ssdnm
środowiskowe
studia doktoranckie
z nauk matematycznych

Robert Szczelina

Uniwersytet Jagielloński

A review of current methods and trends in gene regulatory networks inference

Praca semestralna nr 2

(semestr letni 2011/12)

Opiekun pracy: Jerzy Tiuryn

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

A review of current methods and trends in gene regulatory networks inference

Robert Szczelina

July 3, 2012

Abstract

Gene Regulatory Networks (GRNs) play an important role in the process of understanding complex biological systems. Inferring, or 'reverse-engineering', gene networks can be defined as a process of identifying gene interactions from experimental data through computational analysis. High-throughput microarray technologies are usually used to acquire data for this process. As amount of data produced by microarrays is often too large to be analyzed directly, big effort was put to develop automatic methods for identifying gene interactions. With rapid growth of the computing power many methods have been introduced based on strong mathematical and computer science theoretical principles. Recent studies however show that we still lack trustworthy algorithms that can detect complex network structures even when very rich datasets are available. This problem is even more visible in context of real biological experiments as they are limited by technology costs and amount of work needed to obtain high quality data. Thus, the inference of GRNs of living organisms is still an open question. In this work we try to summarize state-of-the-art methods and we will give short description of recently proposed improvements.

1 Introduction

In recent years we are experiencing rapid growth of new gene expression data. This growth rate is even faster that the well known Moor's law for computing power[6]. This huge amount of data cannot anymore be tackled without automatic simplification that would allow scientist to understand what is really happening at the genetic level of a given organism. One of the most studied concepts for such simplification is *Genetic Regulatory Network* (GRN), a concept based on a strong mathematical theoretical basis: The Graph Theory. One of the main problems that GRN theory is aiming at is the reconstruction of a network of interactions between genes from given expression datasets.

Many methods for automatic GRN inference were proposed and we are going to examine some of them. They differ on the level of an underlying theoretical model and a type of input on which they work. Also the output may be specific for each method as no coherent definition of GRN exists [16].

Since we do not fully understand the laws governing a process of genetic regulation and as genes interact in more complex way than it is modeled with GRNs, we need a verifiable method for assessment of the new algorithms before we apply them in analysis of a data from real biological experiments. Recently, there has been organized a series of challenges called *DREAM*. Those challenges give scientists opportunity to test their methods in unified way[18] [12]. In this document we will briefly review the state-of-the-art methods for GRN inference along with the results of the DREAM3 challenge and the current trends in overcoming issues identified after this challange.

This paper is organized as follows. In the section 2 we will give short explanation of theory of GRNs and the notation we will use throughout the article. We will also explain various kinds of experimental data that can be acquired during genomic experiments, also we will briefly introduce the method for the *in silico* genetic network simulation. In the section 3 we will summarize current methods that are used for GRN inference and their underlying mathematical models. We will discuss their strengths and weaknesses based on results from the DREAM3 challenge. Finally, in the section 4, we will examine current trends for GRN inference methods and discuss how they can potentially improve the quality of the predicted network.

2 Notation and definitions

In this section we will describe basic definitions connected with GRNs and their inference methods.

2.1 Gene Regulatory Networks

Genetic regulation is a very complicated process, even in a single cell organism. It consist of several levels of interacting chemical molecules and compounds. The genetic level consist of concentrations of RNA transcribed from DNA. RNA then is translated into proteins by ribosomes. Proteins interact with themselves at the proteome level creating protein complexes. Some proteins are transcription factors which contribute to repress or increase RNA transcription by binding at specific sites in DNA. Some of the proteins are also catalysts for reactions on the metabolic level. Metabolites may interact with genes and proteins, for example providing energy for transcription and translation processes. Extracellular interactions may be also important. To understand such complicated regulation we need to simplify it. Generally, this is done by projecting all interactions onto the genetic level [16] and studying only those 'theoretic' interactions among genes. This interpretation is basis for the definition of the term *Gene Regulatory Networks.*

There are plenty of definitions based on a way one may project complex real model and how one may choose important interactions from it[16]. In this article however we will assume a simple, yet flexible definition of GRN:

Definition 1 Gene Regulatory Network *(GRN) is a directed graph* (V, E) *where* V *is a set of gene activities (or simply genes) which we call vertices and* E *is a set of directed edges,* $E \subset V \times V$*, where an edge* $e = (v, u) \in E$ *represents a casual interaction between genes* v *and* u *(that is gene* v *is a direct cause for changes in expression level of* u*)*

There are plenty of other kinds of genetic networks, varying in the definition of vertices and edges, the information they contains and the type of data used to construct them. For example *Co-expression Networks* (CENs) contain only undirected edges. In CENs an edge $\{u, v\}$ represents information that presence of u and v is highly correlated. *Transcription Factor Networks* (TFNs) are centered on transcription factors and are build using ChIP-on-chip experiments data[1]. For an extensive comparison between various definitions of GRNs please consult aforementioned [16].

