An algorithm to learn causal relations between genes from steady state data: simulation and its application to melanoma dataset

Xin Zhang[1], Chitta Baral[1], Seungchan Kim[1,2]

[1]Department of Computer Science and Engineering, Arizona State University
Tempe, AZ 85287, USA
[2]Translational Genomics Research Institute
445 N. Fifth Street, Phoenix, AZ 85004, USA

Abstract. In recent years, a few researchers have challenged past dogma and suggested methods (such as the IC algorithm) for inferring *causal* relationship among variables using steady state observations. In this paper, we present a modified IC (mIC) algorithm that uses entropy to test conditional independence and combines the steady state data with partial prior knowledge of topological ordering in gene regulatory network, for jointly learning the causal relationship among genes. We evaluate our mIC algorithm using the simulated data. The results show that the precision and recall rates are significantly improved compared with using IC algorithm. Finally, we apply the mIC algorithm to microarray data for melanoma. The algorithm identified the important causal relations associated with WNT5A, a gene playing an important role in melanoma, verified by the literatures.

1 Introduction

The recent development of high-throughput genomic technologies like cDNA microarray and oligonucleotide chips [14] empowers researchers in new ways to study how genes interact with each other. This has led to researchers using mathematical modeling and *in-silico* simulation study to analyze the interaction structure unambiguously and predict the network dynamic behavior in a systematic way [4, 19].

In previous studies on cellular response of genotoxic damage [10] and melanoma data [2, 11, 17], coefficient of determination (CoD) is used to infer gene network structure. CoD provides a normalized measure of the degree to which target variables can be better predicted using the observations in a feature set than it can be in the absence of observations. While CoD provides useful information for network connectivity, the relationships identified via CoD does not necessarily imply causal relations. Bayesian network model, which represents statistical dependencies, also has been proposed to discover interactions between genes [6, 22, 24]. Based on Bayesian model there are some other network inference methods that were evaluated by applying them to biological simulation with known network topology [18]. Bayesian network model considers the maximum likelihood of the observed data given a structure. It is a model associated with statistical probability that does not infer the real *causal* relationships. Other than Bayesian network model, Gardner et al. [5] proposed a linear dynamic network model to

infer a gene network from steady state measurement. In addition to above models, several other probabilistic models have been proposed to learn gene networks with multiple data types [7, 15].

In fact none of the earlier research on learning gene networks considers the *causal* relationship between genes. A few researchers in [12, 13, 21] have suggested methods (for example, the IC algorithm) to learn the causal relationships between variables with steady state data, but not in biological domain. The assumption of causal theory is that the distribution of the dataset is faithful[1]. However, the under Boolean gene network model faithful function (we will define formally later) does not implies faithful distribution. Therefore learning gene causal connections with IC algorithms does not yield good result when the distribution of the dataset is not faithful.

In this paper, we present a new algorithm – modified IC (mIC) algorithm for learning causal relations between genes using additional knowledge of topological ordering[2]. We implement the algorithms using entropy to test conditional independence of the genes, and evaluate the causal learning approaches using simulation with the notions of precision and recall. We show that the precision and recall rates of the estimated gene network are significantly improved when using our mIC algorithm than original IC algorithm. In the end, we apply mIC algorithm to gene expression profile for a melanoma data with partial ordering information to learn gene regulatory network. The result shows that the important causal relationships associated with WNT5A gene are identified using mIC algorithm, and those causal connections have been confirmed in the literatures.

2 Learning Gene Causal Relationship

In this section we give a brief background on the difference between simple predictive relationships and causal relationships, and the basic intuition behind learning causal relationship from steady-state data. We also have an introduction to our simulation methodology.

