Statistic Models in Biological Network Analysis

WANLU LIU and AUSTIN QUACH, University of California, Los Angeles

As the development of high throughput technology, it becomes more and more important for us to look at the behavior of genes, proteins and metabolites at whole genome level. Instead of studying a single gene, by understanding biological process at network level, we can shed light on some new mechanisms in both developmental processes and diseases. Modeling of these networks is an important challenge in the post genomic era. Several statistical models have been applied to analysis the biological networks including logic-based models, continuous models and correlation-based models. In this review, we focus on the different methods for reconstructing biological networks and analysis of their functionality.

Categories and Subject Descriptors: I.5.1 [**Networks Analysis**]: Models—*Biological Networks*

General Terms: Network Analysis

Additional Key Words and Phrases: Biological Network, Statistic Models, Network Construction, Network Analysis

1. INTRODUCTION

In the post-genomic era, genome-wide experiments have become commonplace and methods to interrogate such datasets are becomingly increasingly important. One approach to analyzing these large datasets is to model the experimental observations using a network approach. In a process known as (re)construction, inference, identification or reverse engineering, model parameters are fit to the data yielding a defined network that can then be analyzed to gain higher level insights into the complex molecular biology. Network biology is an expansive field and there are many different methods that are employed. In this survey paper we briefly review the frameworks of several popular modeling methods including logical models, boolean networks, Bayesian networks, weighted correlation networks, differential equations, and some basic network analysis concepts.

2. STATISTICAL MODELS IN NETWORK CONSTRUCTION

2.1 Logical Model

Logical model is the simplest method for constructing network and was first introduced by Kauffman and Thomas[Glass and Kauffman 1973][Thomas 1973]. Logical models represent the local state of each node in the system (for example, genes and proteins) at any time as a discrete level, and the temporal development of the system is often assumed to occur synchronously. Thus, with logical models, the researchers can only rely on the qualitative knowledge. Several statistical method have been applied in this type of discrete and logic-based models including Boolean networks, and Bayesian networks.

2.1.1 *Boolean Networks.* **Boolean networks** are a dynamic model of synchronous interactions between nodes in a network. The nodes can attain two alternative states: active(1) or inactive(0); in this way, the network is able to show the expression state of a gene or protein and the state of each node in the network is updated synchronously according to a Boolean function (Figure 1 A).

A Boolean Network (BN) could be described as a directed graph $G(X, E)$ where the nodes, $X = x_1, x_2, \ldots, x_n$ are a set of binary-valued Boolean variables and a vector of Boolean functions $f = (f_1, \ldots, f_n)$. At certain time point, the states of each nodes represent the states of the network, given by the vector $S(t) = (x_1(t), x_2(t), \ldots, x_n(t))$. The Boolean functions are decided by the parents of the nodes. In the next time point, the states of all nodes are updated synchronously according to their respective

Boolean functions:

$$x_i(t+1) = f_i(x_i1(t), x_i2(t), \ldots, x_il(t))$$

2.1.1.1 *Limitations and Extensions.* In order to study the dynamic of cell-cycle regulation in yeast, Li *et al.* [Li et al. 2004] constructed a Boolean network based on pervious literature. The models generated trajectories with a high degree of overlap, most of which led into a path that corresponded to the cell-cycle phases of yeast. In addition, most small changes in the network did not significantly change its dynamic behavior, indicating that it is robust.

Boolean network is limited in several aspects such as its discreteness and synchronism. In reality, the gene expression levels are always continuous. Thus the assumption of the discretization of the original data often reduces its ability to capture some real signals. In addition, the updates of the network states in this model are synchronous, whereas biological networks are typically asynchronous. Finally, despite their simplicity, only small nets can be reverse engineered with the current state-of-the-art algorithms.

2.1.2 *Bayesian Network.* **A Bayesian network** is a representation of a joint probability distribution. This representation consists of two components: a directed acyclic graph(DAG) G, whose vertices correspond to the random variables $X_1, \ldots, X_n$, and θ describing a conditional distribution for each variable, given its parents in G. Together, these two components specify a unique distribution on $X_1, \ldots, X_n$.

The graph G represents conditional independence assumptions that allow the joint distribution to be decomposed, economizing on the number of parameters. The graph G encodes the Markov assumption(*): each variable X_i is independent of its non-descendants, given its parents in G.
By applying the chain rule of probabilities and properties of conditional independencies, any joint distribution that satisfies (*) can be decomposed into the product form

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa^G(X_i)) \tag{1}$$

where $Pa^G(X_i)$ is the set of parents of X_i in G.

