Discovery of Gene-Regulation Pathways using Local Causal Search

Changwon Yoo and Gregory F. Cooper

Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh PA

ABSTRACT

This paper reports the methods and results of a computer-based algorithm that takes as input the expression levels of a set of genes as given by DNA microarray data, and then searches for causal pathways that represent how the genes regulate each other. The algorithm uses local heuristic search and a Bayesian scoring metric.

We applied the algorithm to induce causal networks from a mixture of observational and experimental gene-expression data on genes involved in galactose metabolism in the yeast Saccharomyces cerevisiae. The observational data consisted of gene-expression levels obtained from unmanipulated "wild-type" cells. The experimental data were produced by deleting ("knocking out") genes and measuring the expression levels of other genes. We used this data to evaluate several variations of the local search method. In each evaluation, causal relationships were predicted for all 36 pairwise combinations of nine key galactose-related genes. These predictions were then compared to the known causal relationships among these genes.

INTRODUCTION

Caausal 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[1,2] and gene pathway (network) methods[3-7]. Dutilh[8] gives a short review of genetic networks. A more thorough review of genetic networks based on biological context was published by Smolen, et al.[9]. Wessels, et al.[10] conducted a limited comparison study of selected continuous genetic network models[3,11,12]. Unlike these previous methods, we use a method that models a latent variable implicitly and experimental interventions explicitly when evaluating hypotheses about causal relationships. In this paper, we extend our scoring method[18] by introducing a local search method. Many Bayesian network-structure search methods have been introduced and evaluated[13-15]. Most of the search methods use global search, i.e., a search that considers all modeled variables at a time, to search for the best structure that fits the data. Because of the large search space of the global structures, this makes the global search method less efficient in learning local (e.g., pairwise) causal relationships.

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*[16]. Our analysis focuses on pairs of genes (X, Y) and attempts to determine whether the expression level of gene X has a causal influence on the expression of gene Y. As a representation of causation, we use causal Bayesian networks. These networks include variables that represent expression levels of measured genes, as well as latent (hidden) variables that represent unmeasured quantities, such as the cellular levels of proteins and small molecules[17]. We introduce several variations of our local causal search method and compare their pathway predictions with the pathways that have been established in the biological literature[18].

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[19]. Figure 1 illustrates the structure of a hypothetical causal Bayesian network structure that contains five nodes. The probabilities associated with this causal network structure are not shown.

The causal network structure in Figure 1 indicates, for example, that the *Gal4* gene can regulate (causally influence) the expression level of the *Gal3* gene, which in turn can regulate the expression level of *Gal5* gene.

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).*

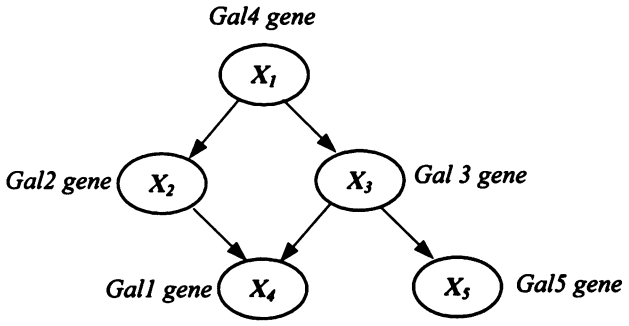

Figure 1. A causal Bayesian network that represents a portion of the gene-regulation pathway for galactose metabolism in yeast.

The causal Markov condition permits the joint distribution of the n variables in a causal Bayesian network to be factored as follows[19]:

$$P(x_1, x_2, \ldots, x_n \mid K) = \prod_{i=1}^{n} P(x_i \mid \pi_i, K) \qquad (1)$$

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.

We introduce 6 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. We understand that modeling all nodes will decrease the structure and parameter prediction errors, but as the first step of modeling a latent variable, we only look at pairwise relationships between two nodes (X and Y) and a latent variable H.

