Send Orders for Reprints to reprints@benthamscience.ae

Current Genomics, **2015***, 16,* **3-22** **3**

Reverse Engineering of Genome-wide Gene Regulatory Networks from Gene Expression Data

Zhi-Ping Liu*

Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China

Abstract: Transcriptional regulation plays vital roles in many fundamental biological processes. Reverse engineering of genome-wide regulatory networks from high-throughput transcriptomic data provides a promising way to characterize the global scenario of regulatory relationships between regulators and their targets. In this review, we summarize and categorize the main frameworks and methods currently available for inferring transcriptional regulatory networks from microarray gene expression profiling data. We overview each of strategies and introduce representative methods respectively. Their assumptions, advantages, shortcomings, and possible improvements and extensions are also clarified and commented.

Keywords: Gene expression data, Genome-wide inference, Computational model, Transcriptional regulatory network, Reverse engineering.

1. INTRODUCTION

Transcriptional regulation plays crucial roles in protein synthesis and its dynamical responses to internal and external signals, such as development processes and environmental stimuli [1, 2]. The temporal and spatial levels of mRNA and ultimately protein abundance are actually controlled by transcriptional regulations in a cell [3]. A regulation system consisting of genes, RNAs, proteins, and other molecules constructs the complicated regulatory interactions during sequentially transcriptional, post-transcriptional, translational and post-translational processes, which structure into multiplex networks [4]. A transcriptional regulatory network generally refers to regulatory activities between regulators, e.g. transcription factors (TFs), and their targets, e.g. genes [1, 5]. A gene's transcription will be initialized or terminated by the TF proteins binding to its promoter region generally at the 5' upstream of the transcription start site. To some degree, the final expression abundance is mainly determined by the activation or repression of their regulatory relationships [2, 6, 7]. Without distinguishably considering the physical regulations, a gene regulatory network refers to a collection of gene-gene interactions corresponding to such regulatory relationships through their products, and the interactions in gene regulatory network denote this kind of regulations. In contrast, a transcriptional regulatory network represents the physical bindings and direct regulatory interactions between regulators and their targets [8]. It contains more concrete and specific regulatory information between TFs and genes. From a systematic perspective, genome-wide transcriptional regulatory networks in cells control gene expression dynamically and precisely in response to biological context specificities [9].

Identifying transcriptional regulatory networks is of paramount importance from deciphering transcriptional mechanisms to uncovering potential drug targets [10, 11]. Various network reconstruction methods have been proposed and they can be generally categorized as 'bottom-up' and 'top-down' methods. The traditional gene knockout experiments can be categorized as bottom-up methods, which firstly identify the detailed regulations between TFs and targets individually, and then summarize all these regulations to form a regulatory network. The genetic relationships between genes can be detected from the effected genes after knocking out some gene [12-14]. And a global gene regulatory network can be built up after collecting these experimentally identified genetic interactions. Alternatively, top-down methods refer to the emerging systems biology approaches of identifying the global regulatory interactions systematically and in parallel. They firstly acquire many potentially regulatory interactions and then validate each of them by additional experiments. For instance, ChIP-Seq technology makes the genome-wide identification of protein-DNA interactions possible [15, 16]. The regulatory elements of DNA-binding proteins such as TFs are identified from massively parallel sequencing [17]. A genome-wide regulatory network is then drafted from these identifications. The details of TF-target binding event in specific conditions are often checked by further experiments [18]. Microarrays are another type of systematic expression monitoring technologies, which measures the amount of mRNA produced during transcription by hybridization [19, 20]. The reconstruction or inference of regulatory network from microarray gene expression data is often called a reverse engineering process, which backwardly reasons the regulatory system from its observational behavior [21]. Recently, the reverse engineering of

*Address correspondence to this author at the Department of Biomedical Engineering, Shandong University, Jinan, Shandong 250061, China;
Tel: 86-531-8839-2280; Fax: 86-531-8839-2205; E-mail: zpliu@sdu.edu.cn

1875-5488/15 $58.00+.00

transcriptional control network from microarray data becomes very popular to revealing genome-wide regulations [21-25]. Numerous computational strategies have been proposed to reconstruct large-scale gene regulatory relationships from expression profiles [26-29]. Several papers [30-33] have summarized and compared the available strategies from different perspectives. For instance in [30], Emmert-Streib and colleagues presented a systematic overview and comparison study of the network inference methods. They conceptually categorized the existing methods from statistical learning perspective. In this review, we focus on these available computational methods by highlighting their assumptions, advantages, weaknesses, possible improvements and future research directions individually.

Computational methods of inferring transcriptional regulatory networks from expression data are highly motivated by the availability of genome-wide expression profiling data [34-37]. The activities of gene regulation are closely related to gene expression levels [6, 38]. Gene expression profiles of time series or perturbations indicate the dynamics and differences of genes and then imply the causal regulatory possibilities between them. Moreover, the individual gene pairs between regulators and target genes should also be considered with cooperative and systematic perspectives, such as co-regulations, competitive regulations of activators and repressors, and indirect genetic regulations [9, 37, 39]. A global transcriptional regulatory network is embedded with high interacting affinities between regulators and targets, which can be learned from transcriptomic data. And the details of individual regulatory events are hypothesized to be validated by further experiments [13, 40]. The top-down method generates a global view of regulatory relationships in form of network illustrating the context-dependent scenario of regulations. Existing computational methods of inferring regulatory networks are all to formulate the regulations into certain models with these measured expression values [23, 26, 27].

In this review, we firstly formulate the reverse engineering of transcriptional regulatory networks from transcriptomic profiles into a general framework, and then review the major available strategies developed to address this problem, e.g., correlation-based methods, Boolean network methods, Bayesian network methods, differential equation methods, and knowledge-based methods of integrating and evaluating prior regulations. We focus on introducing the assumptions and main ideas behind these strategies and their approximations in the modeling of regulatory systems. Then the current research directions and alternatives of deciphering regulatory network from expression data are discussed. A brief vision of reconstructing transcriptional regulatory networks from high-throughput expression profiling dataset is then concluded.

2. FRAMEWORK OF REVERSE ENGINEERING

The surge of microarray technologies provides unprecedented opportunities to measure genome-wide gene expression simultaneously [19]. Various strategies have been developed to infer the regulatory architectures from their corresponding gene expression profiles for transforming experimental data into regulatory knowledge [22]. The inferred networking linkages represent the regulatory relationships among these measured genes.

(Fig. **1**) illustrates the general framework of the reverse engineering of transcriptional regulatory networks from gene expression data. Essentially, transcriptional regulatory network reconstruction is to identify physical and genetic regulatory relationships between TFs and target genes from their expression profiles. Without distinguishing the difference between TF and its own gene, gene regulatory network is often used as an approximation to the transcriptional regulatory system. Since the abundance of TF protein is often not available, it is approximated by its gene's expression. Specifically, a transcriptional regulatory system is represented by a network, whose nodes refer to regulators and target genes and whose edges indicate their regulatory interactions. As shown in (Fig. **1A**), from microarray gene expression data, such as profiles of time-series physiological processes or perturbation experiments of gene knockout or RNA interface, we reversely engineer the network structures and parameters, e.g., regulatory logic, causality and strength, from the measured gene expressions by developing models and algorithms. The measured genes are those nodes in the regulatory network, and the linkages and related parameters can be identified from the patterns underlying the gene measurements. The regulatory network and expression data are often represented by regulatory matrix $\mathbf{A}$ and expression matrix $\mathbf{X}$, respectively, i.e.,

$$\mathbf{A}=\begin{pmatrix} a_{11} & \cdots & a_{1q} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pq} & \cdots & a_{pn} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nq} & \cdots & a_{nn} \end{pmatrix},$$

$$\mathbf{X}=\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{pmatrix},$$

where entry a_{pq} is the regulatory interactions between the p-th gene and the q-th gene ($1 \le p, q \le n$), and entry x_{ij} represents the gene expression value of the i-th gene ($1 \le i \le n$) at the j-th experiment ($1 \le j \le m$). It is noted that j refers to a sample or a time point with specific phenotype meaning. The process of reverse engineering is to determine the unknown elements of matrix $\mathbf{A}$ from the known $\mathbf{X}$, which is a reverse strategy for reconstructing the underlying regulatory relationships of the system.

As illustrated in (Fig. **1B**), there are four levels of clarity for the elements of $\mathbf{A}$, which answer different questions about the regulatory parameters respectively. Suppose there are two genes, G_1 and G_2. From the available gene expression data $\mathbf{X}$, Level I inference is to determine whether there

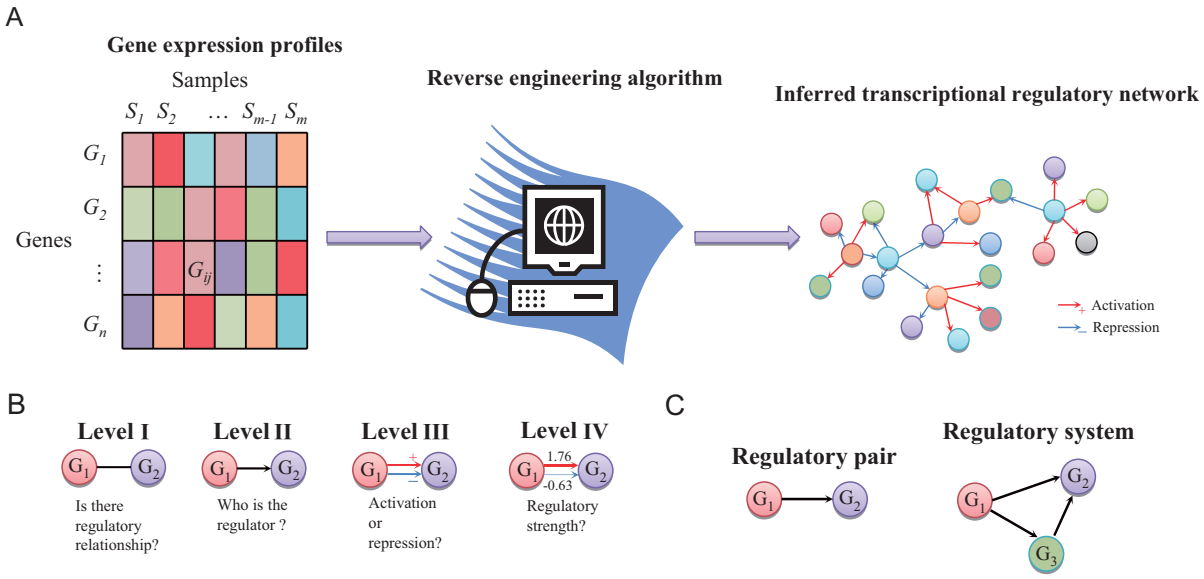


Fig. (1). The general framework of reverse engineering transcriptional regulatory networks. (**A**) The framework of inferring regulatory network from gene expression profiles. There are various sample types of gene expression data, such as condition-specific, perturbation and time series data. A reverse engineering algorithm takes the input of the gene expression profiles and outputs the inferred gene regulatory relationships in form of a network. (**B**) The interrelated four levels of regulatory parameter information should be determined in the reverse engineering. The algorithm addresses the gene regulatory questions at one or several combined levels. (**C**) The regulatory pair and system in the modeling. The decision-making of regulatory relationship of an individual pair is in an isolated manner. However, the regulatory system consists of complicated regulations of combination and cooperation, such as the indirect regulation from gene G_1 to gene G_2 conditioned upon gene G_3, which needs to be modeled in a systematic manner.

is a regulatory connection between G_1 and G_2 from data $\mathbf{X}$. Let a_{12} and a_{21} represent the regulatory interactions from G_1 to G_2 and that from G_2 to G_1, respectively. Level I is to determine whether $a_{12,}a_{21} = or \neq 0$. The binary decision makings build the fundamental architecture of these regulations from gene expression data. Then, when we identify the causal influence from the regulator of TF G_1 to its target gene G_2, Level II inference determines the edge direction and causality in the regulatory network, i.e., $a_{12} \neq 0, a_{21} = 0$. In certain conditions or states, TF might activate or repress the transcription of a target gene, and the concentration of the target is then increased or decreased accordingly. The edge orientation underlying the Level Ⅲ regulatory relationship contains the type information of activation and repression, i.e., $a_{12} > 0, a_{21} = 0$ when G_1 activates G_2, and $a_{12} < 0, a_{21} = 0$ when G_1 represses G_2. More specifically, when we identify the regulation strength from G_1 to G_2 in the Level IV inference, such as $a_{12} = 1.76$ *or* $a_{21} = -0.63$, it provides concrete regulatory weight of its transcriptional dynamics. Level I inference is to reconstruct gene regulatory interactions, while the other inference levels contain more detailed information about transcriptional regulatory interactions, such as regulator and target, activation and repression, and concrete regulatory strength. The strong or weak regulation can then be relatively assessed when all the real numbers of regulatory strengths are determined. (Fig. **1C**) shows the direct modeling of the regulation in an isolated gene pair and in a simple regulatory system respectively. The left graph refers to the regulation between G_1 and G_2, while the right one shows the direct causality from G_1 to G_2 and the indirect influence transferring from G_3. When the system contains a large number of genes, it is apparent that they are needed to be modeled in a systematic manner.

