A Comparison of Models for Gene Regulatory Networks Inference

Blagoj Ristevski[1], Suzana Loskovska[2]

[1]Department of Information Systems Management
Faculty of Administration and Information Systems Management
St. Kliment Ohridski University – Bitola, Republic of Macedonia
blagoj.ristevski@uklo.edu.mk
[2]Department of Computer Science and Informatics
Faculty of Electrical Engineering and Information Technologies
Ss. Cyril and Methodius University – Skopje, Republic of Macedonia
suze@feit.ukim.edu.mk

Abstract. Gene regulatory networks are complex networks composed of nodes representing genes, transcription factors, microRNAs and other components or modules and their mutual interactions represented by edges. These networks can reveal and depict the fundamental gene regulatory mechanisms in the cells. In this paper we compare the obtained results of gene regulatory networks inference from gene expression microarray data. We have used dynamic Bayesian networks, Boolean networks and graphical Gaussian models as models for network inference. We applied three different size gene expression datasets simulated using a simple autoregressive process. After network inference, we compared the values of the area under ROC curve (AUC) as a validation measure. Some directions for further improved approach for GRNs reconstruction which will include prior knowledge are proposed at the end of this paper.

Keywords: bioinformatics, gene regulatory networks, Bayesian networks, graphical Gaussian models, Boolean networks, area under ROC curve.

1 Introduction

The complex networks composed of genes, proteins and other components regulate the functions and development of the cells through their interactions. The gene regulatory networks (GRNs) provide an understandable view for gene regulatory mechanisms and can uncover the reasons for many diseases. GRNs components are nodes which represent the genes, metabolites, proteins or modules, and edges which correspond to the direct and indirect interactions between nodes. Genes as key components in the GRNs are DNA segments which present fundamental heredity units of every living organism.

The central dogma in molecular biology is presented by two processes: *transcription* and *translation*. In the process of transcription a gene is transcribed into mRNA and after that proteins are produced by translation. When the protein is

synthesized the corresponding coding gene is expressed. The gene expression levels correspond to the approximate number of produced RNA copies from the corresponding gene, which means that gene expression is related to the amount of produced proteins. The microarray technology provides gene expression data as an observation of gene expression under specific experimental conditions or different time points [8].

The inferring or reconstructing of gene regulatory interactions from experimental data is called GRNs inference. There are many models for inferring of GRNs such as Boolean networks, Bayesian networks, dynamic Bayesian networks, graphical Gaussian models, Petri networks, linear and nonlinear differential and difference equations systems, information theory approach, state space models and fuzzy logic models.

The remainder of this paper is organized as follows. In Section 2 we describe Boolean networks. In the third section we present the graphical Gaussian models and their assumptions and usage in the networks inference. We also describe the partial correlation coefficients and their significance for network inference. Bayesian Networks and dynamic Bayesian networks (DBNs) are presented in the following section. The area under ROC curve as a validation measure is described in Section 5. In Section 6 we describe the simulation of artificial gene expression data used for GRNs inference and the obtained inferred networks and the AUC values are shown, too. The concluding remarks are given in the last section.

2 Boolean Networks

Boolean Networks model is a simple model for GRNs inference, consisting of a set of nodes and edges. The nodes represent genes whereas the edges between the nodes correspond to the gene interactions. In Boolean networks, gene expression levels are discretized and presented by two levels states. The genes which have expression levels above a certain threshold are represented by state 1 and the other genes by state 0.

The graph representing a Boolean network gives information about the connection between genes, but it is not sufficient for understanding the all dependencies between genes. The main goal of the reverse engineering in Boolean networks is finding a Boolean function of every gene in the network, so that discretized values of gene expression can be explained by the model. But, the small changes in the gene expression levels cannot be covered by two levels discretization, which leads to information loss. Another shortcoming of Boolean networks is the super-exponential number of all possible networks depending on the number of genes n and it is equal to 2^{2^n}.

REVerse Engineering Algorithm (REVEAL) based on Boolean networks has been introduced by Liang et *al*. (1998) [9]. This algorithm constructs a Boolean network of given expressed gene data by setting the gene in-deegree. If n is the number of nodes and k is the value of in-degree of the genes, then the number of all possible networks can be computed by the Eq. 1:

$$\left(2^{2^k}\frac{n!}{(n-k)!}\right)^n \quad (1)$$

REVEAL extracts minimal network structures using the mutual information approach from the state transition tables of the Boolean network.

3 Graphical Gaussian Models

Graphical Gaussian models (GGMs) are commonly used as a method for GRNs reconstruction based on gene expression data and they are very computationally efficient [3]. GGMs as graphical probabilistic models can identify conditional independence relations among the nodes. They make an assumption that the input gene expression data follow a multivariate Gaussian distribution [6].