2.1 Learning causal relationship with steady state data

To understand the difference between the simple and causal relationships between variables consider the propositions *rain* and $falling\_barometer$ from an example in [12]. When one observes that they are either both true or both false one concludes that they are related. One would then write $rain = falling\_barometer$. But neither *rain* causes $falling\_barometer$ nor vice-versa. Thus if one wanted *rain* to be true, one could not achieve it by somehow forcing $falling\_barometer$ to be true. This would have been possible if $falling\_barometer$ caused *rain*. We say that the relationship between *rain* and $falling\_barometer$ is correlation, but not cause. In the context of genes and proteins if one would like to turn on a gene,

[1] A probability distribution P is a faithful/stable distribution if there exist a directed acyclic graph (DAG) D such that the conditional independence relationship in P is also shown in the D, and vice versa.

[2] A topological ordering is an ordering among vertices of a DAG such that all edges are from vertices labeled with a smaller number to vertices labeled with a larger number. Knowledge about topological ordering between genes can be obtained if partial information about the pathways in which the genes (or their products) are involved is known; and also from existing knowledge about homologous genes in other organisms.

which cannot be achieved directly, through other genes one would need to know the causal connection between the genes. Thus knowing the causal relationship is very important.

The question then is how to obtain (learn or infer) causal relationship between genes. In wet-labs this can be done by knocking down the possible subsets of genes of a given set and studying its impact on the other genes in the set. This is of course not easy to obtain when the number of genes in the set is more than a handful. An alternative approach is to use time series gene expression data. Unfortunately such data can only be obtained for cells of particular organisms such as yeast. For human tissues high-throughput gene expression data is only available in the steady state observation. Thus the question that begs is how to infer causal relationship between genes from steady state data.

For long it was thought that one can only infer correlations and other statistical measures such as conditional independence from steady state data and there is no way to infer causal relationship from such data. In recent years some researchers have challenged this view and have suggested methods, while not specially for the gene expression data, to infer causal information. The idea is generalized by Pearl in [12] and Spirtes et al in [21], and an Inductive Causation (IC) algorithm is presented where causal relation between variables is learned or inferred by first analyzing independence and dependence between variables and then constructing minimal and stable causal influence graphs that satisfy the independence and dependence information.

The idea behind the inference of causality from steady-state data is based on the principle of finding the simplest explanation of observed phenomena [12]. The causal relationship between a set of genes can be expressed using a causal model which consists of a causal structure (a directed acyclic graph, or a DAG), and parameters that define the value of one node in terms of the value of its parents (in the DAG). The causal theory has an assumption on the distribution called **stability** or **faithfulness**[3] [12]. The assumption is that all the independencies in distribution P are stable, that is P is entailed by a causal structure of a causal model regardless of the parameter. However in microarray dataset, the distribution might not be faithful. Hence the performance of IC algorithm is not good (w.r.t precision and recall) for inferring causal relationship in this case. We propose an modified IC (mIC) algorithm that uses entropy to test conditional independence and combines the steady state data with prior knowledge of gene topological ordering to jointly learn the causal relationship between genes.

2.2 Modelling and simulation of a causal Boolean network

In order to evaluate the performance of the causal algorithms, we perform sets of simulations. We apply Boolean network model, originally introduced by Kauffman [8, 9], for modeling gene regulatory networks. Although Boolean network cannot model quantitative concepts, it provides useful insights in network dynamics [1]. There are two main objectives in modeling and simulation of data-driven Boolean network for the genetic regulatory systems. First, we need to infer the model structure and parameters (rules) from observations such as gene expression profiles. Second, we can explore the dynamic behaviors of the system driven by the inferred rules through simulation.

[3] A DAG G and a distribution P are *faithful* to each other if they exhibit the same set of independencies. A distribution P is said to be faithful if it is faithful to some DAG.

In the simulation, we construct a Boolean network model as a directed acyclic graph (DAG), and obtain the steady-state observations. The model contains n nodes with binary values. The state space has a total of 2^n states. Theoretically there are 2^{2^k} possible functions for a Boolean network, where k is the number of predictors. Among the 2^{2^k} functions, many do not actually reflect the influence of predictors. For example, assume that a gene g_i has two causal parents g_1 and g_2, and a Boolean function f determines the state of g_i at next time step with $g_i = f(g_1, g_2) = (g_1 \wedge g_2) \vee (g_1 \wedge \neg g_2)$. The function f is one of the 2^{2^2} functions, but can be simplified as $g_i = f(g_1, g_2) = g_1$. In this case, function f does not reflect the causal influence of one of its causal parents g_2. Therefore, we define the concepts of *influence* and *proper Boolean function* and only use such functions in our simulation.