The intrinsic difficulties of transcriptional regulatory network reverse engineering come from several sources. Mathematically, one difficulty is the so-called curse of dimensionality, i.e., $n >> m$ in the formation of expression matrix $\mathbf{X}$. For intensive cost, there are often a few samples (m) of microarray that have been experimented, while thousands of genes (n) have been tested simultaneously in each experiment [41]. From the statistical learning perspective, it is hard to infer a reliable solution of gene regulations from expression data [27]. Moreover, genome-wide regulatory networks tend to be sparse [34-36, 42], all of which result in the high likelihoods to achieve false positive regulations or low likelihoods to achieve false negative regulations [34-36]. Biologically, gene regulation is a complicated physiological process that contains some important steps, such as TF selectively binds to the upstream of the transcription start sites of certain genes to initialize the transcription. Thus, we often model the regulatory system by simplifying some mechanisms, such as cooperation or competition of the TF regulators [43]. Furthermore, the real environment of gene regula-

tion is very dynamic with respect to temporal and spatial features. For example, the up-regulation of one gene encoding a TF can sequentially affect its downstream targets and some regulations can only take place in particular cell types [44, 45].

The reconstructed regulatory network is a graphical representation of transcriptional topology of both *trans*- and *cis*-regulations [46]. The static network structure is usually not efficient to describe the three-dimensional regulatory contexts in cells [47]. Moreover, the epigenetic regulations, such as DNA methylation [48], histone modification and nucleosome positioning [49], strongly influence transcriptional concentrations [50]. miRNAs are also regarded as crucial regulators in the post-transcriptional regulations [51]. The multiplex, hierarchical, heterogeneous regulatory processes are intensely cooperative to generate gene expression levels of mRNA abundance detected by microarray. At the same time, the microarray technique of measuring gene expression is still in its maturing period. The sample preparation, such as cell numbers [52], as well as data preprocessing alternatives including probeset design, background correction and normalization [53, 54], highly affect the quantitatively measured values. Furthermore, the cognate mRNA level is used to represent TF activity in the reverse engineering. The abundance mismatch between mRNA and protein also interfere with the inference of the regulation system [55]. These obstacles challenge the perfect reconstruction of regulatory relationship from expression data.

To address these difficulties of reverse engineering regulatory networks, numerous efforts have been devoted and many substantial regulations have been discovered by *in silico* methods and validated by traditional experiments [30-33, 56]. An international competition named DREAM (Dialogue for Reverse Engineering Assessments and Methods) has been initialized to catalyze the quantitative modeling of transcriptional network inferences [57, 58]. For evaluating the reconstruction performances, several types of measures have often been utilized, e.g., general statistical measures, functional consistency measures and network-based measures [30]. For widely-used statistical measures, the evaluations are often implemented by opening the expression profiling dataset and blinding the benchmarked network structure. After the transcriptional regulatory interactions are inferred from the data by some proposed method, the assessments are performed by comparing the identification results with the benchmarked network [23]. Compared to true regulations, these measures are employed to evaluate the predictions, e.g., sensitivity, specificity, accuracy, F-measure, and Matthews correlation coefficient [30, 59]. The tradeoffs between sensitivity and specificity are often presented by the receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) is often calculated for assessment [60]. Currently, many methods for reverse engineering regulatory networks have been available [22, 27, 30]. Instead of introducing them individually, we categorize them into several main streams of strategies and introduce their main ideas and philosophies.

3. EXISTING METHODS

Due to the difficulties mentioned above, the transcriptional regulatory network inferences are far from accurate and perfect [61], and almost all available methods have their own advantages and drawbacks [27, 61]. We summarize them into the following five categories, namely correlation-based methods, Boolean network methods, Bayesian network methods, differential equation methods, and integrative prior knowledge-based methods.

3.1. Correlation-based Methods

The first endeavor to identify the regulatory relationships in thousands genes measured in microarray is to investigate their pairwise correlations. If gene X highly coexpresses with gene Y, that is to say, when gene X's expression grows up, gene Y's expression grows up or down simultaneously, then the association between the two genes can be detected and modeled by some methods. The regulation can be inferred according to their transcriptional dependence. For multiple genes, clustering is often employed to identify the coexpressed genes [62, 63]. The genes in the same clusters or groups characterize similar expression patterns during physiological processes. They are often assumed to be regulated by the same or related TFs. Two correlation measures are widely used to detect the associated gene pairs, i.e., correlation coefficient [64] and mutual information [65].

The most popular linear correlation between two variables is Pearson's correlation coefficient (PCC). Suppose gene X and gene Y have a series of m measurements X_i and Y_i, where $i = 1, 2, ..., m$, then the PCC r between X and Y is estimated by the sample correlation coefficient, i.e.,

$$r_{XY} = \frac{\sum_{i=1}^{m}(X_i - \bar{X})(Y_i - \bar{Y})}{(m-1)S_X S_Y} = \frac{\sum_{i=1}^{m}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{m}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}},$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y, and S_X and S_Y are the sample standard deviations of X and Y. WGCNA (Weighted Gene Coexpression Network Analysis) is a representative method of building the gene coexpression regulatory network by employing PCC [66]. (Fig. **2**) shows its general framework [67]. Firstly, a clustering method such as hierarchical clustering is implemented to group thousands of genes into some clusters. In each cluster, the highly coexpressed genes are linked by correlation values. For example, when r_{XY} exceeds a defined threshold such as $|r_{XY}| > 0.8$, a functional linkage between X and Y is created in the resulting coexpression network. After the pairwise functional implications between any two genes are identified, a genome-wide network is built up. The simplicity underlies the method that makes it popular to analyze gene expression data, especially to build gene coexpression relationships [68]. Beyond the linear correlation metric of PCC, some rank-based correlations such as Spearman's correlation are also employed to detect the relationship between genes [69]. These correlations replace gene expression values to their relative

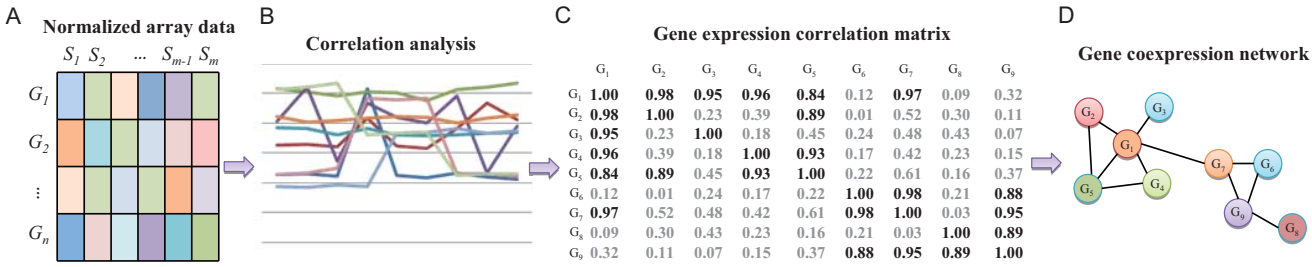


Fig. (2). The framework of building gene coexpression regulatory network [67]. (**A**) The array data. (**B**) The correlation analysis of these genes. (**C**) Pairwise gene correlation matrix. The bold numbers are those over a defined threshold 0.80. (**D**) The built gene coexpression network.

ranks and then calculate the correlation coefficient between the two ranking lists.

Mutual information (MI) is often employed to measure the non-linear gene expression associations between pairs of genes [65, 70]. Generally, MI is an information-theoretic measure of the mutual dependence between two random variables. For two genes X and Y, it is defined as

$$I(X,Y) = -\sum_{X_i \in X, Y_j \in Y} p(X_i,Y_j) \log \frac{p(X_i,Y_j)}{p(X_i)p(Y_j)},$$

where two gene expression values construct two vectors, in which the elements $X_i (i = 1,2,...,m), Y_j (j = 1,2,...,n)$ denote their expression values in different samples respectively. $p(X_i)$ and $p(Y_j)$ are the marginal probabilities of each discrete value X_i in X and Y_j in Y, respectively. $p(X_i,Y_j)$ is the joint probability of X_i and Y_j. High MI value indicates that there may be a close relationship between the two genes, while low MI value implies their independence [60].

MI has been widely used to identify transcriptional regulatory relationships from gene expression data [71]. The quick and accurate estimation of MI is a crucial step in the reverse engineering because computing pairwise MI is non-trivial and quite time-consuming [72]. Similar to the PCC-based framework shown in (Fig. **2**), the available approaches compute the pairwise MI between all gene pairs and construct an association matrix. RN (Relevant Network) chooses the gene pairs when its MI value exceeds a given threshold of significant value [65, 70]. ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) implements the data processing inequality on each connected gene triplet to remove the least significant edge in the MI relevant networks [73]. CLR (Context Likelihood of Relatedness) transforms the MI values into *z*-scores and connects the genes by employing a background sensitive estimator [74]. MRNET (Maximum Relevance Network) is built on the MI-based mRMR (minimum redundancy maximum relevance) feature selection method [75]. MINET presents a software package of MI estimators for inferring large-scale transcriptional regulatory networks [76]. By implementing these MI-based methods, some important transcriptional regulations have been revealed and validated [77, 78].

Unlike PCC and MI, maximum information correlation (MIC) is proposed to detect the strength of any type of linear or nonlinear correlations between genes [79]. MIC adopts binning as a scheme to apply MI to calculate the association between gene variables. It is defined as

$$MIC(X;Y) = \max_{|X||Y|<B} \frac{I(X,Y)}{\log_2(\min(|X|,|Y|))},$$

where $I(X,Y)$ is the MI of X and Y. $|X|,|Y|$ are the numbers of X bins and Y bins divided, and the total number of bins $|X||Y|$ is constrained to be less than some number B. MIC defaults $B = M^{0.6}$ and M is the sample size [79]. Although the effectiveness of MIC is controversial [80], it devotes an effort to identifying diverse types of gene relationships and indicates the importance of an association metric to identify genetic relationships [81].

The correlation or coexpression is a fundamental strategy to identify the regulatory relationships at the former Level Ⅰ and Level Ⅳ inferences (Fig. **1**) and should be improved to be more reasonable in the reverse engineering [82]. Although it is found that the genes in the same grouped clusters tend to have similar functions, these genes might have no direct interactions with each other, and there is no any information to distinguish causal regulators and responsive targets. The built network is not directed (Level Ⅱ) and without the causality of functional linkages (Level Ⅲ) [83], though it can be determined by additional information, such as annotated TFs [77]. Moreover, the clustering methods such as hierarchical clustering are highly dependent on the threshold chosen to cut the hierarchical tree (dendrogram). The number of clusters and chosen distance metrics also highly affect the resulting networks [81]. It is often assumed that there is modularity property in coexpression regulatory networks, which means dense connections between the genes within the same modules but sparse connections between genes in different modules [68]. The clusters form the building blocks of genome-wide regulatory networks. The linkages between these modules are often omitted in these available methods [84]. These functional linkages indicate the crosstalk and functional cooperation between these modules upon certain conditions [67, 85-87].

Another important issue of this type of methods is the isolated modeling of individual gene pairs as shown in (Fig.

1C). The regulatory effect from G_1 to G_2 can also be transferred from G_3. The indirect regulations highly bias the inferred results [88, 89]. We should consider the degree of association with the removal of the effects from indirect regulations by controlling one or several other genes. Partial correlation coefficient can be employed to quantify the association between two genes when conditioning on other gene or genes [88]. For instance, conditioning on a gene or gene set Z, partial correlation $r_{XY \cdot Z}$ between gene X and gene Y is to measure the exact correlation between the parts of X and Y that have no relationship with Z. The order of partial correlation coefficient is determined by the number of conditioned genes. Obviously, the mentioned PCC is the zeroth-order partial correlation coefficient. Theoretically, it can be raised to any arbitrary order. The first-order and second-order partial correlation is defined as $r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$ and $r_{XY \cdot ZQ} = \frac{r_{XY \cdot Z} - r_{XQ \cdot Z} r_{YQ \cdot Z}}{\sqrt{(1 - r_{XQ \cdot Z}^2)(1 - r_{YQ \cdot Z}^2)}}$, respectively. In practice, it is difficult to calculate high-order partial correlation coefficient because of the curse of dimensionality. It is often estimated by developing some specific computation techniques in the reverse engineering of regulatory networks [89].

Similarly, conditional mutual information (CMI) measures the conditional dependency between two genes given other gene or gene set. The CMI of genes X and Y given Z is defined as

$$I(X_i, Y_j \mid Z_k) = \sum_{X_i \in X,\ Y_j \in Y,\ Z_k \in Z} p(X_i, Y_j, Z_k) \log \frac{p(X_i, Y_j \mid Z_k)}{p(X_i \mid Z_k) p(Y_j \mid Z_k)}$$

CMI has been applied to reconstruct genome-wide regulatory networks [90-92]. The recently proposed MIC is also expected to be extended to calculate the conditional and partial versions for detecting more delicate and meaningful associations between genes [93].

Based on CMI, we proposed a reverse engineering method [60] by utilizing path consistency algorithm [94] to remove the edges with conditional independent correlation from the network. (Fig. **3**) shows the general framework of our PCA-CMI method. The main idea of PCA-CMI is to eliminate the edges with independent correlations recursively, i.e., from low to high order independent correlation until there is no edge that can be removed. Firstly, we began with a complete graph, in which all the possible regulations among these genes are contained. Secondly, for adjacent gene pair i and j, we calculated MI $I(i, j)$, i.e., zeroth-order CMI. We removed the edges between genes i and j if they have low or zero MI values. Thirdly, for adjacent gene pair i and j, we computed the first-order CMI $I(i, j \mid k)$ conditioned on their adjacent gene k. We removed the edge between them if they have low or zero CMI. The next step is to identify higher order CMI until there are no more adjacent edges to be eliminated [60]. Since it is also time-consuming to calculate CMI [60, 90], in our proposed algorithm, with the assumption of Gaussian distribution, CMI is estimated with Gaussian kernel probability density estimator [56].