The probabilistic nature of Bayesian network allows it to handle noise inherent in both biological processes and follow-up microarray or sequencing experiments. The gene expression profiles could provide a complete joint distribution of gene expression level, while a Bayesian networks expand the joint probability in terms of simpler conditional probabilities. There are several statistical methods for learning can be applied to the network including BNarray[Chen et al. 2006], B-course[Myllymäki et al. 2002], BNT[Murphy et al. 2001] and Werhli's implementation of BN[Werhli et al. 2006].

A major advantage of Bayesian network models is the ability to learn from observed data. Bayesian networks can capture linear, non-linear, combinatorial, stochastic and other types of relationships among variables. They are suitable for modeling gene networks because of their ability to represent stochastic events, to describe locally interacting processes, to handle noisy or missing biological data in a principled statistical way and to possibly make causal inferences from the derived models. Hence, Bayesian networks, including their variants Dynamic Bayesian networks, Gaussian networks, Module networks, mixture Bayesian networks and state-space models (SSMs), etc., have become widely used tools for regulatory-network modeling.

2.1.2.1 *Limitations and Extensions.* Bayesian network have been applied in several experiment data. For example, *Nir Friedman* have used bayesian networks to inferring cellular networks[Friedman

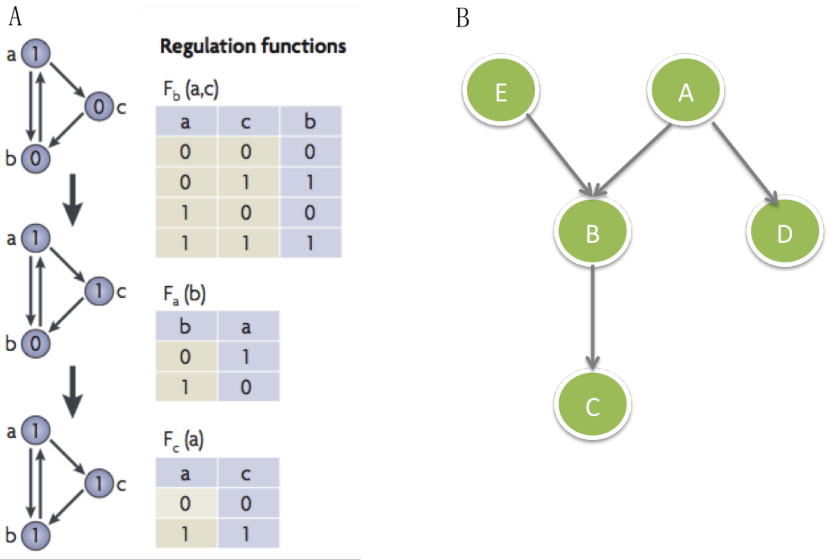

Fig. 1. An example of logical based models. (A) A Boolean network. each nodes in the network a, b and c can be in state 0 or 1. State transition follow the boolean regulation function on the right which describe the rule of the model. (Figure modified from [Karlebach and Shamir 2008]) (B) A simple Bayesian network structure. This network structure implies several conditional independence statements: $I(A; E), I(B; D|A, E), I(C; A, D, E|B), I(D; B, C, E|A), I(E : A, D)$. The network structure also implies that the joint distribution has the product form $P(A, B, C, D, E) = P(A)P(E)P(B|A, E)P(C|B)P(D|A)$

2004]. They first built an unconstrained acyclic network where each gene can have a different regulator set from the experiments of Pe'er et al. [Pe er et al. 2001] Then, by applying bayesian networks and bootstrap estimates of the pre-built network they are able to identify some subnetwork of genes with high confidence relations and biological manfulness among them.

Bayesian network work effective in dealing with noise, incompleteness and stochastic aspects of gene regulation.One limitation of Bayesian networks for modeling genetic networks is that these models must be in the form of directed acyclic graphs and, as such, are not able to represent feedback control mechanisms. Dynamic Bayesian networks, on the other hand, are Bayesian networks that are capable of representing temporal processes[Kim et al. 2003; Zou and Conzen 2005] that may include such feedback loops, but their benefits are hindered by the high computational cost required for learning the conditional dependencies in the cases where large numbers of genes are involved.

