A survey of models for inference of gene regulatory networks

Blagoj Ristevski

Faculty of Administration and Information Systems Management
St. Kliment Ohridski University – Bitola
Ul. Partizanska bb. 7000, Bitola, Republic of Macedonia
blagoj.ristevski@uklo.edu.mk

Received: 18 September 2012 / **Revised:** 10 June 2013 / **Published online:** 25 September 2013

Abstract. In this article, I present the biological backgrounds of microarray, ChIP-chip and ChIP-Seq technologies and the application of computational methods in reverse engineering of gene regulatory networks (GRNs). The most commonly used GRNs models based on Boolean networks, Bayesian networks, relevance networks, differential and difference equations are described. A novel model for integration of prior biological knowledge in the GRNs inference is presented, too. The advantages and disadvantages of the described models are compared. The GRNs validation criteria are depicted. Current trends and further directions for GRNs inference using prior knowledge are given at the end of the paper.

Keywords: gene regulatory networks, gene expression, transcription factors, reverse engineering, GRNs validation.

1 Introduction

Many biological, physiological and biochemical molecular processes occur simultaneously in the cells. Regulation of these processes is performed by inherited information contained in the organisms genome. Inference of the mutual interactions between numerous components of biological systems based on available experimental data on DNA, RNA, proteins and metabolites interactions is needed for clearer representation and understanding of the regulatory mechanisms. These components and their mutual interactions compose complex networks named as *gene regulatory networks* (GRNs). There are two approaches for modeling of GRNs [1]:

- Mechanistic (or physical) networks, which use data from protein-DNA and protein-protein interactions and therefore they are usually dubbed transcription or protein networks. The goal of these static networks is to uncover molecular interactions on physical level.
- Influence networks, which refer to the reconstruction of GRNs based on gene expression data and the inferred networks refer to gene-gene interactions.

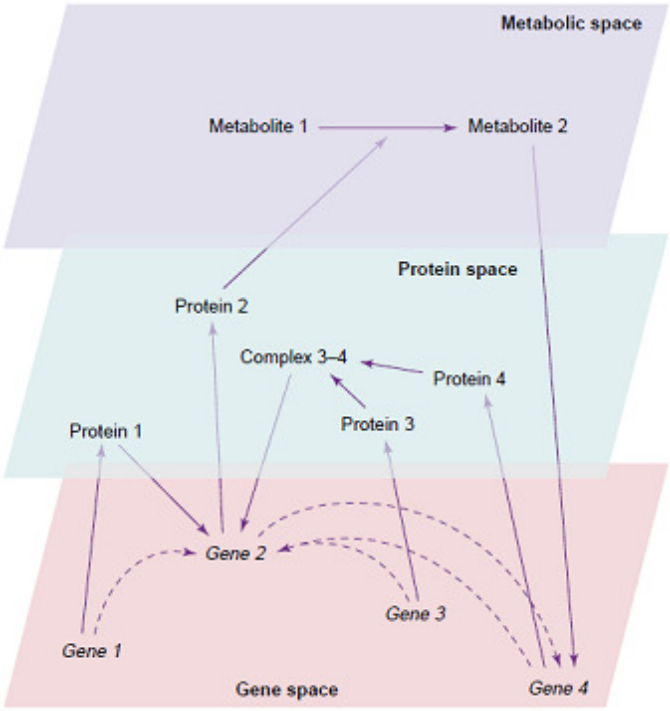


Fig. 1. Projection of GRNs in different spaces [2].

The GRNs structure is depicted by graph consisted of *nodes* representing the genes, proteins, metabolites, their complexes or even modules and *edges* which represent direct or indirect interactions between nodes. Proteins and metabolites appear as hidden variables and GRNs are inferred only from gene expression data as observable variables. These hidden variables can model unobserved effects that cannot be measured. Fig. 1 presents the projection of interactions from the space of metabolites and proteins in genes space. Dashed lines represent gene interactions and the full lines represent the interactions among genes, proteins, metabolites and their complexes [2].

This paper systematizes the different inferring GRNs models such as Boolean, Bayesian, dynamic Bayesian and relevance networks and other models, and compares their advantages and disadvantages.

This article is organized as follows. In Section 2, the biological and experimental backgrounds of the gene regulation are presented. In Section 3, the computational backgrounds of the inference of GRNs are described. Section 4 is devoted to description of Boolean networks, Bayesian networks and dynamic Bayesian networks, differential and difference equations system models and association networks and their usage for GRNs inference. In addition, several other models are briefly described in this section. Current trends and a new-proposed model for GRNs inference by integration of prior knowledge

is described in Section 5. Section 6 describes the commonly used validation criteria for validation of the GRNs models. In Section 7, GRNs models comparison based on several different attributes is shown. Finally, in Section 8, the concluding remarks and future works in the GRNs inference are given.

2 Biological and experimental backgrounds

Genes are fundamental physical and functional inheritance units of every living organism. The coding genes are templates for synthesis of proteins. Other genes might specify RNA templates as machines for production of different types of RNAs.

The process in which DNA is transcribed into mRNA and proteins are produced by translation represents the well-known central dogma in molecular biology. The first stage is transcription, then the second stage – translation of mRNA into a sequence of amino acids that compose the protein. When a protein is produced, the corresponding coding gene is expressed.

The gene expression levels indicate the approximate number of produced RNA copies from corresponding gene, which means that gene expression level corresponds to the amount of produced proteins. DNA microarray technology is used to obtain gene expression data experimentally.

One of the most important regulatory functions of proteins is transcription regulation. Proteins, which bind to DNA sequences and regulate the transcription of DNAs and gene expression, are called transcription factors (TFs). TFs can inhibit or activate gene expression of the target genes [3].

Besides gene expression data, other data such as protein-DNA, protein-protein interaction data and microRNAs should be considered for revealing gene regulatory mechanisms.

Only a small part of RNAs is coding RNAs whereas the bigger part from genome of eukaryotes is transcribed into non-coding RNAs. In the last few years, several small non-coding RNAs such as microRNAs and siRNAs are revealed [4]. The length of nucleotide thread in microRNAs is about 18–25 nucleotides [5]. MicroRNAs cause transcription cleavage or translation repression by connecting to their target mRNA [6]. MicroRNAs regulate expression by more than 30% of coding genes [7, 8]. Beside TFs, microRNAs are in mutual interaction with more cis-regulatory elements. Similarly to TFs, genes also contain binding sides for other TFs that may be targeted by microRNAs. Thus, the mutual influence between microRNAs and TFs makes microRNAs important components in the gene regulation.