From a regulatory system perspective, linear regression methods identify the associations among genes comprehensively [95, 96]. Compared to the former correlation or partial correlation based methods, the regression methods model each gene by multiple predictors. They associate the expression of one gene to all the genes in the whole system and then identify these predictors by variable selection. So the cooperative regulatory relationships among genes can be identified simultaneously. Let Y denote a gene and $R = (X_1, X_2, ..., X_r)$ be the gene set potentially regulate gene Y. Their relationship is modeled by a linear function, i.e., $Y = \beta_0 + \sum_{j=1}^{r} \beta_j X_j$. The ordinary least squares, partial least squares and maximum likelihood methods can then be used to estimate the parameters of the linear system [97, 98]. Under the parsimony assumption, a regulatory network tends to be sparse [34, 36, 42]. Some variable selection method such as LASSO [99] and elastic net [100] are often employed to recognize the crucial regulators by the regularization techniques [101]. Specifically, LASSO minimizes the residual sum of squares subject to a bound on the L_1-norm of the coefficients, i.e.,

$$\hat{\beta}_{lasso} = \arg\min_{\beta \in R^r} \left((Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\|_1 \right),$$

where $\lambda > 0$ and $\|\beta\|_1 = \sum_{j=1}^{r} |\beta_j|$. Obviously, some coefficients may be shrunken to zero and the global linkages (coefficients) between these genes can be then inferred. We can find that the causal relationships or directions between these genes are embedded in the regression model. Regression combined with variable selection formulates the regulations into a systems biology approach to reconstructing the underlying genetic interactions from expression profiles. Apparently, regression-based methods achieve a sparse regulatory network and perform the four levels of regulation inferences shown in (Fig. **1B**). For time course expression data, the vector autoregressive model is also employed to specify the gene expression value by a linear regression of those of earlier time points [97]. Similarly, Granger causality is modeled to learn time-lagged regulatory networks from time-course gene expression data [102, 103].

3.2. Boolean Network Methods

One of the main-stream strategies to reverse engineering transcriptional regulatory networks is based on Boolean networks. Boolean models treat the genes in a regulation system as logical elements [104]. It assumes that a single gene can be represented by a Boolean variable denoting whether it is

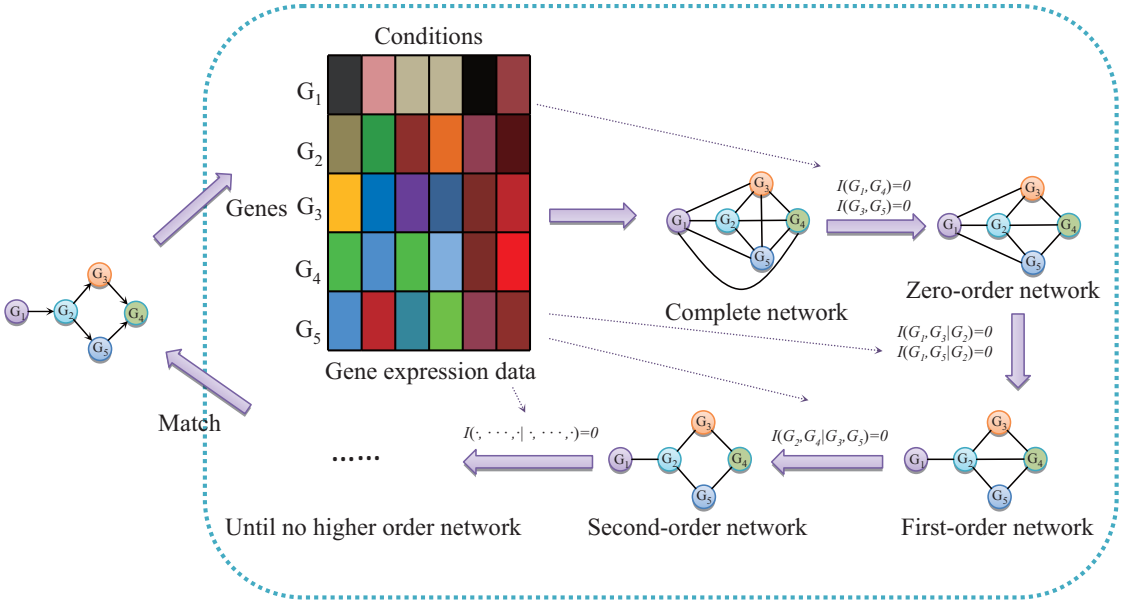


Fig. (3). The reverse engineering diagram of PCA-CMI (path consistency algorithm based on conditional mutual information) [60].

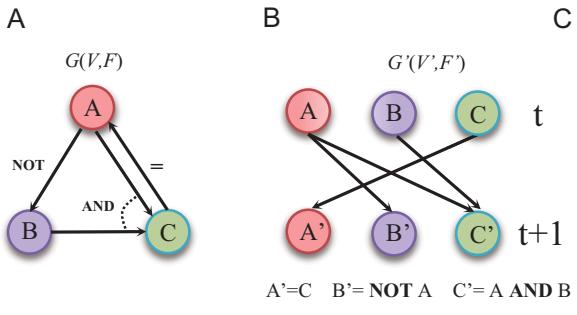
expressed or not. The wiring of an element to one another corresponds to functional linkages between genes, and the Boolean rules determine the result of a regulatory signaling transduction given a set of input values [105, 106]. Boolean network provides a simple decision-making model of describing the regulatory mechanisms in a transcriptional system [104, 107, 108].

Specifically, a Boolean network is a directed graph $G(V,F)$, where the set $V=\{X_1,X_2,...,X_n\}$ of nodes representing genes. (Fig. **4A**) shows a simple example. For each node $X_i \in V$, $i \in \{1,2,...,n\}$, a Boolean function $f_i = f_i(X_{i_1},...,X_{i_k}) \in F = \{f_1, f_2,..., f_n\}$ is associated with it individually. The inputs of f_i are from the specified parent nodes $X_{i_1},...,X_{i_k}$ in V to each node X_i. The variable X_i is Boolean and its value is often denoted as 0 or 1 which corresponds to the logical value True or False respectively. The logic operators 'AND', 'OR', and 'NOT' are employed to define the Boolean operations in these genes [107]. At any given time t, an expression pattern of V names a state of a Boolean network, i.e., $S(t)=(X_1(t),X_2(t),...,X_n(t))$. The state at time point $t+1$ is determined by Boolean functions F from the state $S(t)$, i.e., $S_{t+1}(X_i)=f_i(S_t(X_{i_1}),...,S_t(X_{i_k}))$. The states of all nodes are updated according to their respective Boolean functions and all states' transitions together correspond to a state transition of the regulatory network.

For representing the state transition, it is convenient to build a corresponding wiring diagram $G'(V',F')$ of a Boolean network G as shown in (Fig. **4B**) [106, 109]. For each node $V_i \in V$, let $V_{i_1},...,V_{i_k}$ be the parent nodes of V_i in $G(V,F)$. By introducing an additional node V_i', we link an edge from V_{i_j} $(1 \le j \le k)$ to V_i'. Then $V'=\{V_1,...,V_n,V_1',...,V_n'\}$ in the resulting network. Apparently, the expression pattern of the additional node set $\{V_1',...,V_n'\}$ is determined by $V_i' = f_i(V_{i_1},...,V_{i_k})$ individually and corresponds to the regulatory network state at the next time point. If we regard the expression patterns of the set $\{V_1,...,V_n\}$ as the input of F, the expression patterns of $\{V_1',...,V_n'\}$ are the output as shown in (Fig. **4C**).

The reverse engineering of a Boolean network is to infer the Boolean functions F at these nodes from expression data. When F is known, the underlying network topology of regulations can be built spontaneously. An exhaustive search is to try out all Boolean functions on all $\binom{n}{k}$ combinations of k out of n genes. It is known to be an NP-complete problem and takes exponential time in the inference [105, 106]. So it is often tractable by employing certain computational techniques to avoid exponentially searching a consistent network structure with the observational data. When multiple network structures are found to be consistent with the gene expression data, more scoring metrics and assumptions can be defined to select one suitable regulatory architecture [26, 110].

Boolean network is a fundamental model of genetic system which identifies the network structure from a systematic perspective. It fulfills the Levels I, II and III inferences of gene regulatory networks. The dynamic property and the simplicity in understanding and analyzing make it an attractive model of regulatory network reverse engineering. However, the binary and synchronous (i.e., the state of all genes updates to the next one at the same time) assumptions are not consistent with the true biological system [111]. To address these limitations, the discretization strategies and Boolean models have been extended in various ways to make them more biologically realistic and computationally tractable



t			t+1		
INPUT			OUTPUT		
A	B	C	A'	B'	C'
1	1	1	1	0	1
1	1	0	0	0	1
1	0	1	1	0	0
1	0	0	0	0	0
0	1	1	1	1	0
0	1	0	0	1	0
0	0	1	1	1	0
0	0	0	0	1	0

Fig. (4). An example of Boolean network. (**A**) A Boolean network $G(V,F)$. (**B**) The corresponding wiring graph of $G(V,F)$. (**C**) The logic operations and state transition table. The possible input at time point t and the corresponding output at time $t+1$ are listed in the table. Boolean network models the regulatory relationships in the logical operating scheme [106].

[26]. With the availability of gene expression data with larger sample size and higher quality, there have been approaches to introducing stochasticity to these models, such as probabilistic Boolean networks [112-114] in which the state transition diagram is stochastic. The generalized Boolean network models also try to cope with the shortcomings by enabling more sophisticated forms of logical update which allows asynchronous transition of elements [115].

3.3. Bayesian Network Methods

Definition. *Bayesian network is a directed acyclic graph (DAG) representing a set of random variables and their joint probability distribution together with the family of conditional probabilities induced by the graph* [116, 117].

Bayesian network is a typical probabilistic graphical model of causal inference in statistics. The general idea of learning Bayesian network structure from data is to evaluate each network structure with respect to the given data by defining a scoring function and to identify the optimal one according to the score [118]. The structure represents the conditional independence of these variables that facilitate their joint distribution to be decomposed. The graph G is often assumed to follow the Markov property that each gene X_i is independent of its non-descendents, given its parents in G. By applying the chain rules of probability and the properties of conditional independency, the joint distribution on genes $X_1, X_2, ..., X_n$ can be uniquely represented by the product form $P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid parent\{X_i\})$,

where $parent\{X_i\}$ is the set of parents of X_i in G. In this way, each Bayesian network specifies the joint probability distribution over all genes down to the conditional distributions of the genes X_i given their parents. As shown in (Fig. **5A**), gene *D* is dependent on gene *A* and gene *E*, and independent on the other gene or genes. The global network probability is determined by the dependence structure between multiple interacting components.

The graphical representation consists of two distinct parts in reverse engineering transcriptional regulatory networks. The first component $G(V,E)$ is a DAG representing the causal relationships of regulations (i.e., edges of set E) among a set of genes (i.e., nodes of set V). An edge exists from gene *A* to gene *B* if and only if *A* is a direct regulator of *B*. The second component is a set of parameter θ, which describes a conditional probability distribution of each gene, given its parent regulators. Taken together, the two components specify a probability distribution over the set of genes in V, i.e., the network structure of regulations. Often, Bayesian scoring metric is derived to evaluate the posterior probability of a graph G given the gene expression data $\mathbf{D}$, i.e.,

$$\begin{aligned} S(G:\mathbf{D}) &= \log P(G \mid \mathbf{D}) \\ &= \log \frac{P(\mathbf{D} \mid G)P(G)}{P(\mathbf{D})} \\ &= \log P(\mathbf{D} \mid G) + \log P(G) + C, \end{aligned}$$

where C is a constant which can be ignored [119]. In a Bayesian network framework, the calculation of the log marginal likelihood $\log P(\mathbf{D} \mid G)$ involves the probability of the data over all possible parameters θ assigned to G. It is an NP-hard problem to select the maximum scored network structure given the data [117, 118]. Thus, the most probable network structure is generally implemented by approximating the posterior probabilities of the regulatory combinations heuristically [37, 118]. Bayesian network model becomes appealing for modeling causal relationships between these genes by selecting the most likely causalities in form of a DAG [9, 29, 119]. Some techniques have been developed to narrow down the search space to a tractable size. As an assumption, the basic form of Bayesian network cannot handle cyclic regulations and the temporal dynamic regulatory relationships [117]. Other alternatives have been proposed to extend the applicability of Bayesian network modeling, such as dynamic Bayesian network [120-124], module network [84] and state-space model [121, 125].

Based on the framework of Bayesian network, dynamic Bayesian network (DBN) introduces the time concept and models a stochastic temporal process of a set of random variables over time series [121-123]. It has been employed to describe the qualitative nature of the dependencies that exist between genes in a temporal process. The structure of a

DBN is assumed to perform regulatory functions over discrete time points indexed by $t \in \{1,.....,T\}$. Similar to the assumptions in Bayesian network, let $X^t = (X_1^t,...,X_n^t)^T$ be the gene expression vector of n genes at time t. For the time points $\{1,...,t,...,T\}$, under the first-order Markovian assumption, i.e., X^{t+1} is independent of $X^{t'}$ for $t' < t$ given X^t, we thus have

$$P(X^1,...,X^t,...,X^T) = P(X^1)\prod_{t=1}^{T}\prod_{i=1}^{n} P(X_i^t \mid parent\{X_i^t\})$$

in the time-course gene expression data [123]. As illustrated in (Fig. **5B**), the underlying acyclic graph in Bayesian network can now be permitted to contain cycles. DBN model can explore the general network structure of gene regulations and overcome the shortcomings of the acyclic assumption and static network structure in Bayesian network learning models. A more complicated time-varying DBN model of describing the time-evolving network structures underlying the time series is also developed [126].

