Published in final edited form as:
Stat Methods Med Res. 2009 December ; 18(6): 595–617. doi:10.1177/0962280209351890.

Reconstructing transcriptional regulatory networks through genomics data

Ning Sun and **Hongyu Zhao**
Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA

Abstract

One central problem in biology is to understand how gene expression is regulated under different conditions. Microarray gene expression data and other high throughput data have made it possible to dissect transcriptional regulatory networks at the genomics level. Owing to the very large number of genes that need to be studied, the relatively small number of data sets available, the noise in the data and the different natures of the distinct data types, network inference presents great challenges. In this article, we review statistical and computational methods that have been developed in the last decade in response to genomics data for inferring transcriptional regulatory networks.

1 Introduction

The central dogma of molecular biology is DNA→RNA→Protein, where DNA is transcribed to RNA, which is subsequently translated into protein to perform diverse functions. Transcription regulation is a fundamental biological process and much effort has been made to identify important players in this process and to understand how they work with each other to regulate gene expression levels. Before transcription of a specific gene is initiated, its regulatory region is bound by one or more transcription factors (TFs), which recognise their targets through specific sequences, often called binding motifs. After the binding of TFs to the regulatory region of a gene, they recruit the transcriptional machinery to initiate the transcription process. After a gene is transcribed, the product messenger RNA (mRNA) is transported to the cytoplasm of a cell, where it is used as a template for translation. Chromatin structure also plays a key role in transcription regulation and is under many levels of tight regulation.[1,2] After transcription, mRNA decay and silencing are examples of yet further regulation of gene expression.[3] Therefore, the transcript abundance in a cell is the result of joint actions of many regulators. Because of its central role in molecular biology, delineating transcriptional regulatory networks (TRNs) has been a very active research area and much progress has been made in recent years as a result of the availability of genome-wide gene expression data and other high throughput data. In this review, we discuss statistical and computational methods that have been developed to infer TRNs. We will primarily focus on inferring the regulatory targets of different TFs through the analysis of different types of genomics data.

In the simplest case, a TRN can be represented by a $G \times R$ matrix, where G is the number of genes considered and R is the number of TFs involved in regulation. An entry of 1 in the (i, j) position indicates a regulatory relationship between the j-th TF and the i-th gene, and an

Address for correspondence: Ning Sun, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA. ning.sun@yale.edu.

entry 0 indicates no regulation between them. Under this formulation, the goal of a TRN inference is to infer which elements in this matrix are 1. This matrix can be made more general to allow distinction between activation and repression and the joint action from multiple TFs. Furthermore, the strength of regulation may also be included in this matrix, that is, a 0 element indicates regulation relationship, but non-0 elements can take on real values that quantify the regulatory strength of a TF on its target genes.

For the purpose of inferring TRNs, the most relevant type of genomics data are location data including ChIP-chip data using either gene-based microarrays[4,5] or tiling arrays,[6,7] and ChIP-seq data.[8,9] ChIP stands for chromatin immunoprecipitation, which is a protocol to separate the truncated DNA sequences that bind to specific protein from DNA suspensions. The DNA sequence samples are then amplified and measured through microarray or sequencing technique with different genome coverage. Each ChIP-chip or ChIP-seq experiment involves one TF and the results give information about the binding locations in the genome for this specific TF either *in vivo*[4,5] or *in vitro*.[10,11] The binding locations can then be associated with genes according to genome annotation. From the experimental results, for each location in the genome, the data are either summarised as a statistical significance level for binding between a TF and this location, or an intensity level that quantifies the strength of binding. In their analysis of nine TFs involved in the yeast cell cycle, Simon *et al.*[12] showed that the observed binding patterns for these nine TFs can largely explain the expression profiles of their target genes. For example, the binding targets of TFs that are active in the early phase of the cell cycle also tend to have their peak expression levels in the early phase. Therefore, it may be possible to infer TRNs directly from ChIP-chip data. For example, for a given TF, we can select a statistical significance threshold, say 0.001, and view all genes that have an observed statistical significance for binding for this TF less than the threshold as regulatory targets for this TF. It is apparent that such a thresholding rule will lead to both false-positive and false-negative results due to the inherent noises in ChIP-chip data collection. If the threshold is set too stringently, we will miss too many regulatory targets, whereas there will be many false-positive results if the threshold is set too high. The newly developed ChIP-seq has the potential to outperform ChIP-chip to provide more accurate results on TF-DNA binding locations. Additionally, three other concerns make this naïve TRN method less appealing. First, the observed binding information does not imply regulatory relationship between a TF and a gene that it is binding to, even when the results are highly significant. Gao *et al.*[13] estimated that only 58% of the genes whose promoters were bound by a TF are its true regulatory targets. Second, TF binding is a dynamic process, and a TF can have different targets at different time points in a time course experiment and/or under different conditions. If we draw conclusions on the regulatory targets for a TF based on one or a few ChIP-chip experiments, we will miss many true targets and also include many false targets for this TF under other conditions even if the experiments are done perfectly. Third, TRNs involve combinatorial effects of multiple TFs whereas ChIP-chip data only involve one TF per chip, so joint regulation from multiple TFs needs to be inferred indirectly.[14]

In addition to location data, other types of genomics data offer valuable information on TRNs, with the most commonly available and studied type being microarray gene expression data. Compared to the relatively limited amount of ChIP-chip or ChIP-seq data, there is a very large amount of gene expression data in the literature,[15–18] and there are many public repositories for expression data. It is easy to understand how gene expression data can help delineate TRNs. If a set of genes are under the control of the same TFs, they should exhibit similar expression profiles over time or across different conditions. If we can identify a gene cluster based on their similar expression profiles, we may hypothesise that they should share similar regulation patterns and use this rationale to improve TRN inference. Alternatively, if we can either observe or infer the activity levels of a TF across a

set of conditions, we may infer its regulatory targets by selecting those genes showing similar (maybe time delayed) expression profiles. These ideas have been formalised mathematically in the literature and we will discuss these methods below.

Another type of data commonly used in inferring TRNs are DNA sequences. As TFs recognise their targets through specific sequences, genes sharing common sequences in their regulatory regions are more likely to be under similar regulation. This logic has been extensively used to infer TF binding motifs. On the other hand, if the binding motif for a TF is known, a gene whose regulatory region contains one or more instance of this motif is more likely to be the regulatory target of this TF. Putative regulatory targets only identified through motif matching are known to lead to many false-positive and false-negative results as many genes whose regulatory regions contain the motifs are not the regulatory targets and many regulatory targets do not have known binding motifs in their regulatory regions. This lack of sensitivity and specificity through sequence and motif analysis alone is the result of many unknown factors and mechanisms for gene expression regulation. Nevertheless, sequence information does provide valuable information complementary to ChIP-chip and expression data for TRN inference.

Apart from the above three data types commonly used in TRN inference, other high throughput data also offer useful information, such as mRNA decay data[19] and chromatin modification data.[2] In principle, joint analysis of all data types under a single unified framework should be the most informative way for TRN inference. In this article, we review various approaches that have been developed for reconstructing TRNs based on one or more data types discussed above. We will focus on the basic modelling assumption and main ideas in inferential strategies without going to detailed implementations and parameter inference. The technical details can be found in the original articles. Because TRN inference research is highly inter-disciplinary and fast evolving, we only provide a partial review of what are available with a focus on those methods with a strong statistical flavour.

2 Methods based only on gene expression data

Since microarray technology was first developed for expression analysis, gene expression microarrays were the only genomic array type available before other platforms, e.g. ChIP-chip data. Even today, most genomics data are collected from gene expression microarrays. Therefore, considerable efforts were made to deduce genetic networks from gene expression data alone. In this review, we begin with the methods that use only gene expression data for network inference. In this context, most methods consider general relationships among genes, without specific reference to TFs. Therefore, the inferred network should be more appropriately described as gene regulatory networks (GRNs).[20]

For gene expression data, there are broadly two types of experiments. In the first type, gene expression profiles are followed across a number of time points. For the second type, a number of perturbations or a number of individual samples are studied but only one observation is made for each perturbation/individual. Some published methods can be applied to both types of gene expression data, whereas specific methods have been developed to deal with only one type, e.g. time course experiments. In our following discussion, we call these types time course and steady state data.

