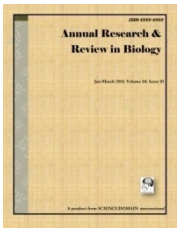

Annual Research & Review in Biology
4(4): 577-601, 2014

SCIENCEDOMAIN *international*
www.sciencedomain.org

Biological Network Inference: A Review of Methods and Assessment of Tools and Techniques

Jimmy Omony[1*]

[1]*University of Groningen, Molecular Genetics department, P.O. Box 11103.9700 CC, Groningen, Netherlands.*

Author's contribution

The author (JO) conceived the idea, wrote, read and approved the manuscript.

Review Article

Received 30th June 2013
Accepted 16th October 2013
Published 9th November 2013

ABSTRACT

The topic of reconstruction of genetic networks is of great interest to the scientific community today – particularly those in the biological sciences. Essentially the need for network reconstruction is motivated by the need to find relationships between regulation mechanisms for genes, the need for discoveries in medicine, drug and pharmaceutical industry, the need for improved agricultural crops. All this requires a concerted effort from multi-disciplinary sciences, e.g. physics, mathematics, biology and chemistry – which have led to disciplines such as Systems Biology and Bioinformatics. Mathematical and statistical modeling has particularly been very instrumental for engineering and software development has been very useful in biological networks inference. Sometimes the link between theory, modeling and data acquisition is unclear. The goal in this article is to discuss tools and techniques for biological network inference and the areas of application. The pros and cons of network reconstruction methods are also provided. The number of scientific articles on network inference is overwhelming. Additionally, there is a dilemma in methodology choice, which is attributed to the scarcity of novel ways to compare the performance of the existing methods on experimental data. Applications of data visualization tools, modeling and simulation, data analysis and storage are given.

Keywords: Genetic network reconstruction; dynamic modeling; parameter identification.

**Corresponding author: Email: jimmy.omony@gmail.com;*

1. INTRODUCTION

The topic of biological network inference is of great interest and its history dates back to as far back as the 1960s [1] and the work on graph theory which is instrumental for studying network structures was pioneered as early as1959 [2]. Since then, network inference has been vastly explored using genomic data with numerous analytic and numerical approaches [3,4,5,6]. We have also in the past few decades seen the publication of numerous papers on biological systems including research and review articles providing insight into the applications and challenges of network reconstruction, see. e.g. [5,7,8,9,10,11,12,13]. A look at literature reveals an enormous list of genetic network reconstruction (GNR) methods that have been proposed and used to reverse engineer networks in various model organisms, e.g. the *E. coli* bacteria [14][15]plants (especially the model plant *Arabidopsis* [16,17]) or in humans to represent biological signaling networks, such as the tumor suppressor protein p53 which regulates gene activity in cell growth and death [18,19].

By definition, a genetic regulatory network refers to genes which code for transcription factor (TF) proteins connected to their respective target genes. Elsewhere, a genetic network has been defined as a group of genes in which individual genes can change the activity of other genes [20]. Specific mathematically formulated definitions are sometimes used depending on whether the regulation mechanisms in a network are directed or undirected. Most if not all GNR methods still fall short of perfection in the total recovery of the ”true” network structures, e.g. [21,22,23,24,25,26,27,28,29]. It still remains a challenge to find a clear guide on the choice of methods, especially for those with little or no experience in GNR. Not all methods are equally powerful or applicable under the same conditions, there are circumstances when one modeling approach (or algorithm) performs well and in some cases performs poorly. The performances are based on the training datasets used for the model calibration. Here, a discussion on the available methods for GNR is provided. This discussion is aimed at enlightening and guiding those with interest in networks inference. The theories and working principles of the formalisms are not given in this review article, instead focus is on classifying the suitability and applicability of the methods. An assessment of opinions from scientific articles on biological network reconstruction in the fields of Systems and Synthetic biology, Bioinformatics, Biotechnology, Mathematics and Computing science is made. Various terms have been used in literature to refer to studies involving biological networks, these terms are: GNR, network inference, network identification and reverse engineering of networks.

Studies of biological systems have for long been plagued by lack of data, making the use of *in silico* studies with synthetic data a common practice. Modeling and simulation enables rigorous probing of network dynamics prior to validation with experimental data. Mathematical modeling is a powerful tool for testing hypotheses that might be difficult to assess otherwise. The comparison of model predictions to experimental data enables validation of current knowledge. Similarly, a poor match in model predictions to current knowledge triggers a need to bridge the gap in knowledge. Modeling enables *in silico* testing and validation of experiments that cannot be done *in vitro*. The availability of relatively low cost, high through-put genomic data has significantly increased model validation and hypothesis testing. GNR algorithms are divided into two categories, namely: the discrete state and continuous state approach. In discrete state based approaches, each node in a network is considered to have a small number of discrete states and the regulatory interactions between nodes (gene) are described using logical functions (typically, derived from a combination of the Logical conjunction (AND), Logical disjunction (OR), Inverter

(NOT) and Exclusive or (XOR) operators). Historically, the use of logical operations stems from the need to describe biological processes using mathematical models.

In the continuous-state approach, messenger ribonucleic acid (mRNA) and protein levels are considered as continuous functions in time. GNR can be done at various levels namely: (i) genetic networks, (ii) metabolic networks, and (iii) protein networks. Although integrating these levels into a single network is a nontrivial task, there is progress, resulting from a concerted effort from inter-disciplinary research. Therefore, information at the gene expression level is useful for projecting biochemical networks [30]. In a directed network, the edges are considered to be an ordered pair of vertices from one node to another. In an undirected network the edges are unordered. Signaling pathways are a good example of directed networks since they contain all genes or proteins being represented as nodes. In a directed network, the flow of information is from one gene to another while in an undirected network the flow of information is not specified. Meanwhile, protein–protein interactions form an undirected network with the proteins as nodes and pair-wise interaction between proteins being represented with edges. These concepts can be found in books on Graph theory (see e.g. [31,32]).

In literature there are numerous summaries of the requirements for good network inference. Most of the opinions converge around a couple of fundamental necessities. A good example of a discussion of such opinions can be found in the work of Kitano [33] which summarizes the tasks required for proper understanding of biological systems. These tasks include: (i) system structure (topology) identification, (ii) system behavior (network dynamics) analysis, (iii) systems design, and (iv)systems control. Most biological networks are dynamic rather than static, which makes the dynamic approaches preferable over the static ones [34]. In general, there are two components of network inference, namely: (i) structure identification which refers to the determination of relationships between the genes in a network of interest and their corresponding transcripts and (ii) the quantification of the relationship between these genes (or transcripts) – a process that is referred to as parameter estimation. These two processes are closely interconnected and should not be confused to refer to the same thing. The challenges associated with modeling of network dynamics and parameter identification are discussed in this review article. Avast number of scientific articles and information storage systems on biological networks and genomic datasets have sprung up in the last two decades. It is therefore useful to have guiding documents like this review paper on the state-of-the-art methods, currents trends and emerging challenges associated to network inference. Recently, major progress has been made following scientific meetings like the Dialogue for Reverse Engineering Assessment and Methods (DREAM) challenge [35]. The DREAM challenge is a project that aims to fairly compare the strengths and weaknesses of network reconstruction methods. It also aims to validate the reliability of the models in the various situations in which they are used.

A number of interesting papers have since come out of the DREAM series challenge, e.g. DREAM1 [36], DREAM2 [37,38] and DREAM3 [35,39]. Details of the DREAM series challenges are not discussed here and instead its benefits are stressed. For instance in DREAM3, Yip et al. [35] performed *in silico* studies aimed to reconstruct networks from two types of data, i.e. gene expression profiles in the ”deletion data” and time series gene expression trajectories after initial data perturbations. They used deletion data to detect direct regulatory activities and perturbation to enrich data that aids identification of weak and complex regulation mechanisms. One of the interesting approaches that have been exploited in network inference is that network excitation – also known as perturbation studies. This excitation basically consists of introducing some kind of disturbance or external

signal to a targeted component in a network or pathway and then assessing the resultant impact on the other network components. This often involves experimental approaches to obtain information on networks, i.e. excitation of an existing unmodified network and the excitation of a modified network in a bid to exclude certain specific pathways. Excitations of networks have in principle been approached in various ways, e.g. the input-output analysis following excitation of an intact network – an approach that is very much of interest to systems and control engineers. On the other hand, biologists have exploited alternative strategies aimed at obtaining extensive insight by using experimental modification. In this work, the terms time series and time course data often refer to the same process.

2. STATE-OF-THE-ART IN NETWORK INFERENCE

Good network inference requires proper planning and execution of an experiment, thereby ensuring quality data acquisition. Optimal experimental design (OED) in principle refers to the use of statistical and or mathematical concepts to plan for data acquisition. This must be done in such a way that the data information content is enriched, and a sufficient amount of data is collected with enough technical and biological replicates where necessary. These requirements are necessary to ensure that the data quality does not compromise whatever analytic approach is used for the network reconstruction and parameter estimation. To provide some insight into the various components and requirements for proper network inference, an overview is given in Fig. 1.