They might have activate or inhibitory effect on gene expression, although earlier it was supposed that they might have only the role of inhibitors. MicroRNAs have an important role in many diseases such as cancer, cardiovascular, neurological, rheumatic, infectious and metabolic diseases [7–10]. Ripoli et al. had proposed fuzzy logic approach to reveal microRNAs role to gene expression regulation [11].

To discover the transcription factor binding sites (TFBSs) locations on the genome for particular proteins and to reveal protein-DNA interactions, chromatin immunoprecipitation (ChIP) is used [12]. ChIP-chip technology uses ChIP with hybridization microarrays

(chip) to identify the protein binding sides and their locations throughout the genome. In ChIP-chip technology, short DNA sequences as probes are used. A population of immunoprecipitation – enriched DNA fragments is produced and enrichment of each probe from produced population is measured [13]. ChIP-Sequencing (ChIP-Seq) technology, unlike ChIP-chip technology, uses secondary sequencing of DNA instead of microarray [12].

The integration of abovementioned different types of biological data can significantly improve the inference of GRNs [14].

3 Computational backgrounds of GRNs inference

Theoretical studies of GRNs have started in the 1960s. The appearance of experimental technologies for studying mechanisms that regulate gene expression such as DNA microarrays, ChIP-chip and ChIP-Seq has provided large amounts of gene expression, protein-protein and protein-DNA interaction data. Because the experimental wet-lab technologies cannot measure mutual influences among all genes from one organism's genome simultaneously, computational methods are applied to infer and reveal mutual gene interactions.

In the past decade, several models for GRNs inference have been developed, based on the basic reverse engineering methods. However, these models work only with certain data types and inferred networks do not largely match the real regulatory mechanisms. This shortcoming is a motivation for developing of new models that can include prior knowledge and would be able to integrate heterogeneous data. Such inferred GRNs could depict gene regulatory mechanisms more accurately.

Numerous models such as Boolean networks, Bayesian networks, dynamic Bayesian networks, graphical Gaussian models, Petri networks, linear and nonlinear differential and difference equations, information theory approach, state space models, fuzzy logic models and many other models are used to reconstruct GRNs.

Finding more accurate and reliable GRNs structures from gene expression data is a problem of machine learning known as structure learning of graph models. GRNs learning is a big challenge that merges learning techniques from artificial intelligence with statistics, bioinformatics and functional genomics.

By GRNs inference, several properties of GRNs should be considered such as sparseness, scale-free topology, modularity and structurality of inferred networks [1].

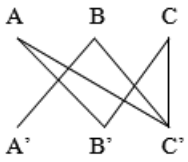
The GRNs should be sparse. In other words, a limited number of genes regulates genes. Some genes in the network called "hubs" can regulate many genes, i.e. the outdegree of the nodes is not limited. Another important feature is the scale-free GRNs topology. Scale-free networks have the power distribution function of the connectivity degree [15]. This property provides the robustness of the networks regarding the random topology changes. Structures with small connectivity follow the regulatory hierarchy. The structurality allows network decomposition into basic modular elements composed of several genes, called network motifs [1]. The network modularity refers to the existence of clusters of highly coexpressed genes and genes with similar function.

4 Gene regulatory networks models

4.1 Boolean networks

One of the simplest models of GRNs is the model based on Boolean networks. The genes are represented by nodes and the edges between nodes representing the interactions between genes. In Boolean networks, gene expression levels are discretized and presented by two-states levels. The state of the genes that have expression levels above a certain threshold is 1, otherwise 0.

The wiring diagram shown in Fig. 2(a) presents connection between genes, but it is not sufficient for understanding logical dependencies between genes. The aim of the reverse engineering in Boolean networks is to find Boolean functions of every gene in the network, so discretized values of gene expression can be explained by the model, shown in Fig. 2(b). Another way of representing Boolean networks is by state transitions table, presented in Fig. 2(c).



(a)

A'=B

B'=A or C

C'= (A and B) or (B and C) or (A and C)

(b)

inputs			outputs		
A	B	C	A'	B'	C'
0	0	0	0	0	0
0	0	1	0	1	0
0	1	0	1	0	0
0	1	1	1	1	1
1	0	0	0	1	0
1	0	1	0	1	1
1	1	0	1	1	1
1	1	1	1	1	1

(c)

Fig. 2. A Boolean network represented by: (a) a wiring diagram, (b) Boolean functions and (c) a state transition table.

The small changes in gene expression time series data cannot be covered by two-level discretization, because it leads to information loss. Thus, inferred regulatory networks can be unrealistic. Another shortcoming of Boolean networks is the super-exponential number of all possible networks. If n is the number of genes, then the number of Boolean functions is super-exponential and equal to 2^{2^n}.

Several extended models based on Boolean networks have been proposed. A REVerse Engineering ALgorithm (REVEAL) constructs a Boolean network of given expressed gene data by setting the in-degree value of genes k [16]. The algorithm extracts minimal network structures by using the mutual information approach from the state transition tables of the Boolean network. If n is the number of nodes, the number of all possible networks can be calculated by:

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

REVEAL can be applied to gene expression data, discretized on multiple discretization levels. On the other hand, multiple discretization levels increase the number of possible state transitions. Thus, the number of all possible networks will be much greater than the number of networks derived from two level Boolean networks and it is calculated by Eq. (1). REVEAL has better inference capabilities when the value of in-degree k is smaller. For greater in-degree k, it is necessary to perform parallel processing or to increase the efficiency of the search space of possible networks [16].

The models based on Boolean network simplify the structure and dynamics of gene regulation. They are deterministic, i.e. the state space is limited and that the networks reach the steady state or enter into dynamic attractor [17]. The inferred networks provide only a quantitative measure of gene regulatory mechanisms.

Another model is the probabilistic Boolean networks model [18]. This model can be considered as a model composed of several Boolean networks, which work simultaneously, but all networks share information about the whole system states. When a network transits to a next state, the remaining networks are synchronized.

4.2 Bayesian networks

Bayesian networks (BNs) are among the most effective models for GRNs inference. A Bayesian network is a special graph model defined as a triple (G, F, q), where G denotes the graph structure, F is the set of probability distributions and q is the set of parameters for F [19]. The graph structure G is consisted of a set of n nodes $X_1, X_2, \ldots, X_n$ and a set of directed edges between nodes. The nodes correspond to the random variables and directed edges show the conditional dependences between the random variables.