3.4. Differential Equation Methods

Differential equation formalisms including ordinary and partial differential equations have been widely used to describe and simulate dynamical systems in science and engineering. The powerful mathematical methods have been implemented to model the biochemical systems of metabolic processes and kinetic dynamics of genetic regulation processes [25, 26]. The regulatory interactions in form of network are revealed by the differential and functional relations between the time-dependent concentration variables [36, 127]. Here, we mainly introduce the ordinary differential equation (ODE) models in modeling transcriptional regulatory network. Partial differential equation (PDE) models contain the similar framework as ODE with more dynamic dimensions beyond the time in ODE [26]. ODE models directly consider the time differentiation and then the dynamics and causal relationships can be simultaneously identified in the four inference levels (Fig. **1**) of reverse engineering regulatory network.

In ODE models, the change rate of gene expression of a component in a regulatory system is modeled as a function of the concentrations of all the components. Mathematically, the general ODE model can be formulated as

$$\frac{dX}{dt} = F(t, X, \theta),$$

where $X = X(t) = (X_1(t),..., X_n(t))^T$ represents the gene expression values of genes $1,...,n$ at time point $t, t \in [t_0, T], 0 \le t_0 \le T < \infty$. The causal effects of gene expressions are embedded in the ODE system [128, 129]. Function F describes the relationship between the first order derivative of X and the concentration of genes in the regulatory system. It is a linear or nonlinear function that describes the relationships between the change rate concentration of genes and their causal regulators. Specifically, a linear ODE model can be written as

$$\frac{dX_i}{dt} = \beta_{i0} + \sum_{j=1}^{n} \beta_{ij} X_j(t),\ j = 1,...,n,$$

where β_{i0} is the intercept and $\beta = \{\beta_{ij}\}_{i,j=1,...,n}$ denotes the regulation effects of genes in the regulatory system on the rate of expression change of the i-th gene.

The problem of network reconstruction from data is then transformed to identify the parameters in the ODE system. Traditionally, the least squares method and likelihood-based methods are implemented to find these parameters [27, 34]. Various techniques have also been employed to evaluate them [41, 130]. However, these methods are not effective for reverse engineering genome-wide regulatory networks. We and Lu *et al.* [128, 129] proposed an integrative pipeline to address the problem by introducing a two-step paradigm to identify these parameters effectively. The first step is to fit the mean curves of the gene expressions and then to estimate the derivative value $\frac{dX_i}{dt}$ respectively, i.e., $y_{ik} = \hat{M}_i'(t_k^*)$, $z_{jk} = \hat{M}_j(t_k^*), i, j = 1,...,n; k = 1,...,m$, where $\hat{M}_i'$ is estimated continuously from the mean curve, t_k^* is one of the m set time points in range $[t_0, T]$. Thus, the regulatory system becomes the following pseudo-regression model, i.e.,

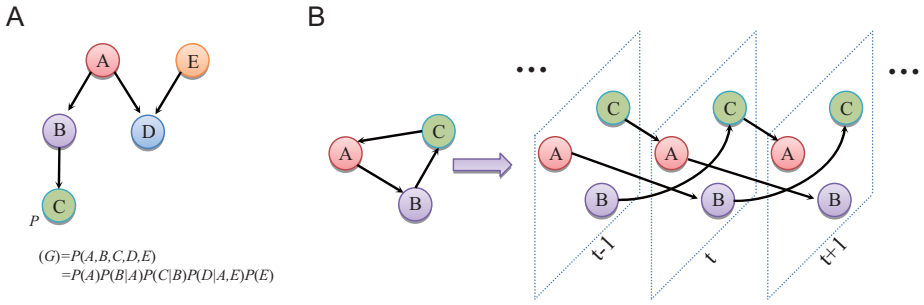


Fig. (5). The graphical representation of Bayesian network and dynamic Bayesian network. (**A**) An example of a Bayesian network. By recursive decomposition, the joint probability distribution of the network is $P(G) = P(A,B,C,D,E) = P(A)P(B \mid A)P(C \mid B)P(D \mid A,E)P(E)$. The conditional independence simplifies the conditional probability distributions of these nodes in the decomposition. (**B**) The graphical representation of a dynamic Bayesian network (DBN). The static and dynamic representations are shown respectively. Assuming the temporal regulations are from time t to $t+1$, cyclic structures are apparently permitted in the DBN framework.

$$y_{ik} = \beta_{i0} + \sum_{j=1}^{n} \beta_{jk} z_{jk} + \varepsilon_{ik},$$

where ε_{ik} is the error term of estimation. Based on the parsimony assumption, the second step is to conduct the variable selection and estimation procedure by a regularization framework, such as LASSO [99] and SCAD [131], to shrink the variables as optimally as possible. The regulatory network is then reconstructed from the data when we identify the parameters of the formulated linear regression system. Original methods [128, 129] include a clustering procedure to divide these genes into groups with similar expression profiles, which helps to build a genome-wide network and simultaneously avoid the identifiability problem [132].

ODE is a directed network model and the dynamic feature of regulations is automatically and naturally quantified. In ODE models, gene regulations are modeled by derivative equations, which quantify the change rate of gene expression of one gene (dependent variable) in the system as a function of expressions of all related genes (independent variables) that refer to its regulators. In a transcriptional regulatory system, it is TFs that regulate the gene transcriptional processes. The abundance of TF proteins is the real independent variables. We usually have no such information and simply use the TF genes' expression as approximation. Under such assumption, the reverse engineering of regulatory network becomes inferring the parameters of some specified functions such as the former linear function from gene expression data [128]. According to the differences between a mathematical modeling perspective and a statistical perspective lying in the network inference [133], ODE is to model the regulatory system but not to directly infer the regulatory network. The derivation equations are firstly assumed to describe the functional relationships among genes and their products. Then, the statistical techniques such as parameter estimation and variable selection are implemented to infer the regulatory architectures [128]. The resulting nonzero regulatory linkages construct a regulatory network. Time delay of the activation and self-degradation can also be flexibly integrated in the dynamical system by introducing certain terms in the differential equations, such as $t-\tau$, where τ denotes a time delay and $-\beta_i X_i(t)$ for the i-th gene's self-regulation [26]. Compared to the former regression methods of modeling the mRNA concentrations of individual components in the system, ODE describes the derivatives of their concentrations. The strategies of parameter estimation are similar to each other.

3.5. Knowledge-based Methods

With the essential difficulties in the reverse engineering of regulatory networks, purely data-driven method is very difficult to identify genuine transcriptional regulations. It is hard to promise the effectiveness and efficiency of the reverse engineering only from gene expression profiles [22, 27, 134]. There are urgent requirements to develop novel methods that can utilize expression data in some alternative manners. At the same time, various prior knowledge of gene regulations from literature and genomic datasets can provide additional functional linkage information between genes, such as documented regulations [135, 136], TF binding sequence motifs in promoter region [45], ChIP-Seq data of protein-DNA binding [137] and protein-protein interactions [59]. These prior knowledge can be integrated together with gene expression data to identify transcriptional regulatory networks. Theoretically, the resolution space can be narrowed down to improving the identification significantly [138-140]. So it guides the inference in right direction and helps remove false positives in the predictions [141, 142]. Knowledge-based methods fall into two subcategories, the combination of prior knowledge and the evaluation of prior knowledge. We review them individually as follows.

3.5.1. Combining Prior Knowledge

The combination of prior knowledge is often implemented on the former reviewed reverse engineering methods. Bayesian network is one of the rational models to integrate prior knowledge in a principled manner to increase the inference reliability [140, 142]. According to the Markov assumption, the probability of a network structure can be decomposed as

$$P(G) = \prod_{i=1}^{n} P\left(X_i \mid parent\{X_i\}\right),$$

where $parent\{X_i\}$ is the parents of X_i in the DAG. The probability of a local regulatory structure $P\left(X_i \mid parent\{X_i\}\right)$ is then calculated according to the structural knowledge priors,

$$P\left(X_i \mid parent\{X_i\}\right) = \prod_{Y \in parent(X_i)} P(Y \rightarrow X_i) \prod_{Y \notin parent(X_i)} P(Y \otimes X_i).$$

The decomposition facilitates to incorporate the prior knowledge about regulatory structure into the network inference. Various techniques have been proposed to calculate these probabilities, i.e., $P(Y \rightarrow X_i)$ and $P(Y \otimes X_i)$, as accurately and effectively as possible. Following a framework of statistical physics, [139] and [143] proposed an energy function to introduce the prior knowledge from multiple sources into the reverse engineering of regulatory network. Their main idea is to express the available prior knowledge in terms of network energy. Specifically, the prior knowledge about the regulatory relationship between gene i and gene j is represented by p_{ij}, $p_{ij} \in [0,1]$. Network energy of a network is then defined on the biological prior knowledge matrix. Then, a prior distribution over network structures is obtained by means of a Gibbs distribution [139]. The parameter of this distribution represents the weight associated with the prior knowledge relative to the gene expression profiles. In this way, the prior knowledge is integrated into a Bayesian network framework to learn the regulatory network structure. They achieved higher performance of inference in both simulated and real data [139, 143].

Based on an ODE model, we proposed a method of linear programming (LP) to integrate prior knowledge in the

reverse engineering of regulatory network [138]. The main idea is to build an LP model to minimize the association gap between gene expression data and network structure with constraints of the priori of regulatory relationships, and then to solve the LP to obtain the integrated regulatory network.

Specifically, given an experiment with n genes and m samples, the gene expression matrix is $\mathbf{X} = (x_{ij})_{n\times m}$, where x_{ij} is the expression level of the i-th gene in the j-th sample. We employed an ODE model to quantify the rate of change of gene expression as a function of the expression of other genes [138]. Due to the unclear structures of regulatory system and data scarcity [41, 95, 138], we used the simplest linear additive models:

$$\dot{X_i}(t) = -\lambda_i X_i(t) + \sum_{j=1}^{n} a_{ij} X_j(t) + b_i(t) + \varepsilon_i(t),$$

for $i = 1, 2, ..., n$, where the state variable $X_i(t)$ is the mRNA concentrations of gene i at time point t, λ_i is the self-degradation coefficient, b_i is the external stimuli, which is set to 0 when there is no external input, and ε_i represents the error and noise. a_{ij} describes the type and strength of the effect of gene j on gene i, whose positive, zero or negative values indicate the activation, naught or repression regulatory relationships between them respectively. For simplicity, we set $b_i = 0$. Hence, the equations can be described as:

$$\dot{X}(t) = \mathbf{A}X(t) + \boldsymbol{\varepsilon},$$

where $X(t) \in R^n, t = 1, ..., m.$ After we approximated $\frac{dX_i}{dt}$ by $\frac{\Delta X_i}{\Delta t} = \frac{X_i(t+1) - X_i(t)}{\Delta t}$ and neglected the error part, the linear additive model becomes

$$\Delta X(t) = \mathbf{A}X(t),$$

where $t = 1, ..., m-1$. Instead of solving the equations by singular value decomposition (SVD) technique [41, 95, 127, 138], we derived a sparse regulation network [36, 56] based on an LP model. At the same time, more and more prior knowledge of gene regulatory network can be obtained from various sources. For example, if we know that gene i and gene j are interactive with the rule that i activates j, such priori should be guaranteed in the inference procedure and the inferred network should contain such information as "i activates j".

In our LP model [138], the objective function is to minimize the number of gene connections to realize the sparseness of the inferring network, and the constraints are the linear additive equations and the prior knowledge of some local network structures. The model is described as

$$Min\ F(\mathbf{A}_{n\times n})$$
$$s.t.\quad \mathbf{A}_{n\times n} X_{n\times m} = \Delta X_{n\times m.}$$

There are $n \times n$ variables $\mathbf{A}_{n\times n}$ and $n \times m$ constraints. It is equivalent to solve a canonical LP:

$$Min\ F(\mathbf{A}^1, \mathbf{A}^2)$$
$$s.t.\quad \mathbf{A}^1 X - \mathbf{A}^2 X = \Delta X$$
$$\mathbf{A}^1 \geq 0, \mathbf{A}^2 \geq 0\ (\mathbf{A}^1 - \mathbf{A}^2 = \mathbf{A}).$$

Clearly, there are $2n \times n$ variables and $n \times m$ constraints. In the canonical form, the linear objective function can be defined as:

$$F(\mathbf{A}) = F(\mathbf{A}^1, \mathbf{A}^2) = C^1 \mathrm{o}\mathbf{A}^1 + C^2 \mathrm{o}\mathbf{A}^2.$$

The sparseness and the prior knowledge for regulatory network are represented in the objective function and in the constraints of the LP model, respectively. When we let $C^1 = \mathbf{1}$ and $C^2 = \mathbf{1}$, the objective function becomes

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^1 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2.$$

Hence, a sparse regulatory network is achieved from gene expression data by minimizing these regulatory strength coefficients with the constraints of the prior knowledge about the gene relationships. Generally, there are three kinds of the prior knowledge about the functional relationship between gene i and gene j; i activates/represses j ($i \rightarrow_+ j / i \rightarrow_- j$), i has no any relationship with j ($i \otimes j$), and i has some relationship with j, but unclear of positive or negative regulation ($i \leftrightarrow_? j$). These prior knowledge are reflected in the constraints by the defined rules. If gene i is an activator of gene j ($i \rightarrow_+ j$), we set $a_{ij}^1 - a_{ij}^2 > 0$ as a constraint in our LP model. Conversely, if gene i represses gene j ($i \rightarrow_- j$), we set $a_{ij}^2 - a_{ij}^1 > 0$. If gene i has no any relationship with gene j ($i \otimes j$), we set $a_{ij}^1 - a_{ij}^2 = 0$ as a constraint. If it is unclear which one is an activator or repressor ($i \leftrightarrow_? j$), we set the constraint as $a_{ij}^1 + a_{ij}^2 > 0$ and $a_{ij}^1 + a_{ij}^2 < 0$. By solving the two LP models with the two constraints respectively, we selected the sparser solution as the inferred network [138].