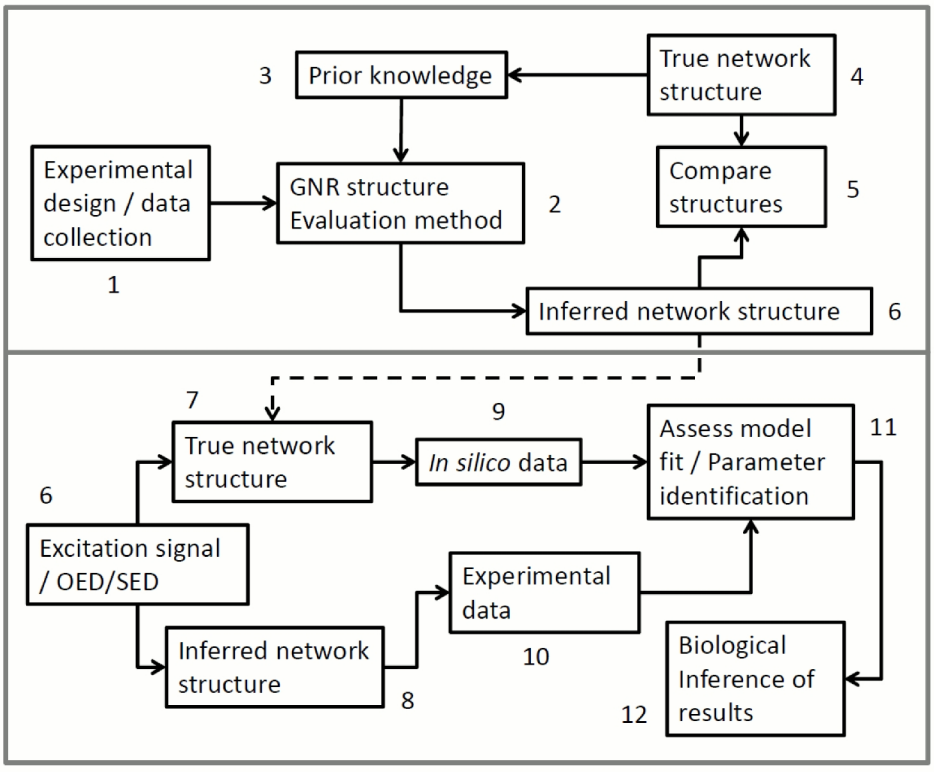

Fig. 1. A schematic representation of the steps involved in network inference

The *Prior knowledge* constitutes literature information from scientific publications, biological databases and expert knowledge on the subject. The performance of a GNR method is judged by how well the *Inferred network* matches the *True network* structure. Starting with a given dataset and *Prior knowledge* of a network, often the goal is to infer the *true* network structure. Generally, testing of a reverse engineering method is done by generating virtual data from an assumed *True* system – which is of course not the true system itself.The term OED represents Optimal Experimental Design. Image adopted from PhD thesis, Omony [40].Numerous network inference formalisms have been proposed in literature, e.g. differential equations, hybrid models, regression models, Bayesian models and neural networks. These approaches require large amounts of data and (super-)fast computers. The pros and cons of the methods used in GNR are given in Table 1. Gathering sufficient prior knowledge from literature and databases is a necessary but time-consuming and tedious process. Once the *true* network topology is known, a performance comparison of reverse engineering approaches can be made (Fig.1). Successful validations require well-designed experiments, high quality data and robust identification procedures.

2.1 Overview of Network Inference

Fig.1 shows the main steps in network inference. One goal is to compare how well the true network approximates the inferred network and the other is to derive meaningful insights from such inferences. Approximations are often sufficient for inferring network dynamics for high levels of identification accuracies. Additional comparisons as to whether two network structures match can be done by assessing the presence of corresponding edges in the *true* and *identified* network structures (step 4 and 6). Many networks have a high level of uncertain information such as the node dynamics and unknown network topological structure. Identification of structures or pathways that are considered consistent and experimentally verifiable in the laboratory is essential in biological sciences. The recent decade has seen huge advances made on network structure prediction. The discovery of a network structure and the interaction mechanisms between the genes helps us: (i) understand the dynamic interaction between the genes, (ii) make predictions of the future expression values and the time trajectories of all genes in a network, and (iii) identify the biological function of a gene, e.g. in relation to drug discovery and disease studies. Readers withkeen interest on this issue are referred to [41,42,43,44].

There are many stages involved in network inference, as depicted in Fig. 1 (steps 1 to 12). These steps entail other elements that play important roles in the network inference. The goal of this paper is not to provide details of the individual steps but to show the interplay between the various steps. Overall, each of the steps has to be carefully considered if the *Biological inference of results* at step 12 is to be meaningful. This is because of the data quality, experimental design, prior knowledge, model and/or network inference formalism used all play a part in network inference. Recently, interesting progress has been made on experimental design (step 6), e.g. [45,46,47]. Good experimental design practice enriches data information content. There is evidence that studies from well-designed experiments significantly improve the network inference, as well as save time and money [47]. First, an overview of the most popular methods for network inference is provided, this is followed by an assessment of the suitability of the method and then circumstances under which it can be used. The working principles of the methods are based on a diversity of approaches, e.g. probability theory, others are stochastic based and again others are deterministic in nature.

Table 1. Summary of pros and cons for GNR methods. These methods and associated literature references showcase the success and challenges faced on the various GNR methods as applied to various model organisms

Methods used		Advantages (pros)	Disadvantages (cons)	Model organism
Differential or difference equation models		Reliably suitable for time course experiments (TCEs) data for small number of genes and conditions. Implementation simplicity.	Unsuitable for large number of genes and leads to under determined models, so unsuitable for models with many parameters [48,49].	*A. niger* [47], *E. coli*[14][50], *B. subtilis* [51], *S. cerevisiae* [52].
Random Boolean Networks		Suitable for TCEs and is used together with other clustering algorithms.	Discretization complexity; decision boundary problem, e.g. critical cut-off p-values.	*E. coli* [53] dataset; often also validated with *in silico* datasets or simulations.
Machine learning approach (MLA) (supervised and unsupervised)	Bayesian Networks (BNs)	Suitable for TCEs, yields reliable networks [54,55,56] and sub-networks [57].	Cyclic regulations in networks are not possible [56,58]. Involves using numerous assumptions some of which are not robust neither adequate.	*E. coli* [59], using simulated expression data [60], yeast cell cycle data [61].
	Dynamic Bayesian Networks (DBNs)	Suitable for TCEs and yields reliable networks [54,55,56]. Cyclic regulations in networks possible [54,58]. Can be easily used to model feedback loops in a network.	Lack of a systematic approach to determine a biologically relevant transcriptional time-lag [61]. Implementation complexity, high computational costs [61].	*S. cerevisiae* [54,62,63], *E. coli* [56], *S. cerevisiae* cell cycle data [3], yeast cell cycle data [61].
	Neural Networks	Effective depending on the data structure and dimensions; reliability of the training dataset problem at hand. Dependent on data structure and classifier function used – i.e. the level of robustness of a classifier function.	Unsuitable for TCEs. Complexity in choice of classifier functions and decision boundaries determination, handling dimensionality reduction.	Budding yeast cell data [64], data from *S. cerevisiae* and *E. coli* [15], tested on simulated and *E. coli* data [65].
	Graphical Gaussian Models(GGMs)	Available open-source software, operate on fairly simple principles, needs no data discretization, little/no need for prior knowledge.	Unsuitable for large or highly connected networks, unable to infer causal relations. Does not infer indirect interaction from hidden state variables.	Using *in silico*(synthetic) data [34], *E. coli* dataset [66], *Arabidopsis* dataset [17].

2.2 Probabilistic Bayesian Network Formalism

Bayesian networks (BNs) are defined by graphical structures which are a family of conditional distributions and a set of corresponding parameters. Together they represent a joint distribution for a set of random variables – the random variables being gene expression. BNs can be learned from a prior known network structures using well known sparse training datasets. BNs were first introduced by Kauffman [67]. This approach is based on conditional dependencies between sets of variables (see [7] for a detailed review). Applications of Dynamic Bayesian Networks (DBNs) can be found in the work of Kim et al. [54] and Zou and Conzen [61]. The DBN as an extension of the BN incorporates time dynamics into the GNR. Unlike simple BNs, the DBN can model cyclic regulation in genetic networks [62].

Cyclic regulation refers to the regulatory effect of a gene starting from one particular gene to another gene or group of genes in some kind of sequence, and finally ending back at the gene from which the regulation of transcription was initiated (Fig. 2B). Cyclic regulation can be positive or negative depending on if the overall impact on transcription of the initial gene of interest is increased (up-regulated) or reduced (repressed). These collective up-regulatory or down regulatory effect can arise from any of the other genes involved in the cyclic loop during regulation – some might be repressors while others might be activators, or even exhibit self-regulation. Cyclic regulation does occur in many biological networks although in some systems unraveling its existence can be a challenge due to indirect loops between the various network components. Additional applications can be found in the works of Werhli et al. [68], Spirtes et al. [58] and Pearl [69]. According to Murphy and Mian [70], BNs are highly stochastic and have been shown tobe suitable for modeling with noisy transcription data [3]. The Bayesian formalism is efficient but requires many working assumptions, good network structural prior knowledge and readily handles missing values in microarray datasets [71].