If there is a directed edge from the node X to the node Y, which is denoted as $X \to Y$, then X is a parent of Y, denoted as $pa(Y)$, and Y is a child of X. If the node Z can be reached by following a directed path starting from node X, then the node Z is a descendant of X, and X is ancestor of Z. Nodes and edges together have to make a directed acyclic graph (DAG). One directed graph is acyclic if there is no directed path

$X_1 \to X_2 \to \cdots \to X_n$ such as $X_1 = X_n$, i.e. there is no pathway that begins and ends at the same node.

The joint probability distribution of all nodes is calculated by the following equation:

$$\mathbf{P}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}\big(X_i \mid pa(X_i)\big). \tag{2}$$

The multipliers from Eq. (2) are called local probability distributions. The factorization of the joint probability distribution on multipliers provides its easier computation as a product of simpler conditional probability distributions.

Structure and parameter learning accompany the inference of GRNs. The aim of structure learning is finding network structure that fits best the real regulatory interactions. However, in these networks also, there is a super-exponential dependence of the number of possible DAGs on the number of nodes n. For a given network structure, the parameter learning includes estimation of the unknown model parameters for each gene. This learning is performed by determining of the conditional dependencies between network nodes. Because the BNs inference is an NP-hard problem, BNs are the most suitable when they are applied to small networks consisted of tens to hundred genes [20].

It is possible to infer GRNs by BNs based on static, dynamic, discrete or continuous gene expression data. If the node states are continuous, then network inference is more difficult to carry out because of the additional complex calculations.

BNs can deal with stochastic nature of gene expression data and incomplete and noisy data, too. The main problem with Bayesian networks learning is the higher number of genes compared to the number of conditions and incapability to capture feedback loops that exist in the real GRNs.

Friedman et al. in [21] have introduced a framework for discovering of interactions between genes based on microarray data using BNs. This method models each variable with conditional probability distribution function related to other variables. In the proposed approach, two comparative experiments are conducted for different probability distributions: multinomial distribution and linear Gaussian distribution. The main shortcomings of this model are search heuristics performed without constraints on the search space and non-using prior biological knowledge.

4.2.1 Dynamic Bayesian networks

BNs can represent probabilistic relations between variables without time lags and their drawback is that they cannot deal with time series data [22]. However, interactions in the real GRNs do not occur simultaneously, so there is a particular time lagging.

Another disadvantage of BNs is that they cannot represent real biological systems, where exists mutual interactions among entities of biological systems, i.e. feedback loops that exist among genes in the GRNs [23].

These shortcomings make BNs inappropriate for GRNs inference from time series gene expression data, where it is necessary to include dynamic (temporal) features of gene regulation. Thus, BNs are extended to model time features by introduction of dynamic

Bayesian networks (DBNs). It is assumed that the changes in time series gene expression data occur in a finite number of discrete intervals T. Let $X = \{X_1, X_2, \ldots, X_n\}$ is a set of time dependent variables and $X_i[t]$ is a random variable representing the value of X_i at the time point t and $0 \leqslant t \leqslant T$. A DBN is a Bayesian network that contains the T random vectors $X_i[t]$ [24], an initial BN, a transition BN consisted of transition DAG $G_{\rightarrow}$ and transition probability distribution $P_{\rightarrow}$:

$$P_{\rightarrow}\big(X[t+1] = x[t+1] \;\big|\; X[t] = x[t]\big). \tag{3}$$

The joint probability distribution of the DBN is computed by:

$$\mathbf{P}\big(x[0], \ldots, x[T]\big) = P_0\big(x[0]\big) \prod_{t=0}^{T-1} P_{\rightarrow}\big(x[t+1] \;\big|\; x[t]\big). \tag{4}$$

From Eq. (4) for each x at each time point t, the following is obtained:

$$\mathbf{P}\big(x[t+1]x[0], \ldots, x[t]\big) = \mathbf{P}\big(x[t+1] \;\big|\; x[t]\big). \tag{5}$$

Equation (5) denotes that the value of the variables at time point t depends on the values of variables at the moment $t-1$ and other information is not required, i.e. the processes described by DBNs have Markov property [25].

For probabilistic inference in DBNs, the standard algorithm used in BNs inference can be used, too. However, in the case of large time series data, DBNs learning becomes too complex.

The DBNs are effective for GRNs inference when they are combined with other types of biological data. An example for that is the proposed method that integrates gene expression data with prior biological knowledge about TFBSs using DBNs and structural EM algorithm [26].

It is shown that high order DBNs can be used for modeling of time lag gene regulatory interactions based on time series gene expression data [27].

Figure 3 shows a DBN that describes cyclic regulation between gene 1 and gene 2, although the graph does not contain obviously cyclic pathway.

In [28] a manner how DBNs can be applied for network inference and how they can be learned; their relationship with the HMM, Boolean and stochastic Boolean networks and DBNs with continuous variables is shown. The Boolean networks, linear and nonlinear equations models can be considered as a special case of DBNs.

To overcome the high complexity needed for GRNs inference by DBNs, a model with constraint of the potential regulators has been proposed. This constraint considers those genes that have changes in gene expression level at the previous or at the same time points regarding their target genes [29]. The proposed model uses the time lag of changes in expression levels at regulator and target genes, which increases the accuracy of the inferred networks. The time points of initial over- or under- regulation of the genes are determined. The genes with changes in previous and current time points are denoted as potential regulators to those genes with expression changes in the following time points. In such a way, a subset of potential regulators for every target gene is chosen.

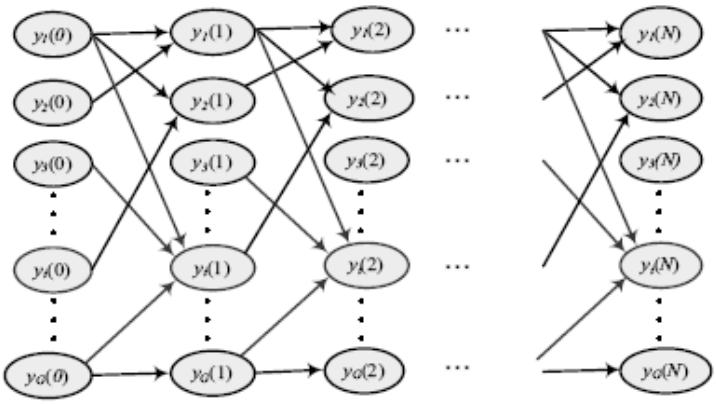


Fig. 3. A GRN inferred by DBNs based on the input time series gene expression data for G genes and N time points.

