Learning Gene Regulatory Network from Microarrays
Based on Bayesian Network

Yu HONG¹,a, Xiao-Hu HUANG¹,b,*, Xiao-Ling XIA¹,c, Jia-Jin LE¹,d

¹School of Computer Science and Technology,
Donghua University, Shanghai, China
aandrehong@mail.dhu.edu.cn, bhtigerdhu@163.com,
csherlysha@dhu.edu.cn, dilejiajin@dhu.edu.cn
*Corresponding author

Keywords: Microarrays, Gene Regulatory Network, Bayesian Network.

Abstract. The development of modern medical technology gives a big boost to analyze microarrays and construct gene regulatory network. To get a clean and straightforward gene regulatory network from enormous microarray expression data, we propose a learning algorithm based on Bayesian network. We use K-Medoids to obtain medoid genes and greedy search with BDe to construct Bayesian network structure related to the certain disease. Experimental result on REGED0 demonstrates that the gene regulatory network of lung cancer we get from microarrays contains representative genes and reveals core regulatory relationships with high effectiveness and efficiency.

Introduction

With the development of modern medical technology, biomedical scientists can obtain the microarray expression level of thousands of genes related to a certain disease at a low cost, which gives a big boost to the research of gene regulatory systems and gene regulatory networks. Nowadays, it is of essential importance to identify the relationships among genes and construct the network of gene regulatory from microarrays. This issue brings up new challenges for existing mathematical models and analysis methods in this area, since the observation data are vary large-scale and the number of variables is enormous. Therefore, how to make effective and efficient expression and reasoning of gene regulatory network from microarrays is the primary problem for biomedical researchers.

Boolean network [1] is a useful but coarse-grained model to indicate gene activation by “1” and gene inhibition by “0”. It shows global logical relationships among genes by Boolean expressions. T. Akutsu, S. Miyano and S. Kuhara [2] give a mathematical proof that if the indegree of each node is bounded by a constant, only $O(\log n)$ state transition pairs are necessary and sufficient to identify Boolean network of $n$ nodes correctly with high probability. But the relationships among genes in real life are complicated and far away from logical rules. Therefore, Boolean network is only a simplified model to express inheritance and interaction of genes.

Differential equation is a linear model based on recurrent neural network [3]. It is more refined to infer time constants, weights and bias terms of genes by genetic algorithm. T. Chen, H.L. He and G.M. Church [4] model both transcription and translation by kinetic equations with feedback loops from translation products to transcription. T. Akutsu, S. Miyano and S. Kuhara [5] infer S-systems (synergistic and saturable systems) from time series data where S-systems are based on a particular kind of nonlinear differential equation. But differential equation has large amount of calculation and lacks of ability to resist noise in data.

Bayesian network [6] is a perfect compromise between qualitative Boolean network and quantitative differential equation. Bayesian network can not only reflect quantitative relationships among variables but also be very intuitive with the natural union of probability theory and graph
theory. Finding relationships among genes from microarrays can be regarded as learning a Bayesian network structure which best represents the observation data.

In this paper, we propose a gene regulatory network learning algorithm based on Bayesian network. To avoid high complexity of network structure due to huge amount of genes, we reduce the dimension of microarray expression data by K-Medoids at the very beginning. Then, we choose BDe as the scoring method and greedy search as the searching algorithm to learn the structure of gene regulatory network that best represents the observation data. Experimental result on REGED0 demonstrates that our algorithm can obtain representative genes and uncover core causal relationships among them, thus becoming an effective and efficient algorithm for gene regulatory network.

**Bayesian Network**

Bayesian network, proposed by Pearl J [6], is an effective graph model in the area of Knowledge Discovery in Database (KDD) and Decision Support System (DSS). In recent years, Bayesian network also becomes a popular model for expression and reasoning of uncertain knowledge in gene regulatory. A Bayesian network consists of two parts: a network structure and a Conditional Probability Table (CPT). The network structure is a directed acyclic graph (DAG) composed of vertexes and arcs. Each vertex represents an event or random variable abstracted from real-life problems. An arc connecting two vertexes represents direct causal relationship between them. Further more, the direction of arcs means causes producing effects. Specially, there are no cycles in the structure. Fig. 1 shows a typical Bayesian network structure of metastatic cancer from Reference [7]. A CPT is defined for a set of random variables to demonstrate marginal probability of each variable with respect to the others. Fig. 2 shows the CPT of metastatic cancer, where each variable has two values: 1(e.g. 1) and 0(e.g. 0).

![Figure 1. Structure of metastatic cancer.](image1)

![Figure 2. CPT of metastatic cancer.](image2)
As we can see above, the structure of Bayesian network indicates causal relationships network among the variables and CPT quantifies these relationships by conditional probability, which makes Bayesian network a rigid but straightforward model compared with neural network. So how to get a Bayesian network for expression and reasoning of uncertain knowledge in gene regulatory? Analogously, Bayesian network learning includes structure learning (discovering the structure from observation data) and parameter learning (learning CPT from both observation data and structure). Structure learning is more important and more complicated than parameter learning. Generally, there are three ways to learn the structure of Bayesian network based on independence test, score and mixture respectively. The algorithm based on score includes two parts: scoring method and searching method. In the next section, we will propose a Bayesian network structure learning algorithm for gene regulatory network with BDe as the scoring method and greedy search as the searching method by reducing dimension first.

