

Evaluating Two Approaches to Extracting Gene Regulatory Networks: Bayesian Networks and Association Rule Mining

Zan Huang¹, Jiexun Li¹, Jie Xu¹, Ritu Pandey², Hsinchun Chen¹

Artificial Intelligence Lab¹

Department of Management Information Systems

The University of Arizona

Tucson, Arizona 85721, USA

{zhuang, jiexun, jxu, hchen} @eller.arizona.edu

Arizona Cancer Center,

University of Arizona,

Tucson, Arizona 85724, USA

ritu@email.arizona.edu

ABSTRACT

Advances in microarray technologies have enabled simultaneous measurement of expression levels of thousands of genes, creating new opportunities and challenges for gene expression data analysis. Several recent studies have proposed to extract gene regulatory relations from microarray data with a wide range of techniques. However, because of the dimensionality problem in microarray data, most existing studies have included only a small number of genes. There is also a lack of evaluation of the extracted networks. Both problems have limited the practical value of the gene regulatory network analysis. In this paper, we present two algorithms for large-scale gene regulatory network analysis: an information-theory-based Bayesian network algorithm and a modified association rule mining algorithm. We also present two types of evaluations of the resulting networks: a simulation-based evaluation and an empirical evaluation. Six simulated gene expression datasets based on three pre-defined regulatory network models and two real datasets (a *Saccharomyces cerevisiae* dataset and a *Homo sapiens* dataset) were used in the evaluation study. The simulation-based evaluation results indicated that the two techniques could extract 30% - 60% correct relations when relation direction was not considered. The empirical evaluation showed that the extracted networks generally failed to identify regulatory relations reported in the literature. However, more than 50% of the extracted relations reflected gene co-occurrence patterns in the literature, and a small set of relations appeared to domain scientists to be potentially correct and interesting.

Keywords

Gene regulatory network, Bayesian network, Association rule.

1. INTRODUCTION

Recent advances in microarray technologies have made possible large-scale gene expression analyses based on simultaneous measurements of thousands of genes. Many data mining techniques (e.g., clustering and classification) have been employed to uncover the biological functions of genes from microarray data. Recently, a reverse engineering approach has been used to extract gene regulatory networks in order to reveal the structure of the transcriptional gene regulation processes.

The general goal of gene regulatory network analysis is to extract pronounced regulatory relations (e.g., activation and inhibition) between genes by examining the global gene expression patterns. The resulting network of regulatory processes may help researchers form new hypotheses about the behavior of biological systems and assist with the design of further experiments. Many studies have proposed various network extraction approaches. However, a common problem in these studies is that they include only a relatively small number of genes. This is mainly because of the inherent dimensionality problem in microarray data, which usually contain an insufficient number of samples of a large number of genes. To enable this type of analysis to capture the complexity of the biological systems, scalable techniques need to be developed to extract regulatory networks that contain a large number of genes. Another problem in previous studies is the lack of empirical evaluation of the gene regulatory

networks generated using various approaches. It is unclear whether resulting networks provide valuable information about the behavior of biological systems.

In this paper, we present two scalable gene regulatory network extraction algorithms: an information-theory-based Bayesian network algorithm and a modified association rule mining algorithm. To assess the usefulness of the two techniques, we conducted a simulation-based evaluation and an empirical evaluation.

The remainder of the paper is organized as follows. Section 2 reviews the literature on existing approaches to extracting gene regulatory network and their limitations. Section 3 addresses the representation issues and presents the algorithmic details of the two proposed techniques. Section 4 presents the two types of evaluation studies. We conclude the paper and identify future directions in Section 5.

2. LITERATURE REVIEW

Gene regulatory network analysis is aimed at identifying regulatory relations among genes. In such analysis, gene interactions are viewed as signaling processes that form a complex feedback network. Information for the construction and maintenance of this signaling system is stored in the genome. The DNA sequence codes for the structure and molecular dynamics of RNA and proteins in turn determine biochemical recognition of the signaling processes. The regulatory molecules that control the expression of genes are themselves the products of other genes [18]. Thus, genes turn each other *on* and *off* within a proximal genetic network of transcriptional regulators [25]. This genetic network provides a partial picture of the complex biological processes. Many other factors (e.g., proteins and metabolites) also play important roles in this signaling system and are largely hidden from observation. However, the aggregated effects reflected in gene expression patterns still reveal valuable information about the underlying processes. To identify such gene regulatory processes, researchers represent genes as activators and inhibitors of each other and have applied various approaches.

A traditional approach to identifying gene regulatory relations is the “knockout” experiment. In such an experiment, the gene expression level of a particular gene is lowered while all other conditions are kept constant. The differences in gene expression levels of other genes are used to infer the underlying regulatory relations. Usually, this approach can reliably uncover regulatory relations among a small number of genes, but is difficult to scale up for regulatory networks consisting of hundreds of genes. This is because of the sheer number of possible combinations of experimental manipulations that are needed to reveal the complete regulatory network.

Recent development of microarray technologies has made it possible to measure expression levels of thousands of genes simultaneously. This has enabled researchers to use reverse engineering approaches to trace back to gene regulatory network structures.

2.1 Extracting Gene Regulatory Networks from Microarray Data

The earliest models proposed in the literature are *discrete models*. Examples are Boolean regulatory networks in which the gene expression levels are represented as 0 (not expressed) or 1 (expressed) [16, 18, 35]. These models are based on the notion that biological networks can be represented by binary, synchronously updating switching networks. However, since variables in real biological systems change continuously in time, many continuous models have been recently proposed.