An effective algorithm for structure learning – extended K2 algorithm is proposed. This algorithm called KN algorithm serves for learning of large BNs [30]. KN algorithm introduces sorting of the genes to improve the efficiency of the large network inference. The examination of the efficiency of the proposed algorithm is performed by carrying out Monte Carlo simulation regarding to the greedy hill-climbing algorithm.

Another algorithm BOLS (Bayesian Orthogonal Least Squares) for reverse engineering of GRNs is proposed in [31]. This algorithm combines the orthogonal least squares, second-order derivatives for network pruning and Bayesian model. The obtained network is sparse, in which limited number of genes regulates every gene and the number of false inferred edges is small.

4.3 Differential and difference equations models

The concentration of RNAs, proteins and other metabolites changes over time. Therefore, to describe gene regulation, differential equations might be an appropriate model [32]. Ordinary differential equations (ODEs) systems use continuous gene expression data directly and can easily model positive and negative feedback loops.

The main constraint of the model based on ODEs is the assumed constant or linear changes of the concentration of regulators, although there are actually non-linear time changes.

The dynamics of gene expression data is presented by the following differential equation:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x, p, u, t), \tag{6}$$

where $x(t) = (x_1(t), x_1(t), \ldots, x_n(t))$ is a vector of gene expression data for n genes at time t, f is the function that describes the changes of variables x_i depending on the model

parameters p and external perturbations u. The aim of GRNs inference is to determine the function f and parameters p given the measured signals x and u at the time t [1].

By solving Eq. (6), more solutions can be obtained, so the structure and parameters identification of the model requires identification of the function f based on prior knowledge or approximations. The function f can be linear or nonlinear. Although the function f is nonlinear, for simplification, it is assumed that is linear and Eq. (6) is transformed into the following equation:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_{j=1}^{N} w_{ij}x_j + b_i u, \quad i = 1, \ldots, N, \tag{7}$$

where w_{ij} are elements of weight matrix W, and parameters b_i determine the external disturbance u to gene expression. This model also is called a model of regulatory matrices composed of weight coefficients w_{ij}, which interpret the regulatory dependences. If the weight coefficient has positive value, then the corresponding gene has activating role and if the weight coefficient is negative, then that gene has the role of inhibitor. If the weight coefficient value is zero, then genes do not interact mutually. In the linear models, the inference from small number of samples is easier to carry out.

The identification of the function f and the parameters in the nonlinear models is difficult because the number of samples in gene expression data is smaller than the number of genes and finding the numerical solutions is more difficult.

One way to model the changes of gene expression is by S-systems with activating and inhibitory components, described by:

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = \alpha_i \prod_{j=1}^{N} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{N} X_j^{h_{ij}}, \tag{8}$$

where α_i and β_i are positive constants and h_{ij} and g_{ij} are kinetic exponents [1]. In these models, there are many parameters whose identification requires carrying out numerous experiments, and therefore often approximations are made by differential equations.

An optimized model for GRNs inference that uses known biological prior knowledge from available databases for genome, proteome, transcriptome and scientific publications has been proposed in [33]. This model is based on differential equations, from which particular solutions are obtained by singular values decomposition. The obtained result is optimized using mathematical programming.

One special case of differential equations system is the model of pairwise linear differential equations, proposed in [34]. This model is based on the assumption that the gene regulation can be represented by pairwise linear equations. The model uses gene expression data and neglects posttranscriptional regulation.

Beside ODEs, difference equations model for GRNs inference is used, too. Unlike the differential equations models that deal with continuous variables, the variables in the difference equations model are discrete. Discretization of the gene expression data leads to information loss [32]. However, difference equations are more suitable when time series

gene expression data are available. The changes of gene expression data are described by the following equation:

$$\frac{x_i(t+\Delta t)-x_i(t)}{\Delta t}=\sum_{j=1}^{N} w_{ij}x_j(t)+b_iu, \quad i=1,\dots,N. \tag{9}$$

Difference equations model can be reduced to a system of linear algebraic equations, which can be solved by linear algebra methods [1].

4.4 Association networks

Association networks (or relevance networks) are static networks that can describe the possible structure of the GRNs. They can be applied for steady-state and time series gene expression data. Association networks are represented by undirected graph. If two genes are connected by an edge, then it is not possible to determine which gene is regulated and which one is regulating. To determine which genes are coexpressed and between which genes should be an edge, it is necessary to apply similarity metrics such as Pearson coefficient or mutual information and additionally to set a threshold. The higher the threshold is, the sparser inferred GRN is.

Although the relevance networks do not determine the directions of the edges in the networks, they are suitable for inference of large GRNs because of their low computational complexity [1]. The directions of the regulations can be determined by computation of the similarity between genes and their possible regulators and with additional knowledge for TFs.

Proposed algorithm ARACNE is based on mutual information between gene expression data [35]. It defines the edges in the network as statistical dependences by which the directed regulatory interactions using data for TFs and their TFBSs can be identified. Using ARACNE, the number of falsely predicted gene interactions in the networks reduces significantly. The complexity of this algorithm is $O(N^3+N^2M^2)$, where N is the number of genes, and M is the number of samples. The low complexity makes this algorithm to be suitable for inference of large GRNs [35].

Graphical Gaussian models (GGMs) use partial correlation coefficients to determine the conditional dependencies between genes and can determine directed and undirected edges in the network [36]. GGMs can distinguish direct or indirect interactions between genes, unlike the correlation networks where the edges present correlation between genes.

Let X is a gene expression data matrix with n rows and p columns, where n is the number of experimental conditions and p is the number of genes. The data from the matrix X are assumed to follow the normal distribution $N_P(\mu,\Sigma)$, where $\mu=(\mu_1,\dots,\mu_p)^{\mathrm{T}}$ is the vector of means and $\Sigma=(\sigma_{ij})_{1\leqslant i,j\leqslant p}$ is a positive definite covariance matrix. By decomposition of the matrix Σ, two parts are obtained: variance components σ_i^2 and Brevis–Pearson correlation matrix $P=(\rho_{ij})$. The partial correlation matrix $Z=(\zeta_{ij})$ is composed of the correlation coefficients between any two genes i and j with respect to all other genes. The matrix Z is related to the inverse matrix P of the standard correlation

coefficients. Their relationship is computed by the following equations [37]:

$$\Omega = (\omega_{ij}) = P^{-1} \tag{10}$$

and

$$\zeta_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}. \tag{11}$$

In Eq. (10) the covariance matrix Σ can be used instead of correlation matrix P. Partial correlation coefficients ζ_{ij} are correlation coefficients of conditional bivariant normal distributions of the genes i and j. Two variables distributed by the normal distribution, are conditionally independent if and only if their partial coefficients are zeros. The conditional independence of the random variables is determined by the zeros in the inverse correlation matrix Ω.