Most GRNs display substantial non-trivial topological features with patterns of connection between their elements that are neither purely regular nor purely random. One of the most notable example is a scale-free topology[4]. It is also believed that GRNs are in general sparse. Those features are connected with the robustness of biological networks [9]. However, some articles points out that this may not be the case and in fact the assumptions about sparseness and scale-free nature of GNRs should be treated with more reserve [17] [16]. This is important because many methods for GRNs inference relay on strong assumptions about properties of the real underlying model. This is also important in the process of generating realistic *in silico* experiments for assessment [11] of inference methods.

One fact that we cannot neglect is that all biological networks may contain loops, which are essential for the homeostasis of the organism and differentiation among cells. Therefore any method for GRN inference should take into account possible existence of multiple feedback loops.

2.2 Data for inference

As we have seen in the previous section, the type of data used to reconstruct a network is often inextricably linked with the network itself. It is because the main problem which the theory of GRNs is trying to solve is how to infer important causal interaction from a given set of biological experiments. As those experiments always involve measurements of concentrations of some chemical quantities, those quantities are often the basis for the set of vertices of the network.

In recent years there were proposed many methods for measurement of gene expression levels. Those methods are based on our understanding what

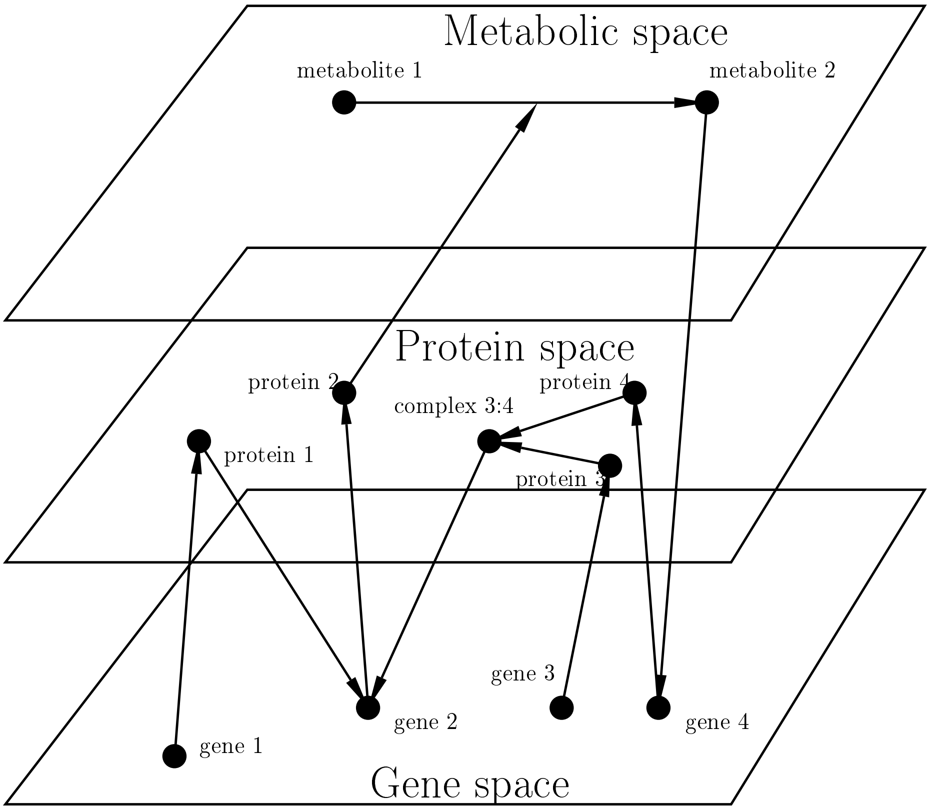

Figure 1: An abstract depiction of cellular physiology, with expressed genes that are transcribed to proteins, proteins binding together to form complexes, proteins and complexes affecting transcription of genes, proteins catalyzing reactions between metabolites. This picture is based on figures found in [16]