Rogers and Girolami [60] used Bayesian regression to infer sparse genetic networks and noted that the likelihood for observing false edges remains high, especially for data measurement noise levels higher than 10%. The term false edges here refers to a prediction that a relationship (or connection) exists between two nodes in a network and yet in reality such a relationship does not exist. They observed that typically precision levels drop to about 10% and that for each true connection there are 9 false connections. This highlights the complications that arise from parameter variations in modeling genetic networks. Identifying which approach is best suited for a given dataset is nontrivial since the efficiencies may vary across datasets [72]. The problem with BNs is that for a large number of variables and relatively small number of samples and/or experimental conditions ($n \gg N$)where n – number of genes and N - number of time points. Generally, the theoretical concepts are challenging and the model complexity and implementation difficulty increases with increasing network size. Incorporating additional information about TFs, signaling molecules or candidate regulators in BNs reduces modeling complexity. This increases the likelihood of obtaining a reliable biological inference.

Data clustering using statistical methods is the commonest way of visualizing data – particularly for uncovering closely related variables and patterns in high dimensional datasets. Data clustering might provide a useful means of extracting qualitative information on gene co-expression between genes using a large dimensional dataset, but it does not provide information on the directionality of regulation between genes. Such information on directionally can be obtained using BNs and DBNs. To ensure that such interactions are unraveled accurately, the last decade has seen advancements made towards reducing the noise levels in genomics datasets. There is therefore hope that with the availability of high

quality and high dimensional data from methods such as RNA sequencing (RNA-Seq), the accuracy and precision with which genetic networks will be reconstructed irrespective of the approach used for the reconstruction will be greatly improved (as was recently demonstrated for co-expression networks [73]).

2.3 Regression-Based Methods

Nonlinear regression involves techniques such as polynomial regression, Spline regression, Gauss-Newton and other iterative numerical techniques. Data clustering alone is insufficient for the determination of the kinetic parameters required in such models; more sophisticated mathematical tools are required for the determination of such parameters [30]. The effectiveness of partial least squares regression in reconstruction of association networks was shown by Pihur et al. [74], especially in networks where directionality in regulation between two genes is not that essential, e.g. if a gene is up or down regulated, if the two genes have similar expression patterns then they are most likely to belong to the same network module – and hence similar function. These are referred to as undirected networks since they involve network structures in which: only edges between nodes are required, how the genes are regulated, key TFs, and the existence of self-regulation (see Fig. 2A). Association networks are particularly useful for uncovering groups of genes that are co-regulated either in time or under a specific experimental condition. Generally, genes with similar expression patterns or those that are regulated by the same TF are considered to be functionally related, hence, belong to the same sub-network cluster. The clustering of genes into modular units can be uncovered using regression methods.

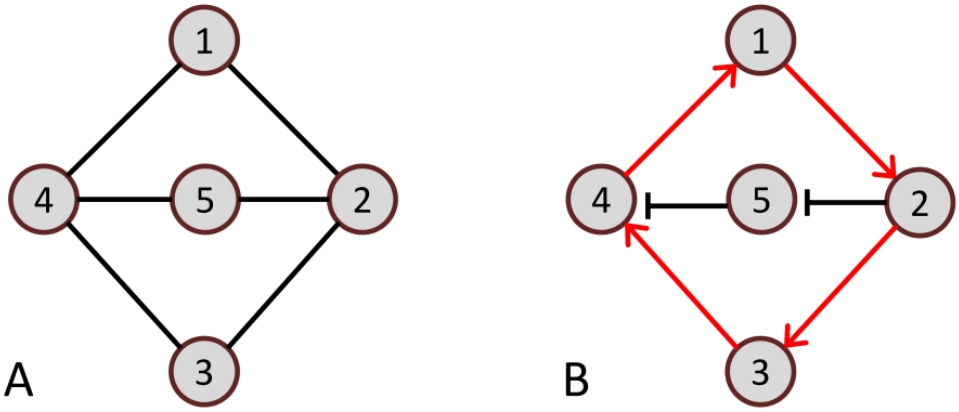

Fig. 2. Network structures and regulation mechanisms. A: A simple representation of an undirected (association) network with 5 genes (1 to 5) indicated in gray nodes, B: Directed network with the 5 genes as in A, the pointed arrows (→) represent activation, blunt arrow heads (—|) represents repression. The loop from gene 1→2—|3→4→1 is an example of cyclic regulation (red arrows) in a biological network

The combination of BNs and nonlinear regression is promising for GNR. Gardner et al. [22] developed the Network Identification by multiple Regression (NIR) algorithm; this algorithm uses steady state RNA measurements from transcriptional perturbation experiments. Though it requires network prior knowledge, the NIR is effective for small scale microbial gene networks. The algorithm was tested on simulated and real data for nine genes in *E. coli*

and about 50% of the network edges were correctly recovered. The 9 gene network was part of the Son of Seven less (SOS) system in the larger *E. coli* network. A different algorithm (Time-Series Network Identification, TSNI) yielded a similar result as NIR [75].Perturbation data enables investigation of how the gene expression changes from its steady state value. Network perturbation enables assessment of which genes, group of genes and/or sub-units (modules) in a network that influence the expression levels of the other genes. Perturbing a network ensures that fair comparisons can be made between datasets from two experimental conditions. Steady state perturbation data basically gives insight into what the expression values of specific genes would be without any changes in the environment as a result of, e.g. stressors like heat shock, change in pH and extreme (low or high) salt concentrations.

Time course Auto Regression models coupled with the Granger causality in GNR have been used for GNR. An extension of this model is the Auto Regressive Integrated Moving Average (ARIMA). The notion of Granger causality was first coined by Wiener [76]and Granger [77]. It is based on the concept that there exists a causal effect from one time series to another, if and only if the prediction of the first time series is improved based on knowledge of the second. The Granger causality measure enables determination of causal relation between two signals and it also determines direct or indirect causality [78][79][80][81][82]. However, much still remains to be done in the study of biological networks using AR-e Xogenous models (ARX) and using the more extended ARIMA models. It was shown that successful network inference can be achieved using Granger causality [77] and partial correlation analysis based methods, e.g. [83,84,85]. These methods do not infer causality between nodes. Another approach is to use association based methods – also referred to as Relevance networks [86].

2.4 Boolean Networks

The use of discrete models in biology dates as far back as the 1940's [67,87]. In discrete mathematics and computer science, discrete time models are viewed as computing machines [88]. Boolean networks are considered as qualitative descriptions of gene regulatory interactions. The first computational methods for genetic network inference were the Boolean and random Boolean network approaches [89,90,91,92]. Boolean functions map state variables at a time point $t$to $t+1$. This formalism performs best for a small number of genes. Boolean networks work on the assumption that transcript production and mRNA degradation are controlled by switch-like processes. This approach uses discretized data; hence there is a risk of information loss. However, according to Rocke and Durbin [93] the use of Boolean formalisms should be treated with caution given the relatively large noise levels in microarray data. For a brief overview into Boolean networks, let $x_i$be an n– dimensional binary vector that represents the state of a network of $n$genes. Boolean functionsgenerallyhave two states, "ON"(1) and "OFF"(0). Given a genetic network of size n, these Boolean functions take on a total of$2n$possible states. In this formalism, a Boolean function of the other target genes is assigned to each gene $i$in a network. This function predicts the state of a target gene at a point in time leading to an enormous number of Boolean functions. Individuals with particular interest on details of the theoretical working principles are advised to look at literature [94,95,96,97].

Huang et al. [98] considered the binary approximation of transcription to be an over-simplification. Many biological phenomena are portrayed as continuous. Nevertheless, numerous studies have demonstrated that binary (and ternary) discretization sometimes

yield reliable results. Using the binary approximation, Huang and co-workers also showed that the Boolean formalism yields biologically meaningful results. Though computationally costly, using ternary and higher order discretization levels has the potential to reveal more characteristics of a biological network [99]. Liang et al. [6] proposed the REVerse Engineering ALgorithm (REVEAL) for large-scale Boolean network reconstruction. Boolean networks are capable of capturing the dynamic behavior in complex systems when used with high through-put microarray data. Random Boolean networks realistically capture essential network characteristics [67,98,100,101]. According to Shmulevich and Zhang [102], this ability to realistically capture crucial features of networks justifies using the Boolean formalism for network inferences.