To infer a GRN by the GGMs from data set, the correlation matrix P is estimated by unbiased sampling of the covariance matrix:

$$\hat{\Sigma} = (\hat{\sigma}_{ij}) = \frac{1}{n-1}(X - \bar{X})^{\mathrm{T}}(X - \bar{X}) \tag{12}$$

The estimation of partial correlation coefficients is computed by Eq. (10) and Eq. (11) from sample correlation matrix. The elements from estimated correlation matrix $\hat{Z}$, which differentiate from zeros, are determined by statistical tests. The network inference ends with a visualization of correlation structure by a graph, whose edges correspond to nonzero partial correlation coefficients.

The main disadvantage of the described classical GGMs is that they can be applied when the number of experimental conditions n is greater than the number of genes p, because they cannot calculate the partial correlations. Also the existence of an additional linear dependence between the variables leads to multicollinearity, and the commonly used statistical tests for GGMs selection are valid only for data with large number of samples [36]. In the case when $p > n$ covariance matrix is not positive definite, so its inverse matrix cannot be found.

Therefore, an estimation of the covariance matrix is performed by shrinkage estimators to obtain positive definite covariance matrix, and thus its inverse matrix could be found [38, 39]. The present edges in the graphs are determined by model selection of the network graphs.

4.5 Other models for GRNs inference

Besides the above mentioned models, numerous models for GRNs reconstruction are proposed.

The Collateral-Fuzzy Gene Regulatory Network Reconstruction (CF-GeNe), proposed by Sehgal et al., applies collateral assessment of the missing values [40]. This model uses the fuzzy nature of gene coregulation, which is determined by fuzzy c-means clustering algorithm. This clustering algorithm allows genes to belong to several clusters, i.e. biological processes. CF-GeNe can deal with missing values, noisy data and it determines the optimal number of clusters.

Fujita et al. have proposed a model of GRNs using sparse autoregressive vector in [41]. This model can infer the gene regulations in case when the number of samples is less than the number of genes without using prior knowledge and it can handle the feedback loops.

The linear model in the finite state space infers gene regulations including discrete and continuous aspects of the gene regulation [42]. This model assumes that gene activity is determined by the state of the TFBSs, each binding sites can be located in one of the final number of states, genes may be inhibited or they can have some activity and the state of the binding sides depends on the TFs concentration.

Li et al. had proposed another model that uses the state space with hidden variables for the GRNs reconstruction [43]. This model is dynamic and consists of observations and states. The observations $(O_1, O_2, \ldots, O_T)$ are generated from the state $(S_1, S_2, \ldots, S_T)$ according to the formal model:

$$S_t = AS_{t-1} + W_t, \qquad O_t = BS_t + V_t, \tag{13}$$

where A denotes the transition state probability $P(S_t \mid S_{t-1})$ from state at time $t-1$ to t, B denotes the probability $P(O_t \mid S_t)$ of observation to be determined by the state in the same time point. The W_t and V_t represent the disturbances of the states and the observations, respectively. This model can be considered as a subtype of DBNs. The hidden variables include the regulatory motifs such as feedback loops and auto-regulation, thus this model gives a significant contribution to the existing models.

A qualitative model for GRNs reconstruction that uses Petri networks is proposed in [44]. This model, which is based on Boolean networks, uses minimization logic to transform Boolean rules into Petri networks. It overcomes the super exponential number of states in the Boolean networks depending on the number of nodes.

An optimized model for GRNs reconstruction based on differential equations has been proposed. This model includes prior knowledge and it is suitable for inference of small networks [33].

For hierarchical reconstruction of GRNs, Lee and Yang had proposed a model that uses the clustering of gene expression data in [45]. It provides inference of regulatory mechanisms for large-scale networks. This method uses the recurrent neural networks to represent GRNs and applies the learning algorithm to update the important network parameters in discrete time steps.

Another method called FBN, applies the clustering of gene expression data for obtaining modules (clusters) and infers the gene regulations between clusters [46]. This method uses fuzzy clustering to reduce the search space for Bayesian networks learning.

In [47], a feature dependency analysis across samples is performed in order to determine regulators (miRNAs and TFs) that significantly describe common and subtype-specific gene expression changes. To rank subtype-specific features, a score based on increase in squared loss on samples, which belong to a subtype excluding the regulator from the learned model, is used.

Liao et al. had developed a data decomposition method – NCA (network component analysis) for reconstruction of regulatory signals and control strengths using partial and qualitative network connectivity information [48]. This method is applied to transcription regulatory network.

5 Current trends – GRNs inference by integration of prior knowledge

The GRNs inference based on gene expression data is a very complex and difficult task, particularly because the present biological and technical noise in microarray data should be considered. In addition, the number of experiments or conditions is less than the number of genes whose expressions are measured. Such shortcomings of the microarray data lead to insufficient precision and accuracy of inferred networks. To increase the accuracy and precision, application of other types of biological data and prior knowledge such as knowledge obtained from scientific papers, protein-DNA interactions data and other available databases is needed [49, 50]. Biology capabilities to elucidate complex systems come from the extended power to include prior knowledge and complementary and various data types [51].

The method suggested by Li in [52] combines qualitative and quantitative biological data for prediction of GRNs. This method uses parallel processing and multiprocessor system to speed up the structural learning of Bayesian networks.

Based on comparison of the inference capabilities in [51,53], Ristevski and Loskovska in [54] have suggested a novel model for GRNs inference which performs in two stages. They have chosen the GGMs in the first phase of the proposed model, because they are a good base for uncovering the "hub" genes. The GRNs structure G can be represented by an adjacency matrix. The adjacency matrix entries G_{ij} can be either 1 or 0, which refers to the presence or absence of a directed edge between ith and jth node of the network G, respectively. As a result of the first phase of the proposed model, a matrix of prior knowledge $Gprior$ is obtained, whose elements are computed by:

$$Gprior_{ij} = \begin{cases} \frac{1}{2}\frac{|pcor_{ij}|-pcor_{\min}}{pcor_{\max}-pcor_{\min}} + \frac{1}{2}, \\ 0 \quad \text{if } |pcor_{ij}| < pcor_{\min} \text{ or edge direction is from } j \text{ to } i, \end{cases} \tag{14}$$

where $pcor_{\max}$ and $pcor_{\min}$ are the minimum (set threshold) and maximum partial correlation coefficient, respectively [54]. The obtained matrix of prior knowledge $Gprior$, whose entries $Gprior_{ij} \in [0, 1]$, presents a basis for the second phase of the proposed model.