The nodes represent genes, and the edges represent conditional dependence relations between nodes. The absence of an edge between two genes means that the corresponding genes are conditionally independent given other genes in the model.

Let Y be the input gene expression data matrix with *G* columns, corresponding to the number of genes, and with *N* rows which correspond to the number of samples (time series data points or other experimental conditions) [3]. It is supposed that matrix Y follows a multivariate normal distribution $N_G(\mu,\Sigma)$, where $\mu=(\mu_1,...,\mu_G)'$ is the mean vector, and $\Sigma=(\sigma_{ij})_{1\le i,j\le G}$ is the positive definite covariance matrix. $\sigma_{ij}=\sigma_i\sigma_j$ are covariance parameters between genes *i* and *j*, and σ_i^2 are related to the variance terms for gene *i*. The estimation of the covariance matrix of the data distribution is a base for the GGMs inference.

First, in the GGM inferencel, to make a reliable estimation of the partial correlation matrix $\tilde{P}=(\tilde{\rho}_{ij})$ is required [4]. This matrix is related to the inverse matrix of the covariance matrix Σ. The straightforward estimator is given by the following Eq. 2:

$$\tilde{r}_{ij}=-\frac{\hat{\omega}_{ij}}{\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}} \quad (2)$$

where

$$\hat{\Omega}=(\hat{\omega}_{ij})=\hat{\Sigma}^{-1} \quad (3)$$

The partial correlation coefficients $\tilde{r}_{ij}$, which describe the correlation between nodes/genes Y_i and Y_j conditional dependent on all other nodes in the network, are measures of the direct interactions among nodes/genes [5]. Partial correlation between two genes measures the degree of correlation remaining after removing the effects of the other genes which differs from Pearson correlation coefficients [1] [6].

The above mentioned procedure is appropriate when *N* is larger than the number of genes *G*, otherwise the covariance matrix is not positive-definite and its inverse matrix cannot be found. In microarray data, the sample size *N* is much smaller than the number of genes *G*. For that reason it is suggested to use a shrunk estimate of the covariance matrix. The goal is to construct well conditioned positive definite matrix,

so that the matrix can be inverted. If λ is a shrinkage coefficient so that $0 \le \lambda \le 1$, then shrunk covariance matrix Σ^* is computed by following Eq. 4

$$\Sigma^* = \lambda T + (1-\lambda)S \tag{4}$$

where $\hat{S}$ is the estimated empirical covariance matrix. The shrinkage parameter λ is chosen to minimize the mean-square error and it is determined analytically given by Eq. 5.

$$\lambda^* = \frac{\sum_{i \neq j} \mathrm{var}(r_{ij})}{\sum_{i \neq j} r_{ij}^{\,2}} \tag{5}$$

After computing the partial correlation coefficients $\tilde{r}_{ij}$, the distribution of $|\tilde{r}_{ij}|$ is checked and the edges with significantly small values of $|\tilde{r}_{ij}|$ are removed from the network [2].

The second stage of the GRNs inference is model selection – assigning statistical significance to the edges from the GGM network.

4 Bayesian Networks

Bayesian networks (BNs) are a special type of graph model defined as a triple (G, F, θ), where G denotes the graph structure, F is the set of conditional probability distributions, and θ is the set of parameters for the graph structure [10]. The structure of the graph G consists of a set of n nodes x_1, x_2, ..., x_n and a set of directed edges between the nodes. The nodes correspond to the random variables and the directed edges show the conditional dependences between the variables (genes).

A directed edge from the node X to the node Y is denoted as $X \rightarrow Y$ which means that X is a parent of Y denoted as $pa(Y)$. Edges and nodes and edges together have to create a directed acyclic graph (DAG).

The joint probability distribution is given by Eq. 6:

$$p(x) = \prod_{i=1}^{n} p\left(x_i \middle| x_{\{1,\cdots,i-1\}}, \theta, G\right) \tag{6}$$

If pa_i denotes the parent nodes of the node x_i which means that the state of each variable x_i depends on the states of its parent pa_i:

$$p(x) = \prod_{i=1}^{n} p\left(x_i \middle| pa_i, \theta, G\right) \tag{7}$$

BNs can deal with noisy and stochastic nature of gene expression data and with incomplete knowledge about the system. The small number of data points (samples) and large number of genes are common problems for BNs learning. Another disadvantage is that feedback loops cannot be captured, although they exist in the GRNs. BNs represent probabilistic relations between genes at the same time and they cannot represent the time relationships between variables

To overcome these drawbacks of BNs, dynamic Bayesian networks (DBNs) are used to model gene regulations. DBNs can deal with stochastic variables, time series gene expression data, feedback loops, missing values, hidden variables and can include prior knowledge [11]. The hidden nodes (variables) can capture effects that cannot be directly measured in a microarray experiment.

