigation, we believe this trend, e.g., $P(E_6) > 2 \cdot P(E_3)$, might be promising to predict feedback mechanisms.

We also have looked into the conditional probabilities among predicted causal relationships in Table 1 to further classify it into excitatory or inhibitory causal relationship. For example, in Table 2 we show conditional probabilities of $P(\text{gene}_3 \mid \text{gene}_1)$. It seems the relationship $\text{gene}_1 \rightarrow \text{gene}_3$ is inhibitory because $P(\text{gene}_3 = \text{high} \mid \text{gene}_1 = \text{low}) = 0.778$ is about nine times more likely than $P(\text{gene}_3 = \text{low} \mid \text{gene}_1 = \text{low}) = 0.083$. From GS (see Figure 4), indeed *gene_1* inhibits *gene_3* through *gene_4*.

6. DISCUSSION

We have shown that EquLIM can be applied to an actual microarray dataset, predict causal relationship, and further infer whether relationships are excitatory or inhibitory. However,

based on the results we report here, causal predictions are tedious -- because of latent variables and feedback mechanisms, need more data, e.g., data from gene knock out experiments, and require better causal discovery algorithms.

We believe including more random variables in G will enable us to cast a larger net and possibly discover causal relationships with more significant posterior score. Also, modeling real valued random variables and modeling time, e.g., using dynamic Bayesian networks, might be a natural extension of EquLIM to analyze microarray data.

Acknowledgments

This research was supported by NIH grant P20RR017670 from NCRR.

References

1. Spirtes, P.; Glymour, C.; Scheines, R. Causation, prediction, and search. 2 ed.. Cambridge, MA: MIT Press; 2000.
2. Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*. 1995; 20:197–243.
3. Mani S, Cooper G, Spirtes P. A Theoretical Study of Y Structures for Causal Discovery. *UAI*. 2006
4. Heckerman, D. Proceedings of the Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann: 1995. A Bayesian approach to learning causal networks.
5. Karp PD. Pathway databases: A case study in computational symbolic theories. *Science*. 2001; 293:2040–2044. [PubMed: 11557880]
6. Gifford DK. Blazing pathways through genetic mountains. *Science*. 2001; 293:2049–2051. [PubMed: 11557882]
7. Mjolsness E, DeCoste D. Machine learning for Science: State of art and future prospects. *Science*. 2001; 293:2051–2055. [PubMed: 11557883]
8. Kitano H. Systems Biology: A Brief Overview., March 1, 2002. *Science*. 2002; 295:1662–1664. [PubMed: 11872829]
9. Friedman N. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*. 2004; 303(5659):799–805. [PubMed: 14764868]
10. Wille A, et al. Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology*. 2004; 5(11):R92.1–13. [PubMed: 15535868]
11. Beal M, et al. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*. 2005; 21(3):349–356. [PubMed: 15353451]
12. Yoo C, Cooper G. An Evaluation of a System that Recommends Microarray Experiments to Perform to Discover Gene-Regulation Pathways. *Journal of Artificial Intelligence in Medicine*. 2004; 31:169–182. [PubMed: 15219293]
13. Cooper, GF.; Yoo, C. Proceedings of the Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann: 1999. Causal discovery from a mixture of experimental and observational data.
14. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*. 1992; 9:309–347.
15. Pearl, J. Probabilistic Reasoning in Intelligent Systems. In: Brachman, RJ., editor. Representation and Reasoning. San Mateo, CA: Morgan Kaufmann: 1988.
16. Yoo, C.; Cooper, G. Center for Biomedical Informatics Research Report CBMI-173. Pittsburgh, PA: Center for Biomedical Informatics; 2001. Causal discovery of latent-variable models from a mixture of experimental and observational data.
17. Yoo, C.; Thorsson, V.; Cooper, GF. Pacific Symposium on Biocomputing. Maui, Hawaii: World Scientific; 2002. Discovery of a gene-regulation pathway from a mixture of experimental and observational DNA microarray data.

18. Yoo C, Brilz E. Local Causal Discovery Algorithm using Causal Bayesian networks. Journal paper to be submitted. 2007
19. Yoo, C.; Cooper, G. AMIA. San Antonio, Texas: 2002. Discovery of gene-regulation pathways using local causal search.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
A hypothetical gene-regulation pathway with manipulation.

Figure 2.
Six Local Causal Hypotheses

Figure 3.

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

Figure 4.

The generating structure (GS) of the five-gene-network data of the DREAM competition. → and -| represent excitatory and inhibitory relationships respectively.

Table 1

EquLIM causal prediction results on DREAM five-gene-network dataset. See Figure 2 for definition of six hypotheses E_1 through E_6 .

(a) Causal relationships ($> 1/3$)	
Relationship (E_2, E_4 , or E_5)	$P(\text{Relationship})$
$gene_1 \rightarrow gene_3$	0.7936

(b) Causal independence relationships ($P(E_3) < 1/3$)			
Relationship (E_3 or E_6)	$P(\text{Relationship})$	$P(E_6)$	$P(E_3)$
$gene_1 \text{ } gene_3$	0.0915	0.0463	0.0452
$gene_3 \text{ } gene_4$	0.6363	0.4361	0.2002
$gene_3 \text{ } gene_5$	0.9068	0.6535	0.2533
$gene_2 \text{ } gene_3$	0.9148	0.6371	0.2777

Table 2

Conditional probability, $P(\text{gene_3} \mid \text{gene_1})$, from predicted causal relationships from Table 1(a).

		<i>gene_1</i>		
		low	no change	High
<i>gene_3</i>	low	0.083	0.833	0.083
	no change	0.146	0.521	0.333
	high	0.778	0.111	0.111