2.1 Relevance networks

One intuitive method for network reconstruction is to examine the degree of association of expression profiles between each pair of genes. A strong association suggests either two genes are under similar regulatory control or one gene is involved in the regulation of the other gene. A network can thus be built based on pairwise dependencies,[21] and this network

is sometimes called relevance network (RN). Practical issues involve (1) the choice of association measures, which can be based on standard Pearson correlation coefficient, its various transformations, or more robust mutual information measures, and (2) the labelling of the network, whether to keep all the edges or only focus on those passing a specific threshold. In this context, a sub-network where all gene pairs have high association values is sometimes called a module,[22] and it has been found that these modules studied as a unit can lead to interesting biological findings. One drawback of this approach is that pairwise dependency may be due to either direct regulatory relationship or indirect relationship through other genes. Therefore, if the primary objective is to identify genes that directly interact with each other, the networks constructed purely from pairwise analysis may contain many false-positive results. One simple way to reduce the number of false positives is to study association between two genes in the presence of one or more other genes. The basic idea is that if both genes are regulated through a third gene, then conditional on the third gene, the first two genes will no longer be associated. This idea can be realised in different forms. ARACNE developed by Margolin *et al.*[23] uses the data processing inequality

$$I(g_1, g_3) \leq \min[I(g_1, g_2), I(g_2, g_3)],$$

where $I(g_1, g_3)$ is the mutual information between g_1 and g_3 to remove dependency due to other genes such as g_2. A similar idea using conditional correlation coefficients was used by Rice *et al.*[24] and Wille and Buhlmann.[25]

2.2 Gaussian graphical models

If we consider all the genes simultaneously instead of two or three at a time, we may model the joint distribution as a multivariate normal with its mean and covariance matrix. Although the multivariate normality assumption may be violated for real data, it does provide a compromise between real data complexity and statistical and computational feasibility. In this formulation called Gaussian graphical models (GGMs), the associations between the genes are described by a covariance matrix $\Sigma = \{\sigma_{ij}\}$, where σ_{ij} is the covariance between genes i and j. One attractive property of GGMs is that it is straightforward to calculate the conditional correlation coefficient between genes i and j conditional on all other genes. This conditional association is defined as partial correlation coefficient $\rho_{ij} = -\sigma^{ij}/(\sigma^{i}\sigma^{jj})^{1/2}$, where σ^{ij} is the (i, j)-th element in Σ^{-1}, the inverse of the covariance matrix Σ. Σ^{-1} is often called the precision or concentration matrix. Two genes in a GGM are connected by an edge if ρ_{ij} is not equal to 0, and the edges are undirected. This is in contrast to RNs, which can be built from the Σ matrix. Therefore, if we have access to the true covariance matrix describing pairwise associations between gene pairs, the GGM formulation will allow us to distinguish gene pairs that directly interact from those that indirectly interact through other genes. However, the matrix Σ is unknown and has to be estimated from the observed data. In typical microarray studies, G is usually in the order of thousands whereas the number of experiments is usually much smaller. Therefore, the rank of Σ is much smaller than its dimension, and this matrix is singular. Early on, Toh and Horimoto[26] proposed to reduce the dimensionality of the matrix Σ by first clustering genes into clusters and then only evaluating the relationships among gene clusters. It is apparent that there will be significant information loss in this process and we can no longer refer to individual genes, which are the primary targets of network inference.

Under a GGM, we can consider one gene at a time and model the dependency between this gene and other genes through regression, where the dependent variable is the expression level of a specific gene and the independent variables are the expression levels of other genes. More specifically, we have the following regression model:

$$m_i = \sum_{j=1, j\neq i}^{G} \beta_{ij} m_j + \varepsilon_i, \quad (1)$$

where m_i is the expression level for the i-th gene, β_{ij} is the regression coefficient of the j-th gene on the i-th gene and ε_i is the noise term. This regression model is over-parameterised in our context as the number of predictors, i.e. genes, is generally much larger than the number of observations. However, because a biological network is usually sparse, that is, the number of regulators for a gene is limited, this sparsity assumption can be used for statistical inference.[27–34] For example, Meinshausen and Buhlmann[29] proposed to consider one gene at a time and perform regression through regularised regression methods such as LASSO.[35] The objective function has the following form:

$$\hat{\beta}_i^{\lambda} = \arg\min_{\beta_i} \left\{ \left\| m_i - \sum_{j=1, j\neq i}^{G} \beta_{ij} m_j \right\|_2^2 + \lambda \sum_{j=1, j\neq i}^{G} |\beta_{ij}| \right\},$$

where $\hat{\beta}_i^{\lambda}$ is the regression coefficient vector for gene i, and λ is the penalty term that balances between model fit and model complexity. This procedure often leads to a model with a few non-zero regression coefficients. The interpretation of these parameters is that those genes with non-zero regression coefficients have an impact on the i-th gene. This regression analysis can be repeated for every single gene treated as the dependent variable to construct the network. Because each gene is fitted separately, the resulting network structure may not be symmetric, e.g. it may happen that $\beta_{ij} = 0$ but $\beta_{ji} \neq 0$, leading to interpretation problems. The sparsity regression can also be accomplished through specifying sparse priors within the Bayesian setting.[27,31] Other methods based on the regression setting include approaches from iterative greedy algorithms and combinatorial optimisation algorithms.[36,37] Gardner *et al.*[38] addressed the sparsity problem by considering a fixed number of TFs at a time and then selecting the subset with the best fit to the experimental data.

Amore direct way of imposing sparsity constraint on the whole network was proposed by Peng *et al.*[30] Their objective function is:

$$\frac{1}{2}\left(\sum_{i=1}^{G} w_i \left\| m_i - \sum_{j=1, j\neq i}^{G} \rho_{ij} \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}} m_j \right\|^2 \right) + \lambda \sum_{1\leq i<j\leq G} |\rho_{ij}|,$$

where $\beta_{ij} = \rho_{ij}(\sigma^{jj}/\sigma^{ii})^{1/2}$ is the regression coefficient, w_i is the weight for gene i, which may be related to the variance of the noise for the i-th gene, and λ is again the penalty term that balances between model fit and model complexity across all the genes. The resulting matrix should retain symmetry under this approach. The authors argued that the general sparsity constraint on the whole network leaves room for specific genes to have many connections, an attractive feature to accommodate the presence of so-called hub genes that have many interactions. Similarly, a threshold gradient descent method was proposed by Li and Gui[28] to estimate the precision matrix.

Yeung *et al.*[39] proposed to use singular value decomposition (SVD) in this context for TRN inference. In the case that external perturbations are used, they wrote their system in the following matrix form:

$$\Delta \boldsymbol{M}_{G\times K} = \boldsymbol{B}_{G\times G} \boldsymbol{M}_{G\times K} + \boldsymbol{E}_{G\times K} + \boldsymbol{\varepsilon}_{G\times K},$$

where $\Delta \boldsymbol{M}_{G\times K}$ is the change of expression levels for all the genes, represented by the G rows, after each of the K perturbations represented by the columns, $\boldsymbol{M}_{G\times K}$ is the collection of all the expression levels, $\boldsymbol{B}_{G\times G}$ is the regression coefficient matrix, $\boldsymbol{E}_{G\times K}$ summarises the external perturbations and $\boldsymbol{\varepsilon}_{G\times K}$ is the noise matrix. Standard regression does not apply here because more genes than conditions are studied. With SVD, $\boldsymbol{M}^T$ can be written as $\boldsymbol{M}^T_{K\times G}=\boldsymbol{U}_{K\times G}\boldsymbol{D}_{G\times G}\boldsymbol{V}^T_{G\times G}$, where $\boldsymbol{U}_{K\times G}$ and $\boldsymbol{V}_{G\times G}$ are orthogonal and $\boldsymbol{D}_{G\times G}$ is a diagonal matrix. The pseudo-inverse of $\boldsymbol{M}^T$ can be calculated as $\boldsymbol{V}_{G\times G}\boldsymbol{D}^{-1}_{G\times G}\boldsymbol{U}^T_{G\times K}$. The family of solutions for $\boldsymbol{B}_{G\times G}$ has the form of $(\Delta \boldsymbol{M}_{G\times K}-\boldsymbol{E}_{G\times K})\boldsymbol{U}_{K\times G}\boldsymbol{D}^{-1}_{G\times G}\boldsymbol{V}^T_{G\times G}+\boldsymbol{C}_{G\times G}\boldsymbol{V}^T_{G\times G}$, where $\boldsymbol{C}_{G\times G}$ is an arbitrary $G\times G$ matrix. Yeung *et al.*[34] proposed to select among these solutions the sparsest one as the inferred network.