3.5.2. Evaluating Prior Regulations

Due to the complexity of gene regulation and the difficulty of network inference from expression profiles, reverse engineering cannot easily identify genuine regulatory relationships [27, 134]. An amount of knowledge about gene regulations has been deciphered by decades of endeavors [41, 144]. Alternatively, we can evaluate the knowledge-based gene regulations documented in literature and databases and filter out the activated regulations in certain biological conditions and phenotypes. The screening evaluation

procedure provides direct evidence for highlighting the condition-specific regulatory network in biological system [91, 134, 144]. Based on the available or predefined regulatory networks, the consistency between architecture and expression are measured, and the most rational network structure with the expression data can be revealed [145, 146]. In the evaluation strategy, each of the reference networks is assessed by measuring the correspondence between network structures and gene expression profiles. The comparison of matching significance in these knowledge-based regulatory networks can identify the responsive regulatory networks of certain conditions and phenotypes.

Network structure determines the regulatory functionality and robustness [147, 148]. The new 'forward-like' engineering of matching network structure with gene expression data provides more alternatives to investigate the regulatory relationships. The original paper in this direction was published in [144]. The authors proposed a Gaussian graphical model to represent the causal relationships of regulatory network architecture and defined a graph consistency probability to measure the goodness of fitting between network and data. However the directed acyclic graph assumption limits its generality and applicability. Collaborating with the senior author of the original work, we introduced a DBN model to handle general regulatory networks [134]. Specifically, by recursive factorization, the joint probability distribution of a certain directed network architecture is represented as a product of the individual density functions conditioned on their parent variables [134, 144], i.e.

$P(G) = P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid parent\{X_i\})$ in graph G. Let $X^t = (X_1^t, ..., X_n^t)^T$ be the gene expression of n genes at time point t. Thus, for $t \in \{1, ..., T\}$, under the first-order Markovian assumption that X^{t+1} is independent of $X^{t'}$ for $t' < t$ given X^t, we have $P(X^1, ..., X^t, ..., X^T) = P(X^1)\prod_{t=1}^{T}\prod_{i=1}^{n} P(X_i^t \mid parent\{X_i^t\})$ in the time course data. Assume

$$X^{t+1} = \mathbf{A}X^t + \boldsymbol{\varepsilon},$$

where

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix},$$

where a_{ij} is the regulatory coefficient of $X_j^t \rightarrow X_i^{t+1}$. $\boldsymbol{\varepsilon}$ is the error vector and $\boldsymbol{\varepsilon} \sim N(0, \sum)$ with $\Sigma = diag(\sigma_1^2, ..., \sigma_n^2)$. According to linear assumption [125, 134], the log-likelihood function

$$\ln f(X^1, ..., X^t, ..., X^T) \propto -\frac{1}{2}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left[\left(X_i^{t+1} - \sum_{j=1}^{n} a_{ij}X_j^t\right)^2 + \ln(2\pi\sigma_i^2)\right].$$

Although the binary regulatory relationship between gene i and gene j is available in the priori, the details of activation ($i \rightarrow_+ j$), repression ($i \rightarrow_- j$), no regulation ($i \otimes j$), as well as the Level Ⅳ of regulatory strength are unknown, especially in specific conditions. So we employed a quadratic programming (QP) to calculate the likelihood value by optimizing the coefficients $a_{ij}, (i, j = 1, ..., n)$, i.e.,

$$Max -\frac{1}{2}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left[\left(X_i^{t+1} - \sum_{j=1}^{n} a_{ij}X_j^t\right)^2 + \ln(2\pi\sigma_i^2)\right]$$

$$s.t. \quad a_{ij} \geq 0 \quad if\ i \rightarrow_+ j$$
$$a_{ij} \leq 0 \quad if\ i \rightarrow_- j$$
$$a_{ij} = 0 \quad if\ i \otimes j$$
$$i, j = 1, ..., n.$$

The constraints in the QP represent the regulatory strength between i and j. Based on the log-likelihood value, the significance of a network architecture was evaluated by a random sampling process [134, 145, 146]. For each regulatory network, we randomly generated N (e.g. 2000) networks by rewiring the same number of regulations in the nodes of the evaluating network. An empirical *p*-value is calculated to evaluate its statistical significance, i.e.,

$$p - value = \frac{\{The\ number\ of\ L(R) > L(G)\}}{N},$$

where R is a random network, $L(\cdot)$ is the maximum log-likelihood value of the random network R and the evaluating network G. The evaluation provides a powerful alternative to identify responsive regulatory networks in certain dynamics of environment and condition [134].

Apparently, the knowledge-based regulatory relationships among these genes are not complete and the reference network library should be as complete as possible. To the ends, [149] and [150] have developed methods to integrate inference and evaluation in the same framework by completing the gene network with modifications so that the resultant network achieves more consistency with the gene expression data. The missing regulations can be identified from initial incomplete prior network. Due to the difficulties of pure data-driven inference of regulatory network, the alternatives of combining prior knowledge and evaluating prior gene regulations show promising research directions to investigate transcriptional regulatory network from gene expression data [134].

4. DISCUSSION AND CONCLUSION

In this review, we summarized the state-of-the-art methods of reverse engineering transcriptional regulatory networks from gene expression data and categorized them into several general frameworks, i.e., correlation-based methods, Boolean network methods, Bayesian network methods, differential equation methods and knowledge-based methods. (Table **1**) lists these strategies and their typical methods. Some methods implement hybrid models and employ several computational techniques to reversely engineer regulatory networks [41, 83]. These methods such as REVEAL [105], BC3NET [151] and GENIE3 [152] can be classified into multiple categories. For simplicity, we only categorized

Table 1. Some available strategies and their representative methods for inferring regulatory networks from gene expression profiles. Their supporting websites and original publications are also shown. Some R packages (http://cran.r-project.org) for Bayesian learning and differential equation parameter identification are also shown. In each category, the methods are ordered alphabetically.

Category	Method	Website	Reference
Correlation-based methods	ANOVA	http://www2.bio.ifi.lmu.de/~kueffner/anova.tar.gz	[155]
	ARACNE	http://wiki.c2b2.columbia.edu/califanolab/ index.php/Software/ARACNE	[56, 73]
	CLR	http://cran.r-project.org/web/packages/parmigene	[74]
	C3NET	http://cran.r-project.org/web/packages/c3net/index.html	[71]
	GLMNET	http://cran.r-project.org/web/packages/glmnet/	[99, 100]
	grangerTlasso	http://www.biostat.washington.edu/~ashojaie/	[103]
	MINET	http://cran.r-project.org/web/packages/minet/	[76]
	MRNET	http://penglab.janelia.org/proj/mRMR/	[75]
	ParCorA	http://www.comp-sys-bio.org/software.html	[88]
	PCA-CMI	http://csb.shu.edu.cn/subweb/grn.htm	[60]
	Relevance Network	http://buttelab.stanford.edu/start	[65, 70]
	Schafer and Strimmer	http://strimmerlab.org/software.html	[89]
	Simone	http://cran.r-project.org/web/packages/simone/	[156]
	Stuart *et al.*	http://cmgm.stanford.edu/~kimlab/multispecies/	[64]
	WGCNA	http://labs.genetics.ucla.edu/horvath/ CoexpressionNetwork/Rpackages/ WGCNA	[66, 68]
Boolean network methods	Akutsu *et al.*	http://www.bic.kyoto-u.ac.jp/takutsu/members/takutsu/	[106, 109]
	Antelope	http://turing.iimas.unam.mx:8080/AntelopeWEB/content/about.jsp	[157]
	BoolNet	http://cran.r-project.org/web/packages/BoolNet	[158]
	BooleSim	https://github.com/matthiasbock/BooleSim	[159]
	Handorf and Klipp	http://code.google.com/p/libscopes/wiki/Paper2011	[160]
	Modent	http://acgt.cs.tau.ac.il/modent/	[161]
	REVEAL	Not available	[105]
	Shmulevich *et al.*	http://shmulevich.systemsbiology.net/	[112, 113, 115]
Bayesian network methods	ARTIVA	http://cran.r-project.org/web/packages/ARTIVA/index.html	[162]
	BC3NET	http://cran.r-project.org/web/packages/bc3net/index.html	[151]
	Beal *et al.*	http://www.cse.buffalo.edu/faculty/mbeal/	[125]
	BNFinder	http://bioputer.mimuw.edu.pl/software/bnf	[163]
	BNLEARN	http://cran.r-project.org/web/packages/bnlearn	[164]
	BNT	http://code.google.com/p/bnt/	[120]
	Frideman *et al.*	http://www.cs.huji.ac.il/labs/compbio/expression/	[117, 119]
	GeneNet	http://cran.r-project.org/web/packages/GeneNet	[83]
	G1DBN	http://cran.r-project.org/web/packages/G1DBN/index.html	[165]
	GlobalMIT	https://code.google.com/p/globalmit	[166]
	Module network	http://ai.stanford.edu/~erans/module_nets/	[84]
	TESLA	http://sailing.cs.cmu.edu/tesla/index.html	[126]
	SSM	http://www.chems.msu.edu/groups/chan/ssm.zip	[167]

(Table 1) contd….

Category	Method	Website	Reference
Differential equation methods	Chen *et al.*	Not available	[127]
	deSolve	http://cran.r-project.org/web/packages/deSolve	[168]
	D'haeseleer *et al.*	Not available	[95]
	D-NetWeaver	https://cbim.urmc.rochester.edu/software/d-netweaver/	[128, 129]
	GRNInfer	http://doc.aporc.org/wiki/Software	[41]
	Inferelator	http://bonneaulab.bio.nyu.edu/software.html	[154]
	Tegner *et al.*	http://www.bu.edu/bme/people/primary/collins/	[34, 36]
	TRNInfer	http://www.sysbio.ac.cn/cb/chenlab/software.htm	[153]
	Wahde and Hertz	http://www.nbi.dk/~hertz/	[169]
Knowledge-based methods	Banjo	http://www.cs.duke.edu/~amink/software/banjo	[142]
	BNP	http://research.bioe.bilgi.edu.tr/bnp/	[170]
	Greenfield *et al.*	http://bonneaulab.bio.nyu.edu/software.html	[171]
	Hill *et al.*	http://mukherjeelab.nki.nl/DBN	[172]
	Linear programming	http://doc.aporc.org/wiki/Software	[138]
	Liu *et al.*	http://doc.aporc.org/wiki/Software	[134]
	Network energy	Not available	[139, 143]
	Network Screening	http://www.molprof.jp/~horimoto/	[144]
	PLASSO	http://nba.uth.tmc.edu/homepage/liu/pLasso	[173]
Miscellaneous methods	GENIE3	http://homepages.inf.ed.ac.uk/vhuynht/software.html	[152]
	Neural network	http://www.me.chalmers.se/~mwahde	[174]
	Petri net	http://dnagarden.hgc.jp/en/doku.php/software	[175]
	Supervised learning	http://cbio.ensmp.fr/sirene	[176, 177]
	TIGRESS	http://cbio.ensmp.fr/tigress	[178]

them into one of them. For instance, REVEAL also employs mutual information technique beyond Boolean network, so it can also belong to correlation-based methods. In (Table **1**), some methods such as TRNInfer [153] and Inferelator [154] reconstruct the four levels of transcriptional regulatory relationships, while others such as PCA-CMI [60] and WGCNA [68] generally identify gene regulations without direction information.

After the emergence of high-throughput microarray techniques, great efforts have been undertaken to infer transcriptional regulatory networks from gene expression profiles. Because of the complexity of gene regulations, it is still a challenging task to infer genome-wide regulatory networks from expression data by mathematical modeling [179]. Various computational methods have been proposed to interpret gene expression data and decipher the regulation mechanism of controlling gene expression. The reviewed methods are very useful for providing the quantitative models of harnessing the perturbation and time series of gene expression datasets and identifying the causal relationship of transcriptional regulations. In turn, the endeavors of coupling the regulatory interaction between genes imply the paramount importance of gene regulations in the study of genomics and genetics. It is difficult to assess these methods and select the best one that supersedes all the others by defining some benchmark standards [23]. The details and assumptions in the modeling of real regulation systems as well as the gene expressions in specific conditions and phenotypes determine the superiority of each method. The simple model as Boolean network can reveal critical implications in transcriptional regulation systems [107].

Besides the methods reviewed above, some other methods such as supervised learning [176, 177], feature selection [152, 178], neural network [174] and Petri net [175] methods have also been proposed to address the problem of learning transcriptional regulatory network from gene expression data. Most of these miscellaneous methods are heuristic for mining the relationship between genes from expression profiles. The availability next generation sequencing (NGS) technologies, e.g., RNA-Seq [180], can generate tran-

scriptomic data of higher quality. Theoretically, these reviewed methods can be extended easily to reverse engineering transcriptional regulatory networks from RNA-Seq data. At the same time, identification of the causal regulatory mechanism of gene expression dynamics from gene expression data is constrained from the assumptions and approximations in the models. For instance, time delay between the activation of a TF and its downstream target genes widely exists in the regulatory relationships, which has not been well considered in the available methods [123, 181]. Also, the dynamics of regulation has not been modeled sufficiently, i.e., the regulation strength between TF and targets are always time-varying with temporal features [126]. The reviewed methods can be extended to integrate these important regulatory features into the models of causal regulatory relationships between genes. The reverse engineering methods will become more and more sophisticated for modeling transcriptional regulatory systems as comprehensively as possible.