2.2 Continuous Models

2.2.1 *Differential Equations.* **Differential equations (DEs)** are used for the quantitative modeling of complex dynamical systems where a system of defined DEs forms a continuous and deterministic model capable of describing non-linear and emerging phenomena [Aluru 2005]. In the application of DEs on gene expression, they are used as rate equations to model the rate of change in the expression of gene x_i as a function of the current levels of expression of other genes (and possibly other quantified factors such as environmental inputs). The general form of each DE for the n genes is:

$$\frac{dx_i}{dt} = f_i(x_{i1}, x_{i2}, ..., x_{ij}) \tag{2}$$

where each x_j is a continuous function representing the expression level of gene j; $f_i(\ldots)$ quantifies the combined effects of the regulators of x_i subsuming all of their biochemical and molecular effects; $x_{i1}, x_{i2}, \ldots, x_{ij}$ is the is a subset of all gene expression functions $x_1, x_2, \ldots, x_n$ that influence the expression of x_i. $f_i(\ldots)$ can be thought of as a function at that takes the input regulator gene levels $x_{i1}, x_{i2}, \ldots, x_{ij}$ and produces an output rate for gene i.

In addition to the variables x_j which represent the gene j expression levels, there will be many free parameters depending on the form of the chosen $f_i(\ldots)$. The parameters can be approximated even if the analytical solutions are unknown using various numerical differential equation solvers. The specific forms of the node functions $f_i(\ldots)$ depend on the considerations of the model. Realistically these functions should be nonlinear to accurately represent natural biomolecular phenomena such as saturation. Sigmoid functions are commonly used such as the *squashing function*: $f_i(x(t)) = 1/(1 + e^{-(\alpha_j x(t)+\beta_j))}$ where the constants α_j and β_j are gene specific parameters that regulate the rapidity of the expression response. Other more complicated nonlinear functions can be used such as Savageau's S-Systems[Kikuchi et al. 2003].

2.2.2 *Linearized Additive Models.* The simplest interesting form of $f_i(\ldots)$ is the linear additive model, also known as the weight matrix model, in which the rate equation becomes:

$$\frac{dx_i(t)}{dt} = ext_i(t) + w_{i1}x_1(t) + ...w_{in}x_n(t) \tag{3}$$

where $ext_i(t)$ indicates possible external control on gene i such as experimental perturbation; w_{ij} are the parameters representing the linear effect of gene n on gene i. Additional possible terms can model other influences such as intrinsic rate of degradation or environmental effects.

Linear additive DE models have several advantages. They intuitively represent the regulatory influence of gene j on gene i, i.e. the magnitude and sign of w_{ij} corresponds to the strength and direction of the regulatory influence. The linearized system is easily represented such that the coefficients represent the independent effects of gene j on gene i: $w_{ij} = \delta f_i/\delta x_j$. When the experimental system is in a steady state or equilibrium where $dx_i/dt = 0$ and is brought to that steady state slowly, the linear approximation holds as accurate. Further, linearization allows for the parameters w_{ij} to be easily fit to the data using linear algebra methods.

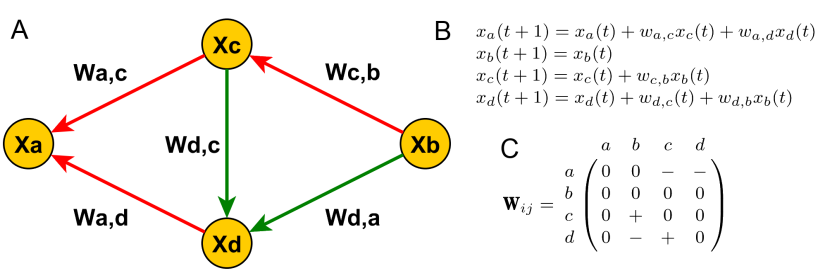

Fig. 2. (A)Graph drawing of a toy gene regulatory network. (B)System of differential equations representing the toy gene regulatory network. (C)Matrix representation of the toy gene regulatory network.