Boolean networks only allow for qualitative rather than quantitative inferences. Steggles et al. [103] showed that Boolean networks fail to capture vital network dynamics. To provide insight into the complexities with random Boolean networks, suppose that ℓrepresents the network connectivity number, n–number of nodes. By assumingthat we have ℓ levels, then the number of possible states for such a network is 2^{ℓ} with a total possible combination of $2^{2^{\ell}}$logical functions. It therefore follows that for each node has a total of $C(n,\ell)$ possible un-ordered combinations for ℓedges. Here the symbol$C$is used to refer to a combinatorial. The number of possible networks for a given set of parameters is $\left(2^{2^{\ell}} n \times C(n,\ell)\right)^{\ell}$ (proof not given here, interested readers can look at [104]). The higher ℓgets the more complex the topology of the network becomes. Boolean networks use discretized data which to some extent subjects the formalism to information loss from the data discretization. The dimensionality challenges highlight the complexity of the Boolean formalism in studying biological networks.

2.5 Ordinary Differential Equation (ODE) Formalism

2.5.1 Variants of the ODE formalism

ODEs are efficient for modeling small-sized networks [105], but face the problem of computational time complexity for large dimensional networks [106][107]. De Hoon et al. [108] illustrated the efficiency of ODEs with real data from *B. subtilis*. By using ODEs, network dynamics can be studied prior to parameter identification [109]– usually through modeling and simulations. Simulation is a useful way to predict systems behavioral dynamics and the way the network components (e.g. mRNA and protein concentrations) vary over time. One of the most popular differential equation formalism for network inference is the S-system [110,111,112,113,114,115]. They are advantageous in terms of system analysis and control design since it allows for convenient use of analytical and computational methods. Steady-state evaluation, control analysis and sensitivity analysis of a given system can be established mathematically using S-system parameters [116,117].

A major disadvantage of S-systems is that it requires a large number of parameters, (i.e. $2\kappa(\kappa+1)$, κbeing the number of state variables) to describe a network further posing a challenge during parameter estimation. The proof of how the number $2\kappa(\kappa+1)$ is arrived at is not given here but further insight on its derivation can be found in [116,117].Another bottleneck lies in the parameter estimation which is also discussed in a subsequent section of this paper. Kabir et al. [118] used Linear Time Invariant models to infer biological network structures and parameter estimation using synthetic data. Zhan and Yeung [119] proposed a method that combines spline theory with Linear Programming and Nonlinear Programming.

They used enzyme kinetics models to describe the network dynamics and study systems parameter sensitivity. ODEs are used with time course perturbation data and knock out data.

2.5.2 Mathematical formulation

A popular representation for modeling biological networks is the transcription-translation model:

$$\begin{cases} \dot{x}_i = \phi(f_i(z_1; k_{i1}, h_1), \ldots, f_i(z_m; k_{im}, h_m)) - k_{id}x_i, \\ \dot{z}_i = \psi(x_i; r_i) - \eta_i z_i, \qquad \text{given } x_i(0), z_i(0) \end{cases}$$

This model formalism arises from the central Dogma of molecular biology in which the DNA is transcribed to mRNA and then translated into proteins.The nonnegative constants $k_{id}$and $\eta_i$represent the mRNA and protein degradation parameters, respectively; h_l $(l = 1,\ldots,i,\ldots,m)$ are Hill coefficients, m– number of TFs for gene i and n– number of genes. The vector-valued functions $f_i \in \{f_i^-, f_i^+\}: \Re \to \Re$describe the gene regulation in time; $f_i^-$and $f_i^+$are repressing and activating Hill functions, respectively. These functions describe the dependence of the mRNA concentration on the protein levels z_l. The mRNA synthesis function ϕconsists of sums or products of f_i. The specific formulation of these functions is based on the specific molecular regulatory mechanism, which TFs are involves, the target genes of interest, the presence or absence of a feedback, feed-forward mechanism, time delay etc.

The translation function $\psi$of mRNA $x_i$to protein $z_i$is often considered to be linear. When the protein $z_l$has no effect on the mRNA levelsx_l, then the corresponding term in the model is set to zero, i.e. $f_i = 0$.Occasionally, people ignore what is happening to state variables that might be crucial for determining the network dynamics and instead focus mainly on the observable components which are easily quantifiable. However, understanding the particular roles of any hidden state variables might be vital in explaining certain peculiar behavioral dynamics of a network, thereby, reducing the reliance on considering the model as a black box with hidden variables. The most common forms of $f_i$are Hill functions and Michaelis-Menten functions [120,121,122]. The parameters$[k_{i1}, \ldots, r_i, \eta_i]^T$can be estimated using the Maximum Likelihood approach [123,124,125]. A good illustration of the working principles of ODEs in network inference can be found in the work of Polynikis et al. [126]. They used Hill functions in ODEs by exploiting analytic approaches such as steady state analysis by investigating how the concentration of mRNAs and proteins change in time for a given network.

Big networks require a large number of ODEs and parameters. This further complicates the network inference and increases the risk of obtaining biased parameter estimates. The computational requirements for such networks are enormous and quickly scale up with network size. ODEs effectively handle small-dimensional networks and irregularly sampled data, for instance by using Kalman Filters [127]. Commonly, perturbation data obtained as a result of variations in a factor that influences gene expression is used for the ODE formalism. In perturbation experiments in which gene expression is followed in time, the perturbation is performed at time zero. After this the cells are allowed sometimes to recover and regain their steady states. Overall, differential equations model network dynamics in fine details and lead to biological realistic inferences. Additionally, on average it requires a small amount of data and uses both discrete and continuous data, hence it is flexible and convenient.

2.6 Hybrid Models

The hybrid formalism combines continuous models of slowly changing metabolite concentrations with discrete representations of the model components, particularly those that are changing in a switch-like fashion between two states. Hybrid models, though rarely used in reverse engineering biological networks, have been shown to successfully yield interesting results. Hybrid models enable creation of quantitatively accurate representations of the concentrations of metabolites in a cell [123]. This formalism is based on a combination of methods that embrace discrete-continuous modeling. Zhang et al. [128] presented a novel network inference method by integrating gene expression data and gene functional category information. Their network inference approach consisted of two parts, namely: (i) module selection, and (ii) network inference. The first of these parts uses optimal modules through fuzzy c–mean clustering and incorporating gene functional category information, while the latter uses a hybrid of particle swarm optimization and recurrent neural network (PSO-RNN) methods during the network inference. The latter method was tested on real data. This demonstrates the applicability of hybrid models in network inference. In this setting the RNN is the model formalism and the PSO refers to the parameters estimation approach. The specifics of the theoretical framework can be found in [128].

In contrast to the approach of Zhang et al. [128] discussed above, Fernandez et al. [129] focused on reducing computation time, increasing the efficiency and robustness of their hybrid based optimization routine. Their approach integrates aspects of experimental design by evaluating the Fisher Information Matrix and effectively handles data measurement noise and partial observations in data. Additionally, local and global model identifiability is also tested in the same approach. Hybrid systems require good prior knowledge of a biochemical system. Prior knowledge of network pathways aid setting up mathematical models upon which experimental designs are based.

2.7 Supporting Tools and Network Information Level

The singular value decomposition (SVD) [130,131,132], independent component analysis [133] and principal component analysis [134] when used with ODEs have aided successful network inference. However, this combination does not guarantee total recovery of the true genetic network structure [135,136]. The use of SVD and ODEs can be improved by exploiting the network structural prior knowledge. This is mainly applied to linear systems of equations for which the dynamics of gene expression data is portioned into a noise-free and data measurement noise. Such a partition enables acquisition of an analytic solution and an approximation of a numerical solution to the system of equations under consideration. It is not possible to directly apply the analytic solution of the linear system since gene expression data because of data measurement noise. In such cases using SVD ensures that an accurate solution to the linear system of equations as demonstrated in [130]. For most TCEs the number of transcripts out-weighs the number of the time points, so the SVD can be used to circumvent this problem [135].

Generally, reverse engineering networks is performed at two information levels, low and high. Bayesian and Neural networks are examples of high-level methods. Low-level methods include ordinary, partial and delay differential equations. The choice of a level depends on the research objective in consideration. Often distinctions are made between the forward and reverse engineering [137]. Achieving high true positive (TP) values (≈100%, nearly all connections/edgesbetween nodes in a network are correctly predicted) from biological

inference without necessarily compromising the levels of false discoveries is a nontrivial task. Likewise, with large datasets, realistic values for TPs are ≈90%. Higher TP values lead to higher false discovery rates, which particularly holds true for large networks.

2.8 Data Requirements, Data Sampling Strategies and Precision

Continuous dynamic models are suitable for modeling gene transcription data which is known to be a continuous process in time. When modeling with sampled data, the challenge is to keep the number of data samples within cost-effective values. Optimal data sampling criteria maximize cost-effectiveness [138,139,140,141,142]. Likewise, large data measurement errors, low resolution and small dimensional datasets can negatively impact meaningful network inference. For instance Bourque and Sankoff [143] showed that for relatively low data measurement noise levels (<2.5%) and $n \leq 12$, false negative rate (FNRs) and false positive rates (FPRs) of as high as 30% are to be expected for a network with low gene transcription correlation values. These findings were validated using both synthetic and true biological networks. In GNR, using time course data, a large number of time points in relation to model fitting to data are associated to low error rates (false negatives, FN and false positives, FP). In such cases the ratio FN/FP seemingly remains invariant of network size.