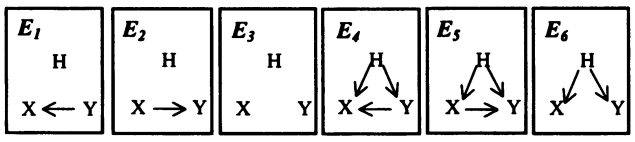

Figure 2. Six Local Causal Hypotheses

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. 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:

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

where the sum is taken over all causal network structures that (1) contain just the nodes in S, and (2) contain a structure E_i^{XY}. Based on the properties of probabilities, the term within the sum in Equation 2 may be rewritten as follows:

$$P(S \mid D, K) = \frac{P(S, D \mid K)}{P(D \mid K)} = \frac{P(S, D \mid K)}{\sum_S P(S, D \mid K)} \qquad (3)$$

Since the probability $P(D \mid K)$ 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 $P(S, D \mid K)$, 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:

$$\begin{aligned} P(S, D \mid K) &= P(S \mid K) P(D \mid S, K) \\ &= P(S \mid K) \int P(D \mid S, \theta_S, K) P(\theta_S \mid S, K) d\theta_S \end{aligned} \qquad (4)$$

where (1) $P(S \mid K)$ 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) $P(D \mid S, \theta_S, K)$ 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) $P(\theta_S \mid S, K)$ 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[13,14]:

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

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[14] with $\alpha_{ijk} = \frac{1}{r_i q_i}$, which is a commonly used non-informative parameter prior for the BDe metric. When a case involves the manipulation of a variable (e.g., the "knocking out" of gene X_i), we do not tally a count for that case in N_{ijk}[20]. See Yoo and Cooper[17] for more information on modeling experimental interventions.

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 in the literature, including methods based on stochastic simulation and on expectation maximization[14]. 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[17].

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 $P(D_o \mid E_5, K)$ 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 $P(D_m \mid E_5, K)$. 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 $P(D \mid E_5, K) = P(D_o \mid E_5, K) \times P(D_m \mid E_5, K)$. 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[17] 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.

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[17] and to yeast DNA microarray data[18]. We have also extended ILVS to create a system called extILVS that scores more than pairwise relationships.[1)]

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=Gal1* and *Y=Gal2*. Then L_1^{XY} could be the causal structure shown in Figure 1. 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 2 by the following equation:

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

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[14]. 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 $S = \{E_i^P \mid P \in \{\text{all pairs in } W\}\}$. We initialize S to a random structure by randomly choosing E_i^P for all P. (Step 2) 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 3) Select the E_j^{P*} that in Step 2 generated the structure with the highest score; update S by substituting E_j^{P*} for E_i^P in S and repeat Step 2 with the new S. Stop the search if either there is no improvement in the structure score or the number of iterations exceeds a user defined limit. We also perturb the structure once the search reaches a local maximum. We provide three different perturbations to avoid a local maximum and they select one of the six hypotheses (from Figure 2) for each of the node pairs (X, Y) according to the following distributions:

- Random Perturbation: $\{P(E_i^{XY} \mid D,K) = 1/6 \mid i=1,2,\ldots 6\}$
- Local Perturbation: $\{P(E_i^{XY} \mid D,K) \mid i=1,2,\ldots 6\}$ calculated by LIM.
- ILVS Perturbation: $\{P(E_i^{XY} \mid D,K) \mid i=1,2,\ldots 6\}$ calculated by ILVS.

We later pair these perturbations to introduce different variations of LIM. Also, two different search methods were implemented:

- Local Search: Iterating Step 2 through Step 3 while forcing S to include pairwise relationships E_1^{XY}, $E_2^{XY},\ldots, E_6^{XY}$ for each of node pairs (X, Y)
- Global Search: Iterating Step 2 through Step 3 with no restrictions on S.

[1)] Detail pseudo code of LIM is available at: http://www.cbmi.upmc.edu/~cwyoo/yoo-thesis.pdf

Example Run of LIM. For example, let us assume there are only five modeled nodes: *U, V, X, Y,* and *Z*. Further assume we are limiting $k = |W| < 4$ and $W=\{X, Y, Z\}$. In Step 1 we randomly initiate a structure, e.g., $S=\{E_1^{XY}, E_2^{XZ}, E_6^{YZ}\}$. In Step 2, we first consider the six different structures derived from S by substituting E_1^{XY} with any of $\{E_1^{XY}, E_2^{XY}, E_3^{XY}, E_4^{XY}, E_5^{XY}, E_6^{XY}\}$. We do the same for E_2^{XZ} and E_6^{YZ}. We evaluate all 18 different structures. In Step 3 we choose the highest scored structure and go to Step 2. Upon reaching a stopping condition, to score $P(E_1^{XY}|S,K)$, for example, we sum all scores of visited structures that include E_1^{XY} and divided 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_1^{XY}.

