Visualization and Structure Learning of Gene Regulatory Networks using Bayesian Networks

Blagoj Ristevski[1], Suzana Loskovska[2]

Abstract - **The cell functions and development are regulated by complex networks of genes, proteins and other components by means of their mutual interactions. These networks are called gene regulatory networks (GRNs). The gene regulatory networks are used to reveal the fundamental gene regulatory mechanisms, to determine the reasons for many diseases and interactions between drugs and their targets, to produce a clear and comprehensible notion for cell regulation,. The introduction of experimental technologies such as microarrays and chromatin immunoprecipitation ChIP-chip, has provided a large number of available datasets related to gene expression and transcription factors (TFs). These datasets are basis for further analysis to reveal the gene regulation mechanisms. We implemented and visualized the dynamic Bayesian network which is able to cope with missing data and can include a prior knowledge about transcription factors. Also, we describe the obtained results and survey the common structure learning algorithms for learning of GRN's structure.**

Keywords - **Gene regulatory networks, Bayesian network, Bioinformatics.**

I. INTRODUCTION

The living cells during their life span carry out many different tasks controlled by the cell genome which is encoded in the DeoxyriboNucleic Acid (DNA) molecule. The genes are transcribed into messenger RiboNucleic Acid (mRNA), and then translated in proteins.

The necessity to generate, analyze and integrate the large scale expression data led to the development of microarray technology [2]. Genes and their products – proteins work coordinately in complex networks. The cell functions and development are regulated by complex networks of genes, proteins and other components by means of their mutual interactions. These networks are called gene regulatory networks (GRNs). The proteins which activate or inhibit the transcription of the other genes are called transcription factors (TFs). Transcription factors are important components in gene regulatory networks. The GRNs are commonly used to study influences between cell components because they provide a clear and understandable notion for cell regulation as well as reveal the fundamental gene regulatory mechanisms and find out the reasons for many diseases.

[1]Blagoj Ristevski is with Faculty of Administration and Information Systems Management, Department of Information Systems Management, St. Kliment Ohridski University – Bitola, Partizanska bb (Kompleks kasarni), 7000 Bitola, Macedonia. e-mail: blagoj.ristevski@uklo.edu.mk

[2]Suzana Loskovska is with Faculty of Electrical Engineering and Information Technologies, Skopje, Department of Computer Science and Computer Engineering, Ss. Cyril and Methodius University – Skopje, Macedonia, e-mail: suze@feit.ukim.edu.mk

Many destructive diseases such as cancer are related to different genetic disorders. Modeling of the GRNs represents one of the most powerful techniques to describe the fundamental cellular mechanisms and associated intracellular and intercellular processes. The goal of many researches which include the experimental and simulating methods is by studying the GRNs to reveal therapeutic and prognostic relevant knowledge about many diseases.

Besides microarray technology, the introduction of other experimental technologies such as chromatin immunoprecipitation ChIP-chip, provides a lot of available datasets related to gene expression and transcription factors. ChIP-chip provides an insight into interaction between transcription factors and promoter region of the gene when it is combined with microarray analysis [5]. These data types are basis for further analysis and they are means of revealing the gene regulation mechanisms and essential knowledge about cell processes on genomic and molecular level.

To represent GRNs, more models are used, such as state space model, Bayesian networks, dynamic Bayesian networks, Boolean networks, linear and nonlinear differential and difference equations model, fuzzy logic model, information theory model, and others models.

Finding out the most reliable and accurate structure of GRNs from high dimensional microarray data is a machine learning problem known as structure learning of graphical models. A subset of the data is used for model fitting and the residual data for the model validation [3]. Cross-validation methods are very useful for validation and training of regulatory networks. But, the obtained networks which fit the best to the training set are overtrained [4]. Such overfitted networks lose their ability to generalize the suitable networks for the data outer the training set.

The remainder of this paper is organized as follows. In the second section we present the models based on Bayesian networks and their advantages and disadvantages. The consequent section describes the dynamic Bayesian networks. A survey of structure learning algorithms of reconstructed gene regulatory networks is given in the Section 4. The following section is devoted to the reconstructed GRN using Bayes Net Toolbox (BNT). The concluding remarks are given in the last section.

II. BAYESIAN NETWORKS

The Bayesian networks are a special case of graph models consisted of two components and based on statistical principles [6]. The first part is a directed acyclic graph *G=(V,E)* where $V = \{1, \cdots, n\}$ is a set of nodes and *E* is a set of edges. Each node $i \in V$ refers to random variable $x_i \in x$ that represents the gene expression - in the case of GRNs. The set

of edges corresponds to the conditional dependence among nodes. The second part of network is a set of conditional probability distributions that describe the conditional probability of each variable (the gene expression).