2.3 Bayesian networks

One extensively studied class of models in causal inference is Bayesian networks (BNs) and it is not surprising that they were proposed very early on to infer gene networks based on gene expression data.[40,41] Within a BN, the relationships among the genes are described by a directed acyclic graph (DAG), where the state of a gene is jointly determined by its parental nodes in a probabilistic manner. For a given DAG, the joint distribution of all the genes can be factored into a product of conditional probabilities. The balance between model fit and model complexity is integrated into the learning for the structure of the underlying DAG for a gene network. BN analysis itself is a very broad and active research field.[42] Because most published articles simply adopted existing methods to analyse gene expression data, we will not attempt to review the general methods related to BNs here, but note a number of limitations in the context of gene network inference. First, different DAGs may give rise to the same likelihood for the observed data, leading to equivalent classes. Therefore, it is impossible to distinguish these DAGs purely from the observed data. Second, different perturbation experiments are treated exchangeably in the analysis although the underlying biological states are clearly different. Third, BNs do not allow feedback loops, a common phenomenon in biological networks. Dynamic Bayesian networks (DBNs), which can alleviate this problem to some extent, have been proposed.[43,44] However, a time homogeneous transition model is normally assumed in DBNs, an oversimplification of the dynamics of the biological process. Fourth, continuous observations need to be discretised for BN analysis. It is often a non-trivial task to decide on the number of levels and the thresholds used for discretisation. Finally, although methods have been developed to incorporate prior information in BN analysis,[45] it is difficult to incorporate some aspects of biological knowledge explicitly, e.g. the kinetic models relating transcription rates with TF activities to be discussed below. The performance of BNs and DBNs was evaluated through simulations.[46] It was found that some local structures of the networks may be recovered and the author also showed the importance of incorporating prior biological knowledge and having the systems studied at appropriate time points, preferentially right after the system is perturbed.

The relative performance of RNs, GGMs and BNs has been compared in the literature.[23,47,48] Both BNs and GGMs had better performance than RNs. It was found that structural perturbations are more informative than dynamic perturbations, e.g. time course experiments. In addition, when a small system was considered, BNs outperformed GGMs.

2.4 Time course microarray gene expression data

The methods discussed above can be applied to both perturbation data and time course data. However, the above methods treat each time point separately, so the time dependencies are not utilised. This may potentially lead to loss of information for TRN inference. Many statistical methods have been developed to analyse time course gene expression data, and

here we focus on those whose primary goal is to infer regulatory relationships among the genes.

Similar to the regression setting discussed above, the following model can be used to describe the dynamics of a gene:

$$\frac{m_i(t+\Delta t) - m_i(t)}{\Delta t} = \sum_{j=1, j\neq i}^{G} \beta_{ij} m_j(t) - d_i m_i(t) + \varepsilon_i(t+\Delta t), \quad (2)$$

where the dependent variable is the expression level change between two observation points instead of the expression level itself, and there is an additional term d_i that characterises the decay rate of the i-th gene. If we ignore the decay rate in this model, the methods discussed in the previous section can be applied here for inference. To take advantage of the fact that expression levels at nearby points tend to be similar, the TSNI algorithm developed by Bansal *et al.*[49] uses splines to smooth the observed time course data first and then apply SVD to the gene expression matrix to reduce the dimensionality of the predictors.

In contrast to linear models analysed in most published articles, Wang[50] considered the following non-linear regression model:

$$\frac{m_i(t+\Delta t) - m_i(t)}{\Delta t} = \left[\alpha_i \prod_j m_j^{w_{ji}}(t) \prod_k m_k^{w_{ki}}(t-\Delta t) \prod_l m_l^{w_{li}}(t-2\Delta t) - d_i m_i(t) \right] + \varepsilon_i(t+\Delta t),$$

where the joint effect from multiple genes are modelled through a multiplicative function and time-delayed responses are allowed by incorporating gene expression levels at previous time points, $t - \Delta t$ and $t - 2\Delta t$. The exponents in the equation, w_{ji}, w_{ki} and w_{li}, quantify the effects of the corresponding genes, and d_i is the mRNA decay rate. This model can be considered a variant of the S-system.[51,52] It is apparent that the model fitting can be challenging for this setup. Indeed, the author only considered a limited set of genes when this model was applied and used the genetic algorithm coupled with Bayesian information criterion (BIC) for model selection.

When data from samples in steady state and time course experiments are both available, Bonneau *et al.*[53] proposed to use a single regression model of equation form (1) to analyse them together. Here is a brief justification. Consider the time course dynamic model of equation form (2). At the steady date, there is no change of expression levels, and we have

$$0 = \sum_{j=1, j\neq i}^{G} \beta_{ij} m_j - d_i m_i.$$

So the regression model can be written as

$$m_i = \sum_{j=1, j\neq i}^{G} \beta_{ij} m_j / d_i,$$

which has similar form to the dynamic model of equation form (2). Bonneau *et al.* also considered alternative forms of the regulation function that may incorporate interactions, e.g. adding another term $\min(m_i, m_j)$, in the model.

Only gene expression data are used for network inference in the methods discussed so far. In this setting, there is an implicit assumption that the observed expression level for a gene can

be used as a surrogate for the TF activity level of this gene so that it can be used as a predictor in regression models. However, it is well known that there is poor correlation between gene expression levels and protein activity levels due to many steps involved from transcription to translation, and then to post-translational modification. Although the above methods have the advantage that they are generic, easily understood, and adapted from established methods, it would seem intuitive that more can be gained by appropriately modelling the underlying biological process and incorporating prior knowledge and different data types in TRN inference.

3 Methods based on combined data analysis

3.1 Clustering-based analysis

Clustering analysis is the most commonly used method for visualising gene expression patterns.[54] Many statistical methods have been proposed to cluster genes. Sets of genes with highly similar expression profiles are often called modules.[22] Genes in the same cluster tend to have similar biological functions, so these clusters can be used to infer the functional roles of un-annotated genes. As TFs are a major regulator of transcription, genes in the same cluster or module are more likely to be under the control of the same or similar TFs. Because TFs recognise their targets through sequence specific binding motifs, the regulatory regions of genes in the same cluster should be enriched for some binding motifs. The discovery of these binding motifs can facilitate the inference of relevant TFs if the binding motifs of the TFs are known from prior experiments. In fact, this was done early on in microarray data analysis, e.g. Spellman *et al.*[18] Since a set of genes may be co-regulated only under a subset of the experimental conditions, a number of methods have been developed to cluster genes under specific conditions,[55–57] and the gene clusters thus identified are sometimes called conditional specific expression modules. The methodologies developed in this context can also be extended to combine data from diverse data sources.[58]

For a given gene cluster or module, it is natural to relate it to the activities of a set of TFs. In the absence of information for TF activities, the corresponding gene expression levels may be used as surrogates. Segal *et al.*[59] developed an iterative procedure for combining gene expression data with regulator activities (estimated by gene expression levels) for both improved cluster analysis and the interpretation of how each cluster is regulated. The basic idea is to iterate between two steps. The first step involves inferring key regulators for each cluster by correlating regulator levels with expression patterns of genes in the cluster across a large number of experimental conditions. In the second step, regulator activities and gene expression data are jointly used to more accurately assign each gene to its corresponding cluster. However, this procedure is limited by the pre-specified number of clusters and the possibly poor correspondence between gene expression levels and regulator activities. A similar approach to incorporate gene expression data and sequence data was introduced by the same authors.[60] In the context of biclustering, Reiss *et al.*[61] developed an iterative algorithm to incorporate sequence information into building biclusters.