To integrate the prior knowledge obtained from first phase, the second phase defines a function $Gprior'$ as a measure of matching between the given network G and the obtained prior knowledge $Gprior$ [50]. The integration of prior knowledge $Gprior$ is according to the prior distribution of the network structure G, which follows Gibbs distribution, given by the following equation [49, 50]:

$$\mathbf{P}(G\beta) = \frac{\mathrm{e}^{-\beta Gprior'(G)}}{Z(\beta)}, \tag{15}$$

where the denominator is normalization constant calculated from all possible network structures Γ by the formula $Z(\beta) = \sum_{G\in\Gamma} \mathrm{e}^{-\beta Gprior'(G)}$. In the second phase of the proposed model, a structure Bayesian learning is carried out using Markov chain – Monte

Carlo simulations [54]. This model has shown even better capabilities of GRNs inference, compared to Boolean networks, DBNs and GGMs in the case when it was applied on simulated datasets, as well as experimental data sets.

Beside gene expression data, the availability of heterogeneous -omics data (transcriptomics, proteomics, interactomics and metabolomics), makes the network inference to become more flexible. Various -omics data reveal different perspectives of regulatory networks. Integration of these data and using prior knowledge can discover a more reliable comprehension of the regulatory mechanisms. Hence, integration of heterogeneous data and prior knowledge still remains challenging and partially unsolved topic in the inference of regulatory networks.

6 Model validation

The validation of inferred GRNs represents an assessment of the quality of the inferred network, compared to the available knowledge in so-called "gold standard" networks such as TRANSFAC [55] and JASPAR [56]. To validate inferred gene regulatory interactions using computational models, wet-lab biological experiments are needed. Commonly used validation criteria are receiver operating characteristic (ROC) curve and area under the ROC curve (AUC).

ROC curves are applied in the GRNs reconstruction for validation of inferred networks. In a graph between two nodes, it might be an edge or it might be no edge, or expressed by the formalism of machine learning, each edge (instance) of the network belongs to either positive (p) or negative (n) class, and classifier outcomes belong to either class p or class n [57, 58].

For a given two-class classifier and test samples, four cases are possible:

- TP (true positive), if the instance is positive and it is classified as positive;
- FN (false negative), if the instance is positive and it is classified as negative;
- TN (true negative), if the instance is negative and it is classified as negative and
- FP (false positive), if the instance is negative and it is classified as positive.

The following rates are defined based on the defined TP, FN, TN and FP rates [59, 60]:

$$tpr = \frac{TP}{P} = \frac{TP}{TP + FN} \quad \text{true positive rate (recall)}, \tag{16}$$

$$fpr = \frac{FP}{N} = \frac{FP}{FP + TN} \quad \text{false positive rate}, \tag{17}$$

$$precision = \frac{TP}{TP + FP}, \tag{18}$$

$$accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FN + FP + TN}. \tag{19}$$

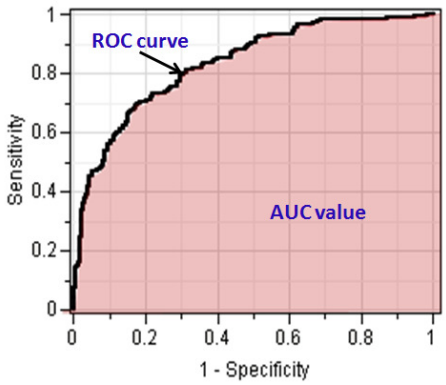


Fig. 4. The ROC curve and AUC value.

ROC curve is a plot of a function where on the x-axis the false positive rate (fpr) – and on the y-axis – the true positive rate (tpr) are applied (Fig. 4). The ROC curve represents the ratio between sensitivity and $(1 - specificity)$ [61]. When the ROC curve is above the line $y = x$, the classification is better.

To facilitate comparison of inference capabilities, instead of ROC curve the area under ROC curve (AUC) can be used. The AUC is the area covered by the ROC curve with the x-axis, shown on Fig. 4. The statistical meaning of the AUC corresponds to the probability that the classifier will rank a randomly selected positive instance higher than a selected negative instance [57].

7 Models comparison

To provide comprehensive description about inference capabilities of GRN models I present a summary of the described models in the previous sections. Models are compared according to several important model attributes. Table 1 systematizes the advantages and disadvantages of the above-described models for GRNs inference. Despite the advantages and drawbacks of described models, inferred network edges, which are not present in the regulatory databases, are indications for further experimental investigation to confirm their presence/absence as true regulatory mechanisms.

Table 1. Atributes comparison between models for GRNs inference.

Model	Advantages	Disadvantages
Boolean network [17]	simplicity; enable analysis of large networks	two-state model; information loss
REVEAL [16]	multi-state model	low number of genes; greatly increase the number of possible state transition

(*continued on next page*)

Table 1. (*Continued.*)

Model	Advantages	Disadvantages
Bayesian network [19, 20]	describe interactions between genes; deal with noisy data	cannot deal with time-series data and feedback regulations; NP-hard learning problem
DBN [24]	time-series data; hidden variables; can use prior knowledge; deal with missing data; continuous and discrete states; stochastic networks; large scale data	accuracy depends on number of selected genes and sampling of time series data; excessive computational time
DBN [29]	increased prediction accuracy; decreased computational time	requires more prior information of the transcription regulation
Optimization model [33]	works with prior knowledge and time-series data; finds solution with biological plausibility and reliability	assumes linearity; small number of genes
Model of GRNs with the sparse vector autoregressive model [41]	can infer GRNs when the number of samples is lower than number of genes without any prior knowledge; deals with feedback loops	reconstructs medium scale GRNs
CF-GeNe [40]	cope with noisy data and missing values	works with clusters obtained from fuzzy c-means clustering
Finite state linear model [42]	combines the discrete and continuous aspects of gene regulation; continuous time	finite state model formalism
State-space model [43]	probabilistic framework to simulate GRNs; hidden variables; determines an optimal threshold value for discretization of the expression data based on prior knowledge	does not include a step to learn the structure
Differential equations model [32]	good performance in case of small scale networks; great physical accuracy; can model negative feedback loops	small numbers of genes; parameters with unknown experimental values are required; difficult to describe non-additive logics in gene regulation; computational intensive
Difference equations model [32]	good performance in case of small scale networks	small numbers of genes; requires parameters with unknown experimental values; discrete model

