Accuracy of the Bayesian Network Algorithms for Inferring Gene Regulatory Networks

Jaakko Niemi
March 14, 2007
# Contents

1   Introduction ................................................................................................................. 3

2   Biological background ................................................................................................ 4

   2.1  Gene expression .................................................................................................. 4

   2.2  DNA microarrays ............................................................................................... 4

   2.3  Gene regulatory networks ................................................................................... 5

3   Bayesian networks ...................................................................................................... 7

   3.1  Structure learning ............................................................................................... 7

   3.2  Algorithms for inferring gene regulatory networks ............................................ 8

4   Data ........................................................................................................................... 10

   4.1  Data set 1 ........................................................................................................... 11

   4.2  Data set 2 ........................................................................................................... 11

5   Results ....................................................................................................................... 12

   5.1  Experiment 1 ..................................................................................................... 13

   5.2  Experiment 2 ..................................................................................................... 15

6   Conclusions ............................................................................................................... 16

References ......................................................................................................................... 18

Appendix 1: The genes in the data set 1 ........................................................................... 20

Appendix 2: The genes in the data set 2 ........................................................................... 21
1 Introduction

The emergence of DNA microarray technology in the 1990’s has made large-scale gene expression level measurements possible. Using the information of the gene expression levels measured during different phases of the cell-cycle the gene regulatory networks can be inferred. Inferring a genetic system, where gene regulatory networks play a large role, is one of the central topics in the genome biology [Stryer, 1995]. Thus, several algorithms that infer gene regulatory networks from experimental microarray data have been developed. These algorithms can be classified into three categories [Wimberley, et al., 2003]

- Algorithms based on Bayesian networks
- Algorithms based on Boolean networks
- Algorithms using stochastic formulas and theories.

However, inferring the structure of a gene regulatory network is a complicated machine learning task. There are previous studies on the performance of inferring algorithms using simulated microarray data [Wimberley, et al., 2003]. The studies found that most of the algorithms did not manage better than if the regulatory connections between gene pairs would have decided randomly. The best algorithms were all based on Bayesian networks, but the performance of these algorithms decreased as simulations became more realistic.

In this study, the accuracy of the inferring algorithms was explored using real microarray data. The study includes some well-managed algorithms from previous studies and some new inferring algorithms. All included algorithms are based on Bayesian networks, because it has been the most promising algorithm class in the previous studies.

First the biological background of the gene expression, DNA microarrays and gene regulatory networks are introduced. Then the theory of Bayesian networks is briefly introduced and after that the Bayesian network inferring algorithms and the data used in this study are represented. Then the performance of the inferring algorithms are presented and discussed. Finally, the conclusions of the study are given.
2 Biological background

2.1 Gene expression

A gene is a sequence of DNA that codes a protein. Gene expression is the process by which a gene's DNA sequence is used for building a protein. Proteins have a major role in the structures and functions of a cell. Gene expression is a multi-step process that begins with transcription of DNA into messenger RNA (mRNA). It is then followed by post transcriptional modification and translation into a protein. After translation the protein is folded into 3d-shape and targeted to its final destination.

2.2 DNA microarrays

A DNA microarray is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip, forming an array for the purpose of monitoring expression levels for thousands of genes simultaneously. RNA is extracted from a cell or tissue sample. Fluorescent tags are enzymatically incorporated into the RNA or can be chemically attached to the RNA. The RNA spot will then fluoresce when examined using a microarray scanner. The fluorescence intensity of each spot is then evaluated in terms of the number of copies of a particular mRNA, which ideally indicates the level of expression of a particular gene.

Nowadays it is possible to measure expression levels for tens of thousands of genes on one microarray. This is a huge possibility for bioinformatics studies. Measuring gene expression using microarrays is relevant to many areas of biology and medicine, such as studying treatments, disease, and developmental stages. One of the hottest topics where DNA microarray techniques are used is the studying of transcriptional regulation of genes.
2.3 Gene regulatory networks

Gene regulation is the cellular control of the amount and timing of appearance of the functional product of a gene, i.e. protein. It is known that some genes regulate other genes, which means that the amount of a gene expressed in a certain time could activate or inhibit the expression of another gene. Gene regulation has an important role on cell functions. It gives the cell control over structure and function, and is the basis for cellular differentiation. The cell cycle is the series of events in a cell between one cell differentiation and the next. Thus, it is the process by which a single-cell fertilized egg develops into a mature organism and the process by which hair, skin, blood cells, and some internal organs are renewed. The cell cycle consists of different stages and the moving from one stage to the other is controlled by gene regulation. Different genes are expressed in different stages of the cell cycle, and the expression of a gene can activate the expression of another gene leading to a transfer from one stage to the next stage, which in turn can inactivate the previously expressed gene. Thus, there are many time dependent regulatory relations between genes.

