19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Evaluation of causal Bayesian network search algorithms using simulated mesotheliomas gene expression data

Changwon Yoo[1] and Meredith Wilcox[1]

[1]Department of Epidemiology and Biostatistics, Florida International University, Miami, FL.
Email: cyoo@fiu.edu

Abstract: To understand the physiology of a complex disease, such as mesotheliomas, it is necessary to learn how the genes that are involved in developing the disease interact with the environment. To this end, statistical methods that can detect these gene-environment interactions will help scientists in detecting causal relationships among genes. These predicted causal relationships among genes can then be later verified through actual laboratory experiments.

In this paper, we have developed a novel causal discovery system that incorporates recent advances in Bayesian network search methods. We introduce a novel algorithm called Equivalence Checking Local Implicit latent variable scoring Method with Markov Chain Monte Carlo (EquLIM-MCMC) search algorithm that extends existing causal Bayesian network discovery algorithms, EquLIM and the Local Implicit latent variable scoring Method (LIM). Markov Chain Monte Carlo (MCMC) search has been shown to be very useful especially in analyzing datasets where the number of input variables greatly exceeds the number of cases that are collected (Friedman and Koller 2000; Hageman, Leduc et al. 2011). More and more datasets that are collected for gene expression studies have thousands of genes' expression levels (input variables) that are measured from tens or hundreds of subjects (cases). Datasets collected in gene-environment interactions studies will show similar trends.

We use LIM with MCMC (LIM-MCMC) and EquLIM-MCMC to analyze purely observational simulated gene expression datasets. To test these algorithms' abilities to detect causal relationships from realistic data, we generate datasets from a gene regulation pathway model of malignant mesothelioma formation proposed by an expert. Using the metrics of Area Under Receiver Operating Characteristic (AUROC) curve, Positive Predictive Value (PPV), and Shannon Entropy, we show that EquLIM-MCMC exhibit clear advantages over LIM-MCMC with causal relationship predictions. EquLIM-MCMC therefore improves over LIM-MCMC's ability in detecting causal relationships in gene networks and gene-environment interactions from presently available observational gene expression data.

Keywords: *Causal Bayesian networks, gene-gene interaction, gene expression study, causal discovery*

1. INTRODUCTION

Discovering causal relationships is the main focus of many scientific studies. While experimental studies (e.g., randomized control trials) are potentially very informative, they may be expensive and/or difficult to conduct the experiments (e.g., knocking out genes in mice). Thus, finding promising causal relationships from observational data – those from retrospective, prospective, case control, and/or longitudinal studies with no interventions – will be helpful. Especially, in gene expression studies, it will be helpful if we can discover gene-gene interactions, i.e., causal relationships among genes, from a study of that only perturbs the environment. Later the gene-gene interactions can be verified in wet laboratories through experimental studies.

Causal modeling is an active field of research in which numerous advances have been made in areas that include Bayesian causal representation, model assessment and scoring, model search, and application to biological networks (Spirtes, Glymour et al. 2000; Yu, Smith et al. 2004; Werhli and Husmeier 2007; Grzegorczyk, Husmeier et al. 2008).

The contribution of the current paper is to investigate promising causal discovery algorithms that identify causal relationships with only observational cases. We also compare the causal discovery algorithms. We introduce a novel pairwise causal relationships scoring method — Equivalence Local Implicit latent variable scoring Method with Monte-Carlo Markov Chain search algorithm (EquLIM-MCMC) — to learn causal networks from observational data. EquLIM-MCMC is based on earlier work (Cooper and Yoo 1999; Yoo and Blitz 2009) by improving the search methods to discover promising causal relationships on observational data alone. However, this paper will give better understanding of the performance of EquLIM-MCMC by comparing it with LIM-MCMC. We investigate EquLIM-MCMC's learning performance using area under receiver operating characteristics (AUROC) curves. This evaluation is a simulation study in which cases were generated from a gene network simulator. We report the result of our analysis in this paper.