Bar-Joseph *et al.*[62] proposed a procedure called GRAM for TRN inference by combining gene expression data and ChIP-chip data. In the first step, they used a stringent criterion to infer the binding targets only for those TF-gene pairs that are highly statistically significant, e.g. p-value less than 0.001. Then, gene expression data are used to define a core expression profile for a set of genes sharing a common set of TFs as their regulators. After the core expression profiles are defined, other genes are included in a transcription module if their expression profile is similar to the core profile and there is some marginal evidence of binding. A conceptually similar approach was proposed by Lemmens *et al.*[63] that also incorporates motif data in the analysis.

Model-based clustering methods have also been proposed to combine gene expression data with other information. A two-component mixture model was proposed by Wang *et al.*[64] to infer true regulatory sites from combined analysis of gene expression data with either sequence data or ChIP-chip data. A sparse regression mixture model was developed by Li *et al.*[65] where the mean expression profile in each cluster is determined by the additive effects of a set of regulators. Liu *et al.*[66] proposed an infinite mixture model to jointly analyse gene expression data and ChIP-chip data. In this work, the expression information for a specific gene i is represented by a vector $\mathbf{m}_i = (m_{i1}, \ldots, m_{iK})$ across a set of K experiments. The ChIP-chip binding data between R TFs and this gene is summarised by $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iR})$, where γ_{ij} indicates statistical evidence for binding from ChIP-chip data. A gene in each cluster is assumed to have a multivariate normal component describing expression levels and a binary vector component modelling binding patterns.

Compared to the methods that rely only on gene expression data for network inference, these cluster-based methods aim to integrate other information, gene annotations, sequence information, or ChIP-chip data, to both make more accurate gene groupings and interpret each cluster as a result of the joint action of a set of regulators. Therefore, not only can we learn which genes belong to the same cluster, we can also explain the regulation mechanism that makes these genes in the same cluster. One major drawback of these methods is that gene expression data are often used as surrogates for TF activities, which may lead to poor performance due to the lack of correspondence between them. In the following, we will review kinetic models relating transcription levels/rates with TF activities, and discuss various methods that have been proposed to infer TF activities from gene expression data.

3.2 Kinetic models for mRNA synthesis

We first consider the Michaelis–Menten kinetics model, which has been used[67] to relate transcription (mRNA synthesis) rate s_i for the i-th gene with a single TF activity r in the following form:

$$s_i = b_i \frac{r}{c_i + r} + a_i, \quad (3)$$

where b_i is the maximum transcription rate, c_i is the half-saturation constant and a_i is the basal transcription rate. The parameters used in this formulation, a, b and c, are TF-gene specific. The value of $1/c$, denoted by γ, can also be interpreted as the ratio of association constant (k_a) and dissociation constant (k_d) between TF and DNA. Let PD denote the concentration of TF-DNA complexes and D denote the concentration of non-occupied sites. DNA occupancy is defined as $O = PD/(PD + D)$. At equilibrium, $k_a/k_d = PD/(D \times P)$, where P is the free TF concentration. Therefore, $O = 1/[1 + (k_a/k_d)/P] = P/[k_a/k_d + P]$. As O is the proportion of the sites occupied by the TF-DNA complexes, it is reasonable to assume that the transcription rate is a linear function of the occupancy, leading to a functional form of Equation (3).[68] The value of O may be estimated from high-throughput TF-DNA interaction data[4,5] or from sequence data[68] based on the binding motif of the TF. To allow for the saturation effect, we can use the following model:

$$s_i = b_i \frac{e^r}{c_i + e^r} + a_i.$$

When there are two TFs, under different assumptions on the cooperativity of the two factors, e.g. independent binding, Nachman *et al.*[69] derived the binding pattern probabilities under the equilibrium state. Then the overall transcription rate can be calculated as the weighted average of these four probabilities, where the weights correspond to the transcription rate at

a specific binding pattern. More general forms are given in Nachman *et al.*[69] Consider the scenario that transcription can only start when a set of R regulators are all present in the promoter region of a gene,

$$s_i=\alpha\frac{\prod_{j=1}^{R}(\frac{r_j}{c_{ij}})^{\beta_{ij}}}{1+\prod_{j=1}^{R}(\frac{r_j}{c_{ij}})^{\beta_{ij}}},$$

where c_{ij} is an association constant for binding between gene i and TF j and β_{ij} measures the cooperativity in binding. Pan *et al.*[70] modelled binding as a function of sequences.

Chen *et al.*[71] considered a simpler additive model of the following form:

$$s_i=c_0+\sum_{j=1}^{R}c_{ij}\frac{e^{r_j}}{1+e^{r_j}}.$$

A similar functional form was used by Chen *et al.*[72] who also included a time delay term from gene expression level to activity level.

Based on equilibrium assumption, Sun *et al.*[73] considered a similar model of the form

$$s_i=\alpha\prod_{j=1}^{R}r_j^{\beta_{ij}},$$

where β_{ij} is the binding affinity between gene i and TF j which can be approximated by the binding intensity measured from ChIP-chip data.

When different TFs can either activate or repress a gene, Yeung *et al.*[39] considered the following model:

$$s_i=\frac{a_i+\sum_{j\in A_i}r_j^{\beta_{ij}}}{1+\sum_{j\in A_i}r_j^{\beta_{ij}}+\sum_{k\in R_i}r_k^{\gamma_{ik}}},$$

where β_{ij} is the activation cooperativity of the j-th TF on the i-th gene, and γ_{ik} is the repression cooperativity of the k-th TF on the i-th gene. When there are no repressors, this model is very similar to the additive model of Chen *et al.*[71]

Porreca *et al.*[74] considered piecewise linear models as a compromise between linear models, which are easy to handle computationally and statistically but may only provide an approximation to real models, and non-linear models, which are more relevant biologically but present significant computationally challenges.

In summary, a variety of kinetics models have been considered in the literature, some based on physical-chemical principles whereas others purely from statistical convenience. A more detailed review on various kinetics models for the synthesis rate of the i-th gene, s_i, can be found in Ben-Tabou de-Leon and Davidson.[75]

The kinetics of mRNA degradation is usually modelled as a first-order differential equation. The combination of mRNA synthesis and degradation depicts the rate of transcription level of a given gene (or $\mathrm{d}m_i/\mathrm{d}t$).