Definition 1. *(Influence)*: Let $z = f(x_1, \ldots, x_n)$ be a Boolean function. We say x_i has an influence on z in the function f if there exists two assignment vectors for $x_1, \ldots, x_n$ that only differ on the assignment to x_i, such that the values of f on those two assignments differ.

Definition 2. *(Proper function)*: We say $z = f(x_1, \ldots, x_n)$ is a proper function if for $i = 1 \ldots n$, x_i has an influence on z in the function f.

We did sets of experiments to show that under non-uniform distribution the proper function is faithful function[4], which entails original causal structure.

The simulation process in this study can be summarized as follows:

- Step 1: Generate M Boolean networks with up to three input causal parents for each node in topological ordering.
- Step 2: For each Boolean network connection, generate random proper Boolean functions for each node.
- Step 3: Assign random probabilities for the root gene (gene with no causal parents).
- Step 4: Given one configuration (fixed connection and functions), run the deterministic Boolean network starting from all possible initial states and get the probability distribution of all possible states.
- Step 5: Collect two hundred data points sampled from the probability distribution.
- Step 6: Repeat Step 3 and Step 5 for all M networks with probability distribution and save the configuration file and the data file.

2.3 Entropy and Mutual Information

Given a probability distribution of a dataset, one needs to compute the conditional independence among genes to find the causal information. Shannon [16] developed the concept of entropy to measure the uncertainty of the discrete random variables. In this paper we calculate entropy H and mutual information I to obtain uncertainty coefficient U to test conditional independence between genes. The uncertainty coefficient U is range from 0 to 1 and defined as follows:

$$U(X|Y) = I(X, Y)/H(X); \tag{1}$$

where $H(X) = -\sum_x p(x) \log p(x)$; $H(X, Y) = -\sum_x \sum_y p(x, y) \log p(x, y)$;
and $I(X, Y) = H(X) + H(Y) - H(X, Y)$ [25].

[4] There exists some special case that under certain non-uniform distribution, proper function might not be faithful function. But those cases are rare and with random generator in the simulation, we may ignore it.

3 Algorithms and Criterion for Inferring a Gene Causal Network

3.1 Modified IC (mIC) algorithm

The IC algorithm [12] examines pairwise conditional independencies between variables to determine the v-structures [5] first, and then applies rules to determine the rest of the network structures. The mIC algorithm is based on the IC algorithm [12], but it incorporates the topological ordering information in the learning step to infer the gene causal relationship from steady state data. It takes as input a probability distribution P generated by a DAG with some gene topological ordering information, and outputs a partially directed DAG. The mIC algorithm is described as follows:

- Step 1: For each pair of gene g_i and g_j in a dataset, test pairwise conditional independence. If they are dependent, search for a set $S_{ij} = \{g_k \mid g_i \text{ and } g_j \text{ are independent given } g_k, \text{ with } i < k < j \text{ or } j < k < i\}$ Construct an undirected graph G such that g_i and g_j are connected with an edge if and only if they are pairwise dependent and no S_{ij} can be found;
- Step 2: For each pair of nonadjacent genes g_i and g_j with common neighbor g_k, if $g_k \notin S_{ij}$, and $k > i, k > j$, add arrowheads pointing at g_k, such as $g_i \rightarrow g_k \leftarrow g_j$;
- Step 3: Orientate the undirected edges without creating new cycles and v-structures.