If x_t^i represents the *i*-th node at time point *t*, the joint probability distribution is given by Eq. 8:

$$p(x_t|x_{t-1}) = \prod_{i=1}^{n} p\left(x_t^i \middle| pa(x_t^i), \theta, G\right) \quad (8)$$

The GRNs inference is followed by structure and parameter learning of the BNs from training data *D* [7]. For given data *D*, the aim is to find posterior distribution of the network structure *M*, and then from this distribution the structure M^* which best fits the data should be found according to Eq. 9:

$$M^* = \arg\max_M \{P(M \mid D)\} \quad (9)$$

For an optimal network structure M^* and given data, it is required to find posterior distribution of parameters q by Eq. 10:

$$q^* = \arg\max_q \{P(q \mid M^*, D)\} \quad (10)$$

The BNs learning is NP-hard task and thus BNs and DBNs are appropriate for inference of small networks [12] because the number of DAGs $G(n)$ super-exponential depends on the number of nodes *n* and it is given by the Eq. 11:

$$G(n) = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} G(n-k) \quad (11)$$

In Table 1 the number of all possible DAGs as a dependence on the number of the graph nodes is shown.

Table 1. A table survey of the number of all possible DAGs depending on the number of nodes.

number of nodes	number of all possible DAGs
4	543
5	29 281
6	3 781 503
8	783 702 329 343
10	$4{,}175099 \cdot 10^{18}$
15	$2{,}377253 \cdot 10^{41}$
20	$2{,}344880 \cdot 10^{72}$
22	$1{,}075823 \cdot 10^{87}$
24	$9{,}435783 \cdot 10^{102}$

5 Validation of Inferred GRNs

To validate obtained results, the inferred network should be assessed in comparison with the referent network. Commonly used criteria for validation are the Receiver Operator Characteristics (ROC) curves and the area under ROC curve (AUC). The ROC curve is a chart of the ratio between sensitivity and (1-specificity), where sensitivity corresponds to a proportion of the actual positives edges which are correctly identified whereas specificity is proportion of negatives edges which are correctly identified [13] [15].

To facilitate the model validation , instead of ROC curve the AUC can be used. The AUC is the area covered by the ROC curve with the x-axis. Bigger value of AUC means better inferred network. The AUC is calculated by integrating the area bounded by the ROC curve and the x-axis [14].

6 Simulated Data and Results

To infer GRNs and then to validate above described models: Boolean networks, GGMs and DBNs, we have simulated artificial gene expression data by a simple first order autoregressive process given by Eq. 12:

$$X(t) = Ax(t-1) + B + \varepsilon(t) \quad (12)$$

where $\varepsilon(t)$ is a vector distributed by zero-centered multivariate Gaussian distribution with diagonal variance matrix.

We have obtained three different size datasets. The first dataset Data1 consists of simulated gene expression data for 5 genes and 50 time points. The dataset Data2 corresponds to 10 genes and 50 time points, and the number of genes in the third dataset Data3 is 15 measured in 100 time points.

The true referent networks and the inferred networks for the three datasets Data1, Data2 and Data3 are illustrated on Fig. 1-Fig.3. The values for AUC as validation criteria are shown tabular on Table 2. These AUC values show that for smaller datasets, Boolean networks model has the best performance in comparison to the GGMs and DBNs. For larger datasets GRNs inference performed by GGMs overcomes the other models.

Table 2. A comparison of the AUC values for three different inference models: GGMs, Boolean network and DBNs.

network inference model	Data1	Data2	Data3
GGMs	0.65	**0.63**	**0.57**
Boolean networks	**0.94**	0.56	0.46
DBNs	0.29	0.15	0.51