Beyond utilizing gene expression data, an important research direction in building transcriptional regulatory network is to predict the interaction between TF protein and DNA by machine learning methods. Currently, one of the most important problems in the predictions is how to effectively formulate a biological sequence with a discrete model or a vector, yet still keep considerable sequence order information. All the existing operation engines, such as covariance discriminant (CD) [182, 183], neural network [184], support vector machine (SVM) [185, 186], random forest [187], conditional random field [188], nearest neighbor (NN) [189]; OET-KNN [190], Fuzzy K-nearest neighbor [191, 192], ML-KNN algorithm [193], and SLLE algorithm [183], can only handle vector but not sequence samples. However, a vector defined in a discrete model may completely lose all of the sequence-order information [194]. To avoid completely losing the sequence-order information for proteins, the pseudo amino acid composition or Chou's PseAAC was proposed [194, 195]. Ever since the concept of PseAAC was proposed in 2001 [194], the approach of representing protein/peptide sequences has been widely used in all the areas of computational genomics [196]. Moreover, the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) was introduced to deal with DNA/RNA sequences in computational genomics, such as for identifying nucleosome positioning [182] and predicting recombination spots [197]. The development of PseAAC for protein sequences and PseKNC for DNA sequences will highly facilitate the prediction of transcriptional regulatory interactions between TF protein and DNA only from sequence information [187, 198, 199].

The challenges of reverse engineering are not only from the information availability, but also from the complexity of regulation system [200]. The measured gene expression levels are not merely determined by the activity of its transcriptional regulators. Post-transcriptional regulations (microRNA silencing [51]) as well as epigenetic modifications (DNA methylation [48] and histone modification [201]) on the gene sequences also highly affect the levels of gene expression. The sequential and combinatorial regulations of gene expression among epigenetic factors, TFs, microRNAs should be considered systematically in reverse engineering regulatory systems when these genomic datasets are available [202]. In the future, the heterogeneous regulatory system with multiple genetic and epigenetic factors should be modeled to integrate transcriptional and post-transcriptional regulations. The integration of genomics, transcriptomics, proteomics datasets, such as ChIP-Seq, protein-binding motifs, gene expression, miRNA abundance, ratios of DNA methylation and chromatin modification, and prior knowledge of regulation, from multiple levels and various aspects of gene regulations provides a possible solution to reconstruct context-specific gene regulations [47]. The networks inferred from various levels can crossly validate each other for accurately identifying gene regulations underlying the whole system. Furthermore, the contradicted identifications in these inferences should be analyzed carefully. They might be caused by the noisy datasets, unrevealed regulatory mechanisms, and specific phenotype associations. Reverse engineering of transcriptional regulatory networks by integrating multiple datasets is a very important research direction [41, 153]. Consistent regulatory relationships at multiple levels shed a brilliant light on the gene expression dynamics in response to various internal signals and external stimuli.

In conclusion, a genome-wide inference of transcriptional regulatory networks from gene expression data provides a promising way to decipher the large-scale causal regulatory relationships among genes. Model-based computational methods of harnessing genomic data facilitate the discovery and revolutionize the research of gene regulation. We summarized the advantages and commented on the improvement possibilities of addressing the disadvantages of these methods individually. The assumptions of modeling the spatial and temporal gene regulations will become more and more reasonable with the accumulation of knowledge about gene regulations. The models will also become more and more close to the real complexity of gene regulation when we obtain better gene expression data with enough sample size and dedicated experiment design, multilevel biological processes, higher quality of expression signals, and systematic perspectives. Knowledge-based methods of integrating existing priori and gene expression seem to be powerful and flexible to decipher the genuine transcriptional control circuits in regulatory systems.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Thanks are due to Drs Songyot Nakariyakul, Xiaoxu Han, Rui-Sheng Wang, Xianwen Ren and Jiguang Wang for their critical comments. This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 31100949 and the Fundamental Research Funds of Shandong University under Grant No. 2014TB006.

REFERENCES

[1] Spitz, F.; Furlong, E.E. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.,* **2012**, *13* (9), 613-626.

[2] Chen, K.; Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.,* **2007**, *8* (2), 93-103.

[3] Levine, M.; Davidson, E.H. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. U S A*, **2005**, *102* (14), 4936-4942.

[4] Orphanides, G.; Reinberg, D. A unified theory of gene expression. *Cell*, **2002**, *108* (4), 439-451.

[5] Babu, M.M.; Luscombe, N.M.; Aravind, L.; Gerstein, M.; Teichmann, S.A. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **2004**, *14* (3), 283-291.

[6] Beer, M.A.; Tavazoie, S. Predicting gene expression from sequence. *Cell*, **2004**, *117* (2), 185-198.

[7] Kim, H.D.; O'Shea, E.K. A quantitative model of transcription factor-activated gene expression. *Nat. Struct. Mol. Biol.*, **2008**, *15* (11), 1192-1198.

[8] de Matos Simoes, R.; Dehmer, M.; Emmert-Streib, F. Interfacing cellular networks of S. cerevisiae and E. coli: connecting dynamic and genetic information. *BMC Genomics*, **2013**, *14*, 324.

[9] Amit, I.; Garber, M.; Chevrier, N.; Leite, A.P.; Donner, Y.; Eisenhaure, T.; Guttman, M.; Grenier, J. K.; Li, W.; Zuk, O.; Schubert, L. A.; Birditt, B.; Shay, T.; Goren, A.; Zhang, X.; Smith, Z.; Deering, R.; McDonald, R. C.; Cabili, M.; Bernstein, B. E.; Rinn, J. L.; Meissner, A.; Root, D.E.; Hacohen, N.; Regev, A. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, **2009**, *326* (5950), 257-263.

[10] Blais, A.; Dynlacht, B.D. Constructing transcriptional regulatory networks. *Genes Dev.*, **2005**, *19* (13), 1499-1511.

[11] Lee, T.I.; Rinaldi, N.J.; Robert, F.; Odom, D.T.; Bar-Joseph, Z.; Gerber, G.K.; Hannett, N.M.; Harbison, C.T.; Thompson, C.M.; Simon, I.; Zeitlinger, J.; Jennings, E.G.; Murray, H.L.; Gordon, D.B.; Ren, B.; Wyrick, J.J.; Tagne, J.B.; Volkert, T.L.; Fraenkel, E.; Gifford, D.K.; Young, R.A. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **2002**, *298* (5594), 799-804.

[12] Bogarad, L.D.; Arnone, M.I.; Chang, C.; Davidson, E.H. Interference with gene regulation in living sea urchin embryos: transcription factor knock out (TKO), a genetically controlled vector for blockade of specific transcription factors. *Proc. Natl. Acad. Sci. U S A*, **1998**, *95* (25), 14827-14832.

[13] Pe'er, D.; Regev, A.; Elidan, G.; Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **2001**,*17* (Suppl 1), S215-224.

[14] Hu, Z.; Killion, P.J.; Iyer, V.R. Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, **2007**,*39* (5), 683-687.

[15] Workman, C.T.; Mak, H.C.; McCuine, S.; Tagne, J.B.; Agarwal, M.; Ozier, O.; Begley, T.J.; Samson, L.D.; Ideker, T.A systems approach to mapping DNA damage response pathways. *Science*, **2006**, *312* (5776), 1054-1059.

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

[17] Johnson, D.S.; Mortazavi, A.; Myers, R.M.; Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*,**2007**,*316* (5830), 1497-1502.

[18] Marson, A.; Kretschmer, K.; Frampton, G.M.; Jacobsen, E.S.; Polansky, J.K.; MacIsaac, K.D.; Levine, S.S.; Fraenkel, E.; von Boehmer, H.; Young, R.A. Foxp3 occupancy and regulation of key target genes during T-cell stimulation. *Nature*, **2007**, *445* (7130), 931-935.

[19] Schena, M.; Shalon, D.; Davis, R.W.; Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*,**1995**, *270* (5235), 467-470.

[20] Hughes, T.R.; Marton, M.J.; Jones, A.R.; Roberts, C.J.; Stoughton, R.; Armour, C.D.; Bennett, H.A.; Coffey, E.; Dai, H.; He, Y.D.; Kidd, M.J.; King, A.M.; Meyer, M.R.; Slade, D.; Lum, P.Y.; Stepaniants, S.B.; Shoemaker, D.D.; Gachotte, D.; Chakrabartty, K.; Simon, J.; Bard, M.; Friend, S.H. Functional discovery via a compendium of expression profiles. *Cell*, **2000**,*102* (1), 109-126.

[21] Hartemink, A.J. Reverse engineering gene regulatory networks. *Nat. Biotechnol.*, **2005**, *23* (5), 554-555.

[22] Marbach, D.; Costello, J.C.; Kuffner, R.; Vega, N.M.; Prill, R.J.; Camacho, D.M.; Allison, K.R.; Kellis, M.; Collins, J.J.; Stolovitzky, G. Wisdom of crowds for robust gene network inference. *Nat. Methods*, **2012**, *9* (8), 796-804.

[23] Marbach, D.; Prill, R.J.; Schaffter, T.; Mattiussi, C.; Floreano, D.; Stolovitzky, G. Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U S A*, *107* (14), 6286-6291.

[24] Csete, M.E.; Doyle, J.C. Reverse engineering of biological complexity. *Science*, **2002**, *295* (5560), 1664-1669.

[25] Kaern, M.; Blake, W.J.; Collins, J.J. The engineering of gene regulatory networks. *Annu. Rev. Biomed. Eng.*, **2003**, *5*, 179-206.

[26] de Jong, H. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **2002**, *9* (1), 67-103.

[27] Bansal, M.; Belcastro, V.; Ambesi-Impiombato, A.; di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **2007**, *3*, 78.

[28] 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* (1), 86-103.

[29] Karlebach, G.; Shamir, R. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell. Biol.*, **2008**, *9* (10), 770-780.

[30] Emmert-Streib, F.; Glazko, G.V.; Altay, G.; de Matos Simoes, R. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Front. Genet.*, **2012**, *3*, 8.

[31] Werhli, A.V.; Grzegorczyk, M.; Husmeier, D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, **2006**, *22* (20), 2523-2531.

[32] Li, H.; Xuan, J.; Wang, Y.; Zhan, M. Inferring regulatory networks. *Front. Biosci.*, **2008**, *13*, 263-275.

[33] Lee, W.P.; Tzou, W.S. Computational methods for discovering gene networks from expression data. *Brief Bioinform.*, **2009**, *10* (4), 408-423.

[34] Yeung, M.K.; Tegner, J.; Collins, J.J. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. U S A*, **2002**, *99* (9), 6163-6168.

[35] Gardner, T.S.; Faith, J.J. Reverse-engineering transcription control networks. *Phys. Life Rev.*, **2005**, *2* (1), 65-88.

[36] Tegner, J.; Yeung, M.K.; Hasty, J.; Collins, J.J. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. U S A*, **2003**, *100* (10), 5944-5949.

[37] Nachman, I.; Regev, A.; Friedman, N. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*,**2004**,*20* (Suppl 1), i248-256.

[38] Luscombe, N.M.; Babu, M.M.; Yu, H.; Snyder, M.; Teichmann, S.A.; Gerstein, M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **2004**, *431* (7006), 308-312.

[39] Kim, H.D.; Shay, T.; O'Shea, E.K.; Regev, A. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science*,**2009**,*325* (5939), 429-432.

[40] Amit, I.; Regev, A.; Hacohen, N. Strategies to discover regulatory circuits of the mammalian immune system. *Nat. Rev. Immunol.*, **2011**,*11* (12), 873-880.

[41] Wang, Y.; Joshi, T.; Zhang, X.S.; Xu, D.; Chen, L. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, **2006**, *22* (19), 2413-2420.

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

[43] Kato, M.; Hata, N.; Banerjee, N.; Futcher, B.; Zhang, M.Q. Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol.*, **2004**, *5* (8), R56.

[44] Fairfax, B.P.; Makino, S.; Radhakrishnan, J.; Plant, K.; Leslie, S.; Dilthey, A.; Ellis, P.; Langford, C.; Vannberg, F. O.; Knight, J. C. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.*, **2012**, *44* (5), 502-510.

[45] Pilpel, Y.; Sudarsanam, P.; Church, G.M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **2001**, *29* (2), 153-159.

[46] Emmert-Streib, F.; de Matos Simoes, R.; Mullan, P.; Haibe-Kains, B.; Dehmer, M. The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Front. Genet.*, **2014**,*5*, 15.

[47] Gibcus, J.H.; Dekker, J. The context of gene expression regulation. *F1000 Biol. Rep.*, **2012**, *4*, 8.

[48] Bock, C. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, **2012**, *13* (10), 705-719.

[49] Segal, E.; Widom, J. What controls nucleosome positions? *Trends Genet.*, **2009**, *25* (8), 335-343.

[50] Jaenisch, R.; Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **2003**, *33* (Suppl), 245-254.

[51] Bartel, D. P. MicroRNAs: target recognition and regulatory func-

tions. *Cell*, **2009**, *136* (2), 215-233.