2. METHODS

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 (Pearl 1988). Using Bayesian Networks, we introduce six equivalence classes (E_1 through E_6) among the structures that are shown in Figure 1. Note that in Figure 1, H is a latent variable and X and Y are nodes (variables). We use these causal hypotheses in EquLIM-MCMC and LIM-MCMC.

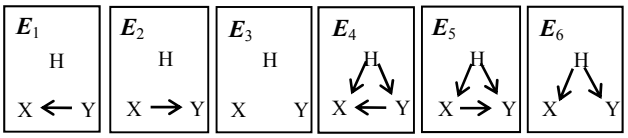

Figure 1. Six Local Causal Hypotheses

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.

2.1. Equivalence Local Implicit Latent Variable Scoring Method with MCMC

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 with MCMC (EquLIM-MCMC).

Implicit Latent Variable Scoring (ILVS) Method. 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. ILVS scores each E_i in Figure 1 by only considering pairwise measured nodes. In earlier studies, ILVS was applied to simulated data (Yoo and Cooper 2001) and to yeast DNA microarray data (Yoo, Thorsson et al. 2002). ILVS is extended (called extILVS) to scores more than pairwise relationships.

Local ILVS Method (LIM) and Equivalence Local ILVS Method (EquLIM). Let L_i^{XY} denote a set of local structures that includes E_i^{XY} and let $L^{XY} = \cup_i L_i^{XY}$. LIM (Local ILVS method) calculates $P(E_i^{XY} \mid D, K)$ 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 **Error! Bookmark not defined.** by the following equation:

$$P(E_i^{XY} \mid D,K) \approx \frac{\sum_{S:E_i^{XY} \text{ is in } T} P(D,S \mid K)}{\sum_T P(D,T \mid K)} \qquad (1)$$

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 (Heckerman, Geiger et al. 1995). LIM 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 $S = \{E_i^P \mid P \in \{\text{all pairs in } W\}\}$. (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 six different structures with extILVS by substituting E_i^P with one of the six hypotheses (from Figure 1) 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 there is no improvement in the structure score (it has reached a local maximum) and 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. In the remainder of the paper, we use Local Structure Size (LSS) to refer the number of nodes in L^{XY}.

EquLIM extends LIM by scoring additional structures in (Step 3) of LIM: when we score six different structures with extILVS by substituting E_i^P with one of the six hypotheses (from Figure 1) for a node pair P in W, we additionally search for the reverse arc structure of each structure and score it (Yoo and Blitz 2009).

LIM with Monte-Carlo Markov Chain Search (LIM-MCMC) and EquLIM with Monte-Carlo Markov Chain Search (EquLIM-MCMC). Both LIM-MCMC and EquLIM-MCMC use ordering search (Friedman and Koller 2000) instead of greedy-hill climbing search for local structure search in (Step 2) and (Step 3) in LIM. For a given data set, the algorithms search for 1) the highest scoring Bayesian order, and 2) the pairwise associations among all variables.

Instead of (Step 2) of LIM, LIM-MCMC and EquLIM-MCMC begin by generating a random order of the variables, i.e. $\{X_1 \prec X_2 \prec \cdots \prec X_{k-1} \prec X_k\}$, where k is the total number of variables within the local structure. The score of the first order, defined as the probability of the data given the order, is then computed as follows:

$$P(D \mid \prec) = \prod_i \sum_{\boldsymbol{U} \in U_{i,\prec}} \{score(X_i, \boldsymbol{U} \mid D)\} = \prod_i \sum_{\boldsymbol{U} \in U_{i,\prec}} \left\{ \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \right\} \qquad (2)$$

where $U_{i,\prec}$ represents the possible parent sets for node X_i; q_i the configurations of a given parent set U for node X_i; r_i the possible values of node X_i; α_{ijk} the prior for node X_i given parent set U_j and value r_k, with $\alpha_{ij} = \sum_k(\alpha_{ijk})$; N_{ijk} the number of cases in D which have value r_k for node X_i and configuration q_j for parent set U, with $N_{ij} = \sum_k(N_{ijk})$; and $\Gamma(\cdot)$ the Gamma function (Friedman and Koller 2000).