In practice the number of data points in a TCE is often low due to financial constraints. Recently, advancements using powerful computational approaches have been made, e.g. the improved self-adaptive Naïve Bayesian Tree (NBTree)[144], in these advancements it was demonstrated that FPRs can tremendously reduce up to acceptable levels. Their work resulted in accuracy levels or TPs of ~99% using the NBTree algorithm. In another algorithm (a noise and redundancy reduction technique improves accuracy of gene regulatory network inference (RARROMI)), proposed by [145], it was demonstrated that it outperformed the then available algorithms by achieving an accuracy level of ~63%. However, such a performance rate might beto some extent dependent on the test datasets used, and is likely to vary for various datasets depending on their quality and dimension. It is therefore important that computational methods and algorithms be trained on a variety of well-known datasets obtained from a diversity of experiments. This is to ensure that the error rates from the network inference can be reported with much bigger certainty and confidence. This highlights the significance of good design strategies to ensure that experimental costs are minimized while still obtaining datasets with high information content.

Kim et al. [146] studied the influence of varying the number of data points on estimated standard errors from a fitted regression model of three classifications, namely: least squares, total least squares and constraint total least squares. Of these criteria, the constraint total least squares approach yielded the best performance. Their findings suggest that a very small number of time points for analysis yields biased results while too many time points is associated with increased financial costs and more experimental time for the data collection process. The data requirement for accurate network inference with $n$genes was investigated by Kanehisa et al. [147]. For a Boolean network with average in-degree connectivity,κ the data requirement scales to $2\kappa + \kappa log(n)$ compared to a fully connected network of $2n$transition states [148]. The in-degree of a node in a genetic network refers to the number of head endpoints adjacent to that specific node. Connectivity is the number of edges linking the nodes in a network.

2.9 Modeling Cycle in Network Inference

The choice of model formalism is important for reverse engineering biological networks, for instance the decision of whether to use a linear or nonlinear models and discrete or continuous models. These choices ensure that the time dynamics of the gene expression are well captured. The problem of network reconstruction is indeed challenging as put forward by D'Alch´e-Buc and Schachter [149]:" As this field of research matures, it is apparent that that there is no one-size-fits-all solution but rather a range of frameworks and methods, each with its specific trade-off between abstraction and tractability, the ultimate test being the ability to answer relevant biological questions." Overall, the steps involved in the modeling cycle are summarized in Fig. 3. Some intermediary steps are skipped from the scheme in this Figure for conceptualization purposes. Step 1 (prior knowledge) involves a thorough literature and database search, or seeking an expert's opinion. Once the information is gathered, it is important to think in advance of a formalism to model your system (step 2).

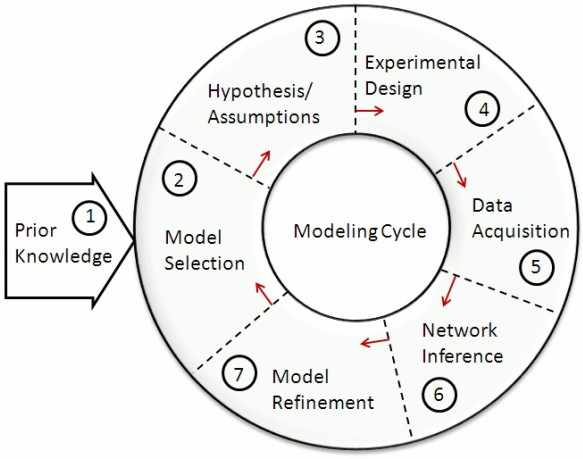

Fig. 3. The Modeling Cycle in network inference: the summary is given in steps 1 to 7.

The starting points for this cycle may vary depending on the situation at hand. Careful handling of each step determines the quality of the network inference. When performing network inference and parameter identification it is vital to clearly state the research hypothesis of interest (step 3). This is closely followed by experimental design (step 4). Getting the experimental design right ensures that high quality data is collected with all the required variables being measured (step 5). Step 6 involves the actual network inference, this process is mathematical rigorous and computationally costly. Upon obtaining the identification results (step 6), there is need to cross-check how well the results meet the expectations. This leads to the model refinement (step 7) and the process is terminated upon obtaining a good model fit to data, otherwise, there is need for model re-adjustment. Between these links small steps exist, all of which contribute to GNR, hence the term Modeling Cycle (Fig.3).

2.10 Parameter Identification

Parameter identification is an important part of network inference and the prediction of network behavior in time. Parameter estimation techniques like linear iterative models, stochastic optimization methods and constrained linear and nonlinear regression models are often used in GNR [100,113,119,123,129,137]. Each approach has its own strengths and weaknesses many of which are strongly linked to the data quality and modeling approach. Genomic, proteomic and other -omic data types are prone to noise and/or have missing data. With recent advancements in high resolution data acquisition techniques, the issue of noisy or missing data is becoming less problematic. Focus has slowly shifted away from obtaining high-resolution data to unraveling the masked information contained in such datasets. Most methods focus on small-sized networks because of the computational challenges associated with larger networks. However, the need to accurately describe molecular mechanisms in biochemical systems cannot be understated. To achieve such high performance descriptions, parameters have to be accurately and precisely identified. Many gradient search optimization procedures often fail to converge to a unique parameter value. Such a failure can bepartly attributed to the presence of correlated parameters and model degeneracy with respect to the cost function. This often compromises the precision with which the parameters are identified, therefore, there is need for alternative solutions to local optimization methods.

Parameter identification, an important aspect of biological network inference, can be studied using sensitivity analysis. Sensitivity analysis is useful for assessing which parameters significantly affect the outputs or measured variables of interest following a network perturbation with some stimuli [6]. Conventionally parameter sensitivity analysis is used as a tool for analysis and design in engineering systems theory. Although it has mostly been applied in physical systems rather than biological systems, its use in the latter has recently increased, especially in the study of complex networks. By using parameter sensitivity analysis, once the most influential parameters are identified the correlation matrix between the parameters is then investigated. Thereafter, the least sensitive parameters can be left out of a model thereby reducing the model complexity and yet retaining its explanatory power. In dealing with model complexity, an intriguing question that comes to mind is how possible is coarse-graining in network? The term course-graining here refers to complexity reduction (see e.g. Erban [150]).For large networks with thousands of genes the number of differential equations required to describe a particular system becomes huge [106]. This implies an increased number of kinetic parameters, e.g. mRNA production and decay rates and Hill constants. In principle, using parameter sensitivity analysis, some of these parameters can be coalesced or dropped from the model – leaving a simpler, yet still powerful model to describe the network dynamics.

3. DISCUSSION AND CONCLUDING REMARKS

The use of OED to enrich data used in GNR has been exploited in engineering. All this shows the worth of using OEDs in improving parameter estimation accuracy and precision. More on how OED can be used to improve parameter estimation can be found in the work of Faller et al. [140]. However, its use is not yet widely embraced in network inference. It would be interesting to see more applications of OEDs in GNR, particularly in parameter estimation problems as suggested in [47]. Progress in the reconstruction of biological networks has been hugely supported by advancements in data acquisition methods. High-throughput technologies have aided quantification of metabolite abundances in cells, e.g. measuring

gene transcription using DNA microarrays or Real Time-quantitative Polymerase Chain Reaction (RT-qPCR) technique. Integrating different omics datasets, e.g. genomic, transcriptomic and proteomic data drastically improves the quality of network inference [151]. The decision of which model formalism to use should be based on data attributes such as: data noise level, data dimension, data type (continuous or discrete, relative or absolute) and the research goal in consideration. Softwares for analyzing and visualizing data have aided network inference – helping improve our understanding of biological pathways, e.g. the Complex Pathway SImulator "COPASI" – abiochemical network simulator developed by Mendes et al. [152] and its earlier version by Hoops et al. [153]; the correlational based approach "Gene Net" by Opgen-Rhein and Strimmer [154]; "Snazer" – a network analyzer and data visualization software by Mazza et al. [155]; "Cell Designer" – a tool for modeling biochemical networks[156][157]. Most of these tools are freely available for download and can be easily installed and run inthe specific software language in which they were written – although some might only run well on some operating systems. Such information is provided in the software user manuals. Prior to data analysis and/or model fitting, it's is advisable to ensure that data collection criterion, experimental conditions, normalization criteria as sound enough to ensure that the results are trustworthy. Preliminary, prior to deciding which approach to use to answer specific research questions or test hypothesis related to network inference, one can start by looking at the summary of advantages and disadvantages of the methods given in Table 1.