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.[16]. The experiments included single gene deletion involving nine of the key genes[2)] that participate in yeast galactose metabolism.

Dataset Preparation. Since the dataset was already normalized[3)], we discretized the dataset according to the distribution of each gene's expression level[18].

Analyses. We applied LIM to every pair of the nine galactose genes. For each gene pair, we used LIM to derive a posterior probability for each of the six causal hypotheses in Figure 2. Since the causal relationships among these nine genes are understood relatively well, we assume that these generally accepted biological relationships are correct and can serve as a standard against which to compare the output of LIM. We implemented two variations of LIM: LIM using Random and Local Perturbation (LRLP); and LIM using ILVS and Local Perturbation (LILP). LRLP was implemented with Local Search (LRLP&Local) and Global Search (LRLP&Global) methods. Later we selected the Local Search method, which outperformed the Global Search method, and implemented with LILP (LILP&Local). All the experiments ran with 10 perturbations for each L_i^{XY} (i=1,2,...,6 and for all node paris (X,Y)). We used the same number of perturbations for the global search. We also set k=6. We also compared the LIM results with the results of ILVS analysis of the same dataset[18].

RESULTS

Comparison of LRLP&Local, LRLP&Global, and LILP&Local is shown in Figure 4. The plot in the figure was taken from the entire run and was sorted by the log scores. Figure 4 shows that LRLP&Local runs with the least number of visited structures but LILP&Local finds the most probable structure (structure with the highest log(P(D|S,K)) among the three methods. Although LRLP&Global considered a relatively large number of structures, it fails to find a structure with a higher log(P(D| than LILP&Local).

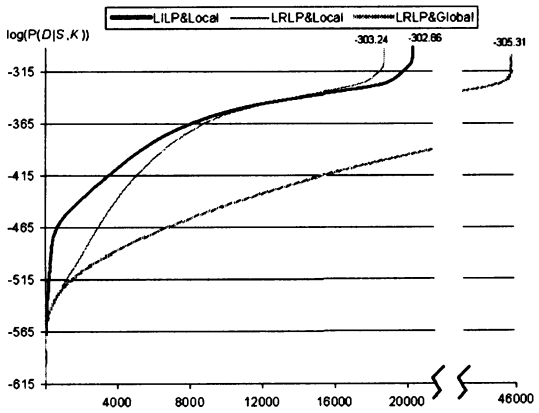

Figure 4. Comparison of LRLP&Local, LRLP&Global and LILP&Local. X-axis represents the number of unique structures visited by LIM and Y-axis represents log(P(D|S,K)). Highest score searched by each method is displayed in the graph.

Table 1. Comparison results: PPV(Positive Predictive Value) and NPV(Negative Predictive Value) of ILVS, LILP&Local, and LRLP&Local. Numbers in the parenthesis represent the total number of cases. Shaded columns in (a) indicate the causal relationship PPV (or NPV), e.g., the first shaded column shows that ILVS predicted 10 out of 12 causal directions correctly; Shaded columns in (b) indicate the confounded relationship PPV (or NPV). ♦ indicates $p < 0.05$ for the null hypothesis that one of the six causal hypotheses was chosen uniformly at random.

	PPV		NPV	
Predictions / Methods	E_1 and E_2	E_4 and E_5	E_3	E_6
ILVS	0.0 (3)	0.4 (9)	0.3 (19)	0.4 (5)
	0.8 (12)♦		0.3 (24)	
LILP&Local	0.5 (4)	1.0 (5)♦	0.2 (22)	0.8 (5)♦
	0.8 (9)♦		0.3 (27)	
LRLP&Local	0.2(5)	0.7(5)	0.2 (22)	1.0 (4)♦
	0.6 (10)		0.3 (26)	

(a) Causal Predictions

	PPV		NPV	
Predictions / Methods	E_4 and E_5	E_6	E_1 and E_2	E_3
ILVS	0.4 (9)	0.4 (5)	0.0 (3)	0.3 (19)
	0.4 (14)		0.2 (22)	
LILP&Local	1.0 (5)♦	0.8 (5)♦	0.3 (4)	0.2 (22)
	0.9 (10)♦		0.2 (26)	
LRLP&Local	0.7(5)	1.0 (4)♦	0.2(5)	0.2 (22)
	0.9 (9)♦		0.2 (27)	