Bins are maintained throughout the search that track the pairwise associations among all variables. For two variables X and Y in which X precedes Y within the order, either the causal relationships $\{E_2^{XY}, E_5^{XY}\}$ or the independent relationships $\{E_3^{XY}, E_6^{XY}\}$ are possible. The bins are updated with each order by placing the appropriate portion of the order score into the respective association bin. The portion placed into the causal and independent bins are calculated as follows:

$$P(R \mid \prec) = \sum_{\boldsymbol{S} \in S_{R,\prec}} P(D \mid \boldsymbol{S}, \boldsymbol{K}) \qquad (3)$$

where $R \in \{\{E_2^{XY}, E_5^{XY}\}, \{E_3^{XY}, E_6^{XY}\}\}$; $P(D|\boldsymbol{S}, \boldsymbol{K})$ is a structure score; $\boldsymbol{S} \in S_{\{E_2^{XY}, E_5^{XY}\}, \prec}$ represents the structures that satisfy the order and in which X is a direct or indirect parent of Y; and $\boldsymbol{S} \in S_{\{E_3^{XY}, E_6^{XY}\}, \prec}$ represents the structures that satisfy the order and in which X and Y are independent.

With appropriate assumptions, we can evaluate $P(D|S,K)$ in Equation **Error! Bookmark not defined.** with the following equation (Cooper and Herskovits 1992; Heckerman, Geiger et al. 1995):

$$P(D \mid S, K) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{4}$$

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 (Heckerman, Geiger et al. 1995) with $\alpha_{ijk} = \frac{1}{r_i q_i}$, which is a commonly used non-informative parameter prior for the BDe metric.

For EquLIM-MCMC the score of the first reverse order is then computed, i.e. the score of the order $\{X_k \prec X_{k-1} \prec \cdots \prec X_2 \prec X_1\}$. The pairwise association bins are updated by adding the pairwise association scores of the first reverse order to the respective bins, which at this point in time contain only those of the first order. This step is not present in LIM-MCMC; the score is not calculated and the bins are not updated for the reverse order. The second order is then generated for both LIM-MCMC and EquLIM-MCMC by switching the positions of two randomly selected variables within the first order. The second order score and its pairwise association scores are computed and the association bins are updated as previously described. For EquLIM-MCMC, the order score of the second reverse order is found as well, and the association bins are updated once again.

To determine whether or not a local move to the second order will be accepted, the order score of the first order is now compared to that of the second. The local move is accepted with probability, $min\left[1, \frac{P(D|\prec_2)}{P(D|\prec_1)}\right]$ (Friedman and Koller 2000). The algorithms continue to generate new orders from the current order until the specified number of orders has been encountered. For every order and in the case of EquLIM-MCMC, its reverse order, scores are computed and the pairwise association bins are updated. With each new order, the probability of acceptance is found and a local move is accepted or denied, as previously described.