3.2 Comparing initial and obtained networks - new definitions for precision and recall

For evaluating the learning results, we define the new notions of precision and recall. In comparing the initial and obtained networks one immediate challenge that we faced is in defining recall and precision for the case where the inferred graph may have both directed and undirected edges. (Note that the original graph has only directed edges.) Intuitively, an undirected edge $A - B$ means that we cannot distinguish the directionality between A and B with given dataset.
To deal with this we define the following six categories: FN (false negatives), TP (true positives), PTP (partial true positives), PFN (partial false negatives), TN (true negatives), FP (false positives), PTN (partial true negatives), and PFP (partial false positives) as follows:

FN = $\{X \rightarrow Y \mid X \rightarrow Y$ is in the original graph and neither $X \rightarrow Y$ nor $X - Y$ is in the obtained graph$\}$
TP = $\{X \rightarrow Y \mid X \rightarrow Y$ is in the original graph and also in the obtained graph$\}$
TN = $\{X \rightarrow Y \mid X \rightarrow Y$ is not in the original graph and neither $X \rightarrow Y$ nor $X - Y$ is in the obtained graph$\}$
FP = $\{X \rightarrow Y \mid X \rightarrow Y$ is not in the original graph and $X \rightarrow Y$ is in the obtained graph$\}$
PFN = PTP = $\{X \rightarrow Y \mid X \rightarrow Y$ is in the original graph and $X - Y$ is in the obtained graph$\}$
PTN = PFP = $\{X \rightarrow Y \mid X \rightarrow Y$ is not in the original graph and $X - Y$ is in the obtained graph$\}$

Now we can define the values AFP (aggregate number of false positive), ATN (aggregate number of true negatives), AFN (aggregate number of false negatives) and ATP (aggregate number of true positives) in terms of above six categories.

[5] A v-structure is of the form $a \rightarrow x \leftarrow b$ such that two converging arrows that the tails of a and b are not connected by an arrow.

$AFN = |FN| + |PFN|/2; \quad ATP = |TP| + |PTP|/2;$
$AFP = |FP| + |PFP|/2; \quad ATN = |TN| + |PTN|/2.$
where $|X|$ is the cardinality of set X. Using the above we can now define Recall and Precision as follows:

$$Recall = \frac{ATP}{(AFN + ATP)}; \quad Precision = \frac{ATP}{(ATP + AFP)}$$

3.3 Precision and Recall with Observational Equivalence

The output of IC algorithm is a pattern, a partially directed DAG, which is a set of DAGs that have equivalence structures. Every edge in the original network is directed, while the edges in obtained graph may be directed or undirected. There might be a case that a directed edge in original graph has a corresponding undirected edge in obtained graph. Therefore with the view of observational equivalence (OE), we should not have penalties for such edges. Here we define the new notions of precision and recall with considering observational equivalence. We transform both original graph and obtained graph into their own observational equivalent classes, called original class and obtained class, using the definition of observational equivalence [12]. Then define the six categories as follows:

FN = $\{(X,Y) \mid X \to Y$ or $X - Y$ is in the original class and neither $X \to Y$ nor $X - Y$ is in the obtained class$\}$
TP = $\{(X,Y) \mid X \to Y$ is in the original class and also in the obtained class or $X - Y$ is in the original class and also in the obtained class$\}$
TN = $\{(X,Y) \mid$ neither $X \to Y$ nor $X - Y$ is in the original class and neither $X \to Y$ nor $X - Y$ is in the obtained class$\}$
FP = $\{(X,Y) \mid$ neither $X \to Y$ Y nor $X - Y$ is in the original class and $X \to Y$ is in the obtained class$\}$
PFN = PTP = $\{(X,Y) \mid X \to Y$ is in the original class and $X - Y$ is in the obtained class or $X - Y$ is in the original class and $X \to Y$ is in the obtained class$\}$
PTN = PFP = $\{(X,Y) \mid$ neither $X \to Y$ nor $X - Y$ is in the original class and $X - Y$ is in the obtained class$\}$

The concepts of the AFN, ATP, ATN, AFP, precision and recall are the same as the ones we defined previous section.