Wessels et al. categorized existing *continuous models* of gene regulatory networks into three types: pair-wise comparison, rough networks, and complex networks [34].

Pair-wise comparison methods construct regulatory networks based on relations between each pair of genes. Two examples of such methods are the Correlation Metric Construction (CMC) [2] and Activation/Inhibition Networks [6]. Such approaches may not reveal complete regulatory structures because they ignore the fact that the expression level of one gene is governed by the combined actions of multiple other genes.

Rough network models represent the output gene expression level as a weighted sum of its input gene expression levels. Wessels et al. used a generalized difference equation, as shown in (1), to characterize this class of models [34].

$$X_i[t+1] = R_i \cdot g\left(\sum_{j=1}^J W_{ij} \cdot X_j[t] + \sum_{k=1}^K V_{ik} \cdot U_k + B_i\right) - \lambda_i \cdot X_i[t] \quad (1)$$

where g refers to the regulation-expression (activation) function, $X_i[t]$ represents the expression of gene i at time instance t , R_i represents the rate constant of gene i , W_{ij} represents strength of control of gene j on gene i , $U_k[t]$ represents k -th external input at time instance t , V_{ik} represents influence of the k -th external input on gene i , B_i represents the base expression level of gene i , and λ_i represents the degradation constant of gene i .

A variety of approaches have been proposed to infer the parameters of such differential or difference equations, including Recurrent Hopfield networks [20], linear programming [6], simulated annealing [19], genetic algorithms [31], and linear regression [10, 24, 32]. When a large number of genes are incorporated into the model and a relatively small number of samples are available, a dimensionality problem inevitably arises due to the large degrees of freedom.

A *complex network* not only models interactions between genes based on measured expression levels but also explicitly models their intermediate products, such as proteins and metabolites [6]. Consequently, these models also require expression levels of intermediate products and typically contain a large number of parameters, making them more difficult to construct than the pair-wise comparison and rough network models.

Murphy and Mian [21] and Friedman et al. [11] have recently proposed to use Bayesian network models to represent and extract gene regulatory networks from microarray data. Bayesian network models can include most previously proposed discrete and continuous models as special cases and allow stochastic mechanisms and hidden variables.

2.2 Limitations of Existing Approaches

Existing gene regulatory network studies have two limitations. First, most approaches proposed in previous studies could model relations among only

a small set of genes. This is mainly because of the dimensionality problem of the microarray data, i.e., the number of genes measured is much larger than the number of available samples. For example, [19] used 8 samples of 27 genes, [31] used 28 samples of 65 genes, and [24] used 18 samples of 45 genes and 14 samples of 113 genes. However, gene interactions form a complex signaling system in which the behavior of any single gene may be affected by many other genes. A study with a limited number of genes reveals only a subset or a local structure of the complete regulatory network. It is desirable to extract the regulatory network including a large number of genes. Only in the recent study by Friedman et al. [11], in which the goal is to identify pronounced gene regulatory relations rather than to specify the exact regulatory network structure, a relatively large number of genes (800) were included in the analysis.

Second, little research has been done to evaluate the various approaches in terms of their abilities to extract known gene relations from microarray data and the abilities to suggest hypotheses about unknown relations.

Both issues have limited the empirical value of gene regulatory network analysis. To address these problems, we present in this paper two techniques that can extract large-scale gene regulatory networks from microarray data: an information-theory-based Bayesian network learning algorithm and a modified association rule mining algorithm. Both approaches were evaluated to assess their empirical value.

3. ANALYTICAL TECHNIQUES

In this section, we first introduce our representation of the gene expression data and then present the two network extraction techniques proposed.

3.1 Gene Expression Representation

We represent the expression level of a gene as a random variable. In microarray data analysis, the expression levels of genes in a test sample (gene chip) usually are compared with a reference sample. A microarray dataset is represented as $O = \{O_1, \dots, O_I\}$, where I is the number of samples (chips), and $O_i = (\langle e_{i1t}, \dots, e_{iNt} \rangle, \langle e_{i1r}, \dots, e_{iNr} \rangle)$, where N is the number of genes, e_{int} is the test expression level of

gene n in sample i and e_{inr} is the reference expression level of gene n in sample i . The gene expression data used for analysis purposes are represented as $X = \{X_I, \dots, X_I\}$, where $X_i = \langle x_{i1}, \dots, x_{iN} \rangle$, and $x_{in} = \log(e_{in}/e_{inr})$.

To keep the network extraction algorithms simple, we represented x_{in} as discrete random variables. Based on discussions with domain scientists, we performed ternary discretization [11, 29] and transformed x_{in} as 1, 0 or -1 as in (2).

$$x_m = \begin{cases} +1, & \text{if } x \geq +\Theta, \text{ over - expressed,} \\ -1, & \text{if } x \leq -\Theta, \text{ under - expressed,} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

where Θ is a predefined threshold.

Expression levels of genes vary from sample to sample due to different experimental treatments or other external factors (e.g., time period). It is these variations that enable us to extract the underlying regulatory network structure that captures the gene interaction patterns.

In the following sections, we present the algorithmic details of the two techniques that can be used to extract large-scale regulatory networks from microarray data.