Example Run of LIM-MCMC and EquLIM-MCMC. Let us assume once again that we have the five modeled nodes: *U, V, X, Y, Z*; and using a LSS of 3, choose $W = \{X, Y, Z\}$. A random ordering of the nodes in W will be generated, for example $\{Z \prec X \prec Y\}$ and the score of the order will be computed. If we are interested in L^{XY}, note that X and Y can have the relationship $\{E_2^{XY}, E_5^{XY}\}$ or $\{E_3^{XY}, E_6^{XY}\}$. The association bins will be updated by adding the scores of the structures that satisfy the order and in which $\{E_2^{XY}, E_5^{XY}\}$ is present into the appropriate causal bin; and those in which $\{E_3^{XY}, E_6^{XY}\}$ is present into the independent bin. For EquLIM-MCMC, the score of reverse order, i.e. $\{Y \prec X \prec Z\}$, will then be computed and the bins will once again be updated. In this case, the scores of the structures that satisfy the order and in which $\{E_1^{XY}, E_4^{XY}\}$ are present will be used to update the causal bin. LIM-MCMC skips this step. The next step for both algorithms involves randomly switching the position of two nodes within the order, say Z and X for example, generating the new order $\{X \prec Z \prec Y\}$. For EquLIM-MCMC, both the second order and its reverse order will be scored, with the association bins being updated for each. For LIM-MCMC, only the score of the second order will be calculated and will contribute to the association bins and not that of its reverse order. If the score of the second order is greater than that of the first, the local move will be accepted. The third order will then be generated from the second order, as the second was from the first. The second order can be accepted even if its score is not greater than the first, if the randomly generated probability is less than or equal to the acceptance probability. If the second order is not accepted, the third order will be generated from the first. These steps will continue until the specified number of orders is visited. The association bins for $\{E_1^{XY}, E_4^{XY}\}$, $\{E_2^{XY}, E_5^{XY}\}$, and $\{E_3^{XY}, E_6^{XY}\}$ will be normalized as they were in EquLIM.

3. EXPERIMENTAL METHODS

If we know the real-world causal relationships among a set of variables of interest, then we could generate simulated observational datasets. Then we can use these datasets as input and let a causal learning method predict the causal structure and estimate the causal parameters that exist among the modeled variables. Since we know the true causal relationships, these predictions and estimates would then be compared to the true causal relationships. However, confident knowledge of underlying causal processes is relatively rare. That is why in this study of causal discovery from observational data, we used as a gold standard a causal model that was constructed by an expert biologist. In particular, we wanted to generate simulated gene expression data from known gene-gene interactions. We started with gene regulation pathways from a model of malignant mesothelioma formation (Murthy and Testa 1999), which was used as the model input for a simulator built for producing high throughput data.

We chose to evaluate LIM-MCMC and EquLIM-MCMC with the TETRAD network simulator, which is open source and runs on the Java platform. TETRAD models microarray-generated data and is able to incorporate measurement noise both from systematic and stochastic sources (Spirtes, Glymour et al. 2008). We generated data from the gene regulation pathways in the model of asbestos-related diseases formation simulator using plausible interactions among eight relatively well studied genes, i.e., *IFNγ*, *NFKB*, *SRA*, *CASP3*, *ICAM1*, *IL6*, *IL1*, and *TNFα*, and then used this data in evaluating the learning method, LIM-MCMC and EquLIM-MCMC by comparing their predicted performance using Area Under Receiver Operating Curve (AUROC), Positive Predictive Value (PPV), and Shannon Entropy.

We have implemented EquLIM-MCMC and LIM-MCMC in R (version 2.12.2). Since EquLIM-MCMC and LIM-MCMC are anytime algorithms, we have let EquLIM-MCMC and LIM-MCMC run for about 4 hours for each dataset. For each dataset, we performed five independent runs of EquLIM-MCMC and LIM-MCMC, thus the experiment ran for a total of 120 hours. We used LSS 5 for this experiment due to the fact that LSS 5 ran for a reasonable time (about 4 hours) to produce stable results.