[52] Loven, J.; Orlando, D. A.; Sigova, A. A.; Lin, C. Y.; Rahl, P. B.; Burge, C. B.; Levens, D. L.; Lee, T. I.; Young, R. A. Revisiting global gene expression analysis. *Cell*, **2012**, *151* (3), 476-482.

[53] Canales, R. D.; Luo, Y.; Willey, J. C.; Austermiller, B.; Barbacioru, C. C.; Boysen, C.; Hunkapiller, K.; Jensen, R. V.; Knight, C. R.; Lee, K. Y.; Ma, Y.; Maqsodi, B.; Papallo, A.; Peters, E. H.; Poulter, K.; Ruppel, P. L.; Samaha, R. R.; Shi, L.; Yang, W.; Zhang, L.; Goodsaid, F. M. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.*, **2006**,*24* (9), 1115-1122.

[54] Irizarry, R. A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y. D.; Antonellis, K. J.; Scherf, U.; Speed, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*,**2003**,*4* (2), 249-264.

[55] Kao, K. C.; Yang, Y. L.; Boscolo, R.; Sabatti, C.; Roychowdhury, V.; Liao, J. C. Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. *Proc. Natl. Acad. Sci. U S A*, **2004**, *101* (2), 641-646.

[56] Basso, K.; Margolin, A. A.; Stolovitzky, G.; Klein, U.; Dalla-Favera, R.; Califano, A. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **2005**, *37* (4), 382-390.

[57] Stolovitzky, G.; Monroe, D.; Califano, A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann. N Y Acad. Sci.*, **2007**, *1115*, 1-22.

[58] Schaffter, T.; Marbach, D.; Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **2011**, *27* (16), 2263-2270.

[59] Liu, Z. P.; Chen, L. Proteome-wide prediction of protein-protein interactions from high-throughput data. *Protein Cell*, **2012**, *3* (7), 508-520.

[60] Zhang, X.; Zhao, X. M.; He, K.; Lu, L.; Cao, Y.; Liu, J.; Hao, J. K.; Liu, Z. P.; Chen, L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, **2012**, *28* (1), 98-104.

[61] Marbach, D.; Prill, R. J.; Schaffter, T.; Mattiussi, C.; Floreano, D.; Stolovitzky, G. Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U S A*, **2010**,*107* (14), 6286-6291.

[62] Ben-Dor, A.; Shamir, R.; Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.*, **1999**, *6* (3-4), 281-297.

[63] Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U S A*, **1998**, *95* (25), 14863-14868.

[64] Stuart, J. M.; Segal, E.; Koller, D.; Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science*,**2003**,*302* (5643), 249-255.

[65] Butte, A. J.; Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **2000**, *5*, 418-429.

[66] Langfelder, P.; Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **2008**, *9*, 559.

[67] He, D.; Liu, Z. P.; Honda, M.; Kaneko, S.; Chen, L. Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. *J. Mol. Cell Biol.*, **2012**,*4* (3), 140-152.

[68] Zhang, B.; Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **2005**,*4*, Article17.

[69] Fujita, A.; Sato, J. R.; Demasi, M. A.; Sogayar, M. C.; Ferreira, C. E.; Miyano, S. Comparing Pearson, Spearman and Hoeffding's D measure for gene expression association analysis. *J. Bioinform. Comput. Biol.*, **2009**, *7* (4), 663-684.

[70] Butte, A. J.; Tamayo, P.; Slonim, D.; Golub, T. R.; Kohane, I. S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. U S A*, **2000**, *97* (22), 12182-12186.

[71] Altay, G.; Emmert-Streib, F. Inferring the conservative causal core of gene regulatory networks. *BMC Syst. Biol.*, **2010**, *4*, 132.

[72] Olsen, C.; Meyer, P.E.; Bontempi, G. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP J. Bioinform. Syst. Biol.*, **2009**, 308959.

[73] Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Dalla Favera, R.; Califano, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **2006**, *7* (Suppl 1), S7.

[74] Faith, J.J.; Hayete, B.; Thaden, J.T.; Mogno, I.; Wierzbowski, J.; Cottarel, G.; Kasif, S.; Collins, J.J.; Gardner, T.S. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **2007**, *5* (1), e8.

[75] Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.*, **2005**, *3* (2), 185-205.

[76] Meyer, P.E.; Lafitte, F.; Bontempi, G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **2008**, *9*, 461.

[77] Carro, M.S.; Lim, W.K.; Alvarez, M.J.; Bollo, R.J.; Zhao, X.; Snyder, E.Y.; Sulman, E.P.; Anne, S.L.; Doetsch, F.; Colman, H.; Lasorella, A.; Aldape, K.; Califano, A.; Iavarone, A. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, **2010**, *463* (7279), 318-325.

[78] Wang, K.; Saito, M.; Bisikirska, B.C.; Alvarez, M.J.; Lim, W.K.; Rajbhandari, P.; Shen, Q.; Nemenman, I.; Basso, K.; Margolin, A. A.; Klein, U.; Dalla-Favera, R.; Califano, A. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat. Biotechnol.*, **2009**, *27* (9), 829-839.

[79] Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science*, **2011**, *334* (6062), 1518-1524.

[80] Kinney, J.B.; Atwal, G.S. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. U S A*, **2014**, *111* (9), 3354-3359.

[81] Song, L.; Langfelder, P.; Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, **2012**, *13*, 328.

[82] D'Haeseleer, P.; Liang, S.; Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*,**2000**,*16* (8), 707-726.

[83] 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.

[84] Segal, E.; Shapira, M.; Regev, A.; Pe'er, D.; Botstein, D.; Koller, D.; Friedman, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **2003**, *34* (2), 166-176.

[85] Liu, Z.P.; Wang, Y.; Zhang, X.S.; Chen, L. Identifying dysfunctional crosstalk of pathways in various regions of Alzheimer's disease brains. *BMC Syst. Biol.*, **2010**, *4* (Suppl 2), S11.

[86] He, D.; Liu, Z.P.; Chen, L. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics*, *12*, 592.

[87] Liu, K. Q.; Liu, Z.P.; Hao, J.K.; Chen, L.; Zhao, X.M. Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics*, **2012**, *13* (1), 126.

[88] de la Fuente, A.; Bing, N.; Hoeschele, I.; Mendes, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*,**2004**,*20* (18), 3565-3574.

[89] Schafer, J.; Strimmer, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **2005**, *21* (6), 754-764.

[90] Liang, K.C.; Wang, X. Gene regulatory network reconstruction using conditional mutual information. *EURASIP J. Bioinform. Syst. Biol.*, **2008**, 253894.

[91] Saito, S.; Hirokawa, T.; Horimoto, K. Discovery of chemical compound groups with common structures by a network analysis approach (affinity prediction method). *J. Chem. Inf. Model*, *51* (1), 61-68.

[92] Frenzel, S.; Pompe, B. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.*, **2007**, *99* (20), 204101.

[93] Speed, T.A correlation for the 21st century. *Science*, **2011**, *334* (6062), 1502-1503.

[94] Kalisch, M.; Buhlmann, P.Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. LearnRes.*, **2005**, *7*, 613-636.

[95] D'Haeseleer, P.; Wen, X.; Fuhrman, S.; Somogyi, R. Linear model-

ing of mRNA expression levels during CNS development and injury. *Pac. Symp. Biocomput.*, **1999**, *4*, 41-52.

[96] Kim, H.; Lee, J.K.; Park, T. Inference of large-scale gene regulatory networks using regression-based network approach. *J. Bioinform. Comput. Biol.*, **2009**,*7* (4), 717-735.

[97] Opgen-Rhein, R.; Strimmer, K. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, **2007**, *8*(Suppl 2), S3.

[98] Pihur, V.; Datta, S. Reconstruction of genetic association networks from microarray data: a partial least squares approach. *Bioinformatics*, **2008**, *24* (4), 561-568.

[99] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, **1996**, *58*, 267-288.

[100] Zou, H.; Trevor, H. Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Series B Stat. Methodol.*, **2005**, *67*, 301-320.

[101] Shojaie, A.; Michailidis, G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, **2010**, *97* (3), 519-538.

[102] Lozano, A.C.; Abe, N.; Liu, Y.; Rosset, S. Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, **2009**, *25* (12), i110-118.

[103] Shojaie, A.; Michailidis, G. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, **2010**, *26* (18), i517-523.

[104] Thomas, R. Boolean formalization of genetic control circuits. *J. Theor. Biol.*, **1973**, *42* (3), 563-585.

[105] Liang, S.; Fuhrman, S.; Somogyi, R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, **1998**, *3*, 18-29.

[106] Akutsu, T.; Miyano, S.; Kuhara, S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.*, **1999**, *4*, 17-28.

[107] Wang, R.S.; Saadatpour, A.; Albert, R. Boolean modeling in systems biology: an overview of methodology and applications. *Phys. Biol.*, *9* (5), 055001.

[108] Ivanov, I. Boolean models of genomic regulatory networks: reduction mappings, inference, and external control. *Curr. Genomics*, **2009**, *10* (6), 375-387.

[109] Akutsu, T.; Miyano, S.; Kuhara, S. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **2000**, *16* (8), 727-734.

[110] Ideker, T.E.; Thorsson, V.; Karp, R.M. Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput.*, **2000**, *5*, 305-316.

[111] Garg, A.; Di Cara, A.; Xenarios, I.; Mendoza, L.; De Micheli, G. Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics*, **2008**, *24* (17), 1917-1925.

[112] Shmulevich, I.; Dougherty, E.R.; Kim, S.; Zhang, W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **2002**, *18* (2), 261-274.

[113] Shmulevich, I.; Lahdesmaki, H.; Dougherty, E.R.; Astola, J.; Zhang, W. The role of certain Post classes in Boolean network models of genetic networks. *Proc. Natl. Acad. Sci. U S A*, **2003**, *100* (19), 10734-10739.

[114] Faryabi, B.; Vahedi, G.; Datta, A.; Chamberland, J. F.; Dougherty, E. R. Recent advances in intervention in markovian regulatory networks. *Curr. Genomics*, **2009**, *10* (7), 463-477.

[115] Shmulevich, I.; Dougherty, E.R.; Zhang, W. From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. *Proc. IEEE*,**2002**,*90* (11), 1778-1792.

[116] Pearl, J. *Probabilistic Reasoning in Intelligent Systems*, San Francisco: Morgan Kaufmann, **1988**.

[117] Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D.Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **2000**, *7* (3-4), 601-620.

[118] Heckerman, D.; Chickering, D.M. Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn*, **1995**, *20*, 197-243.

[119] Friedman, N.Inferring cellular networks using probabilistic graphical models. *Science*, **2004**, *303* (5659), 799-805.

[120] Murphy, K.P.; Milan, S. Modelling gene expression data using dynamic Bayesian networks. *Tech. Rep. MIT Artificial Intelligence Lab.*,**1999**.

[121] Perrin, B.E.; Ralaivola, L.; Mazurie, A.; Bottani, S.; Mallet, J.; d'Alche-Buc, F. Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **2003**, *19* (Suppl 2), ii138-148.

[122] Husmeier, D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **2003**, *19* (17), 2271-2282.

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

[124] Kim, S.; Imoto, S.; Miyano, S. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, **2004**, *75* (1-3), 57-65.

[125] Beal, M.J.; Falciani, F.; Ghahramani, Z.; Rangel, C.; Wild, D.L. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, **2005**, *21* (3), 349-356.

[126] Ahmed, A.; Xing, E.P. Recovering time-varying networks of dependencies in social and biological studies. *Proc. Natl. Acad. Sci. U S A*, **2009**, *106* (29), 11878-11883.

[127] Chen, T.; He, H.L.; Church, G.M. Modeling gene expression with differential equations. *Pac. Symp. Biocomput.*, **1999**, *4*, 29-40.

[128] Wu, S.; Liu, Z.P.; Qiu, X.; Wu, H. Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high-dimensional ordinary differential equations. *PLoS One*, **2014**,*9* (5), e95276.

[129] Lu, T.; Liang, H.; Li, H.; Wu, H. High dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *J. Am. Stat. Assoc.*, **2011**,*106* (496), 1242-1258.

[130] Bonneau, R. Learning biological networks: from modules to dynamics. *Nat. Chem. Biol.*, **2008**, *4* (11), 658-664.

[131] Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **2001**, *96* (45), 1348-1360.

[132] Miao, H.; Xia, X.; Perelson, A.S.; Wu, H. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Rev.*, **2011**, *53* (1), 3-39.

[133] Emmert-Streib, F.; Dehmer, M.; Haibe-Kains, B. Untangling statistical and biological models to understand network inference: The need for a genomics network ontology. *Front. Genet.*, **2014**, *5*, 299.

[134] Liu, Z.P.; Zhang, W.; Horimoto, K.; Chen, L. Gaussian graphical model for identifying significantly responsive regulatory networks from time course high-throughput data. *IET Syst. Biol.*, *7* (5), 143-152.

[135] Salgado, H.; Peralta-Gil, M.; Gama-Castro, S.; Santos-Zavaleta, A.; Muniz-Rascado, L.; Garcia-Sotelo, J.S.; Weiss, V.; Solano-Lira, H.; Martinez-Flores, I.; Medina-Rivera, A.; Salgado-Osorio, G.; Alquicira-Hernandez, S.; Alquicira-Hernandez, K.; Lopez-Fuentes, A.; Porron-Sotelo, L.; Huerta, A. M.; Bonavides-Martinez, C.; Balderas-Martinez, Y. I.; Pannier, L.; Olvera, M.; Labastida, A.; Jimenez-Jacinto, V.; Vega-Alvarado, L.; Del Moral-Chavez, V.; Hernandez-Alvarez, A.; Morett, E.; Collado-Vides, J. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.*,**2012**, *41* (Database issue), D203-213.

