



Review on statistical methods for gene network reconstruction using expression data



Y.X. Rachel Wang, Haiyan Huang*

Department of Statistics, University of California, Berkeley, CA 94720, USA

HIGHLIGHTS

- We review statistical methods for reconstructing gene regulatory networks.
- We discuss statistical and computational challenges in modeling gene interactions.
- For each method we compare their modeling paradigms and data types required.

ARTICLE INFO

Article history:

Received 26 January 2014

Received in revised form

29 March 2014

Accepted 31 March 2014

Available online 12 April 2014

Keywords:

Coexpression networks

Bayesian networks

Dynamic networks

Community detection

Genomic data integration

ABSTRACT

Network modeling has proven to be a fundamental tool in analyzing the inner workings of a cell. It has revolutionized our understanding of biological processes and made significant contributions to the discovery of disease biomarkers. Much effort has been devoted to reconstruct various types of biochemical networks using functional genomic datasets generated by high-throughput technologies. This paper discusses statistical methods used to reconstruct gene regulatory networks using gene expression data. In particular, we highlight progress made and challenges yet to be met in the problems involved in estimating gene interactions, inferring causality and modeling temporal changes of regulation behaviors. As rapid advances in technologies have made available diverse, large-scale genomic data, we also survey methods of incorporating all these additional data to achieve better, more accurate inference of gene networks.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

From ecological networks describing biotic interactions of different species to intricate biochemical networks modeling actions of molecules at a cellular level, the study of biology has seen a fast-expanding effort to analyze individual biological components in the context of large-scale, complex systems with interacting constituents. Most notably, rapid advances in genomic technology have generated an enormous wealth of data on which mathematical and statistical tools can be applied to infer qualitative and quantitative relationships between DNA, RNA, proteins and other cellular molecules. Such a process of reconstructing biochemical networks using genomic data, also known as network inference or reverse engineering, has helped to elucidate the nature of complex biological processes and disease mechanisms in a variety of organisms, bringing us one step closer to understanding how genetic blueprints combined with non-genetic,

environmental factors influence the characteristics of a living system. In particular, comprehending the associations between genotypic and phenotypic characteristics has important ramifications in pathological studies for explaining disease pathways and identifying biomarkers for prognosis and diagnosis.

At a high level, genes, proteins or other metabolites can be conceptualized as nodes and their interactions as edges in a graph. In metabolic networks, reactions are represented as directed edges pointing from reaction substrates to products. While metabolic networks tend to focus on proteins or protein-complexes functioning as enzymes, general protein–protein interaction (PPI) networks are undirected graphs where an edge indicates physical binding between two proteins.

At a more fundamental level, understanding biological processes requires understanding gene regulatory networks since all proteins are encoded by genes. In such a network, transcription factors (TFs), RNA and other small molecules act as regulators to activate or repress the expression levels of genes. Thus gene interactions can occur in the form of direct physical binding of proteins (TFs) to their target sequences, but in a broader sense also include indirect interactions when the expression of a gene influence the expressions of others with regulations caused by

* Corresponding author.

E-mail addresses: rachelwang@stat.berkeley.edu (Y.X.R. Wang), huang@stat.berkeley.edu (H. Huang).

one or more intermediaries. Although experimental evidence can be gathered to search for and verify gene interactions, computational tools utilizing gene expression data offer a much more time and cost efficient way to reconstruct these networks. In the past decade, high quality gene expression data have been made readily available in the form of microarray or RNA-seq data.

The idea of modeling the aforementioned biochemical processes as networks is conceptually appealing as many biologically interesting questions can find their counterparts in graph theory. For example, many biochemical networks demonstrate a high clustering coefficient (Barabási and Oltvai, 2004), indicative of a scale-free topology with a few highly connected nodes, or known as hubs. Comparisons with generative network models that give rise to such a topology can help to explain the evolution of organisms at a cellular level. Another important architectural feature of these networks is modules, where a number of nodes form a densely connected community and have sparser connections with the rest of the network. Community or module detection is of great importance in analyzing biochemical networks since identifying groups of molecules performing a specific cellular function is a key issue in system biology. In a PPI network, highly connected nodes are often proteins interacting as part of a complex or other functional modules, which are fundamental in cellular functions and have been shown to play an important role in disease pathologies (Lim et al., 2006; Soler-López et al., 2011). In gene networks, genes modules are likely to have related biological functions or participate in the same biological pathway.