For experiment l, if we let $x^{(l)}$ and $ext^{(l)}$ be column vectors representing the gene expression functions $x_1, \ldots, x_n$ and the external influence on the individual genes $ext_1, \ldots, ext_n$ respectively, and $\mathbf{X}_{n\times m} = (x^{(1)}, \ldots, x^{(m)})$ and $\mathbf{Ext}_{n\times m} = (ext^{(1)}, \ldots, ext^{(m)})$ be the matrices consisting of the above vectors for all m experiments, and $\mathbf{W}_{n\times n}$ be the matrix of all weights $w_{ij}, 1 \leq i, j \leq n$. then the system of equations

is:

$$\frac{d}{dt}\mathbf{X}_{n\times m} = \mathbf{W}_{n\times n}\mathbf{X}_{n\times m} + \mathbf{Ext}_{n\times m} \tag{4}$$

where $\mathbf{X}_{n\times m}$ is the observed gene expression matrix and $\mathbf{Ext}_{n\times m}$ is the observed external influences matrix. The rate terms $\frac{d}{dt}\mathbf{X}_{n\times m}$ are approximated as zero for steady state or from the time-series data as $\Delta x_i/\Delta t$ where Δt should be small (which may not be the case in practice). In this approximation $\Delta x_i/\Delta t = (x_i(t+1) - x_i(t))/\Delta t$ and the next state of gene i's expression can be expressed as a function of its previous regulators $x_i(t+1) = \Delta t(x_i(t) + \sum_{j=1}^{n} w_{ij}x_j(t))$.

2.2.3 *Network reconstruction with linearized additive differential equations.* With linear additive models, the goal in network reconstruction is basically finding the weight matrix $\mathbf{W}_{n\times n}$ that is most consistent with the observed data. There are n^2 unknowns and nm equations in the system of linearized differential equations, thus depending on the m number of experiments and the r number of linearly independent equations ($\leq nm$), the problem can be classified as exact, overconstrained, or underconstrained. When the system is exact ($r = n^2$) there is one solution; when the system is overconstrained ($r > n^2$) then a solution can be achieved through multiple regression; when the system is under constrained ($r < n^2$), there are many equally good solutions.

In practice, experiments generally have many genes n and few experiments m so $m << n$ and thus $r << n^2$. Choosing from among the number of good solutions however is not a straight-forward task. Typically a solution with special properties or a pseudo-inverse is chosen as a starting point for generating better solutions, e.g. Moore-Penrose pseudo-inverse or the Singular Value Decomposition (SVD)[Weaver et al. 1999]. One method of refinement is known as the *robust regression* optimization technique [Yeung et al. 2002].

Assumptions can be used to reduce the search space. For example, the assumption of network-sparseness, i.e. each gene cannot be regulated by more than a fixed number k of other genes, allows the problem to be translated into a combinatorial problem called Minimum Weight Solutions to Linear Equations which can be polynomially solvable albeit computationally expensive[Chen et al. 2001]. Under the assumption that gene levels can be interpolated between time points, it is possible to get an over-constrained system by generating simulated data points which can then be solved using least squares optimization which is reasonable if the time-points are close enough to capture most of the behavior of the system[D'haeseleer et al. 1999]. Another approach is to cluster the genes to reduce the dimensionality of $\mathbf{W}$; through hierarchical clustering of the genes into co-regulated groups, the system can become over-constrained and thus solvable.

LIMITATIONS: Linear DE models are limited to steady-state or slowly-dynamic systems which must be measured with sufficient time resolution to capture the gene regulatory effects which may be technically difficult. Analytical solutions to DE systems are generally unavailable but there are a variety of methods to address these shortcomings.

2.3 Correlation Based Network

One of the most straightforward ways to explore the gene co-expression network is correlation-based methods. In this type of methods, they usually define a gene co-expression similarity matrix $S = [s_{i,j}]$, where $s_{i,j}$ is the pair-wise transcription correlation coefficients between gene i and j, and S is the correlation matrix. Then either a hard or soft threshold is applied to $s_{i,j}$ to determine the biological meaningfulness of the connections.

2.3.1 *Limitations and Extensions.* These co-expression based methods have been used in several studies and have shown their usefulness in interpreting biological results and identifying important gene modules . Taking weighted gene correlation network analysis (WGCNA) as example[Langfelder

and Horvath 2008], it has been applied to identify several disease-related genes. The WGCNA define co-expression networks as undirected, weighted gene networks. In the network, the nodes represent the genes in the genes expression profile and edges correspond to the pairwise correlations between gene expression. Then a threshold $\beta \geq I$ (soft thresholding) is set to the absolute value of the correlation. In WGCNA, the user is able to choose between unsigned network whose adjancy is $a_{ij} = |cor(x_i, x_j)|^\beta$ or the signed network where the adjacency is defined as $a_{ij} = |(I + cor(x_i, x_j))/2|^\beta$ as well.