3.3 Regulation models

For a given kinetics model, we have the following differential equation describing the dynamics of the gene expression level for the i-th gene:

$$\frac{\mathrm{d}m_i}{\mathrm{d}t} = s_i - d_i m_i + \varepsilon_i,$$

where m_i is the mRNA level of the i-th gene, s_i is the transcriptional synthesis rate, d_i is the mRNA decay rate and ε_i is the noise associated with the system and measurements. This model can be used to describe time course gene expression patterns. If the cells have reached an equilibrium state, we have $s_i - d_i m_i = 0$. Hence the transcript abundance at the equilibrium state m_i is proportional to s_i/d_i. Therefore, we can relate the expression level with TF activities through the kinetics models discussed in the previous section. When a TRN is known, the above models can be used to both estimate TF activities and kinetics parameters[67,69,76,77] from the observed expression data. For example, when the regulatory targets are assumed to be known for a TF, the Michaelis-Menten kinetics model can be used to infer TF activities as well as the kinetic parameters either through maximum likelihood[67] or Bayesian approach.[76] However, care needs to be taken because there may be identifiability problems for some systems.[78]

Because linear models are most commonly used in relating expression levels with TF binding and TF activities, we will focus on linear statistical approaches in the following discussion. In this case, the following model has been used by a number of groups.[13,73,79,80]:

$$\log(m_i) = \sum_{r=1}^{R} a_{ij} \log(r_j) + \varepsilon_i,$$

where m_i is either the absolute or relative observed expression level for the i-th gene, a_{ij} represents the regulatory strength between the j-th TF and the i-th gene (which can be measured from location data) and r_j represents the activity of the j-th TF. When the ChIP-chip data are used as an estimate for the a_{ij}, e.g. the binding intensity is used as a surrogate, the estimates of the TF activities reduce to a regular regression problem.[13,80] Cokus *et al.*[81] studied the dynamics of the inferred TF activities from linear regression models. For time course data, Wang *et al.*[82] proposed to borrow information across time points to have improved TF activity estimates. This is achieved through the following model:

$$m_i(t) = \mu(t) + \sum_{j=1}^{G} r_j(t) x_{ij} + \varepsilon_i(t),$$

where $r_j(t)$ is modelled by natural cubic B-splines, and x_{ij} is the estimated regulation strength between TF j and gene i. These studies found that the inferred TF activities can be substantially different from the expression levels of the same gene, leading to concerns on potential biases in using expression levels as surrogates for TF activities in many of the published studies.

Because physical binding does not imply regulation,[13] TRN inference can be improved by combining ChIP-chip data with gene expression data. Sun *et al.*[73] proposed BEAM, a Bayesian method for integrating ChIP-chip data and expression data, based on the following model:

$$\log\left(m_i/m_i^0\right)=\sum_{r=1}^{R}a_{ij}\log(r_j/r_j^0)+\varepsilon_i, \quad (4)$$

where the log-ratio of the expression levels (m_i and m_i^0) between two conditions is related to the additive effects from a set of R TFs through the log-ratio of the TF activities (r_j and r_j^0) and the regulatory strength a_{ij}. Instead of directly using the binding intensities from ChIP-chip data,[13,80] Sun *et al.* assumed that $a_{ij} = b_{ij}c_{ij}$, where c_{ij} is the unknown but desired regulatory relationship between TF j and gene i. With this formulation, model (4) becomes

$$\log\left(m_i/m_i^0\right)=\sum_{r=1}^{R}b_{ij}c_{ij}\log(r_j/r_j^0)+\varepsilon_i.$$

Under this set-up, gene expression levels m_i and binding intensities b_{ij} are observed, and there is some partial knowledge on c_{ij}. The objective is to infer the true regulatory matrix c_{ij}. For a set of K experiments, the above model can be written in the following matrix form (ignoring the noise terms):

$$\begin{pmatrix} \log\left(\frac{m_1^1}{m_1^0}\right) & \cdots & \log\left(\frac{m_1^K}{m_1^0}\right) \\ \vdots & \vdots & \vdots \\ \log\left(\frac{m_G^1}{m_G^0}\right) & \cdots & \log\left(\frac{m_G^K}{m_G^0}\right) \end{pmatrix}_{G\times K} = \begin{pmatrix} b_{11}c_{11} & \cdots & b_{1R}c_{1R} \\ \vdots & \vdots & \vdots \\ b_{G1}c_{G1} & \cdots & b_{GR}c_{GR} \end{pmatrix}_{G\times R} \begin{pmatrix} \log\left(\frac{r_1^1}{r_1^0}\right) & \cdots & \log\left(\frac{r_1^K}{r_1^0}\right) \\ \vdots & \vdots & \vdots \\ \log\left(\frac{r_R^1}{r_R^0}\right) & \cdots & \log\left(\frac{r_R^K}{r_R^0}\right) \end{pmatrix}_{R\times K}$$

$$\boldsymbol{M}_{G\times K}=\boldsymbol{A}_{G\times R}\Gamma_{R\times K} \quad (5)$$

In this form, the same reference sample is used and the same regulatory matrix $\boldsymbol{A}_{G\times R}$ is assumed across all K experiments. The K columns in $\Gamma_{R\times K}$ summarise the activities of the R TFs in the K experiments.

BEAM assumes that some information is available on the TFs involved in regulation and the availability of ChIP-chip data to provide useful information on the general structure of the TRN. In addition to ChIP-chip data, sequence data can be used to estimate regulatory potential between a TF and a gene. Xing and van der Laan[83] used regression analysis to identify active TFs based on expression data and motifs from known TFs. However, they did not estimate TF activities or regulatory strengths. With sequence information as input for predicting TF binding, an approach similar to BEAM was proposed by Sabatti and James[84] to infer TRNs.

It is apparent that, if only gene expression data are available, it is generally impossible to factor the observed expression matrix $\boldsymbol{M}_{G\times K}$ in the form of Equation (5) without imposing any constraints on $\boldsymbol{A}_{G\times R}$ and $\Gamma_{R\times K}$. Liao *et al.*[79] discussed the constraints on the structure of $\boldsymbol{A}_{G\times R}$, essentially the distribution of the 0 elements in this matrix, needed under which the decomposition is unique and called the analysis as Network Component Analysis (NCA). Compared to Bayesian approaches,[73,84] NCA requires prior knowledge on the connectivity between TFs and genes but does not need inputs on regulatory strengths or TF activities. To accommodate uncertainties in the connectivities, Yu and Li[85] proposed a two-stage constrained space factor analysis where the constraints are derived from ChIP-chip or other data and the algorithm iterates between network configuration estimation and regulation strength estimation. When the number of TFs is large with possible interactions among

them, Boulesteix and Strimmer[86] proposed to use partial least squares in this regression setting to reduce the dimensionality of the predictor space, i.e. the space spanned by the TFs.

Despite somewhat different motivations, these methods all share the same general modelling form, involving three sets of variables: the gene expression levels, the regulation strengths between TFs and their target genes, and the TF activities. Statistical estimation of TF activities is part of the inference procedure. Pournara and Wernisch[87] compared the performance of some methods in this context and concluded that most of the tested algorithms are successful in reconstructing the connectivity structure as well as the TF profile.

Tanay and Shamir[88] considered a different functional form, called a dose-affinity-response function, that relates the response (transcription rate) with the TF doses (activities) and the affinities of TFs to the regulatory region of a gene (regulatory strength). The affinity information can be gathered from the ChIP-chip data and the doses are iteratively estimated from motif information. It appears that only single TFs can be analysed under this approach for practical data and all the variables need to be discretised.

3.4 State-space models

We discussed time course data methods above where only gene expression data are used for inferring relationships among genes. If we consider each gene being regulated by a set of TFs whose activities are unknown, this naturally leads to classical state space models (SSMs).[90] In an SSM, the gene expression levels, $\boldsymbol{m}_t = (m_1(t), m_2(t), \ldots, m_G(t))$, are assumed to be generated from a set of R hidden state variables, $\boldsymbol{h}_t = (h_1(t), h_2(t), \ldots, h_R(t))$, and that the $\boldsymbol{h}_t$ follow the first-order Markov process as follows:

$$\boldsymbol{h}_t = \boldsymbol{T}\boldsymbol{h}_{t-1} + \boldsymbol{w}_t, \boldsymbol{w}_t \sim N(0, \Sigma_w)\ \boldsymbol{m}_t = \boldsymbol{A}\boldsymbol{h}_t + \boldsymbol{v}_t, \boldsymbol{v}_t \sim N(0, \Sigma_v)$$