3.2 Bayesian Networks

3.2.1 Background

Bayesian networks are a special case of a more general class of models called graphical models, in which vertices represent random variables and the absence of an edge between two vertices represents conditional independence between them. Consider a finite set $V = \{V_1, \dots, V_n\}$ of random variables. A Bayesian network representation contains two components: a directed acyclic graph (DAG) G whose vertices correspond to the random variables, and the conditional probability distribution for each variable, given its dependent variables (parents) in G . The graph G specifies the dependency relationships among variables and encodes the *Markov Assumption*: Each variable V_i is independent of its non-descendants given its parents in G .

An example can illustrate the basic idea [11]. Given a Bayesian network specified in Figure 1 for 5

genes: A, B, C, D , and E , this structure specifies the parents for gene B, D and C : $Pa(B) = \{A, E\}$, $Pa(D) = \{A\}$, $Pa(C) = \{A, B, E\}$, where $Pa(X)$ represents the parent vertex set for vertex V . It also implies several conditional independency relationships: $I(A; E)$, $I(B; D | A, E)$, $I(C; A, D, E | B)$, $I(D; B, C, E | A)$ and $I(E; A, D)$, where $I(X; Y | Z)$ represents that X and Y are conditionally independent, given Z . In the gene expression analysis context, it can be interpreted that when gene Z is at a fixed expression level, expression level of gene X does not give any information on the expression level of gene Y and vice versa.

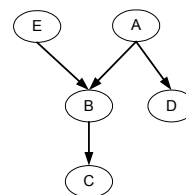


Figure 1. A simple example of Bayesian network

Once G is specified for a gene expression dataset, we can interpret a directional edge from X to Y in G as a statement that X is the “cause” of Y , or expression level of V has an effect on expression level of Y .

3.2.2 Implementation

There are two general approaches to learning Bayesian network from data: the search and scoring methods and the dependency analysis methods [7]. In the first approach, the learning problem is viewed as searching for a structure that best fits the data. Different scoring methods have been applied to determine the fit between the network structure and the data, including the Bayesian scoring method, entropy based method, minimum description length method, etc. Because such search and scoring methods are NP-hard [8], many search heuristics have been adopted. The Bayesian networks methods proposed in the literature of gene regulatory network learning typically used heuristic-based search and scoring methods [11, 21].

The dependency analysis approach tries to discover from data the dependencies among variables and then to use these dependencies to infer the network structure [27, 28, 33]. Different forms of

Conditional Independency (CI) tests have been used to measure the dependency relationships.

The dependency analysis approach is generally more efficient than the search and scoring approach for sparse networks (the number of edges in the graph is relatively small) [7]. In this study, in order to perform large-scale gene regulatory network analysis, we chose to implement the dependency analysis approach for Bayesian network learning. We used the information-theory-based learning algorithm proposed by Cheng et al. [7].

The most significant advantage of this algorithm is that, unlike all other practicable dependency analysis based algorithms, this algorithm can avoid exponential complexity of CI tests. The algorithm is of $O(N^4)$ of CI tests when learning Bayesian networks with completely unknown network structures.

After the network structure has been extracted, we use a simple heuristic to determine the biological meanings of the edges. For an edge that points from gene X to gene Y , we calculate the activation and inhibition measures as defined in (3) and (4) to determine the relation type. Edges with activation (inhibition) measure larger than a specified threshold are labeled as activation (inhibition) relations. The resulting Bayesian network model may contain four types of gene relations: activation, inhibition, causal (when edge direction is determined but cannot be labeled as activation or inhibition) and dependency (when the edge direction cannot be determined).

$$\text{Activation Measure} = \frac{P(Y=1, X=1) + P(Y=-1, X=-1)}{P(X=1 \text{ or } X=-1)} \quad (3)$$

$$\text{Inhibition Measure} = \frac{P(Y=-1, X=1) + P(Y=1, X=-1)}{P(X=1 \text{ or } X=-1)} \quad (4)$$

3.3 Association Rule Mining

3.3.1 Background

Association rule mining was originally proposed for market basket analysis to study consumer-purchasing patterns in retail stores [1]. An association rule is a relationship of the form $A \Rightarrow B$, where A is the antecedent item set and B is the consequent item set. The rule $A \Rightarrow B$ holds in the transaction set D with *confidence* c if $c\%$ of

transactions in D that contain A also contain B . The rule $A \Rightarrow B$ has *support* s if $s\%$ of transactions in D contain both A and B . the goal of association rule mining is to find all the rules that have support and confidence greater than user-specified thresholds.

Association rule mining can be used to extract relationships among multiple genes based on conditional dependency. Association rules not only capture the correlation between genes, but also provide the direction of relationships. Some initial work of using association rules in molecular classification and gene regulatory relation extraction has been reported in [4, 5, 17, 22]. Since microarray data has a very large number of variables (genes), association rule mining method usually generates too many rules for biomedical researchers to explore and analyze. Several rule evaluation operators are reported in [29] in order to resolve this problem of “rule explosion”.

3.3.2 Implementation of a Modified Association Rule Mining Algorithm

Classic association rule mining algorithms can only handle Boolean data, which requires discretization of continuous gene expression values in data preprocessing phase. Becquet et al. applied a binary discretization to gene expression data and set values less than or equal to a threshold to 0 and all values greater than the threshold to 1 [4]. This approach may lead to information loss because binary representation fails to capture the overall gene expression distribution. Alternatively, a ternary discretization approach was used in [5, 29] to convert each gene expression to one of three levels, i.e. under-expressed, normal, or over-expressed. For each gene, three Boolean variables were used to represent the three expression levels respectively. This approach captures gene expression distribution more completely than binary discretization does. However, the increased number of variables to be processed may lead to larger numbers of potential association rules, thereby increasing the difficulty in interpreting and analyzing the rules extracted. It may also reduce the computational efficiency of the mining process.