The pace at which algorithms, software, network analysis and inference tools have developed is astonishing– a development that has facilitated analysis and visualization of high-dimensional datasets. Overall, deriving meaningful inferences from biological data is no easy task. There is need for regular training to update our skills and keep abreast with literature information and latest developments in the field through interaction with experts in conferences, workshops and/or courses on specialized courses on,e.g. network reconstruction and mathematical modeling. Given the vast amount of literature, it is easy to get lost in a sea of information and inexperienced individuals in GNR are advised to seek an expert's opinion where necessary. Availability of relatively cheap high through-put data acquisition techniques has revolutionized research on network inference, e.g. recently, there has been interesting insight into the important role of small RNAs (sRNAs) in posttranscriptional regulation in eukaryotes – particularly in bacteria. The reduced cost of genome sequencing and RNA-Seq data in general has made these advancements in uncovering the roles of sRNAs possible, see [158].Such studies have provided insight into the sRNA functional and regulatory roles. Intensified efforts along this line of research could see interesting discoveries in the line of drug discovery, cancer research and a general improvement in our understanding of posttranscriptional regulatory events. In a nutshell, the most important attributes in network inference are: (i) the ability to model direct and indirect cyclic regulation, (ii) model dynamics, (iii) robustness even for small dimensional datasets, (iv) the ease of implementation and user-friendliness, and (v) the ability to capture reality in life with simple models.

In summary, this review article is mainly concerned with single cells rather than the interaction between cells. Therefore, the work was focused mostly on approaches related to studies on molecular biology and molecular genetics rather than towards developmental biology. During data analysis and network reconstruction it is advisable to pay particular attention to and clearly specify what hypothesis is to be tested, what dataset you intend to use, e.g. gene knock-out data, or time-series data; data quality and dimension; is it discrete or continuous data; what is the noise levels in your data (signal-to-noise ratio), what and how many parameters are required to describe your network or biochemical system; decide if you

are only interested in association between variables or trends in gene expression. It is important to identify appropriate data analysis tools: whether it is freely available (open access) or commercially available software, scrutinize and pay particular attention to what type of data is being used, check if the data are from similar experiments, related strains, ensure that the data is properly scaled and normalized. All the above is in an effort to ensure improved network inference.

ACKNOWLEDGEMENTS

This work was self-sponsored.

COMPETING INTERESTS

The author declares no competing interests.

REFERENCES

1. Lashkari DA, McCusker JH, Davis RW. Whole genome analysis: experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR. Proc Natl Acad Sci U S A. 1997;94:8945-8947
2. Erdos P, Renyi A. On random graphs. Publicationes Mathematicae. 1959;6:290-297
3. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 2000. 2000;7:601-620
4. Chen T, He HL, Church GM. Modeling gene expression with differential equations. Pac. Symp. Biocomp. 1999:29-40
5. Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. Development. 1997;124:1851-1864
6. Liang S, Fuhrman S, Somogyi R. REVEAL, A General Reverse Engineering Algorithm For Inference Of Genetic Network Architectures. In Proc Pac Symp Biocomput World Scientific Publishing Co. 1998;3:18-29
7. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. 2002;9:67–103
8. Gardner TS, Faith JJ. Reverse-engineering transcription control networks. Phys Life Rev. 2005;2:65-88
9. van Someren EP, Wessels LF, Backer E, Reinders MJ. Genetic network modeling. Pharmacogenomics. 2002;3:507-525
10. Schlitt T, Brazma A. Current approaches to gene regulatory network modelling. BMC Bioinformatics. 2007;8:S9
11. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007;3
12. Cho KH, Choo SM, Jung SH, Kim JR, Choi HS, Kim J. Reverse engineering of gene regulatory networks. IET Syst Biol. 2007;1:149-163
13. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature. 2000;5:651-654
14. Zare H, Sangurdekar D, Srivastava P, Kaveh M, Khodursky A. Reconstruction of *Escherichia coli* transcriptional regulatory networks via regulon-based associations. BMC Syst. Biol. 2009;3:39
15. Maraziotis IA, Dragomir A, Bezerianos A. Gene networks reconstruction and time-series prediction from microarray data using recurrent neural fuzzy networks. IET Syst Biol. 2007;1:41-50

16. Keurentjes JJB, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek, L.B, Peeters, A.J.M, Vreugdenhil D, Koornneef M, Jansen RC. Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. PNAS. 2007;108:1708-13
17. Ma S, Gong Q, Bohnert HJ. An *Arabidopsis* gene network based on the graphical Gaussian model. Genome Res. 2007;17:1614-1625
18. Levine AJ, Hu W, Feng Z. The p53 pathway: what questions remain to be explored?. Cell Death Differ. 2006;13:1027-1036
19. Batchelor E, Loewer A, Lahav G. The ups and downs of p53: understanding protein dynamics in single cells. Nat Rev Cancer. 2009;9:371-377
20. Wagner A. Estimating coarse gene network structure from large-scale gene perturbation data. Genome Res. 2002;12:309-315; doi:10.1101/gr.193902
21. de la Fuente A, Brazhnik P, Mendes P. Linking the genes: Inferring quantitative gene networks from microarray data. Trends Genet. 2002;18:395-398
22. Gardner TS, di Bernardo D, Lorenzo D, Collins JJ. Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. Science. 2003;301:102-105
23. Friedman N. Inferring cellular networks using probabilistic graphical models. Science. 2004;303:799-805
24. di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat Biotechnol. 2005;23:377-383
25. Rice JJ, Tu Y, Stolovitzky G. Reconstructing biological networks using conditional correlation analysis. Bioinformatics. 2005;21:765-773
26. de la Fuente A, Makhecha DP. Unravelling gene networks from noisy underdetermined experimental perturbation data. Systematic Biol. 2006;153:257-262
27. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V. The Inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology datasets *de novo*. Genome Biol. 2006;7:R36
28. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5
29. Marbach D, Mattiussi C, Floreano D. Replaying the evolutionary tape: Biomimetic reverse engineering of gene networks. Annals of the New York Academy of Sciences. 2009;1158:234-245
30. Brazhnik P, de la Fuente A, Mendes P. Gene networks: how to put the function in genomics. Trends Biotechnol. 2002;20:467-472
31. Dharwadker A, Pirzada S. Applications of Graph Theory. Proceedings of the institute of mathematics; 2011
32. Wiener A, Polynomials H. Distance in Molecular Graphs - Theory. In: Gutman I, Furtula B, editors. Altenburg, Wiener, and Hosoya Polynomials; 2011. p. 49-70
33. Kitano H. Foundations of Systems Biology, chapter Systems Biology: Toward System level Understanding of Biological Systems. MIT Press. 2001:1-29
34. Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics. 2005;21:754-764
35. Yip KY, Alexander RP, Yan KK, Gerstein M. Improved Reconstruction of *In Silico* Gene Regulatory Networks by Integrating Knockout and Perturbation Data. PLoS One. 2010;5
36. Stolovitzky G, Monroe D, Califano A. Dialogue on Reverse-Engineering Assessment and Methods: The DREAM of High-Throughput Pathway Inference. In: Stolovitzky G and Califano A, editor. ; 2007. p. 11-22

37. Stolovitzky G, Prill RJ, Califano A. Lessons from the DREAM2 Challenges. Annals of the New York Academy of Sciences. 2009;1158:159-95
38. Parisi F, Koeppl H, Naef F. Network inference by combining biologically motivated regulatory constraints with penalized regression. Ann N Y Acad Sci. 2009;1158:114-24
39. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PLoS One. 2010;5
40. Omony J. The Dynamics of the XlnR Regulon in *Aspergillus niger*: a Systems Biology Approach. 2012
41. Wagner A. How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. Bioinformatics. 2001;17:1183-1197
42. Miller MA, Feng XJ, Li G, Rabitz HA. Identifying biological network structure, predicting network behavior, and classifying network state with High Dimensional Model Representation (HDMR). PLoS One. 2012;7:e37664; doi:10.1371/journal.pone.0037664; 10.1371/journal.pone.0037664
43. del Rio G, Koschutzki D, Coello G. How to identify essential genes from molecular networks?. BMC Syst Biol. 2009;3:102-0509-3-102; doi:10.1186/1752-0509-3-102; 10.1186/1752-0509-3-102
44. Cai X, Bazerque JA, Giannakis GB. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS Comput Biol. 2013;9:e1003068; doi:10.1371/journal.pcbi.1003068; 10.1371/journal.pcbi.1003068
45. Steinke F, Seeger M, Tsuda K. Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models. BMC Syst Biol. 2007;1:51; doi:10.1186/1752-0509-1-51
46. Steiert B, Raue A, Timmer J, Kreutz C. Experimental design for parameter estimation of gene regulatory networks. PLoS One. 2012;7:e40052; doi:10.1371/journal.pone.0040052; 10.1371/journal.pone.0040052
47. Omony J, Mach-Aigner AR, de Graaff LH, van Straten G, van Boxtel AJB. Evaluation of design strategies for time course experiments in genetic networks: case study of the XlnR regulon in *Aspergillus niger*. IEEE/ACM Trans Comput Biol Bioinform. 2012;9:1316-1325
48. Zak DE, Gonye GE, Schwaber JS, Doyle III FJ. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an *in silico* network. Genome Res. 2003;13:2396-2405
49. Sachs K, Perez O, Pe'er D, Lauffenburger DA. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science. 2005;308:523-529
50. Nabatame S, Iba H. Estimation of Gene Regulatory Network Using Stochastic Differential Equation Model. 2006
51. de Hoon M, Imoto S, Kobayashi K, Ogasawara N, Miyano S. Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. Proc. Pac. Symp. Biocomp. 2003;8:17-28
52. Vu TT, Vohradsky J. Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of *Saccharomyces cerevisiae*. Nucleic Acids Research. 2007;35:279-287
53. Ropers D, de Jong H, Page M, Schneider D, Geiselmann J. Qualitative Simulation of the Nutritional Stress Response in *Escherichia coli*. INRIA, Rapport de Reacherche. 2004
54. Kim SY, Imoto S, Miyano S. Infering gene networks from time series microarray data using dynamic Bayesian networks. Briefings in Bioinformatics. 2003;4:228-235.