where $\boldsymbol{T}$ is the state dynamics matrix, $\boldsymbol{A}$ is the observation matrix and Σ_W and Σ_V are the covariance matrices for errors associated with the state and observed variables $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$, respectively. This formulation fits naturally to our understanding of transcription regulation because the hidden variables can be interpreted as TFs, the $\boldsymbol{A}$ matrix defines how TFs regulate gene expression levels, and the $\boldsymbol{T}$ matrix defines how TFs regulate among each other over time. The second component of this SSM has the same form as Equation (5), with the difference being that the first component is used to describe the dynamics of the hidden state variables. Because $\boldsymbol{T}$ is time independent, this model implicitly assumes a time invariant regulation pattern among the TFs. If this assumption largely holds, SSMs may allow us to borrow information across time points to make more efficient TRN inference. However, the downside of this model is that if this time homogeneity assumption is seriously violated, the TRN inference can be biased. In this case, we may use non-parametric methods to model TF activities across time,[82] but statistical methods are lacking for model inference in this more general case. A more general model form (called an input driven model[89]) has also been considered in the literature:

$$\boldsymbol{h}_t = \boldsymbol{T}\boldsymbol{h}_{t-1} + \boldsymbol{E}\boldsymbol{m}_{t-1} + \boldsymbol{w}_t, \boldsymbol{w}_t \sim N(0, \Sigma_w),\ \ \boldsymbol{m}_t = \boldsymbol{A}\boldsymbol{h}_t + \boldsymbol{F}\boldsymbol{m}_{t-1} + \boldsymbol{v}_t, \boldsymbol{v}_t \sim N(0, \Sigma_v),$$

where gene expression data are allowed to affect the hidden state variables. As SSMs have been extensively studied in the literature, the model inference can be estimated using established maximum likelihood methods or Bayesian methods. For example, the methods that can be applied to estimate the number of state variables include BIC,[90] cross validation[91] and Bayesian approaches.[92]

In the specific context of TRN inference, the hidden variables have clear biological interpretation as TFs and possibly other regulators. Therefore, with ChIP-chip and sequence data, the structure of the regulation network reflected in $\boldsymbol{A}$ and possibly the $\boldsymbol{T}$ matrix may be partially known. Assuming knowledge on the structure, Xiong and Choe[93] proposed a constrained SSM approach for TRN inference. Similarly, when the connectivity matrix is assumed to be known, Sanguinetti *et al.*[94] used an SSM to model the dynamics of the observed expression data. In this case, the sparsity constraint is directly provided by the connectivity matrix. With prior knowledge on the TFs involved in regulation, SSMs have also been used to infer TF activities.[95]

3.5 mRNA decay data

Although mRNA decay rate, denoted by d_i, in the above discussion is a key component in the kinetics and regulation model, it is not used in most statistical methods. This is partly due to the limited information on the decay rate at the genome level[19] that can be used in the analysis. In addition, at the steady state, the decay rate appears as a factor in the model for the mean gene expression level, so ignoring it does not change the general model form. However, in the analysis of time course data, the decay rate does play an important role and it is conceivable that more accurate TRN inference may be achieved if each gene's approximate decay rate can be incorporated in statistical analysis.

In the context of time course gene expression data analysis, Chen and Zhao[96] used gene expression data and mRNA decay date to estimate transcription rate through simple differential equations. They found that the estimated transcription rates may be more informative on co-regulation and DNA binding motif discoveries. In TRN inference, mRNA decay rates were used by Nachman *et al.*[69] to estimate transcription rate. One caveat in using experimentally derived decay rates is that their values are also context dependent,[97] so it may not add to TRN inference if the expression data and decay data were observed under very different experimental conditions.

4 Discussion

We have reviewed many statistical and computational methods that have been developed in recent years to reconstruct transcriptional regulatory networks from gene expression data, ChIP-chip data, sequence data, and mRNA decay data. Although we have primarily focused on TF-gene interactions in this review, there are many other important regulators involved in chromatin modification and post-transcription regulation. It would be ideal to consider them all under a consistent framework to reconstruct TRNs. For example, Whitington *et al.*[98] found that chromatin information offers accurate prediction of tissue-specific binding sites and such data will undoubtedly lead to more accurate TRN inference.

Because the focus of this current review is the inference of TF targets, a thorough review of the extensive literature on inferring TF binding motifs is beyond the scope of this article. We simply note that a large number of methods have been developed to infer TF binding motifs based on gene expression data, sequence data and ChIP-chip data. Some of these motif-finding algorithms are closely tied to TRN inference. For example, REDUCE[99] regresses gene expression levels on the presence or absence of a putative motif in the promoter regions of all the genes. A functional motif can be identified if there is an association between the observed expression levels and the presence or absence of this motif. In essence, this motif occurrence serves as a surrogate for TF occupancy. The same idea was used to infer binding motifs from ChIP-chip data[100] and combined analysis of gene expression and ChIP-chip data for motif discovery.[101] More accurate motif knowledge can provide more informative prior as well as more effective filtering for TRN inference.

Most TRN work relies on expression studies conducted under controlled experiments. However, the combination of naturally occurring variations observed in individuals in a population sample does represent a rich set of perturbation experiments, though not as well controlled. Such samples have been used to study gene expression profiles across individuals and to identify genetic regions controlling gene expression patterns.[102–104] These studies also yield valuable information on TRNs. However, early results did not show an enrichment for TFs in the identified regions.[105] More recent studies do suggest that more informative regulation results can be derived if different data types are systematically integrated, including binding and protein interaction information.[106]

In addition to collecting more data types, more samples under different conditions are required to reconstruct a global network. How to best perturb the system to facilitate TRN inference is an area that has not received much attention, but see Barrett and Palsson,[107] Tegner *et al.*[108] and Bonneau.[109] This may be due to the limited knowledge on TRNs and the fact that the optimal experiments depend on a good understanding of the true underlying networks. In fact, even when a TF is perturbed, it is often not a trivial task to select the optimal experimental condition to perturb this TF to observe the system's responses.[110]

Essentially, all the methods proposed to date for TRN inference are statistical in nature without invoking detailed mechanistic models for specific genes involved in a TRN. However, with more research done and more knowledge accumulated, there is no question that the field is moving towards a more detailed and mechanism-driven approach to describing the system. For certain pathways, large differential equation systems have been analyzed in the literature.[111] One major limitation in analyzing such systems is that many of the kinetic parameters are either unknown or known only to be within a wide range. The system behaviour may be quite sensitive to the choice of these parameter values, and it is often difficult to infer their values because there may be very limited amount of information in gene expression data.[78,112] Currently, there is still a large gap between statistical (descriptive) and mathematical (mechanistic) approaches to systems modelling and analysis. Although they can be considered complementary at this point, these two approaches will certainly converge in the future, but much effort is needed to make this happen.

One factor that will help to achieve this goal is technology, which keeps evolving at a very rapid pace. Although we have focused on microarray gene expression data, sequence-based expression profiling will provide more comprehensive and accurate assessment of the expression patterns for an organism under a given condition.[113] Due to the nature of sequencing data, the above discussed statistical methods need to be modified even if the same general approach is pursued for TRN inference. In addition, we can obtain data at ever-finer levels, e.g. cell-based assays, and monitor expression patterns both in space and time.

Finally, given the many methods available for TRN inference, it is important to compare the performance of these methods based on both simulated and real data.[114] But very limited studies have been done to date. Although it may well be that different methods suit different systems and data types, it would be helpful to biologists if some general principles and guidelines can emerge from such studies so that they can indeed take advantage of the advances in statistical and computational methods instead of being inundated with different approaches.

Acknowledgments

Supported in part by a Novel Methodologies in Biostatistics Pilot Award from the Yale Center for Clinical Investigation, NIH grants R21 GM 84008 and R01 GM59507.