These relations can be inferred using cell cycle microarray data, which means that the gene expression levels are measured in several different time periods across the cell cycle. From cell cycle microarray data can be seen when, i.e. in which stage of the cell cycle, the genes are expressed and when they are not expressed. From this information it is possible to infer gene regulatory relationships between genes. Based on these relations a gene regulatory network, in which the regulatory relations are easy to see, can be constructed. An example of a gene regulatory network is depicted in Figure 1. Yellow points represent genes and the name of the gene is written in black letters. In the network the regulatory relations between genes are marked with arrows. The direction of the arrow shows which gene regulates which gene. The green color indicates that the regulation is activating and the red color that the regulation is inhibiting. The yellow loops indicate self-inhibition of a gene. The gene is self-inhibited if no other gene inactivates it.
Figure 1. The cell cycle network of the key regulator genes of the yeast [Li, et al., 2004].

Nowadays it is possible to artificially activate or inhibit the expression of a gene. *E.g.* a gene known to be involved in a disease can be inhibited. But before inhibition, it is essential to know is the gene cell cycle regulated, and in which genes it is connected. If the expression of a cell cycle regulated gene is modified, the natural cell cycle may be interrupted and the causes may be serious. Thus, the knowledge of the gene regulatory networks is important.
3 Bayesian networks

Bayesian networks can be used to model gene regulatory networks. A Bayesian network is a directed acyclic graph (DAG) whose nodes represent random variables (genes) and edges represent statistical dependence relations among the variables (regulatory relations among the genes) and local probability distributions for each variable given values of its parents [Jensen, 2001].

If there is an edge from node $A$ to another node $B$, then variable $B$ depends directly on variable $A$ (gene $A$ regulates gene $B$), and $A$ is called a parent of $B$. If for each variable $X_i$, $i \in \{1,\ldots,N\}$, the set of parent variables is denoted by $\text{parents}(X_i)$, then the joint distribution of the variables is product of the local distributions

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid \text{parents}(X_i)).$$

(1)

If $X_i$ has no parents, its local probability distribution is said to be unconditional, otherwise it is conditional. If the variable represented by a node is observed, then the node is said to be an evidence node. In the case of the gene regulatory networks the edges between variables represent connections between genes, and thus now on the word connection is used instead of the word edge.

3.1 Structure learning

In the simplest case, a Bayesian network is specified by an expert and is then used to perform inference. In other applications the task of defining the network is too complex for humans. In this case the network structure and the parameters of the local distributions must be learned from data.

Learning the structure of a Bayesian network is a very important part of machine learning. Assuming that the data is generated from a Bayesian network and that all the variables are visible in every iteration, optimization based search method can be used to find the structure of the network. It requires a scoring function and a search strategy. A
common scoring function is posterior probability of the structure given the training data. The time requirement of an exhaustive search returning back a structure that maximizes the score is superexponential in the number of variables. A local search strategy makes incremental changes aimed at improving the score of the structure. A global search algorithm can avoid getting trapped in local minima.

### 3.2 Algorithms for inferring gene regulatory networks

Bayesian network structure learning algorithms can be used to infer gene regulatory networks from DNA microarray data. In this study six structure learning algorithms were tested using real microarray cell cycle data. The tested algorithms are

- FCI [Spirtes, et al., 2001]
- GES [Spirtes, et al., 2001]
- K2 with Bayesian scoring [Cooper, et al., 1992]
- K2 with BIC scoring [Cooper, et al., 1992]
- MWST [Chow, et al., 1968]
- PC [Spirtes, et al., 2001].

The algorithms FCI, GES and PC are implemented in free software package called Tetrad [Scheines, et al., 1994]. Tetrad can be downloaded from website http:/www.phil.cmu.edu/projects/tetrad. The algorithms PC and FCI use constraint based searches meaning that they start with a fully connected graph, and remove connections if certain conditional independencies are measured in the data. PC algorithm searches for Bayes net when it is assumed there is no latent (unobserved) variables. The algorithm computes many conditional independence tests, and combines these constraints into a DAG. In deciding whether an independence relation holds, the algorithm uses a Table Reducing Chi Square Test. FCI algorithm performs a search similar to PC but allowing that there may be latent variables.
Greedy equivalence search algorithm (GES) is a Bayesian algorithm that searches over equivalence classes of statistical models. It is initialized with equivalence class E containing the empty DAG. Then it repeatedly replace E with the member of E that has the highest score, until no such replacement increases the score.