4. RESULTS

Here we show the average AUROC results of LIM-MCMC and EquLIM-MCMC in analyzing the five independence runs for each six datasets. To calculate the AUROC, we have calculated the AUROC under two prediction categories: *causal*, i.e., $E_1(E_2)$, and *independence*, i.e., E_3, predictions. Since we know the true pairwise causal relationships between all eight genes in the simulator, we used all 28 pairs of genes to calculate whether LIM-MCMC, and EquLIM-MCMC correctly predicted each relationship of the 28 pairs of genes. Also we provide average Causal Prediction Value Rate (CPVR) which calculates the proportion of correctly predicted causal relationships out of total causal relationships in analyzing the five independence runs for each six datasets. We calculate Independence Prediction Value Rate (IPVR) in a similar way. We also calculated Shannon Entropy using three hypothesis posterior probabilities, i.e., $P(E_1^{XY} \mid D_i)$, $P(E_2^{XY} \mid D_i)$, and $P(E_3^{XY} \mid D_i)$, predictions for all node pairs X and Y that have causal relationships with dataset D_i where i represents number of cases, i.e., i = 20, 50, 100, 500, 1000, or 2000.

Table 1 shows that the causal predictions of EquLIM-MCMC outperform those of LIM-MCMC and EquLIM. It also shows that in most cases the independence AUROC averages of LIM-MCMC are better than those of EquLIM-MCMC. This is because EquLIM-MCMC emphasizes in searching for causal relationships not independence relationships. It is also interesting to see EquLIM-MCMC outperform LIM-MCMC with 50 cases since most of the initial high throughput data studies (1) will have small numbers of cases (<100 cases); (2) will be mostly observational data; and (3) will seek novel *causal* relationships.

Table 2 shows there are no significant differences among LIM, EquLIM and EquLIM-MCMC in terms of CPVR and IPVR. However, EquLIM-MCMC better predicts both causal and independence as more data is added. It is also interesting to see that in most of the cases, LIM-MCMC predicts causal relationships with higher confidence, i.e., lower entropy. This is especially promising since most of the initial microarray studies are limited in terms of number of cases (microarray chips).

Table 1. Causal and Independence Predictions of average AUROC of ten independent runs of LIM, EquLIM, and EquLIM-MCMC on six datasets, i.e., dataset with only observational data. Shaded columns represent

Independence Predictions Numbers in the parentheses are standard deviation. The results are shown for Local Structure Size of five.

	Causal Prediction		Independence Prediction	
Algorithm / # of cases	LIM-MCMC	EquLIM-MCMC	LIM-MCMC	EquLIM-MCMC
20	0.872 (0)	0.872 (0)	0.547 (0)	0.547 (0)
50	0.911 (0.010)	0.924 (0.015)	0.510 (0.012)	0.508 (0.016)
100	0.879 (0.029)	0.919 (0.034)	0.559 (0.027)	0.561 (0.026)
500	0.784 (0.069)	0.801 (0.123)	0.596 (0.058)	0.553 (0.053)
1,000	0.779 (0.074)	0.788 (0.075)	0.543 (0.066)	0.502 (0.016)
2,000	0.869 (0.050)	0.879 (0.071)	0.596 (0.078)	0.538 (0.113)

Table 2. The average Causal Predicted Value Rate (CPVR) and Independent Predicted Value Rate (IPVR) of ten independent runs of LIM, EquLIM, and EquLIM-MCMC on six datasets, i.e., dataset with only observational data. Numbers in the parentheses in CPVR column are IPVR. The average Shannon Entropy of Causal Prediction of ten independent runs of LIM, EquLIM, and EquLIM-MCMC on six datasets, i.e., dataset with only observational data. Numbers in the parentheses are standard deviation. The results are shown for Local Structure Size of five.

	CPVR (IPVR)		Shannon Entropy	
Algorithm / # of cases	LIM-MCMC	EquLIM-MCMC	LIM-MCMC	EquLIM-MCMC
20	0.167 (0.300)	0.167 (0.300)	0.428 (0.049)	0.428 (0.049)
50	0.178 (0.140)	0.100 (0.160)	0.403 (0.054)	0.410 (0.053)
100	0.367 (0.160)	0.278 (0.060)	0.364 (0.089)	0.382 (0.080)
500	0.467 (0.220)	0.500 (0.180)	0.219 (0.128)	0.233 (0.125)
1,000	0.611 (0.080)	0.622 (0.140)	0.130 (0.136)	0.165 (0.127)
2,000	0.556 (0.060)	0.567 (0.100)	0.169 (0.137)	0.173 (0.137)