If $x = \{x_1, \cdots, x_n\}$ denotes a set of random variables, G - structure of graph, θ - set of parameters the joint probability distribution is given by Eq. 1.

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

If the pa_i denotes the parent nodes of the node x_i that means the state of each variable x_i depends on the states of its parent pa_i (Eq. 2):

$$p(x) = \prod_{i=1}^{n} p\left(x_i \middle| pa_i, \theta, G\right) \quad (2)$$

Bayesian networks are suitable to show regulatory mechanisms between network components quantitatively as well as qualitatively. The qualitative description of regulatory mechanisms refers to presence/absence of an edge between network nodes whereas quantitative representation is made by a set of conditional probability distributions.

The modeling of the gene regulatory networks is made by structural and parameter learning. The goal of structural learning is to determine topology of network. For a given network structure, the parameter learning includes parameter estimation of unknown model for each gene. This is performed by determination of conditional dependencies between network components.

Bayesian networks can deal with noisy and stochastic nature of gene expression data and with incomplete knowledge about system. The small number of data points (samples) and the big number of genes are common problems for learning Bayesian networks. Another disadvantage is that this model cannot model feedback connections, which exist in the gene regulatory networks.

III. DYNAMIC BAYESIAN NETWORKS

To overcome the problems of Bayesian networks, **dynamic Bayesian networks** are used to model gene regulations. The dynamic Bayesian networks are capable to deal with stochastic variables, time series gene expression data, to include prior knowledge, feedback loops and to handle missing values and hidden variables. The hidden nodes (variables) can capture effects that cannot be directly measured in a microarray experiment.

The joint probability distribution is given by Eq. 3, where x_t^i is the i-th node at time t.

$$p\left(x_t \middle| x_{t-1}\right) = \prod_{i=1}^{n} p\left(x_t^i \middle| pa(x_t^i), \theta, G\right) \quad (3)$$

Dynamic Bayesian networks can apply and learn for real biological data, and there is a relationship to Hidden Markov Model, Boolean networks, stochastic Boolean networks, dynamic Bayesian networks with continuous state and other probabilistic models. The hidden nodes provide a way of linking similar data types and analysis of other network parameters. When some dependency exists between variables in the network, the hidden node can model that dependency.

Relatively low prediction accuracy and excessive computational time are two problems which reduce the performance of the dynamic Bayesian network model. To overcome these problems it is suggested a **dynamic Bayesian network approach** which limits the potential regulators to those genes with either earlier or simultaneous expression changes (up- or down-regulation) in regard to their target genes. Then, the genes with either earlier or simultaneous expression changes are assigned as possible regulators of those genes with a later expression change.

IV. STRUCTURE LEARNING

To choose the most appropriate network for a given dataset it is necessary to carry out a validation of the modeled networks. The precision and the reliability of models predictions are commonly examined in respect with input experimental data during the process of model validation. Structure learning of Bayesian network consists of finding a directed acyclic graph (DAG) that best fits the dataset. The structure learning performs by means of scoring function that evaluates how well the DAG explains the data and then to search for the best DAG that optimized the scoring function. There are two approaches for structure learning: constraint-based and search-based approach. In the constraint-based approach, the algorithm starts with a fully connected graph and removes edges.

The number of directed acyclic graphs as a function of the number of nodes G(n) is super-exponential dependent (Eq. 4)

$$G(n) = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} G(n-k) \quad (4)$$

After modeling of GRNs, the model which provides a good fit to the data should be selected. The criteria such as Bayesian score [8], maximal likelihood, Bayesian information criterion, minimum description length are used for model validation of gene regulatory networks. Model is validated by the *n*-cross validation too, where the input set is divided to *n* parts. The *n*-1 parts are used as a training set, and the remaining as a test set.

Let M denotes the structure of a dynamic Bayesian network, D is the data set, $P(M)$ is the prior probability of the network structure and $P(D|M)$ is its marginal likelihood. θ is a parameter vector of the conditional probability distributions. The marginal likelihood is an average of the likelihood $P(D|M,\theta)$ over all possible parameters associated to the network. The Bayesian score is based on the marginal likelihood of the data is defined as follows:

$$P(D \mid M) = \int P(D \mid M, \theta) P(\theta \mid M) d\theta \quad (5)$$

and provides a matching between model complexity and the data size.