(b) Confounded Predictions

2) Nine galactose genes are *Gal1, Gal2, Gal3, Gal4, Gal5(PGM2), Gal6(LAP3), Gal7, Gal10,* and *Gal80.*

3) The normalization method is available at: www.systemsbiology.org/VERAandSAM/

Summary results of the predictions made by ILVS, LRLP&Local, and LILP&Local are shown in Table 1[4]. Table 1(a) shows the performance of the three methods when causal relationships are (or are not) predicted. LILP&Local performed both better in predicting causal relationships of confounded and not confounded separately and performed comparable to ILVS when the two hypotheses were combined, i.e., E_1 and E_4 were considered equivalent. Table 1(b) shows the performance of the three methods when confounded relationships are (or are not) predicted. LRLP&Local and LILP&Local outperformed ILVS. All the method's NPV(Negative Predictive Value)s were low, which deserves more investigation.

DISCUSSION

We introduced a causal discovery algorithm that can model causal hypotheses with latent variables. LIM extends ILVS by searching through local structures that include the pairwise variables considered in ILVS.

We applied LIM to an available dataset containing gene expression levels from experiments that focused on galactose genes. A variation of LIM that incorporates some aspects of ILVS (LILP) showed better performance than ILVS, especially in predicting confounded relationships, but it needs improvement and further evaluation. Possible reasons for low NPVs of LIM include a small set of samples and limited experimental conditions and variation; for the false positives output by LIM, some may simply be wrong, while others may represent unknown causal relationships within galactose gene regulation.

Future work includes applying LIM as a sub-module in a decision support system that suggests additional DNA microarray experiments to perform that have the highest expected value of information for revealing gene regulation pathways of interest.

ACKNOWLEDGMENTS

We thank Subramani Mani for his helpful discussions.
This research was supported by NSF grant IIS-9812021 and NASA grant NRA2-37143.

REFERENCES

1. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Molecular Biology of the Cell 1998; 9:3273-3297.
2. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proceedings of the National Academy of Sciences 2000; 97:12079-12084.
3. Chen T, Filkov V, Skiena S. Identifying gene regulatory networks from experimental data, ACM-SIGAT The RECOMB, 1999.
4. D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 2000; 16:707-726.
5. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. Journal of Computational Biology 2000.
6. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks, Pacific Symposium on Biocomputing, 2001.
7. Maki Y, Tominaga D, Okamoto M, Watanabe S, Eguchi Y. Development of a System for the Inference of Large Scale Genetic Networks, PSB, Hawaii, 2001.
8. Dutilh B. Gene Networks from Microarray Data, , Literature thesis at Utrecht University, 1999.
9. Smolen P, Baxter DA, Byrne JH. Modeling transciptional control in gene networks - methods, recent results and future directions. Bulletin of Mathematical Biology 2000; 62:247-292.
10. Wessels LFA, Someren EPV, Reinders MJT. A comparison of genetic network models, PSB, 2001.
11. Arkin A, Shen P, Ross J. A test case of correlation metric construction of a reaction pathway from measurements. Science 1997; 277:1275-1279.
12. Weaver DC, Workman CT, Stormo GD. Modeling regulatory networks with weight matrices, Pacific Symposium on Biocomputing, 1999. 123.
13. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 1992; 9:309-347.
14. Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 1995; 20:197-243.
15. Friedman N, Koller D. Being Bayesian about network structure, Proceedings of UAI, 2000.
16. Ideker T, Thorsson V, Ranish JA, et al. Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. Science 2001; 292:929-934.
17. Yoo C, Cooper G. Causal discovery of latent-variable models from a mixture of experimental and observational data. CBMI Research Report CBMI-173. 2001.
18. Yoo C, Thorsson V, Cooper GF. Discovery of a gene-regulation pathway from a mixture of experimental and observational DNA microarray data, Pacific Symposium on Biocomputing, 2002.
19. Pearl J. Probabilistic Reasoning in Intelligent Systems. In: Brachman RJ, ed. Representation and Reasoning. San Mateo, CA: Morgan Kaufmann, 1988.
20. Cooper GF, Yoo C. Causal discovery from a mixture of experimental and observational data, Proceedings of the Conference on Uncertainty in Artificial Intelligence, 1999. Morgan Kaufmann.

[4] Detail results are posted at: http://www.cbmi.upmc.edu/~cwyoo/lim-results.pdf