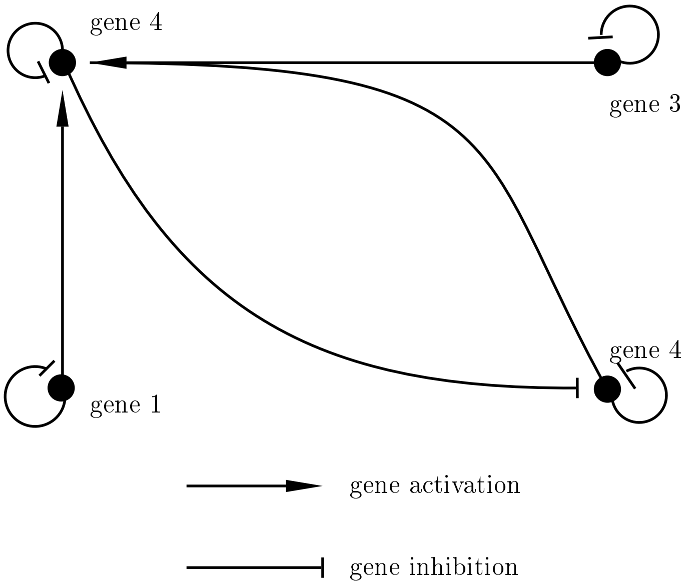

Figure 2: The Gene Regulatory Network resulting from the underlying biological network presented in figure 1 by projecting all interactions to the hypothetic genetic level.

the genes really are, how genes are activated for the transcription process and how the product of transcription, mRNA, is involved in the protein production process called translation.

Up to date the most widely used and available data is an RNA expression profile acquired through a microarray technology. Microarray technologies, as a whole, provide new tools that change the way scientific experiments are carried out. The main advantage of the microarray technology compared with traditional experimental methods is one of a scale. In place of conducting experiments based on results from one or a few genes, microarrays allow for the simultaneous interrogation of thousands of genes or entire genomes [8]. Thus they are an irreplaceable source of data in the process of inferring networks of complex gene interactions and understanding a regulation process in an entire organism.

There are several ways a microarray experiment can be designed for the GRN inference. A measurement of sole concentrations of DNA or RNA in a steady state of the system can be used to infer CENs but it is not sufficient for directed (i.e. causal) interactions. For casual interactions time-series data or perturbation data sets must be acquired. A time-series data experiment uses a small disturbance of the system from a steady state. This is done for example by putting the organism under different environmental conditions or changing initial concentrations of gene expression. Then, during the phase where our system returns to its steady state, a series of microarray measurements of gene concentrations are performed. Such trajectory in the phase space can be used for reasoning about gene interactions which drive the process of returning to the steady state. Perturbation methods, in contrast, are concentrated on 'switching off' single genes (by mutation in DNA) and observing how this action affects expression of other genes.

When it comes to complexity, a sole concentration measurement is the simplest and cheapest method but it also gives very little information. Time series data is a good trade-off between quality of data, feasibility and experiment cost. The best results are obtained by knocking-out a single gene at a time in a series of experiments but it is far more complicated than the previous methods due to the need of carrying separate experiments for each gene in the network. Sometimes this may not be feasible as knocking-out some of the genes may be lethal for the organism under observation.

A second source of data, especially good for identifying interactions in TFNs is a ChIP-on-chip experiment. This is a technique that combines chromatin immunoprecipitation (ChIP) with the previously mentioned microarray technology. Like regular ChIP, ChIP-on-chip is used to investigate interactions between proteins and DNA *in vivo*. Specifically, it allows the identification of the cistrome, sum of binding sites, for DNA-binding proteins on a genome-wide scale [1]. However, this technology is more complicated than sole microarray measurements of differential gene expression levels.

2.3 *In silico* network simulation

The simulated data from *in silico* gene networks is often the only feasible source of test data for systematic performance assessment of inference methods. All aspects of the network and experiments are under full control in a simulation . This allows for characterization of reverse engineering methods for different types of data and levels of noise. Additionally, inferred networks may be compared with perfectly known gold standard which are network models used for numerical simulations.