[136] Bernstein, B.E.; Birney, E.; Dunham, I.; Green, E.D.; Gunter, C.; Snyder, M. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **2012**, *489* (7414), 57-74.

[137] Ren, B.; Robert, F.; Wyrick, J.J.; Aparicio, O.; Jennings, E.G.; Simon, I.; Zeitlinger, J.; Schreiber, J.; Hannett, N.; Kanin, E.; Volkert, T.L.; Wilson, C.J.; Bell, S.P.; Young, R.A. Genome-wide location and function of DNA binding proteins. *Science*, **2000**, *290* (5500), 2306-2309.

[138] Liu, Z.P.; Zhang, X.S.; Chen, L. Inferring gene regulatory networks from expression data with prior knowledge by linear programming. In: *Preceedings of International Conference on Machine Learning and Cybernetics*, Qingdao,China, July 11-14,2010; IEEE:New York, USA, **2010**; pp. 3067-3072.

[139] Werhli, A.V.; Husmeier, D.Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.*, **2007**, *6*, Article15.

[140] Mukherjee, S.; Speed, T.P. Network inference using informative priors. *Proc. Natl. Acad. Sci. U S A*, **2008**, *105* (38), 14313-14318.

[141] Steele, E.; Tucker, A.; t Hoen, P.A.; Schuemie, M.J. Literature-based priors for gene regulatory networks. *Bioinformatics*, **2009**, *25* (14), 1768-1774.

[142] Bernard, A.; Hartemink, A.J. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac. Symp. Biocomput.*, **2005**, *10*, 459-470.

[143] Imoto, S.; Higuchi, T.; Goto, T.; Tashiro, K.; Kuhara, S.; Miyano, S.Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *J. Bioinform. Comput. Biol.*, **2004**, *2* (1), 77-98.

[144] Saito, S.; Aburatani, S.; Horimoto, K. Network evaluation from the consistency of the graph structure with the measured data. *BMC Syst. Biol.*, **2008**, *2*, 84.

[145] Zhou, H.; Saito, S.; Piao, G.; Liu, Z.P.; Wang, J.; Horimoto, K.; Chen, L. Network screening of Goto-Kakizaki rat liver microarray data during diabetic progression. *BMC Syst. Biol.*, **2011**, *5* (Suppl 1), S16.

[146] Piao, G.; Saito, S.; Sun, Y.; Liu, Z.P.; Wang, Y.; Han, X.; Wu, J.; Zhou, H.; Chen, L.; Horimoto, K. A computational procedure for identifying master regulator candidates: a case study on diabetes progression in Goto-Kakizaki rats. *BMC Syst. Biol.*, **2012**, *6* (Suppl 1), S2.

[147] Stelling, J.; Klamt, S.; Bettenbrock, K.; Schuster, S.; Gilles, E. D. Metabolic network structure determines key aspects of functionality and regulation. *Nature*,**2002**,*420* (6912), 190-193.

[148] Barkai, N.; Leibler, S. Robustness in simple biochemical networks. *Nature*, **1997**, *387* (6636), 913-917.

[149] Nakajima, N.; Tamura, T.; Yamanishi, Y.; Horimoto, K.; Akutsu, T.Network completion using dynamic programming and least-squares fitting. *ScientificWorldJournal*, **2012**, *2012*, 957620.

[150] Saito, S.; Zhou, X.; Bae, T.; Kim, S.; Horimoto, K. Identification of master regulator candidates in conjunction with network screening and inference. *Int. J. Data Min. Bioinform.*, **2014**, *8* (3), 366-380.

[151] de Matos Simoes, R.; Emmert-Streib, F. Bagging statistical network inference from large-scale gene expression data. *PLoS One*, **2012**, *7* (3), e33624.

[152] Huynh-Thu, V.A.; Irrthum, A.; Wehenkel, L.; Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **2010**, *5* (9).

[153] Wang, R.S.; Wang, Y.; Zhang, X.S.; Chen, L. Inferring transcriptional regulatory networks from high-throughput data. *Bioinformatics*, **2007**, *23* (22), 3056-3064.

[154] Bonneau, R.; Reiss, D.J.; Shannon, P.; Facciotti, M.; Hood, L.; Baliga, N.S.; Thorsson, V. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.*, **2006**,*7* (5), R36.

[155] Kuffner, R.; Petri, T.; Tavakkolkhah, P.; Windhager, L.; Zimmer, R. Inferring gene regulatory networks by ANOVA. *Bioinformatics*,**2012**,*28* (10), 1376-1382.

[156] Charbonnier, C.; Chiquet, J.; Ambroise, C. Weighted-LASSO for structured network inference from time course data. *Stat. Appl. Genet. Mol. Biol.*, **2010**, *9*, Article 15.

[157] Arellano, G.; Argil, J.; Azpeitia, E.; Benitez, M.; Carrillo, M.; Gongora, P.; Rosenblueth, D.A.; Alvarez-Buylla, E.R. "Antelope": a hybrid-logic model checker for branching-time Boolean GRN analysis. *BMC Bioinformatics*, **2011**, *12*, 490.

[158] Mussel, C.; Hopfensitz, M.; Kestler, H.A. BoolNet--an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*,**2010**, *26* (10), 1378-1380.

[159] Bock, M.; Scharp, T.; Talnikar, C.; Klipp, E. BooleSim: an interactive Boolean network simulator. *Bioinformatics*, **2013**, *30* (1), 131-132.

[160] Handorf, T.; Klipp, E. Modeling mechanistic biological networks: an advanced Boolean approach. *Bioinformatics*,**2011**,*28* (4), 557-563.

[161] Karlebach, G.; Shamir, R. Constructing logical models of gene regulatory networks by integrating transcription factor-DNA interactions with expression data: an entropy-based approach. *J. Comput. Biol.*, **2012**, *19* (1), 30-41.

[162] Lebre, S.; Becq, J.; Devaux, F.; Stumpf, M. P.; Lelandais, G. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Syst. Biol.*, **2010**, *4*, 130.

[163] Wilczynski, B.; Dojer, N. BNFinder: exact and efficient method for learning Bayesian networks. *Bioinformatics*, **2009**, *25* (2), 286-287.

[164] Scutari, M. Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Software*, **2010**, *35* (3), 1-22.

[165] Lebre, S. Inferring dynamic genetic networks with low order independencies. *Stat. Appl. Genet. Mol. Biol.*, **2009**, *8*, Article 9.

[166] Vinh, N.X.; Chetty, M.; Coppel, R.; Wangikar, P.P. GlobalMIT: learning globally optimal dynamic bayesian network with the mutual information test criterion. *Bioinformatics*, **2011**, *27* (19), 2765-2766.

[167] Li, Z.; Shaw, S.M.; Yedwabnick, M.J.; Chan, C.Using a state-space model with hidden variables to infer transcription factor activities. *Bioinformatics*, **2006**, *22* (6), 747-754.

[168] Soetaert, K.; Petzoldt, T.; Setzer, R.W.Solving differential equations in R: package deSolve. *J. Stat. Software*, **2010**, *33* (9), 1-25.

[169] Wahde, M.; Hertz, J. Modeling genetic regulatory dynamics in neural development. *J. Comput. Biol.*, **2001**, *8* (4), 429-442.

[170] Isci, S.; Dogan, H.; Ozturk, C.; Otu, H.H.Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics*,**2014**, *30*(6), 860–867.

[171] Greenfield, A.; Hafemeister, C.; Bonneau, R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, **2013**, *29* (8), 1060-1067.

[172] Hill, S.M.; Lu, Y.; Molina, J.; Heiser, L.M.; Spellman, P.T.; Speed, T.P.; Gray, J.W.; Mills, G.B.; Mukherjee, S. Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics*, **2012**, *28* (21), 2804-2810.

[173] Wang, Z.; Xu, W.; San Lucas, F. A.; Liu, Y.Incorporating prior knowledge into Gene Network Study. *Bioinformatics*,**2013**,*29* (20), 2633-2640.

[174] Wahde, M.; Hertz, J. Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems*, **2000**, *55* (1-3), 129-136.

[175] Matsuno, H.; Doi, A.; Nagasaki, M.; Miyano, S. Hybrid Petri net representation of gene regulatory network. *Pac. Symp. Biocomput.*, **2000**, *5*, 341-352.

[176] Soinov, L.A.; Krestyaninova, M.A.; Brazma, A. Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biol.*, **2003**, *4* (1), R6.

[177] Mordelet, F.; Vert, J.P. SIRENE: supervised inference of regulatory networks. *Bioinformatics*,**2008**,*24* (16), i76-82.

[178] Haury, A.C.; Mordelet, F.; Vera-Licona, P.; Vert, J.P. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst. Biol.*, **2012**, *6*, 145.

[179] Wyrick, J.J.; Young, R.A. Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.*, **2002**, *12* (2), 130-136.

[180] Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **2009**, *10* (1), 57-63.

[181] Bar-Joseph, Z.; Gitter, A.; Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, **2012**, *13* (8), 552-564.

[182] Chen, W.; Lin, H.; Feng, P.M.; Ding, C.; Zuo, Y.C.; Chou, K.C. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One*, **2012**, *7* (10), e47843.

[183] Wang, M.; Yang, J.; Xu, Z.J.; Chou, K.C. SLLE for predicting membrane protein types. *J. Theor. Biol.*, **2005**, *232* (1), 7-15.

[184] Feng, K.Y.; Cai, Y.D.; Chou, K.C. Boosting classifier for predicting protein domain structural class. *Biochem. Biophys. Res. Commun.*, **2005**, *334* (1), 213-217.

[185] Feng, P.M.; Chen, W.; Lin, H.; Chou, K.C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.*, **2013**, *442* (1), 118-125.

[186] Chen, W.;Feng, P.M.; Lin, H.; Chou, K.C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.*, *2***013**, *41* (6), e68.

[187] Lin, W.Z.; Fang, J.A.; Xiao, X.; Chou, K.C. iDNA-Prot: identification of DNA binding proteins using random forest with grey model. *PLoS One*, **2011**, *6* (9), e24756.

[188] Xu, Y.; Ding, J.; Wu, L.Y.; Chou, K.C. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One*, **2013**, *8* (2), e55844.

[189] Cai, Y.D.; Chou, K.C. Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics*, **2004**, *20* (7), 1151-1156.

[190] Shen, H.B.; Chou, K.C. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.*, **2009**, *394* (2), 269-274.

[191] Xiao, X.; Wang, P.; Chou, K.C. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. *Mol. Biosyst.*, **2010**, *7* (3), 911-919.

[192] Xiao, X.; Min, J.L.; Wang, P.; Chou, K.C. iCDI-PseFpt: identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *J. Theor. Biol.*, **2013**, *337*, 71-79.

[193] Wu, Z.C.; Xiao, X.; Chou, K.C. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. Biosyst.*, **2011**, *7* (12), 3287-3297.

[194] Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **2001**, *43* (3), 246-255.

[195] Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **2005**, *21* (1), 10-19.

[196] Shen, H.B.; Chou, K.C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.*, **2008**,*373* (2), 386-388.

[197] Qiu, W.R.; Xiao, X.; Chou, K.C. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.*, **2014**, *15* (2), 1746-1766.

[198] Fang, Y.; Guo, Y.; Feng, Y.; Li, M. Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. *Amino Acids*, **2008**, *34* (1), 103-109.

[199] Zhao, X.W.; Li, X.T.; Ma, Z.Q.; Yin, M.H. Identify DNA-binding proteins with optimal Chou's amino acid composition. *Protein Pept. Lett.*, **2012**, *19* (4), 398-405.

[200] Cheng, C.; Yan, K.K.; Hwang, W.; Qian, J.; Bhardwaj, N.; Rozowsky, J.; Lu, Z.J.; Niu, W.; Alves, P.; Kato, M.; Snyder, M.; Gerstein, M. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput. Biol.*, **2011**, *7* (11), e1002190.

[201] Mendenhall, E.M.; Williamson, K.E.; Reyon, D.; Zou, J.Y.; Ram, O.; Joung, J.K.; Bernstein, B.E. Locus-specific editing of histone modifications at endogenous enhancers. *Nat. Biotechnol.*, **2013**, *31* (12), 1133-1136.

[202] Gerstein, M.B.; Kundaje, A.; Hariharan, M.; Landt, S.G.; Yan, K.K.; Cheng, C.; Mu, X.J.; Khurana, E.; Rozowsky, J.; Alexander, R.; Min, R.; Alves, P.; Abyzov, A.; Addleman, N.; Bhardwaj, N.; Boyle, A.P.; Cayting, P.; Charos, A.; Chen, D.Z.; Cheng, Y.; Clarke, D.; Eastman, C.; Euskirchen, G.; Frietze, S.; Fu, Y.; Gertz, J.; Grubert, F.; Harmanci, A.; Jain, P.; Kasowski, M.; Lacroute, P.; Leng, J.; Lian, J.; Monahan, H.; O'Geen, H.; Ouyang, Z.; Partridge, E.C.; Patacsil, D.; Pauli, F.; Raha, D.; Ramirez, L.; Reddy, T.E.; Reed, B.; Shi, M.; Slifer, T.; Wang, J.; Wu, L.; Yang, X.; Yip, K.Y.; Zilberman-Schapira, G.; Batzoglou, S.; Sidow, A.; Farnham, P.J.; Myers, R.M.; Weissman, S.M.; Snyder, M. Architecture of the human regulatory network derived from ENCODE data. *Nature*, **2012**, *489* (7414), 91-100.

Received on: March 07, 2014 **Revised on: September 05, 2014** **Accepted on: September 05, 2014**