Table 1. (*Continued.*)

Model	Advantages	Disadvantages
Algorithm for reverse engineering with BOLS [31]	reveals GRNs using limited number of experimental data points; deals with noisy data	does not provide the confident levels among interactions within unit network
GGMs [36]	can infer large GRNs	the number of inferred edges depends on the set threshold
Two stage model integrating prior knowledge [54]	integrates prior knowledge; very competitive even better inference capabilities for simulated and microarray gene expression data compared to other models; can deal with feedback loops	no suitable for large scale GRNs inference
Network component analysis (NCA) model [48]	locally accurate and computationally tractable; provides a very good fit to most of the microarray expression data	difficult to measure transcription factor activity (TFA); connectivity between genes and TFs is not atainable for all organisms

8 Conclusions and further works

This survey and comparison of the models for inference of GRNs has shown that there is still a need for development of models that can integrate the available biological prior knowledge and other data such as ChIP-chip, ChIP-Seq and microRNA data. As was shown in this overview, such knowledge significantly improves the accuracy of the inferred networks.

By validation of the inferred networks, the main problem is the lack of "gold standard" networks to which edges the presence/absence of inferred edges is confirmed. Inferred directed edges, which are not present in the available biological regulatory databases, should be clues for further experimental research to confirm their presence/absence as true regulatory interactions. Furthermore, greater efforts should be made toward upgrading of existing databases for regulatory mechanisms between genes, metabolites, proteins, their complexes and other components that take part in the gene regulation. In addition, to characterize the GRNs, combining gene expression microarray analysis and quantitative trait loci (QTL) mapping should be conducted.

References

1. M. Hecker et al., Gene regulatory network inference: Data integration in dynamic models – A review, *Biosystems*, **96**(1):86–103, 2009.

2. P. Brazhnik, A. de la Fuente, P. Mendes, Gene networks: How to put the function in genomics, *Trends Biotechnol.*, **20**(11):467–472, 2002.

3. N.A. Kolchanov et al.,Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes, *Briefings Bioinf.*, **8**(4):266–274, 2007.

4. Z. Shuang, L. Mo-Fang, Mechanisms of microRNA-mediated gene regulation, *Sci. China, Ser. C*, **52**(12):1111–1116, 2009.
5. G. Russo, A. Giordano, miRNAs: From biogenesis to networks, in: Y. Nikolsky, J. Bryant (Eds.), *Protein Networks and Pathway Analysis*, Methods in Molecular Biology, Vol. 563, Humana Press, New York, NY, 2009, pp. 303–352.

6. J.G. Joung, Z. Fei, Identification of microRNA regulatory modules in Arabidopsis via a probabilistic graphical model, *Bioinformatics*, **25**(3):387–393, 2009.

7. Z. Wang, *MicroRNA Interference Technologies*, Springer-Verlag, New York, 2009.

8. C. Li et al., Therapeutic MicroRNA strategies in human cancer, *AAPS J.*, **11**(4):747–757, 2009.

9. A. Drakaki, D. Iliopoulos, MicroRNA Gene networks in oncogenesis, *Current Genomics*, **10**:35–41, 2009.

10. S.K. Shenouda, S.K. Alahari, MicroRNA function in cancer: Oncogene or a tumor suppressor?, *Cancer Metastasis Rev.*, **28**:369–378 2009.

11. A. Ripoli et al., The fuzzy logic of MicroRNA regulation: A key to control cell complexity, *Current Genomics*, **11**(5):350–353, 2010.

12. P.J. Park, ChIP-seq: advantages and challenges of a maturing technology, *Nat. Rev. Genet.*, **10**(10):669–680, 2009.

13. R. Gottardo, Modeling and analysis of ChIP-Chip experiments, in: P. Collas (Ed.), *Chromatin Immunoprecipitation Assays. Methods and Protocols*, Methods in Molecular Biology, Vol. 567, Humana Press, New York, NY, 2009, pp. 133–143.

14. W. Zhao, E. Serpedin, E.R. Douherty, Recovering genetic regulatory networks from chromatin immunoprecipitation and steady-state microarray data, *EURASIP J. Bioinform. Syst. Biol.*, **2008**, 248747, 2008.

15. M. Nicolu, M. Schoenauer, On the evolution of scale-free topologies with gene regulatory network model, *Biosystems* **98**(3):137–148, 2009.

16. S. Liang, S. Fuhrman, R. Somogyi, REVEAL, a general reverse engineering algorithm for inference of genetic network architectures, *Pacific Symposium on Biocomputing*, **3**:18–19, 1998.

17. Y.K. Kwon, K.H. Cho, Analysis of feedback loops and robustness in network evolution based on Boolean models, *BMC Bioinformatics*, **8**, Article No. 430, 9 pp., 2007.

18. I. Shmulevich, E.R. Dougherty, W. Zhang, From Boolean to probabilistic Boolean networks as models of genetic regulatory networks, *Proc. IEEE*, **90**(11):1778–1792, 2002.

19. N. Friedman, M. Goldszmidt, Learning Bayesian networks with local structure, in: *Proceedings of the 12th International Conference on Uncertainty in Artificial Intelligence, Portland, OR, August 1–4, 1996*, pp. 252–262.

20. W.-Po Lee, K.-Cheng Yang, A clustering-based approach for inferring recurrent neural networks as gene regulatory networks, *Neurocomputing*, **71**:600–610, 2008.

21. N. Friedman et al., Using Bayesian networks to analyze expression data, *J. Comput. Biol.*, **7**:601–620.

22. D. Husmeier, R. Dybowski, S. Roberts (Eds.), *Probabilistic Modeling in Bioinformatics and Medical Informatics*, Springer-Verlag, London, 2005.

23. M. Grzegorczyk, D. Husmeier, Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move, *Mach. Learn.*, **71**:265–305, 2008.

24. R.E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, Upper Saddle River, NJ, 2003.

25. L.M. de Campos, A scoring function for learning Bayesian networks based on mutual information and conditional independence tests, *J. Mach. Learn. Res.*, **7**:2149–2187, 2006.

26. Y. Zhang et al. *Inferring Gene Regulatory Networks from Multiple Data Sources via a Dynamic Bayesian Networks with Structural EM*, Springer-Verlag, Berlin, Haidelberg, 2007, pp. 204–214.

27. A. Shermin, M.A. Orgun, Using dynamic Bayesian networks to infer gene regulatory networks from expression profiles, in: *Proceedings of the 2009 ACM symposium on Applied Computing (SAC '09), Honolulu, HI, March 8–12, 2009*, pp. 799–803.