3.4 Comparing the networks based on their transitive closure

There are many ways for comparing the initial and obtained graphs. We discussed the way for comparing two networks directly, with and without observational equivalence. Transitive closure (TC) is another way for graph comparison. Suppose the initial network that we have is $A \to B \to C$ and we obtain the network with the only edge $A \to C$. When comparing the obtained network with the initial network we may not treat $A \to C$ just as a false positive. In fact this obtained network is better than the network that has no edges. To be able to make this conclusion we consider the TC of $\to$ in the initial network and a similar notion in the obtained network. In the obtained network our definition of TC is based on defining two relations: $cc(x,y)$ and $pcc(x,y)$. Intuitively, $cc(x,y)$, denoting x causally contributes to y, is true if there is a directed or an undirected edge from x to y; and $pcc(x,y)$, denoting x possibly causally contributes to y, is true if there is a path from x to y consisting of properly directed edges and undirected edges such that $pcc(x,y) := cc(x,y) \mid pcc(x,z) \wedge pcc(z,y)$

3.5 Steps of learning gene causal relationships

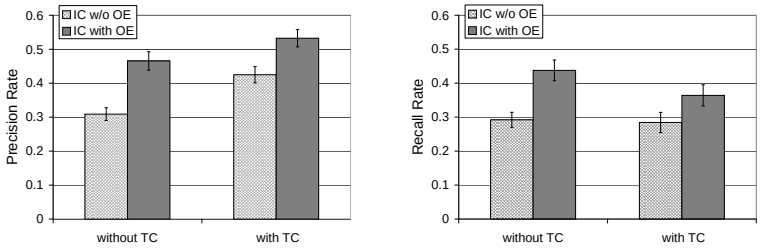
The steps for learning gene causal relationships are as follows:
Step 1: Obtain the probability distribution, data sampling and the topological order of the genes;
Step 2: Apply algorithms such as IC or mIC to find causal relations;
Step 3: Compare the original and obtained networks based on the two notions of precision and recall;
Step 4: Repeat step 1-3 for every random network.

4 Experiments, Results and Discussion

We did two sets of experiments for learning gene causal relationships using the IC algorithm and mIC algorithm. Each experiment contains 100 different randomly generated gene networks (DAGs), each of which contains 10 genes, with topological ordering connected by Boolean proper functions. The distribution of the network is generated based on the probability of the root genes, and Monte Carlo sampling is used to generate 200 samples in a dataset for each network based on the probability distribution. We use the uncertainty coefficient (U) to test the conditional independence in step 1 of the algorithm. We choose the of $U = 0.3$ for pairwise and $U = 0.2$ for triplewise conditional independent test. The threshold cut-off values are based on heuristics that we elaborate in [25].

4.1 Learning with IC algorithm

The first experiment is to use IC algorithm on the learning gene causal relations with steady state data without topological ordering information. The method is applied to derive an obtained graph for every network and then the obtained graphs are compared with their corresponding initial ones. The results with statistical confidence of 95% as the error bar marked are shown in figure 1.



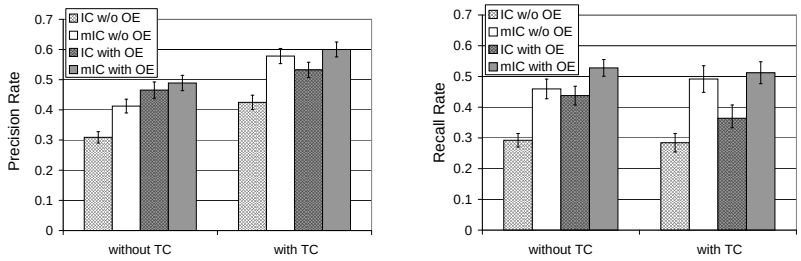
(a) Precision (b) Recall

Fig. 1. Precision and Recall for learning by IC algorithm

Figure 1(a) shows the precisions of the simulation with IC using two notions: with and without OE and TC. The result shows that the precision rate for inferring causal relations in simulation is 0.3 without OE or TC, 0.45 with OE, around 0.4 with TC, and around 0.5 with OE and TC. Figure 1(b) shows that the recall rate is below 0.3 without OE, and around 0.4 with OE.
From the figure we can see that both precision and recall are significantly improved by using the notion of observational equivalence. However the recall rate is still around 0.4. From the above simulation results we can see that IC algorithm is not quite good for learning gene regulatory network using only steady state data.