The algorithms K2 and MWST are implemented in Bayes Net Toolbox [Murphy, 2001]. The toolbox is for MATLAB software and can be obtained from website http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html. The K2 algorithm is a greedy search algorithm. It works as follows. Initially, each node has no parents. It then adds incrementally that parent whose addition increases the score of the resulting structure the most. When the addition of no single parent can increase the score, it stops adding parents to the node. In addition to the search procedure, the user can select the scoring function for two options. Here, both the Bayesian scoring function and the Bayesian Information Criterion (BIC) scoring function are used. The Bayesian score integrates out the parameters i.e. it is the marginal likelihood of the model. The BIC is defined as

\[ BIC = \log(P(D|\hat{\theta})) - 0.5 \cdot d \cdot \log(N) \quad (2) \]

where D is the data, \( \hat{\theta} \) is the maximum likelihood estimate of the parameters, \( d \) is the number of parameters, and \( N \) is the number of data cases.

The maximum weight spanning tree algorithm (MWST) associates a weight to each connection. Here, the weight is the mutual information between the two variables. When the weight matrix is created, a usual MWST algorithm gives an undirected tree that can be oriented with the choice of a root.
4 Data

In this study, real microarray data was wanted to use to test the accuracy of the inferring algorithms. The genome of the yeast is quite small compared to more complex organisms, and the yeast cell cycle is relatively fast, and thus it is studied extensively. A comprehensive yeast cell cycle microarray data [Spellman, et al., 1998] is publicly available in the website http://cellcycle-www.stanford.edu. Thus, we decided to use the yeast cell cycle data for testing the algorithms.

The cell cycle expression level measurements are done with three independent fluorescence measurement methods. Overall, the data set contains 77 mRNA-level measurements for each gene. Each of these 77 measurements is measured in different stage of the yeast cell cycle. The data set includes 6178 genes, almost the whole yeast genome. About 800 of these genes meet the criterion for cell cycle regulation [Spellman, et al., 1998].

It would have been an ideal test situation to take all the genes in the test procedure. Unfortunately, the number of genes is simply too high for the inferring algorithms. The softwares refuse to do the inference with so many variables (i.e. genes). If the inference could have been done, it would have taken too much time, and the final regulatory network would have been very complicated and it would have been difficult to make any conclusions in it. Even if only the 800 cell cycle regulated genes would have taken into the test data set, the inference would have taken too much time and the results would have been difficult to interpret, because of the high amount of the regulatory relations between the genes. Fortunately, the yeast cell cycle is well studied and the studies have exposed that there are smaller amount of key regulator genes that are responsible for the control and regulation of the cell cycle. The interactions between these key regulator genes are well known. According to the literature, we made two data sets including different number of key regulator genes.
4.1 Data set 1

Data set 1 contains 77 time dependent expression level measurements for 19 key regulator genes. The genes, included in the data set 1, are listed in Appendix 1. A gene regulatory network of these key regulator genes have been constructed based on extensive literature studies [Li, et al., 2004]. The network is shown in Figure 1. This is the network the inferring algorithms should be able to infer from the data set 1. The performance of the algorithm is measured comparing the network inferred by the algorithm to the network in Figure 1.

4.2 Data set 2

Data set 2 contains 37 key regulator genes. It includes the 19 genes in the data set 1 and 18 additional cell cycle regulated genes. The additional genes in the data set 2 are listed in Appendix 2. These genes are chosen based on the information of the Comprehensive Yeast Genome Database [Güldener, et al., 2005] and the literature [Futcher, 2002]. The desired network is again the network in Figure 1. The inferred connections among the additional 18 genes are ignored when the performance of the algorithms is measured. The additional cell cycle regulated genes are added to the data set 2 to test how the additional data affect to the performance of the inferring algorithms.
5 Results

In this study two experiments were made. In both experiments the six Bayesian inference algorithms were run ten times. If there was a connection between two genes in over half of the runs, the genes were interpreted as connected. The connections of the network given an algorithm were then compared to the connections of the network in Figure 1. The connections were categorized in four categories:

- correct positive (cp)
- false positive (fp)
- correct negative (cn)
- false negative (fn).