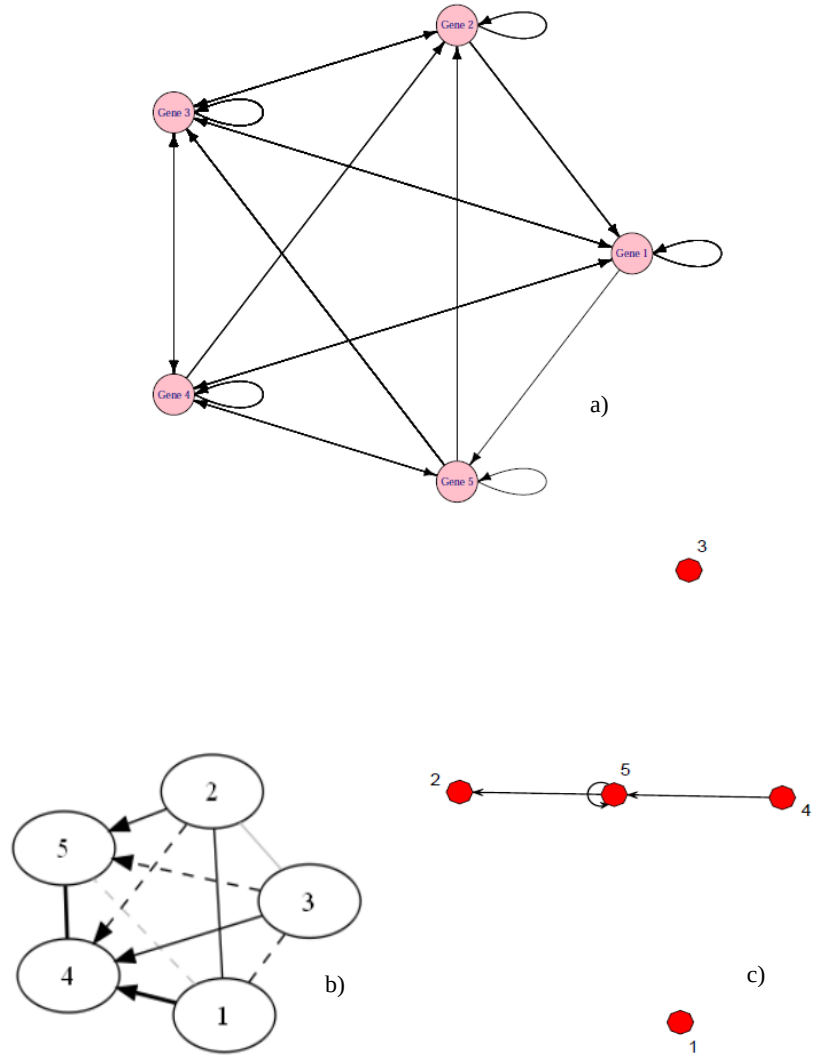

Fig. 1. True and inferred networks a) inferred Boolean network from Data1 b) reconstructed GRNs by GGMs and c) the true referent network corresponds to the Data1.

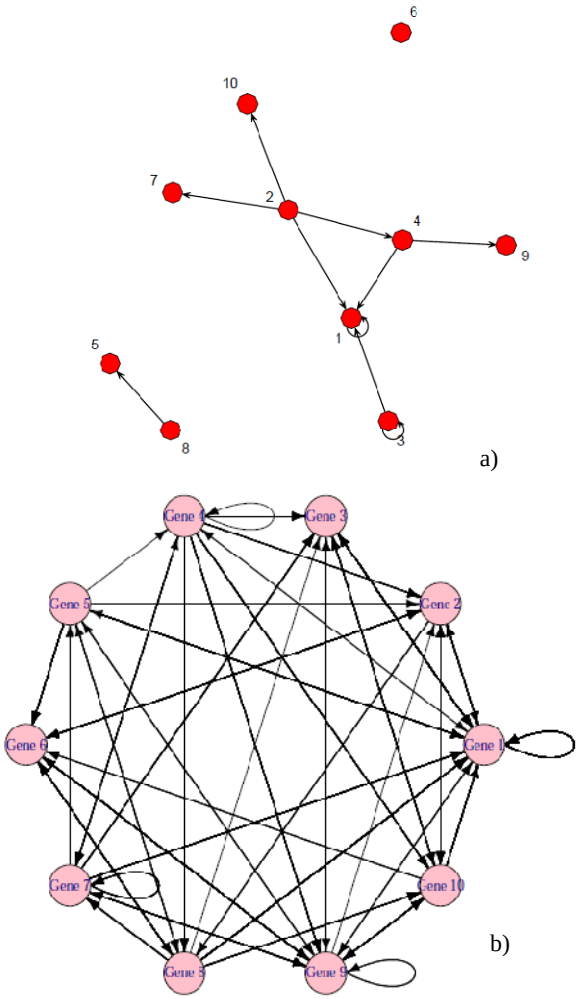

Fig. 2. True and inferred network a) the true referent network corresponds to the Data2 b) inferred Boolean network from Data2.