Proposed Approach

Dimension Reduction by K-Medoids

There are tens of thousands of genes in human genome. Constructing a gene regulatory network using all these genes is unrealistic, since large-sample learning of Bayesian network structure is NP-hard [8]. Even if the network with all details exists, it is extremely difficult to analyze or reason through this complex network. Therefore, it is very necessary to reduce complexity of Bayesian network structure and reveal relationships among some limited but important genes by dimension reduction. Here we use k-medoids [9] as the dimension-reduction algorithm to get medoids in clusters, thus reducing the number of genes and highlight important ones. Algorithm 1 expresses the procedure of K-Medoids.

**Algorithm 1** K-Medoids

**Input** Gene set $G$, number of medoids $k$

**Output** Medoid gene set $V$

1. Choose $k$ genes $\{G_1, G_2, \cdots, G_k\}$ in $G$ at random to be the initial cluster medoids
2. Assign each gene $G_j$ in $G$ to the cluster associated with the closest medoid
3. Recalculate the positions of the $k$ medoids by finding each medoid $G_i$ within each cluster $C_i$ that minimizes $\sum_{G_j \in C_i} d(G_i, G_j)$, where $i = 1, 2, \cdots, k$
4. Repeat Step 2 and 3 until the medoids $\{G_1, G_2, \cdots, G_k\}$ become fixed
5. $V := \{G_1, G_2, \cdots, G_k\}$

We prefer K-Medoids as the dimension-reduction algorithm rather than K-Means [10], the most common clustering algorithm. This is because K-Means is easily affected by extreme values in observation data. Instead of choosing centroid by the mean of coordinates of genes in the cluster, K-Medoids prefers the gene that minimizes the sum of distances in the cluster as the medoid. Iterative computations make the positions of medoids finally become fixed. After clustering the whole gene set into several medoids, we obtain the genes that are most representative and reduce the complexity of gene regulatory network.

Bayesian Dirichlet Equivalent (BDe)

Let $D$ be a database of cases, $B_s$ be the Bayesian network structure that represents $D$, the posterior probability of $B_s$ given $D$ is defined as:
\[ P(B_s | D) = \frac{P(D, B_s)}{P(D)}. \] (1)

\( P(D) \) is independent of network structure \( B_s \), so finding the most probable Bayesian network structure \( B_s \) given a database \( D \) (i.e. finding maximum of \( P(B_s | D) \)) is to find maximum of \( P(D, B_s) \).

However, when we consider a score metric as a measurement of a probable Bayesian network structure \( B_s \), we have to take the current state of information \( \xi \) into consideration. The well-known Bayesian Dirichlet (BD) score [11] is given by:

\[
P(D, B_s | \xi) = P(B_s | \xi) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}
\] (2)

where

\[ N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk} \] (3)

\[ N_{ij} = \sum_{k=1}^{r_i} N_{ijk}. \] (4)

\( n \) is the number of variables, \( r_i \) is the number of states of variable \( x_i \), \( q_i \) is the unique states of the parents \( \Pi_i \) of \( x_i \), \( N_{ijk} \) is the number of cases in database \( D \) in which \( x_i = k \) and \( \Pi_j = j \), \( N'_{ijk} \) is a Dirichlet exponent. \( \Gamma() \) is Gamma function, here for positive integer \( m \) is defined as:

\[ \Gamma(m) = (m - 1)! . \] (5)

Given a domain \( U = \{x_1, x_2, \ldots, x_n\} \) with multinomial parameters \( \Theta_U \), we may state likelihood equivalence in terms of \( \Theta_U \):

**Assumption 1 (Likelihood Equivalence) [11]** Given two network structures \( B_s \) and \( B_{s'} \) such that \( P(B_s | \xi) > 0 \) and \( P(B_{s'} | \xi) > 0 \), if \( B_s \) and \( B_{s'} \) are equivalent, then \( P(\Theta_U | B_s, \xi) = P(\Theta_U | B_{s'}, \xi) \).

The assumption of likelihood equivalence when combined the previous assumptions introduces constraints on the Dirichlet exponents \( N'_{ijk} \). The result is a likelihood-equivalent specialization of the BD metric, which we call the BDe metric:

**Theorem 1 (BDe Metric) [11]** Given domain \( U \), suppose that \( \rho(\Theta_U | B_s, \xi) \) is Dirichlet with equivalent sample size \( N' \) for some complete network structure \( B_s \) in \( U \). Then, for any network structure \( B_{s'} \) in \( U \), there is

\[
P(D, B_s | \xi) = P(B_s | \xi) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}
\] (6)

where

\[ N'_{ijk} = N' P(x_i = k, \Pi_j = j | B_s, \xi). \] (7)
Greedy Search with BDe

Greedy search is a classical heuristic algorithm in Bayesian network structure learning. Here we propose a structure learning algorithm with greedy search as the search algorithm and BDe as the scoring method to learn gene regulatory network. The gene set we have reduced dimension by Algorithm 1 above is considered as the input and parent set of each gene is the output. The procedure is shown in Algorithm 2.