The simulation process is well known and consists of running integration of a set of Ordinary Differential Equations (ODEs) describing the full system, with both RNA and protein concentrations together. For the inference methods only RNA concentrations are provided, exactly like in a real biological experiment. Usually some noise is added to mimic uncertainty in biological experiments. The underlying structure of the mentioned ODE system is given by gold standard genetic network. This is done by interpreting edges of the network as some non-zero parameters defining equations (value of the parameter is usually random). The problem of choosing a right underlying network structure is a very important task, as it may affect how well the various methods will work. It was shown in the DREAM2 Challenge that the winning method had a much better performance on a random Erdös–Rényi network than on a scale-free network[18].

In spite this, Marbach *et. al.*[11] proposed a method based on extracting modules from known biological interaction networks whose structure was validated biologically. Using the yeast transcriptional regulatory network as a test case, they show that extracted modules have a biologically plausible connectivity as they preserve functional and structural properties of the original network. In this way they overcame the problem of uncertainty about topological properties of real biological networks. Their method was selected to generate gold-standard networks for the gene network reverse engineering challenge of the third DREAM conference. For a detailed sample model generation we refer reader to [12] supplementary materials.

3 State-of-the-art methods

In this section we are going to present existing algorithms based on classical and well known principles.

3.1 Ordinary Differential Equations

Ordinary Differential Equations (ODE) are the closest model for simulating complex biological networks as they mimic physical and chemical reaction laws.

The representative algorithms in this category include network identification by regression (NIR), singular value decomposition and regression analysis, mode-of-action by network identification (MNI), time series network identification (TSNI), and "Inferator" [3].

In the simplest, linear, case the underlying model of a GRN with N genes is of the form:

$$x_i' = \sum_{j=1}^{n} a_{ij} x_j + u_i, \quad i \in \{1,..,N\} \tag{1}$$

where $x_i(t)$ is concentration of gene X_i at the time t, a_{ij} define strength of influence of the gene j on the gene i and $u_i(t)$ represents an external perturbation of gene i at the time t. For the steady-state data ($x_i' = 0$) one can rewrite it and solve linear equation as in the NIR algorithm. For a time series data one can approximate the above equation with finite difference scheme based on time points in which the measurements took place. After this discretization we obtain:

$$\frac{x_i(t_{k+1}) - x_i(t_k)}{t_{k+1} - t_k} = \sum_{j=1}^{n} a_{ij} x_j(t_k) + u_i(t_k), \quad k \in \{0,..,T\}, i \in \{1,..,N\} \tag{2}$$

This also gives a system of linear equations. Solution of this system however requires a large amount of measurements T. As this may be unfeasible one may use dimension reduction techniques or a time-series data interpolation to overcome this problem. Such methodology was successfully applied in the TSNI algorithm[2].

Although very good for simulation, ODEs are of limited use in GRNs inference since reconstruction of parameters of the ODE need an enormous computing power and horrendous amounts of data due to the nature of equations. Moreover, searching for the best model requires choosing arbitrary function space, which may not be the best to describe the underlying biological process. Another problem is over-fitting of a model. When taking into account more parameters we obtain models that mimic data almost perfectly but the model is so complex that we cannot make any useful predictions based on it.

3.2 Bayesian Networks

A Bayesian network (BN) is a representation of a joint probability distribution and is defined as follows[14]:

Definition 2 *Bayesian network (BN) is an acyclic directed graph $G = (V, E)$ whose vertices corresponds to set of n random variables $X_1..X_n$, together with conditional distribution for each variable X_i given its parents $pa(X_i)$ in G. The set $pa(X_i) = \{X_j \in V : (X_j, X_i) \in E\}$ denote the* set of parents *of vertex X_i in a directed graph G. We denote the conditional probability of given BN as $P(X_i|pa(X_i))$ and we impose assumption that those distributions fulfill* Markov assumption*: each variable X_i is independent of its non-descendants, given its parents in $pa(X_i)$.*

The definition above implies that the joint probability of $X_i..X_n$ can be decomposed into:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i)) \tag{3}$$

The problem of learning a Bayesian network can be stated as follows. Given a training set $D = \{d_1, .., d_m\}, d_i = (x_1, .., x_n)\}$ of m independent instances of observations of variables $X_i, i = 1..n$, find a network G and corresponding conditional distribution for each variable X_i that best describes D. Many algorithm were developed by focusing mainly on maximizing probability of observing data D given graph G (G is taken from the space of graphs with n vertices):

$$\max_G P(G|D) = \max_G \frac{P(D|G) \cdot P(G)}{P(D)}. \tag{4}$$

In the above equation $P(G)$ is a penalty for choosing a specific graph structure and may be used to favor graphs that conform to known structures (i.e. expert biological knowledge about GRNs). $P(D)$ can be assumed constant not relevant for the score. Finally the most important factor is $P(D|G)$ which represents probability that the given set of data D was generated by the given graph G. In this part of equation one may include penalty for over-fitting model to the data.