One drawback for WGCNA is that it is limited to undirected networks[Langfelder and Horvath 2008] and there are some other methods for orienting edges and constructing directed networks[Opgen-Rhein and Strimmer 2007]. Another limitation for correlation-based model, is that in order to get a relative biological meaningful correlation coefficient from a population, it usually requires large amount of data. Thus, it work efficiently when dealing with large scale data for example projects with several hundreds patients gene expression profile data. However, it does not work well when there are limited informations such as dealing with sample only have biological replicates.

3. NETWORK ANALYSIS

Once a network has been constructed it can be mined for useful knowledge. There are many levels from which this task can be approached; one is to apply metrics on the network which allow for the identification of various node or subnetwork properties which include:

—Degree k_i =number of links connected to node i

—Distance d_{ij}=shortest path between node i and j

—Clustering coefficient $c_i = \frac{2e_i}{k_i(k_i-1)}$ where e_i=the number of existing edges among k_i nodes

—Betweenness $b_l = \sum_{ij} \frac{p_{ij}^{(l)}}{p_{ij}}$ where p_{ij} is the number of shortest paths between nodes i and j, and $p_{ij}^{(l)}$ is the number of shortest paths between i and j that run through node l

Such topological metrics be used to identify interesting nodes based on their centrality (relative importance within the graph). For example, interesting points of genetic control may include high-degree nodes which may represent hub genes which have influence over a large number of neighboring genes, or nodes with betweenness centrality which may represent genes that act as a convergence points for intracellular signaling.

Another approach involves the identification of gene modules through performing clustering methods on the network to yield groups co-regulated genes. These modules often correspond to functional genetic units that are involved in a specific process and thus the function of unknown member genes can be inferred by association. The modular analysis is also valuable in that it decreases the stringency of multiple hypothesis correction normally observed in gene expression association studies by dramatically reducing the number of entities being tested. This reduction in complexity can also facilitate in intuitive interpretation of genome-wide data.

The properties of networks can be studied on a global scale. For instance, a major finding in network biology is that many biological networks can be approximated as having scale-free degree distributions, including metabolic and transcriptional networks[Albert 2005]. Scale-free networks have degree distributions that follow an inverse power law of form $P(k) = Ak^{-\gamma}$ where A and k are constants, $P(k)$ is the degree density function, and γ is a network parameter[Albert and Barabási 2002]. They are known to have properties such as robustness against random failure but also imply weakness toward hub failures.

4. CHALLENGE AND FUTURE

4.1 Challenge

In biological network analysis, the main challenge is the integration of biology and mathematics. First, how to interpret the network results biologically is a key issue in network analysis. For example, from a gene expression network analysis, we identify some potential key drivers in the network. We then should consider whether this result is biologically meaningful based on previous knowledge and next, how can we validate those results biological experiments. For biologists, another key issue is how to choose and apply the right models. Using the right math is really important because different models may lead to totally different conclusions.

In February 2014, Lior Pachter who is a statistic professor at Standford university has published some criticisms in his blog about network analysis. The reason why he made those criticisms is that from the perspective of a statistician, he saw some absolute math errors in some papers that are even published on high impact journals. One of the criticism he made is about the paper published by $Baruch\ Barzel$ and $Albert\ László\ Barbaási$ [Barzel and Barabási 2013]. This paper is about the network link prediction by global silencing indirect correlations. In the context of biology, link prediction refers to identify the indirect interaction between genes that maybe confounded by some indirect effects. For example, if gene A inhibits the expression of gene B while gene B inhibits the expression of gene C, then if the expression of gene A increases, the expression of gene B will decrease and the expression level of gene C might increase. Hence, we might observe the correlation between the expression of gene A and gene C even though there is no direct interaction inside. In their model, they assume that a system of genes is in equilibrium. In order to infer the functional and physical interactions in cell, it relies on both "global response matrix",

$$G_{ij} = \frac{dx_i}{dx_j} \tag{5}$$

where x_i and x_j are the expression levels of genes. The global response matrix is able to captures the change in the node i's activity in response to changes in node j's. However, it's not able to capture the indirect effects. Thus, they focus on the "local response matrix",

$$S_{ij} = \frac{\partial x_i}{\partial x_j} \tag{6}$$

which could eliminate the contribution of confounders. To extract S_{ij} from experimentally accessible G_{ij}, they link the 5 and 6 which lead to,

$$G = SG \text{ (off the diagonal) and } G_{ii} = 1 \text{ for all } i. \tag{7}$$