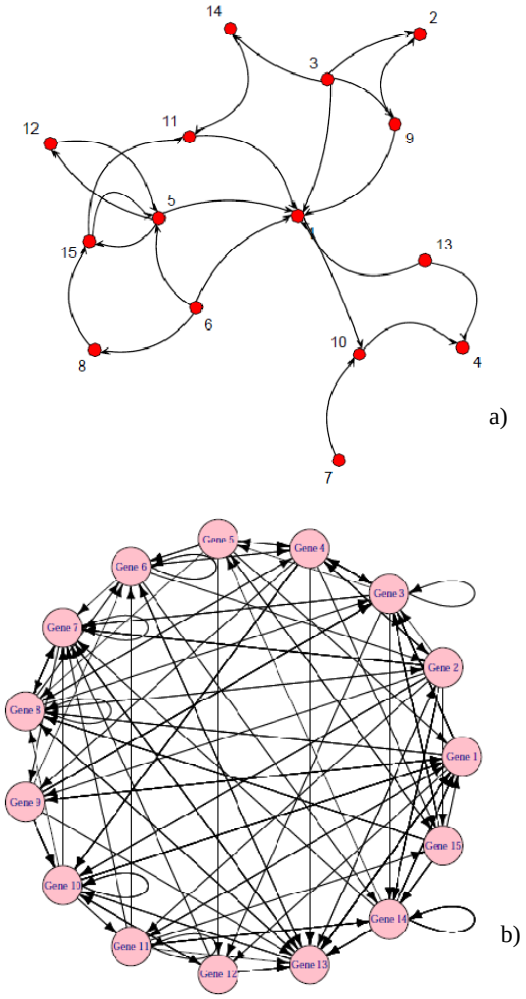

Fig. 3. True and inferred network a) the true referent network corresponds to the Data3 b) inferred Boolean network from Data3.

7 Conclusions

The presented AUC values obtained by GRNs inference using different models and datasets have shown that for datasets containing time series for larger number of genes, GGMs surpass the other network inference models: Boolean networks and DBNs. Only in the case where time series are for small number of genes (in our dataset - 5 genes) the Boolean network model has better inference performance compared to GGMs and DBNs, whereas DBNs model has shown worst inference properties. In accordance to these results we suggest using of GGMs results as prior

knowledge for improved approach for GRNs inference whereas the second inference stage is Markov Chain Monte Carlo (MCMC) simulation method to reconstruct more reliable GRNs.

Reference

1. N. Kraemer, J. Schaefer, A.-L. Boulesteix, *Regularized Estimation of Large-scale Gene Association Networks Using Graphical Gaussian Models*, Technical Report, Department of Statistics, University of Munich, 2009.
2. A. V. Werhli, M. Grzegorczyk, D. Husmeier, *Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks*, Bioinformatics, Vol.22, no. 20, 2006.
3. F. Jaffrezic, G. Tosser-Klopp, *Gene Network Reconstruction from Microarray Data*, BMC Proceedings, 2009.
4. J. Schaefer, R. Opgen-Rhein, K. Korbinian, *Reverse Engineering Genetic Networks using GeneNet Package*, R-News 6/5:50-53, 2006.
5. S. Ma, Q. Gong, H. J. Bohnert, *An Arabidopsis Gene Network Based on the Graphical Gaussian Model*, Genome Research, 2009.

J. Schaefer, K. Strimmer, *Learning Large-Scale Graphical Gaussian Models from Genomic Data*, American Institute of Physics, Vol. 776, 2005.

6. B. Ristevski, S. Loskovska, *Bayesian Networks Application for Representation and Structure Learning of Gene Regulatory Networks*, 12th International Conference on Computers and Information Technology ICCIT '09. , Dhaka, Bangladesh; 2009.
7. B. Ristevski, S. Loshkovska, S. Dzeroski, I. Slavkov, *A Comparison of Validation Indices for Evaluation of Clustering Results of DNA Microarray Data*, Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008, China.
8. S. Liang, S. Fuhrman and R. Somogyi, *REVEAL, a general reverse engineering algorithm for inference of genetic network architectures*, Pacific Symposium on Biocomputing 3, 1998, pp. 18-19.
9. N. Friedman and M. Goldszmidt, *Learning Bayesian Networks with Local Structure*, Proceedings of the NATO Advanced Study Institute on Learning in graphical models, 1998.
10. R. E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, 2003.
11. M. Grzegorczyk and D. Husmeier, *Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move*, Machine Learning, 71: 265-305, 2008.
12. L. M. de Campos, *A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests*, Journal of Machine Learning Researc 7, 2006.
13. T. Fawcett, *An introduction to ROC analysis*, Pattern Recognition Letters 27, 2006, pp. 861-874.
14. C. T. Le, *Introductory Biostatistics*, John Wiley & Sons, Inc., New Jersey 2003.