4.2 Learning with Topological Ordering (mIC)

Since using IC algorithm for learning gene causal network from single type of dataset - steady state data did not show a good result, our hypothesis is that a better way is to use additional knowledge such as gene topological ordering. The second simulation we did is jointly learning the gene regulatory network using mIC algorithm combining steady state observation and the background knowledge of gene topological orders. We then compare the results with the ones learned by IC algorithm as shown in Figure 2.



(a) Precision (b) Recall

Fig. 2. Precision and Recall for learning by mIC with Ordering

Figure 2(a) shows that with statistical confidence of 95% as the error bar marked, the precision rate is significantly improved with mIC algorithm, and the precision rate is around 0.6 with TC and OE. Figure 2(b) are promising since the recall rates of learning with mIC algorithm are significantly improved from less than 0.3 with IC algorithm to greater than 0.45 by applying mIC algorithm, improved more than 50%, both with and without considering TC or OE. If considering the observational equivalence, the recall rate of learning with mIC algorithm has been significantly improved to above 0.5 with or without TC.

5 Applying mIC algorithm on Melanoma Dataset

We finally applied mIC algorithm to a gene expression profile used in the study of melanoma [2]. 31 malignant melanoma samples were quantized to ternary format such that the expression level of each gene is assigned to -1(down-regulated), 0(unchanged) or 1 (up-regulated). The 10 genes involved in this study are chosen from 587 genes from the melanoma dataset that have been studied to cross predict each other in a multivariate setting [11]: *pirin*, WNT5A, MART-1, S100, RET-1, MMP-3, PHO-C, synuclein, HADHB and STC2.

In previous expression profiling study, WNT5A has been identified as a gene of interest involved in melanoma[2], and expression level of WNT5A is closely related with metastatic status of melanoma [23]. It was shown that the abundance of messenger RNA for WNT5A can be significantly distinguished between cells with high metastatic competence versus those with low metastatic competence [2]. Later, it was also proved experimentally that increasing the level of WNT5A protein can directly change the cell metastatic competence [23]. It has been also suggested that controlling the influence of WNT5A in the regulation can reduce the chance of melanoma metastasizing [3].

In this study of set of 10-gene network, we have a partial biological prior knowledge that MMP-3 is expected to be at the end of the pathway. We applied mIC

algorithm using entropy to test conditional independence among those 10 genes with the above prior knowledge to infer the *causal* regulatory network. The learning results are shown in figure 3.

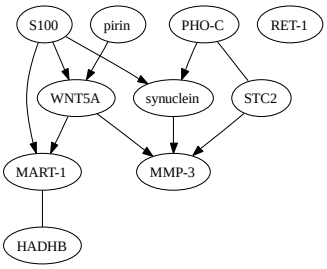


Fig. 3. Learning Melanoma Dataset with prior knowledge that MMP-3 is at the end of gene regulatory network

Figure 3 shows that *pirin* causatively influences WNT5A. This result is consistent with the literature[3] that in order to maintain the level of WNT5A, we need to directly control WNT5A or control WNT5A through *pirin*. The result also shows the causal connection between WNT5A and MART-1 such that WNT5A directly causes MART-1, which has been verified in the literature [20] that WNT5A may actually directly influence the regulation and suppression of MART-1 expression. In Figure 3, there are some causal connects that have not been verified by the scientist yet. However, they are unlikely to be obtained by random chance (see supplement materials http://www.asu.edu/~zhang24/AIME05). mIC algorithm bring up a systematic way to predict the *causal* connections among genes using steady state data with some prior biological knowledge. It could be applied as a guidance for the biologist to verify the causal connections in future experiments.