We also compared EquLIM-MCMC and LIM-MCMC with global network search (Cooper and Herskovits 1992; Heckerman, Geiger et al. 1995) using the same simulated data. The global network search results show high variance and unreliable causal and independence predictions compared to those of LIM–MCMC and EquLIM-MCMC. In conclusion, comparing the results in Table 1 and Table 2 with the global network search shows that LIM-MCMC and EquLIM-MCMC produces more reliable predictions in both causal and independent relationships.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have described causal Bayesian networks structure search algorithm called LIM-MCMC and EquLIM-MCMC. We have generated simulation data that includes no perturbations of a gene network. Better causal prediction abilities can be achieved by perturbations of the gene network. However, with limited budget and time, it is also useful to have an analysis that can predict novel causal relationships from limited resources. We believe EquLIM-MCMC can be useful in such initial analysis of experiments with limited resources. We have also shown that EquLIM-MCMC predicts causal relationships better than LIM-MCMC using a dataset with only observational cases.

Extensions of this work include examining the synergy of case control data in conjunction with observational and experimental data (Cooper 2000), and modeling beyond pairwise causal relationships between the measured variables (Yoo and Cooper 2002). We plan to study the effect of larger LSS (> 5) and plan to apply EquLIM-MCMC to actual experimental datasets from different genomic studies.

REFERENCES

Cooper, G. F. (2000). A Bayesian method for causal modeling and discovery under selection. Proceedings of the Confernce on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Franciso, CA.

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

Cooper, G. F. and C. Yoo (1999). Causal discovery from a mixture of experimental and observational data. Proceedings of the Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann.

Friedman, N. and D. Koller (2000). Being Bayesian about network structure. Proceedings of Uncertainty in Artificial Intelligence.

Grzegorczyk, M., D. Husmeier, et al. (2008). "Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler." Bioinformatics **24**: 2071 - 2078.

Hageman, R., M. Leduc, et al. (2011). "A Bayesian Framework for Inference of the Genotype–Phenotype Map for Segregating Populations." Genetics **187**(4): 1163-1170.

Heckerman, D., D. Geiger, et al. (1995). "Learning Bayesian networks: The combination of knowledge and statistical data." Machine Learning **20**: 197-243.

Murthy, S. and J. Testa (1999). "Asbestos, Chromosomal Deletions, and Tumor Suppressor Gene Alterations in Human Malignant Mesothelioma." Journal of Cell. Physiol. **180**: 150-157.

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

Spirtes, P., C. Glymour, et al. (2000). Causation, prediction, and search. Cambridge, MA, MIT Press.

Spirtes, P., C. Glymour, et al. (2008). The TETRAD Project: Causal Models and Statistical Data.

Werhli, A. V. and D. Husmeier (2007). "Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge." Statistical Applications in Genetics and Molecular Biology **6**(1): Article 15.

Yoo, C. and E. Blitz (2009). "Local Causal Discovery Algorithm using Causal Bayesian networks." Annals of the NY Academy of Science **1158**: 93-101.

Yoo, C. and G. Cooper (2001). Causal discovery of latent-variable models from a mixture of experimental and observational data. Center for Biomedical Informatics Research Report CBMI-173. Pittsburgh, PA, Center for Biomedical Informatics.

Yoo, C. and G. Cooper (2002). Discovery of gene-regulation pathways using local causal search. AMIA, San Antonio, Texas.

Yoo, C., V. Thorsson, et al. (2002). Discovery of a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. Pacific Symposium on Biocomputing, Maui, Hawaii, World Scientific.

Yu, i., V. A. Smith, et al. (2004). "Advances to Bayesian network inference for generating causal networks from observational biological data." Bioinformatics(20): 3594 - 3603.