The goal of the Minimum Descriptive Length criterion (MDL) is to provide an optimal matching between the precision of the data fitting and the complexity of network

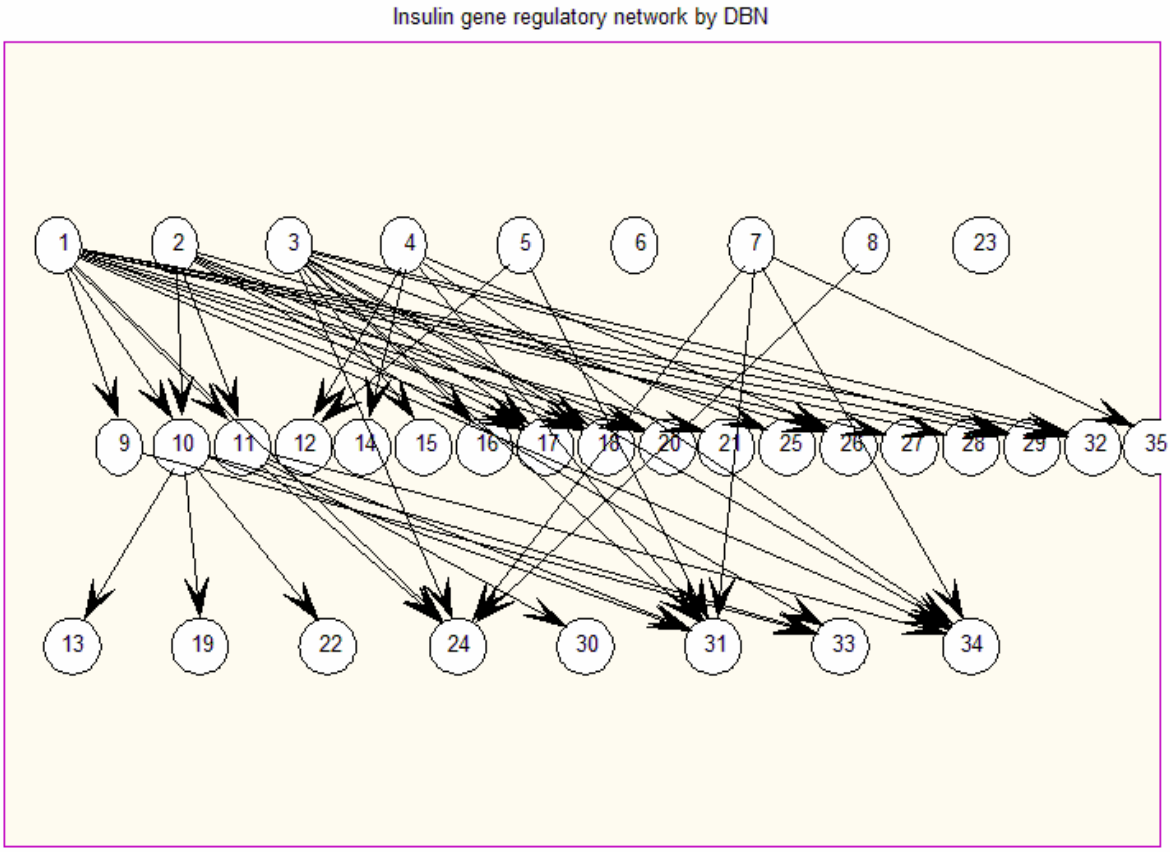

Figure 1. Insulin gene regulatory network with 35 genes.

model. The MDL score consists of model and data set encoding, hence the MDL criterion L is a sum of the network coding length L_M and data set coding lengths given by the following equation:

$$L = L_M + \sum_{j=1}^{m-1} H(x_{j+1}|x_j) = L_M - \sum_{j=1}^{m-1} \log(p(x_{j+1}|x_j)) \qquad (6)$$

where $H(x_{j+1}|x_j)$ is the state transition conditional entropy, $p(x_{j+1}|x_j)$ is the transitional probability and m – number of sample points [9].

At the Bayesian Information Criterion (BIC) n denotes the sample size, k – numbers of parameters and RSS is the residual sum of squares from estimated model. If it is assumed that the distribution of the models errors is normal, then the aim is to minimize BIC, given by Eq. 7.

$$BIC = n \log\left(\frac{RSS}{n}\right) + k \log n \qquad (7)$$

The Maximum Likelihood (ML) criterion expresses the likelihood $L(\theta)$ as a function of the unknown vector parameter θ and aims to find the parameters of all possible values which maximizes $L(\theta)$. The solution can be a function of one or many parameters and often this problem is nonlinear optimization problem.