6 Conclusion

In this paper we presented a modified IC algorithm with entropy that can learn steady state data with gene topological ordering information. We did simulation based on Boolean network to evaluate the performance of the causal algorithms. In the process we developed ways to compare initial networks with obtained networks. From our simulation based evaluation we conclude that (i) IC algorithm does not work well for learning gene regulatory networks from steady state data alone, (ii) a better way for learning the gene causal relationship from steady state data is to use additional knowledge such as gene topological ordering, (iii) the precision and recall rates for mIC algorithm is significantly improved compared with IC algorithm with statistical confidence of 95%. For randomly generated networks, the mIC algorithms work well for joint learning the causal regulatory network by combining steady state data and gene topological ordering knowledge, with precision rate of greater than 60%, and recall rate greater than 50%. We then applied the algorithm to real biological microarray data Melanoma dataset. The result showed that some of the important causal relationships associated with WNT5A gene have been identified using mIC algorithm, and those causal connections have been verified in the literatures.

7 Acknowledgement

This work was supported by NSF grant number 0412000. The authors acknowledge the valuable comments of the anonymous reviewers of this paper.

References

1. Akutsu, T., et al. Identification of genetic networks from a small number of gene expression patterns under the Boolean network models. *PSB*,17-28, 1999.
2. Bittner, M. et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536-540, 2000.
3. Datta, A., Bittner, M. and Dougherty. E. External Control in Markovian Genetic Regulatory Networks. *Machine Learning*, Vol 52, 169-191, 2003.
4. De Jong, H. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of Computational Biology, 2002*. 9(1): p. 67–103.
5. Bernardo, T. et al Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. *Science* 2003.
6. Friedman, N., Linial, L., Nachman, I. and Pe'er D. Using Bayesian Networks to Analyze Expression Data. *RECOMB* 2000.
7. Hartemink, A. et al. Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models. *PSB*, 2002.
8. Kauffman, S.A. Requirements for Evolvability in Complex Systems: Orderly Dynamics and Frozen Components. *Physica D, 1990.* 42: p. 135-152.
9. Kauffman, S.A. The Origins of Order, Self-Organization and Selection in Evolution. *Oxford University Press*, 1993.
10. Kim, S., et al. Multivariate measurement of gene-expression relationships. *Genomics*, vol. 67, pp. 201–209, 2000
11. Kim S., et al. Can Markov Chain Models Mimic Biological Regulation? *Journal of Biological Systems, vol. 10, No. 4* (2002) 337-358.
12. Pearl, J. Causality : models, reasoning, and inference. 2000 *Cambridge, U.K. ; New York: Cambridge University Press.* xvi, 384 p.
13. Scheines, R., Glymour, C. and Meek, C. TETRAD II: Tools for Discovery. *Hillsdale, NJ: Lawrence Erlbaum Associates*, 1994.
14. Schulze, A. and Downward, J. Navigating Gene Expression Using Microarrays - A Technology Review. *Nature Cell Biology, 2002.* 3: p.190-195.
15. Segal, E., et al. From Promoter Sequence to Expression: A Probabilistic Framework. *RECOMB* 2002.
16. Shanon, C. A mathematical theory of communication *The Bell Systems Technical Journal*, 27, 1948.
17. Shmulevich, I. et al Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2): p. 261-74, 2002.
18. Smith, V., Jarvis, E. and Hartemink, A. Influence of Network Topology and Data Collection on Network Inference. *PSB* 2003.
19. Smolen, P. et al. Modeling transcriptional control in gene networks–methods, recent results, and future directions. *Bull Math Biol*, 62(2): p. 247-92, 2000.
20. Sosman, J., Weeraratna, A., Sondak, V. When will Melanoma Vaccines be proven effective? *Journal of Clinical Oncology*, vol. 22, No 3, 2004.
21. Spirtes, P., Glymour, C. and Scheines, R. Causation, Prediction, and Search. *New York, N.Y.: Springer-Verlag*. 2nd Edition, MIT, Press 1993
22. Spirtes, P. et al. Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data. *Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems & Technology*, 2000.
23. Weeraratna, A.T., Jiang, Y., et al Wnt5a signalling directly affects cell motility and invasion of metastatic melanoma. *Cancer Cell*, 1, 279-288, 2000.
24. Yoo, C. and G.F. Cooper Discovery of gene-regulation pathways using local causal search. *Proc AMIA Symp*, 2002: p. 914-8.
25. Supplement materials: http://www.public.asu.edu/~xzhang24/AIME05