Correct positive means that the algorithm inferred a connection between two genes like in Figure 1. False positive means that the algorithm inferred a connection between two genes, although the genes are not connected in Figure 1. Correct negative means that the algorithm did not infer a connection between genes that should not be connected and false negative means that the algorithm did not infer a connection between genes, although they should have been connected. Only the connections between the genes were taken into account, and thus the gene regulatory network in Figure 1 was simplified. The direction of the connection and the difference between positive and negative regulation were ignored. The self-inhibition of the genes was also ignored.

In Figure 1 there are some points that include more than one gene (e.g. Cln1, 2 includes the genes Cln1 and Cln2), and some gene complexes (e.g. the complex SBF consists of the genes Swi4 and Swi6). The data sets, however, include the expression levels for the single genes, not for the gene combinations or complexes. In these situations the genes were counted connected if the algorithm inferred that at least one of the genes in the complex was connected to the other gene (e.g. SBF and Cln1, 2 were counted connected if Swi4 was connected to Cln1 according to the algorithm, although Swi6 was not connected neither Cln1 nor Cln2). The inferred connections inside gene combinations and
complexes were ignored. In the desired network there are 11 points representing genes or gene complexes, and between these points there are overall 55 possible connections, in which 24 are positive connections meaning that the genes are actually connected, and 31 are negative connections meaning that the genes are not connected.

The performance of the algorithms was compared calculating two parameters from the results; error rate ($er$) and correct relations ratio ($crr$). The error rate tells how many percent of the connections are misclassified. The correct relations ratio tells the ratio of the inferred connections that are actual connection and the falsely positively inferred connections. The error rate tells how close the network inferred by an algorithm is to the desired network. The smaller the error rate is, the better the algorithm is. While, the correct relation ratio does not measure the closeness of the entire networks, but it measures the proportion of correct positive connections in all positively inferred connections. The error rate is calculated

\[
er = \frac{fp + fn}{pc},
\]  

where $er$ is the error rate, $fp$ is the number of false positives, $fn$ is the number of false negatives and $pc$ is the number of all possible connections, which here is always 55. The correct relations ratio is calculated

\[
crr = \frac{cp}{cp + fp},
\]  

where $crr$ is the correct relations ratio, $cp$ is the number of correct positives and $fp$ is the number of false positives.

5.1 Experiment 1

In experiment 1 the six Bayesian inference algorithms were tested using the data set 1. The results are shown in Table 1. The error rates varied from 0.44 to 0.53 meaning that there is not much difference between the performances of the algorithms. GES algorithm had the smallest error rate, while MWST was the only algorithm that had error rate
over 0.50 meaning that it inferred more false connections than correct connections. GES algorithm inferred clearly more positive connections than the other algorithms, but the number of false positive connections was also the largest. K2 algorithm with BIC scoring function had the highest correct relations ratio. It inferred 63% of the positive connections correctly. MWST algorithm was again the worst. It was the only algorithm that inferred more false positive connections than correct positive connections having the worst correct relations ratio, 0.43.

Table 1: The results of Experiment 1

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correct pos</th>
<th>False pos</th>
<th>Correct neg</th>
<th>False neg</th>
<th>Error rate</th>
<th>Correct relations ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>GES</td>
<td>15</td>
<td>12</td>
<td>16</td>
<td>12</td>
<td>0.44</td>
<td>0.56</td>
</tr>
<tr>
<td>K2(BIC)</td>
<td>5</td>
<td>3</td>
<td>25</td>
<td>22</td>
<td>0.45</td>
<td>0.63</td>
</tr>
<tr>
<td>PC</td>
<td>6</td>
<td>5</td>
<td>23</td>
<td>21</td>
<td>0.47</td>
<td>0.54</td>
</tr>
<tr>
<td>K2(Bayesian)</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td>23</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>FCI</td>
<td>5</td>
<td>5</td>
<td>23</td>
<td>22</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>MWST</td>
<td>6</td>
<td>8</td>
<td>20</td>
<td>21</td>
<td>0.53</td>
<td>0.43</td>
</tr>
</tbody>
</table>
5.2 Experiment 2

In experiment 2 the six Bayesian inference algorithms were tested using the data set 2, but the studied network is the same as in the experiment 1. The purpose of this experiment was to test how the additional data affects to the inference algorithms. The results of the experiment 2 are shown in Table 2.