55. Nachman I, Regev A, Friedman N. Inferring quantitative models of regulatory networks from expression data. Bioinformatics. 2004;20 Suppl 1:i248-56; doi:10.1093/bioinformatics/bth941.
56. Ong IM, Glasner JD, Page D. Modelling regulatory pathways in *E. coli* from time series expression profiles Bioinformatics. ISMB2002. 2002.
57. Chen X, Chen M, Ning K. BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. Bioinformatics. 2006;22:2952-2954; doi:10.1093/bioinformatics/btl491
58. Spirtes P, Glymour G, Kauffman S, Aimalie V, Wimberly F. Constructing bayesian network models of gene expression networks from microarray data. Proc. Atlantic Symp. Comp. Biol., Genome Information Systems & Technology. 2000
59. Ong IM, Page D. Inferring regulatory pathways in *E. coli* using dynamic Bayesian networks. Technical Report 1426, University of Wisconsin-Madison. 2001
60. Rogers S, Girolami M. A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics. 2005;21:3131-3137; doi:10.1093/bioinformatics/bti487
61. Zou M, Conzen SD. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics. 2005;21:71-79; doi:10.1093/bioinformatics/bth463
62. Zhang Y, Deng Z, Jiang H, Jia P. Dynamic Bayesian Network (DBN) with Structure Expectation Maximization (SEM) for Modeling of Gene Network from Time Series Gene Expression Data. BIOCOMP. 2006:41-47
63. Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. Bioinformatics. 2001;17 Suppl 1:S215-24
64. Soinov LA, Krestyaninova MA, Brazma A. Towards reconstruction of gene networks from expression data by supervised learning. Genome Biol. 2003;4:R6
65. Cerulo L, Elkan C, Ceccarelli M. Learning gene regulatory networks from only positive and unlabeled data. BMC Bioinformatics. 2010;11:228-2105-11-228; doi:10.1186/1471-2105-11-228; 10.1186/1471-2105-11-228
66. Schafer J, Opgen-Rhein R, Strimmer K. Reverse Engineering Genetic Networks using the GeneNet Package. R News. 2006;6
67. Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol. 1969;22:437-467
68. Werhli AV, Grzegorczyk M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics. 2006;22:2523-2531; doi:10.1093/bioinformatics/btl391
69. Pearl J. Causality: Models, Reasoning, and Inference. Cambridge University Press. 2000
70. Murphy K, Mian S. Modelling Gene Expression Data Using Dynamic Bayesian Networks. Technical Report. 1999
71. Heckerman D. Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery. 1997;1:79-119
72. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. PNAS. 2010;107:6286-6291
73. Iancu OD, Kawane S, Bottomly D, Searles R, Hitzemann R, McWeeney S. Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics. 2012;28:1592-1597; doi:10.1093/bioinformatics/bts245; 10.1093/bioinformatics/bts245

74. Pihur V, Datta S, Datta S. Reconstruction of genetic association networks from microarray data: a partial least squares approach. Bioinformatics. 2008;24:561-568; doi:10.1093/bioinformatics/btm640; 10.1093/bioinformatics/btm640
75. Bansal M, Della Gatta G, di Bernardo D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics. 2006;22:815-822; doi:10.1093/bioinformatics/btl003
76. Wiener N. The theory of prediction. In Modern Mathematics for Engineers. Volume 1. Edited by Beckenbach EF. ed. New York: McGraw-Hill; 1956.; 1956
77. Zou C, Ladroue C, Guo S, Feng J. Identifying interactions in the time and frequency domains in local and global networks - A Granger Causality Approach. BMC Bioinformatics. 2010;11:337
78. Granger C. Investigating causal relations by econometric models and cross-spectral methods. Econometrica. 1969;37:424-438
79. Mukhopadhyay ND, Chatterjee S. Causality and pathway search in microarray time series experiment. Bioinformatics. 2007;23:442-449; doi:10.1093/bioinformatics/btl598
80. Chen Y, Rangarajan G, Feng J, Ding M. Analyzing multiple nonlinear time series with extended Granger causality. Phys Lett. 2004;324:26-35
81. Nagarajan R. A note on inferring acyclic network structures using Granger causality tests. Int J Biostat. 2009;5:Article 10-4679.1119; doi:10.2202/1557-4679.1119; 10.2202/1557-4679.1119
82. Fujita A, Severino P, Sato JR, Miyano S. Granger causality in systems biology: modeling gene networks in time series microarray data using vector autoregressive models. Proceeding BSB2010 Proceedings of the Advances in bioinformatics and computatiional biology, and 5th Brazilian conference on Bioinformatics. 2010
83. de la Fuente A, Bing N, Hoeschele I, Mendes P. Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics. 2004;20:3565-3574; doi:10.1093/bioinformatics/bth445
84. Magwene PM, Kim J. Estimating genomic coexpression networks using first-order conditional independence. Genome Biol. 2004;5:R100; doi:10.1186/gb-2004-5-12-r100
85. Reverter A, Chan EK. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics. 2008;24:2491-2497; doi:10.1093/bioinformatics/btn482; 10.1093/bioinformatics/btn482
86. Butte AS, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Proc. Pac. Symp. Biocomput. 2000:418-429
87. Thomas R, d'Ari R. Biological Feedback. CRC, Press, Boca Raton. 1990
88. Hopcroft JE, Motwani R, Ullman JD. Introduction to Automata Theory, Languages, and Computation. Addison Wesley, Boston, MA, USA; 2006
89. Shmulevich I, Dougherty ER, Zhang W. From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. Proceedings of the IEEE. 2002;90:1792
90. Dougherty ER, Shmulevich I. Mappings Between Probabilistic Boolean Networks. Signal Processing. 2003;83:799-809
91. Shmulevich I, Kauffman SA. Activities and sensitivities in boolean network models. Phys Rev Lett. 2004;93:048701
92. Hashimoto RF, Kim S, Shmulevich I, Zhang W, Bittner ML, Dougherty ER. Growing genetic regulatory networks from seed genes. Bioinformatics. 2004;20:1241-1247; doi:10.1093/bioinformatics/bth074
93. Rocke DM, Durbin B. A Model for Measurement Error for Gene Expression Arrays. J. Comput. Biol. 2001;8:557-569

94. Palsson BO. Systems Biology: Properties of Reconstructed Networks. Cambridge university press: UK; 2006
95. Newman MEJ. Networks: An Introduction. Oxford university press: Oxford, UK; 2010
96. Alon U. An Introduction to Systems Biology: Design Principles of Biological Circuits. Boca Raton, FL: Chapman and Hall / CRC Press; 2006
97. Alon U. Network motifs: theory and experimental approaches. Nat. Rev. Genet. 2007;8:450-461
98. Huang S. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. J Mol Med (Berl). 1999;77:469-480
99. Repsilber D, Liljenstrom H, Andersson SG. Reverse engineering of regulatory networks: simulation studies on a genetic algorithm approach for ranking hypotheses. BioSystems. 2002;66:31-41
100. Dimitrova E, Garcia-Puenteb LD, Hinkelmann F, Jarrah AS, Laubenbacher R, Stiglere B, Vera-Licona P. Parameter estimation for Boolean models of biological networks. Theoretical Computer Science. 2010;412:2816-2826
101. Ideker TE, V T, Karp RM. Discovery of regulatory interactions through perturbation: inference and experimental design. Pac. Symp. Biocomput. 2000;5:302-313
102. Shmulevich I, Zhang W. Binary analysis and optimization-based normalization of gene expression data. Bioinformatics. 2002;18:555-565
103. Steggles LJ, Banks R, Shaw O, Wipat A. Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach. Bioinformatics. 2007;23:336-343; doi:10.1093/bioinformatics/btl596
104. Somogyi R, Sniegoski CA. Modeling the complexity of genetic networks: Understanding multigene and pleiotropic regulation. Complexity. 1996;1:45
105. August E, Papachristodoulou A. Efficient, sparse biological network determination. BMC Syst Biol. 2009;3:25-0509-3-25; doi:10.1186/1752-0509-3-25; 10.1186/1752-0509-3-25
106. Bornholdt S. Less Is More In Modeling Large Genetic Networks. Systems Biology (Perspectives), Science. 2005;310:449-451
107. Shi Y, Mitchell T, Bar-Joseph Z. Inferring pairwise regulatory relationships from multiple time series datasets. Bioinformatics. 2007;23:755-763
108. de Hoon MJ, Makita Y, Nakai K, Miyano S. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. PLoS Comput Biol. 2005;1:e25; doi:10.1371/journal.pcbi.0010025
109. Mazur J, Ritter D, Reinelt G, Kaderali L. Reconstructing nonlinear dynamic models of gene regulation using stochastic sampling. BMC Bioinformatics. 2009;10:448
110. Kimura S, Araki D, Matsumura K, Okada-Hatakeyama M. Inference of S-system models of genetic networks by solving one-dimensional function optimization problems. Math Biosci. 2012;235:161-170; doi:10.1016/j.mbs.2011.11.008; 10.1016/j.mbs.2011.11.008
111. Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M. Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics. 2003;19:643-650
112. Thomas R, Mehrotra S, Papoutsakis E, Hatzimanikatis V. A model-based optimization frame-work for the inference on gene regulatory networks from dna array data. Bioinformatics. 2004;20:3221-3235
113. Almeida JS, Voit EO. Neural-network-based parameter estimation in S-system models of biological networks. Genome Inform. 2003;14:114-123
114. Spieth C, Streichert F, Speer N, Zell A. A memetic inference method for gene regulatory networks based on S-systems. 2004:152-157