As the size of the space of graphs grows exponentially with n, finding a global maximum directly by examining each graph is not feasible. It was proven that learning structure of the Bayesian network is an NP-Hard problem. Thus all algorithms for learning BNs rely on heuristics. For example one of the well known methods is finding a local maximum by a greedy, climbing-like algorithm of adding one edge at a time and choosing the high scoring graph for the next iteration.

It is worth noting that the basic definition of Bayesian network does not say anything about causality of the relationship. To overcome this problem we can assume the *Causal Markov Assumption*:

Definition 3 *Bayesian Network satisfies* Causal Markov Assumption *when for all variables X_i is independent of its earlier causes (i.e. variables X_j such that $X_i \in pa(X_j)$) given the values of the variable X_i immediate causes (i.e. $pa(X_i)$) it.*

It can be shown that when the Casual Markov Assumption holds then the network satisfies also the Markov Assumption and edges represent real casual relations. However learning such a network requires more data than learning classical BN. Usually, one needs to have datasets of mentioned earlier gene knock-out experiments [5].

The main problem of Bayesian networks is that they cannot contain cycles (i.e. no feedback loops). This restriction is the principal limitation of the Bayesian network model. Dynamic Bayesian networks overcome this limitation by creating copies $X_i(t)$ of variables X_i for each time t and allowing edges only to go in time non-decreasing directions, i.e. from $X_i(t-1)$ to $X_j(t)$ or $X_i(t)$ to $X_j(t)$ but never from $X_i(t)$ to $X_j(t-1)$. In such a network no cycles are allowed, but after projecting all edges from all time-dependent variables to non-dependent ones we can obtain a graph that can contain cycles. This projection is our GRN. Dynamic Bayesian networks are an extension of Bayesian networks able to infer interactions from a data set D consisting of time-series rather than the steady-state data[3] [21].

3.3 Statistical Methods

Clustering, although not properly a network inference algorithm, is the current method of choice to visualize and analyze gene expression data. Clustering is based on the idea of grouping genes with similar expression profiles in the same cluster[3].

Similarity is measured by a distance metric as, for example, the correlation coefficient among a pair of genes (i.e. Pearson Correlation Coefficient - PCC) defined as:

$$r_{ij} = \frac{\sum_{k=1}^{m}(d_{ki} \cdot d_{kj})}{\sqrt{\sum_{k=1}^{m} d_{ki}^2 \cdot \sum_{k=1}^{m} d_{kj}^2}}, \tag{5}$$

where d_{ki} represents experimentally measured value of the random variable X_i in the k-th experiment. This equation literally is a covariance normalized w.r.t standard deviations of variables given that their mean is 0. For highly correlated variables the value $|r_{ij}|$ should be close to 1 and, conversely, it should be close to 0 for uncorrelated ones. However, the correlation coefficient has a big drawback. It can capture only linear dependence among variables, which may produce huge errors due to fact that the gene regulation is often highly non-linear. Some studies have been proposing recently a new definition of the correlation coefficient that can capture this non-linearity in data[15].

4 New trends

Many new approaches were proposed to address issues that arose in the course of the GRN inference methods assessment.

4.1 PCA-CMI

Much attention was given recently to probabilistic methods based on the information theory as they are capable of inferring realistic networks even from small amounts of available data. An example of such a method is proposed in [19]. It combines a nonlinear criterion for variable independence which is called Conditional Mutual Information (CMI) with Path Consistency Algorithm (PCA).

Mutual Information (MI) is defined as

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \tag{6}$$

where $H(X)$ is entropy of a random variable X defined as

$$H(X) = -\sum_{x \in X} p(x) \log p(x). \tag{7}$$

The $H(X,Y)$ denotes entropy for a joint probability distribution of X and Y. It is easy to see that $I(X,Y) = 0$ for independent variables, so one can use it to determine significant dependencies in the data.