The additional information had different effect on the algorithms. Both K2 algorithms and GES algorithm inferred almost the same amount of positive connections than in experiment 1, while as FCI, PC and MWST algorithms inferred clearly less connections now. The variation of the error rates was same than in the experiment 1, but now the K2 algorithms managed best. The K2 algorithm with BIC score had the smallest error rate 0.44 and the highest correct relations ratio 0.71. The performance of the FCI and PC algorithms decreased dramatically, and they were clearly the worst algorithms in experiment 2 having the highest error rate 0.53 and the lowest correct relations ratio 0.33.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correct pos</th>
<th>False pos</th>
<th>Correct neg</th>
<th>False neg</th>
<th>Error rate</th>
<th>Correct relations ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>K2(BIC)</td>
<td>5</td>
<td>2</td>
<td>26</td>
<td>22</td>
<td>0.44</td>
<td>0.71</td>
</tr>
<tr>
<td>K2(Bayesian)</td>
<td>5</td>
<td>3</td>
<td>25</td>
<td>22</td>
<td>0.45</td>
<td>0.63</td>
</tr>
<tr>
<td>MWST</td>
<td>4</td>
<td>4</td>
<td>24</td>
<td>23</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>GES</td>
<td>12</td>
<td>12</td>
<td>16</td>
<td>15</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>FCI</td>
<td>1</td>
<td>2</td>
<td>25</td>
<td>27</td>
<td>0.53</td>
<td>0.33</td>
</tr>
<tr>
<td>PC</td>
<td>2</td>
<td>4</td>
<td>24</td>
<td>25</td>
<td>0.53</td>
<td>0.33</td>
</tr>
</tbody>
</table>
None of the tested Bayesian inference algorithms can be recommended to use for inferring gene regulatory networks from real gene expression microarray data according to the results. The error rates were in general high and varied from 0.44 to 0.53. If an algorithm would have given an empty network, the error rate would have been 0.44 that is the same rate the best algorithms achieved in the experiments. The worst algorithms managed even worse than if the connection between every gene would have decided randomly (error rate 0.50). If only the inferred positive connections are considered, most of the algorithms inferred more correct positive connections than false positive connections. But all algorithms inferred also false positive connections, and some algorithms inferred more false positive than correct positive connections.

All algorithms were robust when the same data set was used meaning that they gave almost the same network in different runs. When different data sets were used, the algorithms, which start from empty network and then add connections to the network according to a scoring function (K2 and GES algorithms) performed almost similarly. While, the algorithms starting from fully connected graph and then removing connections according to a scoring function (PC and FCI algorithms) inferred less connections when the amount of the data was increased.

It was expected that the performance of the algorithms would be poor when they infer a gene regulatory network from real microarray data. Microarray measurements have a small signal to noise ratio. The noise makes the inference more difficult, and large expression differences are needed to reliably distinguish zero from non-zero correlations between the genes. The gene regulatory network that we used as reference network is not an absolute truth. There could be some connections between genes that are not shown in the network. E.g. the most commonly inferred false positive connection, that almost every algorithm inferred, was Cln 1, 2 – Clb5, 6. These genes are not connected in Figure 1, but according to the Comprehensive Yeast Genome Database [Güldener, et al., 2005] these genes interact. Thus, this connection may be correct positive after all.
The study indicates that the tested Bayesian inferring algorithms are not qualified to infer gene regulatory networks from real microarray data. Overall, the study was highly simplified, because only few key regulator genes were selected to the data sets. In ideal situation the data set would have included the whole yeast genome, and the algorithms would give good results with small error rates. However, this ideal situation is far away from today’s inferring algorithms. Thus, it is recommended that also in future the experimental techniques are used for inferring gene regulatory relations instead of machine learning methods.
References


Appendix 1: The genes in the data set 1

- Cdc14
- Cdc20
- Cdh1
- Clb1
- Clb2
- Clb5
- Clb6
- Cln1
- Cln2
- Cln3
- Fkh1 (part of the SFF complex)
- Fkh2 (part of the SFF complex)
- Mbp1 (part of the MBF complex)
- Mcm1
- Ndd1 (part of the SFF complex)
- Sic1
- Swi4 (part of the SBF complex)
- Swi5
- Swi6 (part of the SBF and the MBF complexes)
Appendix 2: The genes in the data set 2

- Ace2
- Cdc4
- Cdc5
- Cdc6
- Cdc7
- Cdc16
- Cdc23
- Cdc27
- Cdc28
- Cdc34
- Cdc45
- Cdc53
- Clb3
- Clb4
- Dbf4
- Mcm10
- Orc1
- Skp1

Plus all the genes in the data set 1