References

1. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007; 128(4):707–719. [PubMed: 17320508]
2. Schones DE, Zhao K. Genome-wide approaches to studying chromatin modifications. Nature Reviews Genetics. 2008; 9(3):179–191.
3. Halbeisen RE, Galgano A, Scherrer T, Gerber AP. Post-transcriptional gene regulation: from genome-wide studies to principles. Cellular and Molecular Life Sciences. 2008; 65(5):798–813. [PubMed: 18043867]
4. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature. 2001; 409(6819):533–538. [PubMed: 11206552]
5. Ren B, Robert F, Wyrick JJ, et al. Genome-wide location and function of DNA binding proteins. Science. 2000; 290(5500):2306–2309. [PubMed: 11125145]
6. Carroll JS, Meyer CA, Song J, et al. Genome-wide analysis of estrogen receptor binding sites. Nature Genetics. 2006; 38:1289–1297. [PubMed: 17013392]
7. Lee TI, Jenner RG, Boyer LA, et al. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell. 2006; 125:301–313. [PubMed: 16630818]
8. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein–DNA interactions. Science. 2007; 316:1497–1502. [PubMed: 17540862]
9. Robertson G, Hirst M, Bainbridge M, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods. 2007; 4:651–657. [PubMed: 17558387]
10. Liu X, Noll DM, Lieb JD, Clarke ND. DIP-chip: rapid and accurate determination of DNA-binding specificity. Genome Research. 2005; 15(3):421–427. [PubMed: 15710749]
11. Mukherjee S, Berger MF, Jona G, et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nature Genetics. 2004; 36(12):1331–1339. [PubMed: 15543148]
12. Simon I, Barnett J, Hannett N, et al. Serial regulation of transcriptional regulators in the yeast cell cycle. Cell. 2001; 106(6):697–708. [PubMed: 11572776]
13. Gao F, Foat BC, Bussemaker HJ. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics. 2004; 5:31. [PubMed: 15113405]
14. Datta D, Zhao H. Statistical methods to infer cooperative binding among transcription factors in *Saccharomyces cerevisiae*. Bioinformatics. 2008; 24(4):545–552. [PubMed: 17989095]
15. Chu S, DeRisi J, Eisen M, et al. The transcriptional program of sporulation in budding yeast. Science. 1998; 282(5389):699–705. [PubMed: 9784122]
16. Gasch AP, Spellman PT, Kao CM, et al. Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of the Cell. 2000; 11(12):4241–4257. [PubMed: 11102521]
17. Hughes TR, Marton MJ, Jones AR, et al. Functional discovery via a compendium of expression profiles. Cell. 2000; 102(1):109–126. [PubMed: 10929718]
18. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell. 1998; 9(12):3273–3297. [PubMed: 9843569]
19. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. Precision and functional specificity in mRNA decay. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99(9):5860–5865. [PubMed: 11972065]
20. Wang RS, Zhang XS, Chen L. Inferring transcriptional interactions and regulator activities from experimental data. Molecular Cell. 2007; 24(3):307–315.
21. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proceedings of the National Academy of Sciences of the United States of America. 2000; 97(22): 12182–12186. [PubMed: 11027309]

22. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology. 2005; 4 Article 17.
23. Margolin AA, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006; 7(Suppl 1):S7. [PubMed: 16723010]
24. Rice JJ, Tu Y, Stolovitzky G. Reconstructing biological networks using conditional correlation analysis. Bioinformatics. 2005; 21(6):765–773. [PubMed: 15486043]
25. Wille A, Buhlmann P. Low-order conditional independence graphs for inferring genetic networks. Statistical Applications in Genetics and Molecular Biology. 2006; 5 Article 1.
26. Toh H, Horimoto K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics. 2002; 18(2):287–297. [PubMed: 11847076]
27. Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004; 90(1):196–212.
28. Li H, Gui J. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. Biostatistics. 2006; 7(2):302–317. [PubMed: 16326758]
29. Meinshausen N, Buhlmann P. High-dimensional graphs and variable selection with the Lasso. Annals of Statistics. 2006; 34(3):1436.
30. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association. 2009; 104(486):735–746. [PubMed: 19881892]
31. Rogers S, Girolami M. A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics. 2005; 21(14):3131–3137. [PubMed: 15879452]
32. Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology. 2005; 4 Article 32.
33. Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics. 2005; 21(6):754–764. [PubMed: 15479708]
34. Shimamura T, Imoto S, Yamaguchi R, Miyano S. Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. Genome Informatics. 2007; 19:142–153. [PubMed: 18546512]
35. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 1996; 58(1):267–288.
36. Andrecut M, Huang S, Kauffman SA. Heuristic approach to sparse approximation of gene regulatory networks. Journal of Computational Biology. 2008; 15(9):1173–1186. [PubMed: 18844584]
37. Andrecut M, Kauffman SA. On the sparse reconstruction of gene networks. Journal of Computational Biology. 2008; 15(1):21–30. [PubMed: 18257675]
38. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science. 2003; 301(5629):102–105. [PubMed: 12843395]
39. Yeung MK, Tegner J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99(9):6163–6168. [PubMed: 11983907]
40. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. Journal of Computational Biology. 2000; 7(3–4):601–620. [PubMed: 11108481]
41. Hartemink, AJ.; Gifford, DK.; Jaakkola, TS.; Young, RA. Pacific Symposium on Biocomputing. Hawaii: 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks; p. 422-433.
42. Ellis B, Wong WH. Learning causal Bayesian network structures from experimental data. Journal of the American Statistical Association. 2008; 103(482):778–789.
43. Perrin BE, 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–ii148. [PubMed: 14534183]

44. Zou M, Conzen SD. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics. 2005; 21(1):71–79. [PubMed: 15308537]
45. Mukherjee S, Speed TP. Network inference using informative priors. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(38):14313–14318. [PubMed: 18799736]
46. Husmeier D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics. 2003; 19(17):2271–2282. [PubMed: 14630656]
47. Soranzo N, Bianconi G, Altafini C. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. Bioinformatics. 2007; 23(13):1640–1647. [PubMed: 17485431]
48. Werhli AV, Grzegorczyk M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics. 2006; 22(20):2523–2531. [PubMed: 16844710]
49. Bansal M, Gatta GD, di Bernardo D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics. 2006; 22(7):815–822. [PubMed: 16418235]
50. Wang SC. Reconstructing genetic networks from time ordered gene expression data using Bayesian method with global search algorithm. Journal of Bioinformatics and Computational Biology. 2004; 2(3):441–458. [PubMed: 15359420]
51. Savageau MA. Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. Journal of Theoretical Biology. 1969; 25(3):370–379. [PubMed: 5387047]
52. Savageau MA. Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. Journal of Theoretical Biology. 1969; 25(3):365–369. [PubMed: 5387046]
53. Bonneau R, Reiss DJ, Shannon P, et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biology. 2006; 7(5):R36. [PubMed: 16686963]
54. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95(25):14863–14868. [PubMed: 9843981]
55. Cheng Y, Church GM. Biclustering of expression data. Proceedings of the International Conference on Intelligent Systems for Molecular Biology. 2000; 8:93–103.
56. Gu J, Liu JS. Bayesian biclustering of gene expression data. BMC Genomics. 2008; 9(Suppl 1):S4. [PubMed: 18366617]
57. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2004; 1(1):24–45. [PubMed: 17048406]
58. Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(9):2981–2986. [PubMed: 14973197]
59. Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics. 2003; 34(2):166–176. [PubMed: 12740579]
60. Segal E, Yelensky R, Koller D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics. 2003; 19(Suppl 1):i273–i282. [PubMed: 12855470]
61. Reiss DJ, Baliga NS, Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics. 2006; 7:280. [PubMed: 16749936]