For each gene, our algorithm chooses the gene that maximizes the current BDe score as the candidate of its parents at each iteration. If the current new score is higher than the old one, which means it is the most probable parent of this gene, we add it into parent set and update the old score. Finding parents of this gene stops until the BDe score of the gene and its parents no longer increases. Finally, all the parent sets of each gene we get can be the useful information to construct a gene regulatory network.

Algorithm 2 Greedy search algorithm with BDe

Input  Dimension-lowered gene set \( V \)

Output Parent set \( \Pi_i \) of each gene

1. \( \text{For } i := 1 \text{ to } n \)
2. \( \Pi_i := \emptyset \)
3. \( \text{score}_\text{old} := BDe(V, \Pi_i) \)
4. \( \text{FLAG} := \text{TRUE} \)
5. \( \text{While } \text{FLAG} \)
6. \( \text{Let } Z \text{ be the node in } \{ V - \Pi_i \} \text{ that maximizes } BDe(V, \Pi_i \cup \{Z\}) \)
7. \( \text{score}_\text{new} := BDe(V, \Pi_i \cup \{Z\}) \)
8. \( \text{If } \text{score}_\text{new} > \text{score}_\text{old} \)
9. \( \text{score}_\text{old} := \text{score}_\text{new} \)
10. \( \Pi_i := \Pi_i \cup \{Z\} \)
11. \( \text{Else} \)
12. \( \text{FLAG} := \text{FALSE} \)
13. \( \text{End while} \)
14. \( \text{End for} \)

Empirical Study

REGED (REsimulated Gene Expression Dataset) is a genomics dataset on Causality Workbench [12]. REGED0 is the version without manipulation, including 999 features (genes) and 500 training examples. The goal of REGED is to find genes, which could be responsible of lung cancer. The data are “re-simulated”, i.e. generated by a model derived from real human lung-cancer microarray gene expression data. From the causal discovery point of view, it is important to construct gene regulatory network of lung cancer.

Firstly, we use K-Medoids to reduce the dimension of REGED0 and highlight important genes. The number of medoids we set is 9, which means 1 gene out of 111 genes on average. These 9 medoids and the number of genes in each cluster are shown in Table 1. “g789” means NO. 789 gene in REGED0 for example.

<table>
<thead>
<tr>
<th>cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>medoid</td>
<td>g789</td>
<td>g386</td>
<td>g697</td>
<td>g562</td>
<td>g928</td>
<td>g777</td>
<td>g419</td>
<td>g359</td>
<td>g660</td>
</tr>
<tr>
<td>number of genes</td>
<td>249</td>
<td>66</td>
<td>182</td>
<td>157</td>
<td>132</td>
<td>144</td>
<td>40</td>
<td>20</td>
<td>9</td>
</tr>
</tbody>
</table>
Then, we use these 9 main medoid genes to construct gene regulatory network of lung cancer using greedy search algorithm with BDe as the scoring method. We discretize their microarray expression data into discrete values before Bayesian network structure learning. Due to the dimension-reduction step, time complexity of structure learning is greatly reduced. The gene regulatory network of lung cancer is shown in Fig. 3.

![Gene regulatory network of lung cancer.](image)

Fig.3 is a very straightforward Bayesian network for lung cancer. We can see that there are no arcs pointing at or from g386 and g660, which means they are isolated by other genes and have little or none causal relationships with others. Therefore, it is assumed that 66 genes in cluster 2 whose medoid is g386 and 9 genes in cluster 9 whose medoid is g660 are irrelevant to the lung cancer. For the rest 7 medoids, there are 4 arcs pointing from g697 to g359, g419, g777 and g562 respectively and 1 arc pointing from g789 to g697, which means g697 is an extremely important gene in the disease of lung cancer. Thus, the method we propose can select representative genes from enormous microarray gene expression data and construct clean and straightforward gene regulatory network with high effectiveness and efficiency.

**Summary**

To construct clean and straightforward gene regulatory network from enormous microarray expression data, we propose a gene regulatory network learning algorithm based on Bayesian network. Firstly, we use K-Medoids to reduce the dimension of microarrays and obtain medoid genes. Then, we use greedy search with BDe to construct Bayesian network structure related to the certain disease. Experimental result on REGED0 demonstrates that the gene regulatory network of lung cancer we get contains representative genes and reveals core regulatory relationships with high effectiveness and efficiency.

**Acknowledgement**

This research was financially supported by Informatization Development Special Foundation of Shanghai of China under Grants No. 20140314 and Industry Development Special Foundation of Software and Integrated Circuit of Shanghai of China under Grants No. 150657.
References