The innovation of this paper is that they provide an approximation formula for solving S from G,

$$S \approx (G - I + D((G - I)G))^{-1} \tag{8}$$

However, one of the errors Lior Pachter mentioned is that just by applying the Sherman-Morrison formula, the could get the exact solution:

$$S = I - D(1/G^{-1})G^{-1} \tag{9}$$

4.2 Future

Our current picture of how regulation is carried out is probably still missing several significant pieces. As the development of next generation sequencing technology, we are able to get large amount of experiment data. By applying network analysis, we could get a more broad understanding of how

the genes are regulated. In addition to understanding regulation as a stand-alone process, models for interplay of gene regulation with other processes need to be created for example metabolism and cell-cell signaling. The benefits of accurate, large-scale regulatory network models for medicine and biotechnology provide a strong incentive for cooperation between experimentalists and computational scientists.

REFERENCES

Reka Albert. 2005. Scale-free networks in cell biology. *Journal of cell science* 118, 21 (2005), 4947–4957.

Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of modern physics* 74, 1 (2002), 47.

Srinivas Aluru. 2005. *Handbook of computational molecular biology*. CRC Press.

Baruch Barzel and Albert-László Barabási. 2013. Network link prediction by global silencing of indirect correlations. *Nature biotechnology* (2013).

Ting Chen, Vladimir Filkov, and Steven S Skiena. 2001. Identifying gene regulatory networks from experimental data. *Parallel computing* 27, 1 (2001), 141–162.

Xiaohui Chen, Ming Chen, and Kaida Ning. 2006. BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics* 22, 23 (2006), 2952–2954.

Patrik D'haeseleer, Xiling Wen, Stefanie Fuhrman, and Roland Somogyi. 1999. Linear modeling of mRNA expression levels during CNS development and injury.. In *Pacific symposium on biocomputing*, Vol. 4. 41–52.

Nir Friedman. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303, 5659 (2004), 799–805.

Leon Glass and Stuart A Kauffman. 1973. The logical analysis of continuous, non-linear biochemical control networks. *Journal of theoretical Biology* 39, 1 (1973), 103–129.

Guy Karlebach and Ron Shamir. 2008. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology* 9, 10 (2008), 770–780.

Shinichi Kikuchi, Daisuke Tominaga, Masanori Arita, Katsutoshi Takahashi, and Masaru Tomita. 2003. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics* 19, 5 (2003), 643–650.

Sun Yong Kim, Seiya Imoto, and Satoru Miyano. 2003. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Briefings in bioinformatics* 4, 3 (2003), 228–235.

Peter Langfelder and Steve Horvath. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* 9, 1 (2008), 559.

Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. 2004. The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America* 101, 14 (2004), 4781–4786.

Kevin Murphy and others. 2001. The bayes net toolbox for matlab. *Computing science and statistics* 33, 2 (2001), 1024–1034.

Petri Myllymäki, Tomi Silander, Henry Tirri, and Pekka Uronen. 2002. B-course: A web-based tool for Bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools* 11, 03 (2002), 369–387.

Rainer Opgen-Rhein and Korbinian Strimmer. 2007. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology* 1, 1 (2007), 37.

Dana Pe er, Aviv Regev, Gal Elidan, and Nir Friedman. 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17, suppl 1 (2001), S215–S224.

Ren Thomas. 1973. Boolean formalization of genetic control circuits. *Journal of theoretical biology* 42, 3 (1973), 563–585.

Daniel C Weaver, Christopher T Workman, Gary D Stormo, and others. 1999. Modeling regulatory networks with weight matrices.. In *Pacific symposium on biocomputing*, Vol. 4. 112–123.

Adriano V Werhli, Marco Grzegorczyk, and Dirk Husmeier. 2006. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics* 22, 20 (2006), 2523–2531.

MK Stephen Yeung, Jesper Tegnér, and James J Collins. 2002. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences* 99, 9 (2002), 6163–6168.

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