62. Bar-Joseph Z, Gerber GK, Lee TI, et al. Computational discovery of gene modules and regulatory networks. Nature Biotechnology. 2003; 21(11):1337–1342.
63. Lemmens K, Dhollander T, De Bie T, et al. Inferring transcriptional modules from ChIP-chip, motif and microarray data. Genome Biology. 2006; 7(5):R37. [PubMed: 16677396]
64. Wang W, Cherry JM, Nochomovitz Y, Jolly E, Botstein D, Li H. Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(6):1998–2003. [PubMed: 15684073]
65. Li, H. Statistical methods for inference of genetic networks and regulatory modules. In: Emmert-Streid; Dehmer, editors. Analysis of microarray data: network-based approaches. Wiley VCH, Verlag GmbH & Co. KGaA; 2008. p. 143-167.
66. Liu X, Jessen WJ, Sivaganesan S, Aronow BJ, Medvedovic M. Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and ChIP-chip data. BMC Bioinformatics. 2007; 8:283. [PubMed: 17683565]
67. Khanin R, Vinciotti V, Mersinias V, Smith CP, Wit E. Statistical reconstruction of transcription factor activity using Michaelis-Menten kinetics. Biometrics. 2007; 63(3):816–823. [PubMed: 17825013]
68. Liu X, Clarke ND. Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. Journal of Molecular Biology. 2002; 323(1):1–8. [PubMed: 12368093]
69. Nachman I, Regev A, Friedman N. Inferring quantitative models of regulatory networks from expression data. Bioinformatics. 2004; 20(Suppl 1):i248–i256. [PubMed: 15262806]
70. Pan Y, Durfee T, Bockhorst J, Craven M. Connecting quantitative regulatory-network models to the genome. Bioinformatics. 2007; 23(13):i367–i376. [PubMed: 17646319]
71. Chen KC, Wang TY, Tseng HH, Huang CY, Kao CY. A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. Bioinformatics. 2005; 21(12):2883–2890. [PubMed: 15802287]
72. Chen HC, Lee HC, Lin TY, Li WH, Chen BS. Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle. Bioinformatics. 2004; 20(12):1914–1927. [PubMed: 15044243]
73. Sun N, Carroll RJ, Zhao H. Bayesian error analysis model for reconstructing transcriptional regulatory networks. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(21):7988–7993. [PubMed: 16702552]
74. Porreca R, Drulhe S, de Jong H, Ferrari-Trecate G. Structural identification of piecewise-linear models of genetic regulatory networks. Journal of Computational Biology. 2008; 15(10):1365–1380. [PubMed: 19040369]
75. Ben-Tabou de-Leon S, Davidson EH. Modeling the dynamics of transcriptional gene regulatory networks for animal development. Developmental Biology. 2009; 325(2):317–328. [PubMed: 19028486]
76. Rogers S, Khanin R, Girolami M. Bayesian model-based inference of transcription factor activity. BMC Bioinformatics. 2007; 8(Suppl 2):S2. [PubMed: 17493251]
77. Cheng C, Yan X, Sun F, Li LM. Inferring activity changes of transcription factors by binding association with sorted expression profiles. BMC Bioinformatics. 2007; 8:452. [PubMed: 18021409]
78. Cao J, Zhao H. Estimating dynamic models for gene regulation networks. Bioinformatics. 2008; 24(14):1619–1624. [PubMed: 18505754]
79. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(26):15522–15527. [PubMed: 14673099]
80. Zhao H, Wu B, Sun N. DNA-protein binding and gene expression patterns. Science and Statistics: A Festschrift for Terry Speed. IMS Lecture Notes-Monograph Series. 2003; 40:259–274.

81. Cokus S, Rose S, Haynor D, Gronbech-Jensen N, Pellegrini M. Modelling the network of cell cycle transcription factors in the yeast *Saccharomyces cerevisiae*. BMC Bioinformatics. 2006; 7:381. [PubMed: 16914048]
82. Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics. 2007; 23(12):1486–1494. [PubMed: 17463025]
83. Xing B, van der Laan MJ. A statistical method for constructing transcriptional regulatory networks using gene expression and sequence data. Journal of Computational Biology. 2005; 12(2):229–246. [PubMed: 15767778]
84. Sabatti C, James GM. Bayesian sparse hidden components analysis for transcription regulation networks. Bioinformatics. 2006; 22(6):739–746. [PubMed: 16368767]
85. Yu T, Li KC. Inference of transcriptional regulatory network by two-stage constrained space factor analysis. Bioinformatics. 2005; 21(21):4033–4038. [PubMed: 16144806]
86. Boulesteix AL, Strimmer K. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. Theoretical Biology & Medical Modelling. 2005; 2:23. [PubMed: 15978125]
87. Pournara I, Wernisch L. Factor analysis for gene regulatory networks and transcription factor activity profiles. BMC Bioinformatics. 2007; 8:61. [PubMed: 17319944]
88. Tanay A, Shamir R. Multilevel modeling and inference of transcription regulation. Journal of Computational Biology. 2004; 11(2–3):357–375. [PubMed: 15285896]
89. Hirose O, Yoshida R, Imoto S, et al. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. Bioinformatics. 2008; 24(7):932–942. [PubMed: 18292116]
90. Wu, FX.; Zhang, WJ.; Kusalik, AJ. Modeling gene expression from microarray expression data with state-space equations; Pacific Symposium on Biocomputing; 2004. p. 581-592.
91. Rangel C, Angus J, Ghahramani Z, et al. Modeling T-cell activation using gene expression profiling and state-space models. Bioinformatics. 2004; 20(9):1361–1372. [PubMed: 14962938]
92. Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics. 2005; 21(3):349–356. [PubMed: 15353451]
93. Xiong H, Choe Y. Structural systems identification of genetic regulatory networks. Bioinformatics. 2008; 24(4):553–560. [PubMed: 18175769]
94. Sanguinetti G, Lawrence ND, Rattray M. Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. Bioinformatics. 2006; 22(22):2775–2781. [PubMed: 16966362]
95. Li Z, Shaw SM, Yedwabnick MJ, Chan C. Using a state-space model with hidden variables to infer transcription factor activities. Bioinformatics. 2006; 22(6):747–754. [PubMed: 16403793]
96. Chen L, Zhao H. Integrating mRNA decay information into co-regulation study. Journal of Computer Science and Technology. 2005; 20:434–438.
97. Shalem O, Dahan O, Levo M, et al. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. Molecular Systems Biology. 2008; 4:223. [PubMed: 18854817]
98. Whitington T, Perkins AC, Bailey TL. High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. Nucleic Acids Research. 2009; 37(1):14–25. [PubMed: 18988630]
99. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nature Genetics. 2001; 27(2):167–171. [PubMed: 11175784]
100. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nature Biotechnology. 2002; 20(8):835–839.
101. Conlon EM, Liu XS, Lieb JD, Liu JS. Integrating regulatory motif discovery and genome-wide expression analysis. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(6):3339–3344. [PubMed: 12626739]
102. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002; 296(5568):752–755. [PubMed: 11923494]

103. Morley M, Molony CM, Weber TM, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004; 430(7001):743–747. [PubMed: 15269782]
104. Schadt EE, Monks SA, Drake TA, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003; 422(6929):297–302. [PubMed: 12646919]
105. Yvert G, Brem RB, Whittle J, et al. Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nature Genetics. 2003; 35(1):57–64. [PubMed: 12897782]
106. Zhu J, Zhang B, Smith EN, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature Genetics. 2008; 40(7):854–861. [PubMed: 18552845]
107. Barrett CL, Palsson BO. Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. PLoS Computational Biology. 2006; 2(5):e52. [PubMed: 16710450]
108. Tegner J, Yeung MK, Hasty J, Collins JJ. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(10):5944–5949. [PubMed: 12730377]
109. Bonneau R. Learning biological networks: from modules to dynamics. Nature Chemical Biology. 2008; 4(11):658–664.
110. Chua G, Robinson MD, Morris Q, Hughes TR. Transcriptional networks: reverse-engineering gene regulation on a global scale. Current Opinion in Microbiology. 2004; 7(6):638–646. [PubMed: 15556037]
111. Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ. Integrative analysis of cell cycle control in budding yeast. Molecular Biology of the Cell. 2004; 15(8):3841–3862. [PubMed: 15169868]
112. Zak DE, Gonye GE, Schwaber JS, Doyle FJ 3rd. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. Genome Research. 2003; 13(11):2396–2405. [PubMed: 14597654]
113. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009; 10(1):57–63.
114. Camacho D, Vera Licona P, Mendes P, Laubenbacher R. Comparison of reverse-engineering methods using an in silico network. Annals of the New York Academy of Sciences. 2007; 1115:73–89. [PubMed: 17925358]