Without assigning a Boolean variable to each expression level of each gene like in [29], we

simply use one single variable to represent each gene, i.e. for gene X , its expression level can be represented as under-expressed ($X = -1$), normal ($X = 0$) or over-expressed ($X = 1$), respectively. Based on such a ternary discretization, we modified the classic association rule mining algorithm to extract gene activation and inhibition relations from microarray data.

According to biological interpretations, the activation and inhibition relations between two genes, X and Y , can be respectively denoted as follows:

(a) X activates Y ($X \xrightarrow{+} Y$): IF X is over-expressed ($X = 1$), THEN Y is over-expressed ($Y = 1$); IF X is under-expressed ($X = -1$), THEN Y is under-expressed ($Y = -1$)

(b) X inhibits Y ($X \xrightarrow{-} Y$): IF X is over-expressed ($X = 1$), THEN Y is under-expressed ($Y = -1$); IF X is under-expressed ($X = -1$), THEN Y is over-expressed ($Y = 1$).

Based on these notions, support and confidence of association rules are redefined in our modified association rule mining algorithm. Given a gene expression matrix D , the *support* of a single gene X is defined in (5):

$$\text{support}(X) = \frac{\|X=1 \text{ or } X=-1\|}{|D|} \quad (5)$$

where $\|X=1 \text{ or } X=-1\|$ is the number of samples in D in which $X = 1$ or -1 ; $|D|$ is the number of samples in D .

For an itemset of two genes, X and Y , two new measures called *support*⁺ and *support*⁻ are defined in (6) and (7), respectively:

$$\text{support}^+(XY) = \frac{\|X=1 \text{ and } Y=1\| + \|X=-1 \text{ and } Y=-1\|}{|D|} \quad (6)$$

$$\text{support}^-(XY) = \frac{\|X=1 \text{ and } Y=-1\| + \|X=-1 \text{ and } Y=1\|}{|D|} \quad (7)$$

The confidence⁺ of an activation rule and the confidence⁻ of an inhibition are defined in (8) and (9):

$$\text{confidence}^+(X \xrightarrow{+} Y) = \frac{\text{support}^+(XY)}{\text{support}(X)} \quad (8)$$

$$\text{confidence}^-(X \xrightarrow{-} Y) = \frac{\text{support}^-(XY)}{\text{support}(X)} \quad (9)$$

The challenge of this modified association rule mining approach is to generate all rules (both activation and inhibition) that have *support* and *confidence* equal to or greater than a user-specified minimum support (called *minsup*) and minimum confidence (called *minconf*) respectively.

We use an example to illustrate the basic idea behind our method, given transformed expression levels of gene X and Y in six samples (Table 1), *minsup* = 20%, and *minconf* = 60%. Based on our definition of *support* and *confidence*, for rule “ $X \xrightarrow{+} Y$ ”, we have $\text{support}(X) = 100\%$, $\text{support}^+(XY) = 66.67\% > 20\%$, and $\text{confidence}^+(X \xrightarrow{+} Y) = 66.67\% > 60\%$. $X \xrightarrow{+} Y$ is thus extracted as an activation rule.

Table 1. Transformed expression levels of gene X and Y

Gene	1	2	3	4	5	6
X	-1	-1	-1	1	-1	1
Y	-1	-1	-1	0	1	1

4. EVALUATION STUDY

One major goal of this research is to assess the empirical value of employing microarray data to extract gene regulatory networks. In this section, we present two types of evaluation studies: a simulation-based evaluation and an empirical evaluation. The purpose of the simulation-based evaluation is to determine the ability of our techniques to uncover the underlying regulatory network structure. The empirical evaluation, on the other hand, provides direct evidence on the potential of the techniques through identifying regulatory relations that are either known or helpful for developing new hypotheses.

4.1 Simulation-based Evaluation

Simulation approach is frequently used to validate network-extraction techniques (e.g., [7]). With this approach, a network model is usually pre-defined, based either on existing knowledge in an application domain or on a randomly generated structure. Values of the variables in the model are generated based on this network model. Using these simulated data as inputs, network models can be constructed using different network-extraction techniques. The inferential power of the techniques,

i.e. the ability to uncover underlying structures, can then be assessed by comparing the extracted model with the pre-defined model.

The simulation approach has been recently employed to evaluate different approaches to extracting gene regulatory networks. Wessels et al. conducted a comprehensive evaluation of various network extraction techniques using simulated gene expression data [34]. Their study was focused on time-series models with a small number of genes (15 genes). Their major conclusion was that most existing approaches had surprisingly low inferential power.

We specifically incorporated characteristics of gene expression data (e.g., the limited number of samples, the large number of genes, and noise and measurement errors) into the simulation process in order to evaluate the usefulness of the network-learning techniques in deriving gene regulatory networks from gene expression data. The advantage of using simulation data for evaluation is that the underlying network model is known. In contrast, when using real experimental data, large portion of the underlying regulatory network might be unknown, making it difficult to determine the quality of the extracted networks through comparison against the literature or knowledge of domain experts.

4.1.1 Model and Data Simulation

We used a Bayesian network model to represent a gene regulatory network. Thus regulatory relations among genes were described by a directional acyclic graph, in which directional edges represented the regulatory effects of parent (upstream) genes on child (downstream) genes. Expression level of a gene was assumed to be dependent on expression levels of its parent genes. The Bayesian network model captures the stochastic nature of gene regulatory processes and does not require time-series data for learning purpose, which makes it applicable to most microarray datasets.