115. Norman N, Iba H. Inference of gene regulatory networks using S-system and differential evolution. 2005:439-446.
116. Cho DY, Cho KH, Zhang BT. Identification of biochemical networks by S-tree based genetic programming. Bioinformatics. 2006;22:1631-1640; doi:10.1093/bioinformatics/btl122.
117. Voit EO. Computational Analysis of Biochemical Systems. Cambridge University Press; 2000
118. Kabir M, Noman N, Iba H. Reverse engineering gene regulatory network from microarray data using linear time-variant model. BMC Bioinformatics. 2010;11 Suppl 1:S56-2105-11-S1-S56; doi:10.1186/1471-2105-11-S1-S56; 10.1186/1471-2105-11-S1-S56.
119. Zhan C, Yeung LF. Parameter estimation in systems biology models using spline approximation. BMC Syst. Biol. 2011;5:14.
120. Hofmeyr JH, Cornish-Bowden A. The reversible Hill equation: how to incorporate cooperative enzymes into metabolic models. Comput Appl Biosci. 1997;13:377-385.
121. Mendes P, Sha W, Ye K. Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics. 2003;19:ii122–ii129.
122. Fersht A. Enzyme structure and mechanism. In: W. H. Freeman and Company, New York, editor. ; 1985.
123. Singhania R, Sramkoski RM, Jacobberger JW, Tyson JJ. A Hybrid Model of Mammalian Cell Cycle Regulation. PloS Comp Biol. 2011;7:e100107.
124. Fisher RA. On an absolute criterion for fitting frequency curves. Messenger of Mathematics. 1912;41:155-160
125. Ljung L. Systems identification. Theory for the user Prentice Hall PRT. 1999;21.
126. Polynikis A, Hogan SJ, di Bernardo M. Comparing different ODE modeling approaches of gene regulatory networks. J. Theor. Biol. 2009;261:511-530.
127. Kasabov NK, Chan ZSH, Jain V, Sidorov I, Dimitrov DS. Gene Regulatory Network Discovery from Time series Gene Expression Data - A Computational Intelligence Approach. ICONIP. 2004;3316:1344-1353.
128. Zhang Y, Xuan J, de los Reyes BG, Clarke R, Ressom HW. Reverse engineering module networks by PSO-RNN hybrid modeling. BMC Genomics. 2009;10 Suppl 1:S15-2164-10-S1-S15; doi:10.1186/1471-2164-10-S1-S15; 10.1186/1471-2164-10-S1-S15.
129. Rodriguez-Fernandez M, Mendes P, Banga JR. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. BioSystems. 2006;83:248-265; doi:10.1016/j.biosystems.2005.06.016.
130. Chang C, Ding Z, Hung YS, Fung PCW. Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. Bioinformatics. 2008;24:1349-1358.
131. Kaern M, Blake WJ, Collins JJ. The engineering of gene regulatory networks. Annu Rev Biomed Eng. 2003;5:179-206; doi:10.1146/annurev.bioeng.5.040202.121553
132. Laubenbacher R, Stigler B. A computational algebra approach to the reverse engineering of gene regulatory networks. J Theor Biol. 2004;229:523-537; doi:10.1016/j.jtbi.2004.04.037.
133. Lee SI, Batzoglou S. Application of independent component analysis to microarrays. Genome Biol. 2003;4:R76; doi:10.1186/gb-2003-4-11-r76.
134. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A. 2000;97:10101-10106.
135. Yeung MKS, Tegner J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression. Proc. Natl. Acad. Sci. 2002;99:6163-6168.

136. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell Jr JE. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res. 2003;13:1863–1872; doi:doi: 10.1101/gr.997703.
137. Chou IC, Voit EO. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Math. Biosci. 2009;219:57–83; doi:doi: 0.1016/j.mbs.2009.03.002.
138. Casey FP, Baird D, Feng Q, Gutenkunst RN, Waterfall JJ, Myers CR, Brown KS, Cerione RA, Sethna JP. Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model. IET Syst Biol. 2007;1:190-202.
139. Banga. J.R, Versyck KJ, van Impe JF. Computation of optimal identification experiments for nonlinear dynamic process models: a stochastic global optimization approach. Ind Eng Chem Res. 2002;41:2425-2430.
140. Faller D, Klingmuller U, Timmer J. Simulation methods for optimal experimental design in systems biology. Simulation-Transactions of the Society for Modeling and Simulation International. 2003;79:717-725.
141. Gadkar KG, Gunawan R, Doyle III FJ. Iterative approach to model identification of biological networks. BMC Bioinformatics 2005. 2005;6:155.
142. Balsa-Canto E, Alonso AA, Banga JR. An optimal identification procedure for model development in systems biology; 2007.
143. Bourque G, Sankoff D. Improving gene network inference by comparing expression time-series across species, developmental stages or tissues. J Bioinform Comput Biol. 2004;2:765-783.
144. Farid DM, Hoa NH, Darmont J, Harbi N, Rahman MZ. Scaling up Detection Rates and Reducing False Positives in Intrusion Detection using NBTree. World Academy of Science, Engineering and Technology. 2010;40.
145. 145. Zhang X, Liu K, Liu ZP, Duval B, Richer JM, Zhao XM, Hao JK, Chen L. NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. Bioinformatics. 2013;29:106-113; doi:10.1093/bioinformatics/bts619; 10.1093/bioinformatics/bts619.
146. Kim J, Bates DG, Postlethwaite I, Heslop-Harrison P, Cho KH. Least-squares methods for identifying biochemical regulatory networks from noisy measurements. BMC Bioinformatics. 2007;8:8; doi:10.1186/1471-2105-8-8.
147. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006;34:D354-7; doi:10.1093/nar/gkj102.
148. Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Proc. Pacific Symp. Biocomput. 1999;4:17-28
149. DAlche Buc F, Schachter V. Modeling and identification of biological networks. International Symposium on Applied Stochastic Models and Data Analysis. 2005
150. Erban R, Kevrekidis IG, Adalsteinsson D, Elston TC. Gene regulatory networks: a coarse-grained, equation-free approach to multiscale computation. J Chem Phys. 2006;124:084106; doi:10.1063/1.2149854
151. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models-a review. BioSystems. 2009;96:86-103; doi:10.1016/j.biosystems.2008.12.004; 10.1016/j.biosystems.2008.12.004.
152. Mendes P, Hoops S, Sahle S, Gauges R, Dada J, Kummer U. Computational modeling of biochemical networks using COPASI. Methods Mol Biol. 2009;500:17-59; doi:10.1007/978-1-59745-525-1_2; 10.1007/978-1-59745-525-1_2 .

153. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U. COPASI--a COmplex PAthway SImulator. Bioinformatics. 2006;22:3067-3074; doi:10.1093/bioinformatics/btl485.
154. Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol. 2007;1:37; doi:10.1186/1752-0509-1-37.
155. Mazza T, Iaccarino G, Priami C. Snazer: the simulations and networks analyzer. BMC Syst Biol. 2010;4:1-0509-4-1; doi:10.1186/1752-0509-4-1; 10.1186/1752-0509-4-1.
156. Funahashi A, Tanimura N, Morohashi M, Kitano H. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. BIOSILICO. 2003;1:162.
157. Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H. CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. In Proceedings of the IEEE. 2008;96:1254-1265.
158. Modi SR, Camacho DM, Kohanski MA, Walker GC, Collins JJ. Functional characterization of bacterial sRNAs using a network biology approach. Proc Natl Acad Sci U S A. 2011;108:15522-15527; doi:10.1073/pnas.1104318108; 10.1073/pnas.1104318108.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=316&id=32&aid=2478