28. K. Murphy, S. Mian, Modelling gene expression data using dynamic Bayesian networks, Technical Report, Computer Science Division, University of California, Berkeley, CA, 1999.

29. M. Zou, S.D. Conzen, A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data, *Bioinformatics*, **21**(1):71–79, 2005.

30. K. Numata, S. Imoto, S. Miyano, A structure learning algorithm for inference of gene networks from microarray gene expression data using Bayesian networks, *7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE), Boston, MA, 14–17 October, 2007*, pp. 1280–1284.

31. C.S. Kim, Bayesian orthogonal least squares (BOLS) algorithm for reverse engineering of gene regulatory networks, *BMC Bioinformatics*, **8**, Article No. 251, 15 pp., 2007.

32. L.F.A. Wessels, E.P. Van Someren, M.J.T. Reinders, A comparison of genetic network models, *Pacific Symposium on Biocomputing*, **6**:508–519, 2001.

33. J. Li, X.S. Zhang, An Optimization model for gene regulatory networks reconstruction with known biological information, in: *1st International Symposium on Optimization and Systems Biology (OSB'07), Beijing, China, August 8–10, 2007*, pp. 35–44.

34. J. Gebert, N. Radde, G.-W. Weber, Modeling gene regulatory networks with piecewise linear differential equations, *Eur. J. Oper. Res.*, **181**(3):1148–1165, 2007.

35. A.A. Margolin et al., ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics*, **7**(Suppl. 1), Article No. S7, 15 pp., 2006.

36. J. Schäfer, K. Strimmer, Learning large – scale graphical Gaussian models from genomic data, in: *AIP Conference Proceedings, Vol. 776, Science of Complex Networks: From Biology to the Internet and WWW (CNET 2004), Aveiro, Portugal, 29 August – 2 September, 2004*, pp. 263–276.

37. J. Schäfer, K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Stat. Appl. Genet. Mol. Biol.*, **4**(1), Article 32, 32 pp., 2005.

38. A.V. Werhli, D. Husmeier, Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of priori knowledge, *Stat. Appl. Genet. Mol. Biol.*, **6**(1), Article 15, 47 pp., 2007.

39. S. Li, Integrate qualitative biological knowledge to build gene networks by parallel dynamic Bayesian network structure learning, *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2007), Boston, Massachusetts, October 14–17, 2007*, pp. 87–92.

40. M.S.B. Sehgal, I. Gondal, L.S. Dooley, CF-GeNe: Fuzzy framework for robust gene regulatory network inference, *Journal of Computers*, **1**(7):1–8, 2006.

41. A. Fujita et al., Modeling gene expression regulatory networks with the sparse vector autoregressive model, *BMC Systems Biology*, **1**, Article No. 39, 11 pp., 2007.

42. A. Brazma, T. Schlitt, Reverse engineering of gene regulatory networks: A finite state linear model, *Genome Biol.*, **2003**, 4:P5, 31 pp., 2003.

43. Z. Li et al., Using a state-space model with hidden variables to infer transcription factor activities, *Bioinformatics*, **22**(6):747–754, 2006.

44. L.J. Steggles, R. Banks, A. Wipat, Modelling and analysing genetic networks: From Boolean networks to Petri net, Technical Report, No. CS-TR-962, Computing Sciense, University of Newcastle upon Tyne, 2006.

45. F. Wang, D. Pan, J. Ding, A new approach combined fuzzy clustering and Bayesian networks for modeling gene regulatory networks, in: *Proceedings of the First International Conference on BioMedical Engineering and Informatics, Sanya, Hainan, China, 27–30 May, 2008* Vol. 1, pp. 29–33.

46. S. Imoto et al., Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks, in: *Proceedings of the Computational Systems Bioinformatics (CSB '03), Stanford, CA, August 11–14, 2003*, pp. 104–113.

47. M. Setty et al., Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma, *Molecular Systems Biology*, **8**, Article No. 605, 16 pp., 2012.

48. J.C. Liao et al., Network component analysis: Reconstruction of regulatory signals in biological systems, *Proc. Natl. Acad. Sci. USA*, **100**(26):15522–15527, 2003.

49. R. Daly et al., Using higher-order dynamic Bayesian networks to model periodic data from the circadian clock of Arabidopsis Thaliana, in: V. Kadirkamanathan, G. Sanguinetti, J. Noirel (Eds.), *Pattern Recognition in Bioinformatics, Proceedings of the 4th IAPR International Conference (PRIB 2009), Sheffield, UK, September 7–9, 2009*, pp. 67–78.

50. F. Jaffrezic, G. Tosser-Klopp, Gene network reconstruction from microarray data, *BMC Proceedings*, **3**(Suppl. 4), Article No. S12, 4 pp., 2009.

51. T. Ideker, J. Dutkowski, L. Hood, Boosting signal-to-noise in complex biology: Prior knowledge is power, *Cell*, **144**(6):860–863, 2011.

52. J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, *IEEE Trans. Knowl. Data Eng.*, **17**(3):299–310, 2005.

53. B. Ristevski, S. Loskovska, ROC curves comparison of inferred gene regulatory networks, in: *13th International Multiconference: Information Society 2010, Ljubljana, Slovenia, 11–15 October, 2010*, pp. 39–42.

54. B. Ristevski, S. Loshkovska, A comparison of models for gene regulatory networks inference, in: *Proceedings of the 2th International Conference ICT Innovations 2010, Ohrid, R. Macedonia, September 12–15, 2010*, pp. 59–68.

55. A. Sandelin et al., JASPAR: an open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res.*, **32**, D91-4, 2004.

56. T. Fawcett, An introduction to ROC analysis, *Pattern Recognit. Lett.*, **27**:861–874, 2006.

57. B. Ristevski, S. Loskovska, A novel model for inference of gene regulatory networks, *HealthMED*, **5**(6):2024–2033, 2011.

58. B. Ristevski, S. Loskovska, ROC curves comparison of inferred gene regulatory networks, in: *13th International Multiconference: Information Society 2010, Ljubljana, Slovenia, 11–15 October, 2010*, pp. 39–42.

59. E. Wingender et al., The TRANSFAC system on gene expression regulation, *Nucleic Acids Res.*, **29**(1):281–283, 2001.

60. P. Baldi, S. Brunak, *Bioinformatics: The Machine Learning Approach*, The MIT Press, London, 2001.

61. L. Kaderali, N. Radde, Inferring gene regulatory networks from expression data, *Stud. Comput. Intell.*, **94**:33–74, 2008.