When generating gene regulatory models, we considered the following key factors that might have significant impacts on network extraction performance: (1) number of genes; (2) number of source genes (corresponding to the root vertices in

the network model or the genes that are directly affected by the experiment treatments); (3) density of the network model (we used a fan-out factor, defined by the maximum number of decedents or downstream gene of one gene, to control the network density); (4) gene expression level (-1, 0 or 1); (5) gene regulatory relation types (two relation types were included: activation and inhibition); (6) noise level (a noise level represented the uncertainties in regulatory processes; it reflected many limitations of the gene expression data, such as measurement errors and effects of other factors that might have been overlooked).

To translate the gene regulatory networks into a Bayesian network model, we used the conditional probabilities described in (10) and (11) to specify the activation and inhibition relations (assuming that X is the only parent gene of Y).

$$X \xrightarrow{+} Y : P(Y = y | X = x) = \begin{cases} 1-2p, & \text{when } y = x \\ p, & \text{otherwise} \end{cases} \quad (10)$$

$$X \xrightarrow{-} Y : P(Y = y | X = x) = \begin{cases} 1-2p, & \text{when } y = -x \\ p, & \text{otherwise} \end{cases} \quad (11)$$

where $x, y = -1, 0, \text{ or } 1$; $p \in [0, 1/3]$ is the noise level. When $p = 1/3$, the effect of noise dominates the gene expressions, resulting in equal probabilities for the three possible expression values.

We used formula (12) to derive the probabilities of the expression level of a gene (Z) given expression levels of two parent genes (X and Y). The formula can be easily extended to compute probabilities of expression levels of genes given expression levels of multiple parent genes.

$$P(Z = z | X = x, Y = y) = \sum_{x,y=z} P(Z = z | X = x) \cdot P(Z = z | Y = y) \quad (12)$$

where $x, y, z = -1, 0, \text{ or } 1$; $P(Z = z | X = x)$ and $P(Z = z | Y = y)$ are defined as in (10) and (11).

Based on the factors described above, three gene regulatory models were generated using a random procedure. The key factors of the three models are presented in Table 2. Because a gene regulatory network is generally believed to be sparse [3], we set fan-out factors to relatively small numbers. Based on the three models, we employed the

probabilistic logic sampling method proposed by Henroin [12] to generate samples of gene expression data. For each network model, two simulation datasets were generated with sample sizes of 40 and 80. These sample sizes were selected to reflect typical sample sizes of microarray datasets, such as those real datasets used in the empirical evaluation section.

Table 2. Pre-defined regulatory network models for simulation

Network Model	Number of Genes	Number of Source Genes	Fan-out Factor	Noise Level	Number of Relations
G1	10	2	3	0.2	12
G2	50	5	4	0.2	83
G3	200	5	4	0.2	392

4.1.2 Simulation-based Evaluation Results

Gene regulatory network models were generated based on the six simulation datasets using the Bayesian network and association rule techniques. To evaluate the inferred network models, we used an accuracy measure as described in (13).

Network accuracy =

$$\frac{\text{Number of relations within the } N \text{ strongest relations in the inferred model that match with relations in the true model}}{N} \quad (13)$$

where N is the number of relations in the true model.

Three versions of network accuracy measures were computed based on different relation match definitions, including an *exact match*, a *directional match* and a *non-directional match*. The non-directional match is the basic match definition. It only requires the two relations in comparison to involve the same genes. A directional match further requires that the directions of the two relations in comparison are identical. The exact match is the strongest relation match definition. It requires two relations to have identical relation types (activation or inhibition) in addition to satisfy the directional match requirements. The three network accuracy measures could provide a complete picture of how well an inferred model matched the true model.

The accuracy measures of the inferred network models on the six simulation datasets are summarized in Table 3. We denote the accuracy measures based on exact match, directional match and non-directional match as NA-EM, NA-DM, and NA-NM, respectively.

Table 3. Simulation-based evaluation results

True Model	Sample Size	MAR			BN		
		NA-EM	NA-DM	NA-NM	NA-EM	NA-DM	NA-NM
G1	40	33.33%	33.33%	33.33%	33.33%	41.67%	75.00%
G1	80	33.33%	33.33%	33.33%	33.33%	41.67%	75.00%
G2	40	30.12%	30.12%	31.33%	4.82%	12.05%	39.76%
G2	80	33.73%	33.73%	36.14%	3.61%	10.84%	61.45%
G3	40	28.32%	28.32%	30.10%	4.08%	9.69%	34.18%
G3	80	31.12%	31.12%	31.89%	4.85%	10.46%	36.73%

It is shown in Table 3 that MAR results achieved higher exact and directional match accuracies than BN results except for G1 (10 genes). BN achieved higher non-directional accuracies than MAR. These results need to be interpreted with caution. The relatively high exact and directional match accuracy measures of the MAR results were partly due to the fact that for many pairs of genes, relations of both directions were included in the results. In other words, the association rule algorithm included relations with both directions when it was difficult to infer the relation direction. This fact also explains the comparable values for all three measures for MAR results. In this sense, the three measures of the MAR results were more comparable to the non-directional match accuracies of the BN results. From these four measures we can conclude that the two techniques had the potential to extract 30% - 60% correct relations in the underlying models for typical gene expression data analyses (50 or 200 genes, 40 or 80 samples) when relation direction was not considered. Based on the exact match and directional match accuracies of the Bayesian network results, we can also conclude that it was difficult to extract accurate gene regulatory models from gene expression data and that only small numbers of accurate relations could be extracted.