CMI is an extension of the idea of MI to the conditional dependency of X and Y given some variable(s) Z. As one needs to stick with the standard experimental data, in [19] the authors proposed to estimate entropy with a Gaussian kernel probability density estimator. They obtained after some rearrangements:

$$I(X,Y|Z) = \frac{1}{2}\log\frac{\det(C(X,Z))\cdot\det(C(Y,Z))}{\det(C(Z))\cdot\det(C(X,Y,Z))}, \tag{8}$$

where $C(X)$ denotes the covariance matrix of variable X and $\det(\cdot)$ is the matrix determinant.

The Path Consistency Algorithm is a well known algorithm designed to filter graphs from inconsistent edges i.e. those that violate some given constraints [13]. In the context of CMI-PCA those constraints consist of statistical significance of testing whether CMI for an edge (x, y) given other k variables $z_i, i \in \{1..k\}$ is near 0. The case when $k = 0$ is equivalent to the MI approach. For each $k = 1..n - 2$ the graph G_k is produced recursively from the graph G_{k-1} by removing edges that with high probability has CMI equal to zero. The graph G_0 is produced from the full graph (i.e. the graph containing all possible edges between all vertices) by removing edges based on MI criterion. The PCA-CMI methods shows very good performance in detecting false positive forward interactions, thus eliminating source of one of the motif errors described in the DREAM3 review [12]. However, there is some limitation for PCA-CMI. As in the ARACNE algorithm, PCA-CMI cannot infer edge directionality. This limitation can be relieved by a two-tier approach in which an undirected GRN is inferred first using PCA-CMI, and then edge directionality is accessed via some other method.

4.2 Meta analysis

Meta-analysis is a method of statistically combining and analyzing data from separate studies [7]. It was mentioned in the DREAM3 summary that a simple "community prediction" can outperform other approaches. The "Community prediction" was a list of sorted edges w.r.t. the average score from output of several best methods that take part in the challenge. Yet very simple, this method was as good as the best method and has another advantage: it is less prone to the invalid assumptions about the underlying gold-standard network structure that were made in each of the participating method. This is extremely important in the process of GRN inferring from real data as we cannot make any assumption about the underlying biological process.

In [10] authors used a Fishers Inverse Chi-Square meta-test to combine Bayesian Network, Lasso and Dantzig selector predicted networks. Their meta analysis tool outperforms all participants from the DREAM5 challenge 3A competition on all sizes of networks. They however admit that this method still has problems with predicting fan-in degree. Its advantage is that it combines linear (regression: Lasso, Dantzig) and nonlinear methods

(Bayesian Networks), so it overcomes limitations of underlying methods: linear dependence and un-directionality of edges in case of regression methods and the fact that networks produced with BN need to be acyclic. They show that their method proves well in a case where there is limited amount of available data.

There are many more ways one can combine various methods. In [10] authors used a Bayesian network method for learning of parameters in an ODE based model. The main reason for combining those two methods was reduction of computational complexity for large networks that is presented in ODE based models. As mentioned earlier, ODE based inference methods needs much more computational power than machine learning techniques based on heuristic algorithms. Thus DELDBN uses DBN for estimation of non-zero parameters in ODE model with given time series data. The proposed method, DELDBN, showed high accuracy in predicting network structures in an in-vivo benchmark dataset and was successfully applied to human Hela cell time series data. However, the authors of DELDBN used a linear ODE system as an underlying model so their method may have problems with predicting non-linear interactions among genes. Moreover, they do not provide any comparison with currently known methods for GRN inference. This shows that, despite how much attention have been recently put on the development of assessment methods, they are not widely used to compare emerging methods to the state-of-the-art ones.

4.3 Phylogeny information

As mentioned earlier collected data indicate that GRNs are sparse and follow a scale-free edge distribution law. One can try then to come up with a biological mechanism hypothesis which justifies the above observations. One of those mechanisms was proposed in [4]. They conjectured that the scale-free topology emerges from the preferential attachment of interactions by the process of gene duplication. Thus, the scale-free topology may be connected with a known biological mechanism. This mechanism is, in turn, present in phylogeny, as older and more preserved genes will be duplicated more often and thus will have more links.

The above observation is a basis for yet another idea for improvement of GRN inference. There are several works that try to employ phylogenetic information to enhance prediction from standard GRN inference algorithms. Although their results are promising, phylogeny-based GRN inference has inherent weakness: it makes it almost impossible to work with currently available synthetically generated data due to lack of sufficient evolutionary information. Other disadvantages which make it not feasible are as follows: the need for simultaneous creation of networks for many related organisms (thus the need for much more experimental data) and uncertainty in phylogenetic trees (which are also created algorithmically by heuristic algorithms)[20][22].