To obtain a good balance between the accuracy of data fitting and the complexity of gene network models, some of above mentioned criteria are applied. Their application infers GRNs with significant biological performances of inferred interactions between networks elements.

Because of super-exponential dependency between number of directed acyclic graphs and number of nodes, the local or global search algorithms are used (K2, Hill-climbing, MCMC, Structured EM).

In addition to the search procedure, the scoring function should be specified.

The K2 algorithm is a greedy search algorithm that works as follows. In the beginning each node has no parents, it then adds incrementally that parent whose addition most increase the score of the resulting structure. When the adding of no single parent can increase the score, the adding parents to the node stops. It attempts to select the network structure that maximizes the posterior probability of network given the experimental data [1].

Hill-climbing starts at a specific point in space, considers all nearest neighbors, and moves to the neighbor that has the highest score. If no neighbors have higher score than the current point (a local maximum is reached), the algorithm stops.

To evaluate the obtained results, the inferred network structure should be compared with the reference network. The Receiver Operator Characteristics (ROC) curves are used to evaluate inferred network structure quantitatively [7]. The ROC curve is a chart of the ratio between sensitivity and (1-specificity), where sensitivity corresponds to proportion of

actual positives edges which are correctly identified and specificity is proportion of negatives edges which are correctly identified. The ROC curves can be summarized by computing the AUC (Area Under the ROC Curve).

V. RESULTS

To implement dynamic Bayesian gene regulatory network we use insulin gene expression data, Bayes Net Toolbox [10] and Bayesian Network Structure Learning – software package [11]. The obtained gene regulatory network is shown on Fig. 1. From insulin data, we specified the interaction between genes (we take 35 genes) which is in regard to interactions transcription factors - target genes.

VI. CONCLUSION

Besides the amount of microarray data sets, the reconstructing of gene regulatory networks is still a hard and challenging problem. Bayesian networks especially dynamic BNs are powerful tools which provides elucidation of interaction among genes. The main shortcoming of utilized Bayes Net Toolbox and Bayesian Network Structure Learning is their limitation to cope with networks with small number of nodes, especially for structure learning. For large gene networks with hundreds and thousands of genes, these tools are not advisible.

Future tools should be able to deal with large gene networks.

REFERENCES

[1] X.-wen Chen, G. Anantha and X. Wang, An effective structure learning method for constructing gene networks, *Bioinformatics*, Vol. 22 no. 11, 2006, pp 1367-1374.

[2] S. Qing Ye, *Bioinformatics A Practical Approach*. Chapman & Hall/CRC, Taylor &Francis Group, Boca Raton, 2008.

[3] P. Baldi and S. Brunak, *Bioinformatics:The Machine Learning Approach*.The MIT Press, Cambridge, 2001.

[4] H. Lahdesmaki et *al*., On Learning Gene Regulatory Networks Under the Boolean Network Model, *Machine Learning*, 52, 147-167, 2003.

[5] T. I. Lee et *al*., Chromatin immunoprecipitation and microarray-based analysis of protein location, *Nature protocols*, Vol.1 No.2, 2006.

[6] N. Friedman and M. Goldszmidt, Learning Bayesian Networks with Local Structure, *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, 1998.

[7] L. Kaderali and N. Radde, Inferring Gene Regulatory Networks from Expression Data, *Computational Intelligence in Bioinformatics 2008*, 33-74, 2008

[8] G. F. Cooper and E. Herskovits, A Bayesian Method for the Induction of Probabilistic Networks from Data, *Machine Learning*, 9, 309-347, 1992.

[9] W. Zhao et *al*. Inferring gene regulatory networks from time series data using the minimum description length principle, *Bioinformatics*, Vol. 22 no. 17 2006, pp. 2219-2135.

[10] K. Murphy, Bayes Net Toolbox for Matlab, www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html

[11] D. Eaton and K. Murhy, Bayesian Network Structure Learning - A Software Package for Matlab, www.cs.ubc.ca/~deaton/struct/bnsl.html

[12] B. Ristevski and S. Loskovska, Modeling of Gene Regulatory Networks by Boolean Networks, Proceedings of the 11th International Multiconference, Information Society – IS 2008, 97-100, pp. 2008.