We also observed that for the MAR results the effects of the number of genes and samples sizes on the accuracy measures were not significant. However, accuracy measures of the BN results degraded greatly with larger number of genes. The sample size also had a strong effect on accuracy measures for the BN results for G2 (50 samples) and G3 (200 samples).

We are not ready to make general conclusions based on the simulation-based evaluation results

because of the limited simulation datasets used. However, we do think the non-directional match accuracies demonstrate that regulatory network analysis based on observational gene expression data has a practical value for gene researchers. The learning results captured large amounts of information regarding the gene regulatory structures, making the technique quite useful as a data analysis tool for new hypothesis development.

4.2 Empirical Evaluation

Although simulation-based evaluation provides an estimate of the usefulness of the gene regulatory network analysis, much more complex real gene expression data still pose challenges on the empirical value of this approach. Empirical evaluations that compare analysis results based on real experimental data with existing knowledge of gene regulatory relations provides direct evidence on the potential usefulness of the approach. We employed two real microarray datasets, a *S. cerevisiae* dataset and a *Homo sapiens* dataset. The resulting networks were compared with three types of knowledge sources: known gene pathways, gene co-occurrence patterns in the literature, and expert judgments.

4.2.1 Datasets and Data Preprocessing

The *S. cerevisiae* yeast cell-cycle dataset of Spellman et al. [26] has been used frequently in previous microarray data analysis studies, which provide good benchmarks for evaluation. This dataset contains 76 gene expression measurements of the mRNA levels of 6,177 *S. cerevisiae* ORFs. The experiments measured six time series under different cell cycle synchronization methods [11]. The *Homo sapiens* dataset was provided by Arizona Cancer Center, and contained 33 samples (11 cell lines under 3 treatments) of 5,306 human genes in total.

Based on our gene expression data representation, we applied a ternary discretization to both datasets. For the *S. cerevisiae* dataset, our major goal was to compare the network learning results of our techniques with those of the previous studies. We used the same threshold as in Friedman et al.'s research [11], and applied the two techniques on the same 800 genes analyzed in their study. The threshold they used was 0.5 in logarithmic (base 2)

scale. Thus, expression levels with ratio to the reference that were lower than $2^{-0.5}$ were considered as under-expressed; levels higher than $2^{0.5}$ were considered as over-expressed; expression levels between $2^{-0.5}$ and $2^{0.5}$ were considered as normal. For the *Homo sapiens* dataset, we selected a threshold value of 1 in logarithmic (base 2) scale based on discussions with the domain scientists who conducted the experiments and had substantial experiences with the specific dataset. The domain scientists suggested that the network extraction techniques be applied on the 200 genes with greatest expression variations across the 33 samples.

4.2.2 Evaluation Against Known Pathways

Our first empirical evaluation study was intended to compare the regulatory networks generated from our information-theory-based Bayesian networks and association rule mining algorithms with those generated using search and scoring methods in Friedman's research [11]. Since the three resulting networks were all based on the same 800 yeast genes, the overlapping edges could suggest the degree of agreement among these three techniques.

In order to evaluate the abilities of these techniques to extract correct gene regulatory relations, we selected two sets of gene regulatory relations that have appeared in manually drawn regulatory pathways and previous literature as evaluation benchmarks. One of the relation sets was from the manually drawn regulatory pathway maps in Kyoto Encyclopedia of Genes and Genomes (KEGG: <http://www.genome.ad.jp/kegg/>). These maps incorporated existing common knowledge about gene interactions, including many types of relations like phosphorylation, dephosphorylation, ubiquitination, glycosylation, and transcription activation. We selected the relations relevant to gene expressions for evaluation purposes. In total we obtained from these maps 25 gene relations in which both genes were in the 800 genes in our network results. The other relation set is from the Biomolecular Relations in Information Transmission and Expression (BRITE) database. The relations consisted of yeast protein-protein interactions compiled from the literature, and yeast two-hybrid system interactions discovered in [13,

14, 30]. We identified 399 gene expression relations in which both genes were among the 800 genes in the network results.

Table 4. Empirical evaluation against known pathways

(BN1: the Bayesian network results of Friedman et al. [11], BN2: the information-theory-based Bayesian network results, MAR: the modified association rule mining results)

(a) Overlapping relations among three regulatory networks

Top N edges with highest strength	BN1 & BN2	BN1 & MAR	BN2 & MAR
20	5	1	3
50	15	3	5
200	51	25	44
500	175	94	138
1000	279	249	240

(b) Overlapping relations with know pathways

Top N edges with highest strength	Overlap with KEGG (25 relations)			Overlap with BRITE (399 relations)		
	BN1	BN2	MAR	BN1	BN2	MAR
200	0	0	1	0	0	0
500	2	0	2	0	0	0
1000	2	0	3	0	0	0
5000	3	2	4	5	4	7

The overlap analysis results are presented in Table 4a. We selected different numbers of regulatory relations by varying the minimum edge strengths in the three networks. The results indicated that the three regulatory networks had significant amounts of overlap in the extracted relations. More overlaps were observed between the two Bayesian network results. However, all three techniques generally failed to extract the gene regulatory relations appeared in known pathways (Table 4b). Only a small number of overlapping relations were found even when 5,000 relations were selected from each network.