5 Summary

Many of the methods presented in the state-of-the-art section of this work were tested during DREAM3 challenge on *in silico* generated datasets. To the best knowledge of the author, the GeneNetWeaver tool created by the authors of the DREAM project is currently the only widely available tool that allows for systematical assessing and comparison of the performance of various methods. The study of organizers of the DREAM challenge shows that we are still far from discovery of the perfect GRN inference algorithm [12]. Systematic errors of the current methods pop up in predictions of specific network motifs such as feed-forward loops and cascade error. Those errors arise mainly due to inappropriate recognition of indirect regulation.

Some of the new methods are aiming at overcoming those problems while others are trying to adopt an approach proposed by the authors of the DREAM project to build better community predictors based on meta-analysis. Authors of those methods sometimes try to compare their methods to DREAM3 participants but those comparisons are far from complete. As the methods for GRN inference grow in number, we should put more emphasis on systematic and continuous comparison of those methods. Authors of the DREAM challenge offer their tools called GeneNetWeaver for free download, although it may be worth to consider the creation of a similar tool in form of a webservice accessible online for all who work in the field. Such a tool should be able to provide automatic testing, report generation, comparison between methods and test-case editor functionality. With help of such a software the assessing of new methods and their comparison to current standards may become more accessible, less cumbersome and finally may lead to better understanding of strengths and weaknesses of the available GRN inference methods.

References

[1] O. Aparicio, J.V. Geisberg, and K. Struhl. *Current Protocols in Cell Biology, Chapter 17: Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo.* John Wiley & Sons, Inc., University of Southern California, Los Angeles, California, 2004.

[2] M. Bansal, G. Della Gatta, and D. di, Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22:815–822, 2006.

[3] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego di Bernardo. How to infer gene networks from expression profiles. *Mol Syst Biol.*, 122(3), 2007.

[4] A.L. Barabási and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet.*, 5(2):101–130, 2004.

[5] G. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. *Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*, pages 116–125, 1999.

[6] The Economist. Report: Biology 2.0. *The Economist*, 2010.

[7] M. Egger and G.D. Smith. Meta-analysis. potentials and promise. *BMJ*, 315(7119):1371–1374, 1997.

[8] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *DNA Microarray Data Analysis, second edition.* Scientific Computing Ltd, 2005.

[9] R.D. Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Mol Syst Biol.*, 4(213), 2008.

[10] Zheng Li, Ping Li, Arun Krishnan, and Jingdong Liu. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. *Bioinformatics*, 27(19), 2011.

[11] D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J Comput Biol.*, 16(2):229–239, 2009.

[12] Daniel Marbach, Robert J. Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *PNAS*, 107(14):6286–6291, 2010.

[13] R. Mohr and T.C. Henderson. Arc and path consistency revised. *Artificial Intelligence*, 28:225–233, 1986.

[14] D. Peer, I. Nachman, M. Linial, and N. Friedman. Using bayesian networks to analyze expression data. *J Comput Biol*, 7:601–620, 2000.

[15] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting novel associations in large data sets. *Science*, 2011.

[16] Das Sanjoy et al. *Handbook of Research on Computational Methodologies in Gene Regulatory Networks.* 2010.

[17] M.L. Siegal, D.E. Promislow, and A. Bergman. Functional and evolutionary inference in gene networks: does topology matter? *Genetica*, 129(1):83–103, 2007.

[18] G. Stolovitzky, D. Monroe, and A. Califano. Dialogue on reverse-engineering assessment and methods: the dream of high-throughput pathway inference. *Ann. NY Acad. Sci.*, 1115:1–22, 2007.

[19] Zhang Xiujun, Zhao Xing-Ming, He Kun, Lu Le, Cao Yongwei, Liu Jingdong, Hao Jin-Kao, Liu Zhi-Ping, and Chen Luonan. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, 28(1), 2012.

[20] Zhang Xiuwei, Zaheri Maryam, and Bernard M.E Moret. Using phylogenetic relationships to improve the inference of transcriptional regulatory networks. *International Conference on BioMedical Engineering and Informatics*, 2008.

[21] J. Yu, A. Smith, V., P.P. Wang, A.J. Hartemink, and E.D. Jarvis. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20:3594–3603, 2004.

[22] Xiuwei Zhang and Bernard M.E. Moret. *Algorithms for Molecular Biology*, 5(1), 2010.