4.2.3 Evaluation Against Literature Co-occurrence

We also evaluated our regulatory networks extracted from the *Homo sapiens* dataset using our Bayesian network (BN) and modified association rule mining (MAR) algorithms based on a Web database called PubGene (<http://www.PubGene.com>). PubGene creates a gene-to-gene co-occurrence literature network for 13,712 named human genes by automated analysis of titles and abstracts in over 10 million MEDLINE article records [15]. For each of the two human gene regulatory networks (MAR and BN), we chose 50 relations with highest strength.

Since PubGene can search genes only by gene names, genes without a name but only an accession ID cannot be searched in PubGene. Such genes might not have been well studied in previous research. We treated relations involving such unnamed genes as unknown relations. More than half of the remaining relations (19 out of 29 relations from MAR, 17 out of 32 relations from BN) were found in PubGene. Among these matched relations, some are direct relations between two genes. For example, for $MT2A \rightarrow MT1E$, the top 1 relation from MAR, $MT2A$ and $MT1E$ co-occurred in literatures for 11 times, indicating that a biological relation exists between these two genes. The remaining of the matched relations are indirect relations, which means that the pair of genes are indirectly linked to each other through a couple of other genes. For example, the two genes involved in $ARHGDI B \rightarrow CTL2$ are not directly linked in PubGene network, but both associated with a third gene $IL2$. They may also reflect some kind of association between those two genes. The percentage of each type of relations (unknown, direct, indirect, and incorrect) is presented in Table 5.

Table 5. Literature co-occurrence evaluation results

Type of Relation	BN	MAR
Direct Relations	12%	0%
Indirect Relations	26%	34%
Incorrect Relations	20%	30%
Unknown Relations	42%	36%

4.2.4 Evaluation Against Expert Judgments

In the 3rd empirical evaluation study, we asked domain scientists to evaluate gene regulatory relations in the networks extracted from the *Homo sapiens* dataset. A domain scientist determined correctness and interestingness of relations based on her experience as well as additional information about the genes by extensive literature and background information search. A relation was considered as interesting if it suggested potential hypothesis for a biological function.

Currently we have conducted the study with one domain scientist whose expertise is in Human genes. We selected the 100 regulatory relations with highest strength from the Human gene regulatory network.

The overall observation during this study was that the gene regulatory relations generated by our network extraction algorithms contained many genes and relations that have not been well studied. The domain scientist could not evaluate most of the relations, which conformed to the literature co-occurrence evaluation results. However, the domain scientist did make judgments on several regulatory relations. She identified 10 relations as potentially correct, including four activation relations (SDCCAG28→CATX-8, MAP7→CDH1, S100P→CDH1, and AA598508→LCN2) and 6 dependency relations (CAV1-CAV2, EPB72-SGK, EGR1-CASP1, SPOCK-ADAM12, LAMA4-CD44, and MST1-N92646). She also identified six relations as interesting, four of which were among the potentially correct relations and the two additional relations were an activation relation: MME→CAV2 and a dependency relation: H61003-AFAP. She also identified four relations as incorrect and four relations as potentially incorrect.

The evaluation results show that the regulatory network extraction techniques were able to extract some correct and interesting gene regulatory relations from microarray data. At the same time, a certain number of incorrect regulatory relations might also be extracted.

5. CONCLUSION AND FUTURE DIRECTIONS

In this study, we presented two scalable network extraction algorithms: an information-theory-based Bayesian network algorithm and a modified association rule mining algorithm. We conducted two types of evaluations to assess the practical value of these two techniques in helping researchers analyze large amount of gene expression data. In the simulation-based evaluation, the two techniques could extract 30% - 60% correct relations when relation direction was not considered. The empirical evaluation results showed that the extracted networks generally failed to identify regulatory relations reported in the literature. However, more than 50% of extracted relations reflected the gene co-occurrence patterns in the literature, and a small set of relations appeared to domain scientists to be potentially correct and interesting.

Our general conclusion is that regulatory network analysis can capture large portions of the potential regulatory structures behind the gene expression data. The small overlap of the analysis results with the existing literature and domain knowledge indicated that there might be potentials that new regulatory relations could be discovered with assistance of this type of analysis.

We are in the process of making the evaluation studies complete by including more regulatory relations from literature and conduct expert evaluation with more domain scientists. At the same time, we are also working on mechanisms to integrate the regulatory networks from the literature and from gene expression data to enable guided network learning from data and to allow researchers to evaluate the newly identified relations in the context of existing relation structures. We also plan to conduct larger-scale simulation-based evaluation by incorporating richer biological characteristics of the gene regulation process. More network-learning techniques will be evaluated using a large number of simulation datasets to provide practical guidance on conducting effective regulatory network analysis on gene expression data.

6. ACKNOWLEDGEMENTS

This research is supported by the grant: NIH/NLM, 1 R33 LM07299-01, 2002-2005, "Genescene: a Toolkit for Gene Pathway Analysis." We would like to thank Dan McDonald for comments and discussions. We would also like to thank George Watts, Jesse Martinez, Ryan Falsey, and Kerri Kislin from the Arizona Cancer Center for providing datasets, expert evaluations and valuable discussions. We would also like to thank Dana Pe'er from the Hebrew University for providing the regulatory network results for the comparative study.

7. REFERENCES

- [1] Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. in Proceedings of the IACM-SIGMOD International Conference on Management of Data (Washington D.C., 1993), 207-216.
- [2] Arkin, A. P., Shen, P.-D., and Ross, J. Deduction of a complex reaction mechanism from measured time series: Verification of the theory of statistical construction. *Science*, 277, 5330 (1997), 1275.

- [3] Arnone, A., and Davidson, B. The hardwiring of development: Organization and function of genomic regulatory systems. *Development*, 124, (1997), 1851-1864.
- [4] Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.-F., and Gandrillon, O. Strong-association-rule mining for large-scale gene-expression data analysis: A case study on human sage data. *Genome Biology*, (2002),
- [5] Berrar, D., Dubitzky, W., Granzow, M., and Eils, R. Analysis of gene expression and drug activity data by knowledge-based association mining. in *Proceedings of the Critical Assessment of Microarray Data Analysis Techniques (CAMDA '01)* (2001), 25-28.
- [6] Chen, T., Filkov, V., and Skiena, S. Identifying gene regulatory networks from experimental data. in *Proceedings of the RECOMB* (1999), 94-103.
- [7] Cheng, J., Greiner, R., Kelly, J., Bell, D. A., and Liu, W. Learning Bayesian networks from data: An information-theory based approach. *The Artificial Intelligence Journal*, 137, (2002), 43-90.
- [8] Chickering, D. M., Geiger, D., and Heckerman, D. Learning Bayesian networks is NP-hard. Technical Report MSR-TR-94-17, Microsoft Research, Microsoft Corporation, (1994),
- [9] Chow, C. K., and Liu, C. N. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14, (1968), 462-467.
- [10] D'haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. Linear modeling of MMA expression levels during CNS development and injury. in *Proceedings of the Pacific Symposium on Biocomputing '99* (1999), 41-52.
- [11] Friedman, N., Linial, M., Nachman, I., and Pe'er, D. Using Bayesian network to analyze expression data. *Journal of Computational Biology*, 7, (2000), 601-620.
- [12] Henrion, M. Propagating uncertainty in Bayesian networks by probabilistic logic sampling. *Uncertainty in Artificial Intelligence*, 2, (1988), 149-163.
- [13] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98, 8 (2001), 4569-4574.
- [14] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences*, 97, 3 (2000), 1143-1147.
- [15] Jenssen, T.-K., Lagreid, A., Komorowski, J., and Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature genetics*, 28, (2001), 21-28.
- [16] Kauffman, S. *The origin of order: Self-organization and selection in evolution*. Oxford University Press, 1993.
- [17] Kotala, P., Perera, A., Zhou, J. K., Mudivarthi, S., Perrizo, W., and Deckard, E. Gene expression profiling of DNA microarray data using peano count trees (p-trees). in *Proceedings of the Online Proceedings of the First Virtual Conference on Genomics and Bioinformatics* (2001).
- [18] Liang, S., Fuhrman, S., and Somogyi, R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. in *Proceedings of the Pacific Symposium on Biocomputing* (1998), 18-29.
- [19] Mjolsness, E., Mann, T., Castano, R., and Wold, B. From coexpression to coregulation: An approach to inferring transcriptional regulation among gene classes from large-scale expression data. *Neural Information Processing Systems*, 12, (1999), 928-934.
- [20] Mjolsness, E., Sharp, D. H., and Reinitz, J. A connectionist model of development. *Journal of Theoretical Biology*, 152, 4 (1991), 429-454.
- [21] Murphy, K., and Mian, S. Modelling gene expression data using dynamic Bayesian networks. Technical report, Computer Science Division, University of California, Berkeley, (1999),
- [22] Naitou, T., Satou, K., Furuichi, E., Kuhara, S., and Takagi, T. A system for finding association rules from microarray data and public databases. in *Proceedings of the Genome Informatics* (Tokyo, Japan, 2000), Universal Academy Press, 356-357.
- [23] Pearl, J. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.
- [24] Someren, E. v., Wessels, L., and Reinders, M. Linear modeling of genetic networks from experimental data. in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* (2000), 355-366.
- [25] Somogyi, R., and Sniegowski, S. A. Modeling the complexity of genetic networks: Understanding multigenic and pleiotropic regulation. *Complexity*, 1, 6 (1996), 45-63.
- [26] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9, (1998), 3273-3297.
- [27] Spirtes, P., Glymour, C., and Scheines, R. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9, (1991), 62-72.
- [28] Srinivas, S., Russell, S., and Agogino, A., Automated construction of sparse bayesian networks from unstructured probabilistic models and domain information. in Henrion, M., Shachter, R. D., Kanal, L. N., and Lemmer, J. F., (eds.). *Uncertainty in artificial intelligence*, North-Holland, Amsterdam, 1990.
- [29] Tuzhilin, A., and Adomavicius, G. Handling very large number of association rules in the analysis of microarray data. in *Proceedings of the SIGKDD '02* (Edmonton, Alberta, Canada, 2002).
- [30] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Qureshi-Emili, P. P. A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature*, (2000), 623-627.
- [31] Wahde, M., and Hertz, J. Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems*, 55, (1999), 129-136.
- [32] Weaver, D., Workman, C., and Stormo, G. Modeling regulatory networks with weight matrices. in *Proceedings of the Pacific Symposium on Biocomputing* (Hawaii, 1999), World Scientific Publishing Co., 112-123.
- [33] Wermuth, N., and Lauritzen, S. Graphical and recursive models for contingency tables. *Biometrika*, 72, (1983), 537-552.
- [34] Wessels, L., Someren, E. v., and Reinders, M. A comparison of genetic network models. in *Proceedings of the Pacific Symposium on Biocomputing* (2001).
- [35] Wuensche, A. Genomic regulation modeled as a network with basins of attraction. in *Proceedings of the Pacific Symposium on Biocomputing* (1998), 89-02.