In this paper we review methods for the reconstruction of biological networks with an emphasis on gene networks. In reality, the relationships between genes are directional in nature and they can change over time or in response to external stimulus. Therefore when modeling gene networks a researcher is faced with the choice of whether to include extra features such as causality and temporal behaviors into the model. This choice of modeling paradigm is largely dependent on the type and quality of data available, relevant biological questions to be addressed, and statistical and computational considerations. In Section 2, we focus on methods used to reconstruct static gene networks, highlighting the statistical and computational challenges in inferring undirected or directed network edges and identifying tightly connected communities as potential functional groups. In Section 3, we discuss methods that model temporal changes of gene regulations in a dynamic network. In Section 4, we expand on the data type under consideration from gene expression to other types of genomic data. We survey some methods available for integrating the additional information given, and the connection between biochemical networks and disease biomarkers.

2. Static gene networks

2.1. Inferring undirected gene association networks

Gene expression data has the form of a matrix with p genes arranged in rows and their expression levels measured under n experimental conditions in columns. A typical feature of this type of data is their high dimensionality with p much larger than n , posing many estimation and computation challenges. Most methods for inferring edges in gene networks are based on a notion of similarity or coexpression measure. Coexpression is one of the earliest tools used to infer edges in a gene network and is based on the concept of “guilt by association”: genes that have similar expression profiles under different experimental conditions are likely to be co-regulated and hence functionally related.

As we review a number of methods developed for inferring edges in gene networks, we evaluate their advantages and

disadvantages considering both their biological implications and statistical and computational properties.

2.1.1. Correlation-based coexpression networks

Common measures for quantifying the degree of coexpression between two genes include Pearson correlation, rank correlation, Euclidean distance, and the angle between a pair of observed expression vectors (Wen et al., 1998; D'haeseleer et al., 2000; Horvath and Dong, 2008). Since expression data routinely require normalization, Euclidean distance can be sensitive to different scaling methods used, whereas correlation measures are invariant with respect to linear transformations. Rank correlations such as Spearman rank correlation are more robust and less sensitive to outliers compared to Pearson correlation, although some information is lost in the process of converting numerical values to ranks. The angle between two expression vectors can be seen as the geometric interpretation of Pearson correlation, which is 0 when the two vectors are perfectly correlated. Empirical performances of different coexpression measures on simulated and real data have been compared in, for example, Kumari et al. (2012).

Pearson correlation remains the most popular coexpression measure used in the literature (Eisen et al., 1998; Spellman et al., 1998; Stuart et al., 2003; Wolfe et al., 2005). Either hard (Carter et al., 2004) or soft thresholding (Langfelder and Horvath, 2008) is then applied to produce a binary or weighted network. One widely used soft thresholding scheme is proposed by Zhang and Horvath (2005) which raises the absolute values of the correlations to some positive power. A topological overlap matrix can then be computed using the thresholded values to take into account topological properties shared by pairs of nodes for the identification of functional modules later.

A related issue in the computation of correlation measures is the existence of dependency between measurements taken under different experimental conditions for each gene. When unaccounted for, these experimental dependencies can confound gene dependencies and lead to inaccurate estimation of the quantities of interest. One approach is to decouple and remove the experimental dependencies before proceeding to computing correlations or other similarity measures, as demonstrated in Teng and Huang (2009). Their method is based on the assumption that iteratively projecting a centralized expression matrix to its eigenspaces of gene-wise and experiment-wise covariance matrices removes the dependencies both in genes and in experiments. When the measurements are time series, methods which either directly adjust for time lags in correlation inference for pathway regulations (Bickel, 2005), or functional smoothing approaches for comparing time series curves are available (Chen et al., 2001; Filkov et al., 2002).

Model-based methods and/or data-driven developments are also available for measuring coexpression. We only mention a few here. By considering the specific properties of Serial Analysis of Gene Expression (SAGE) data, Cai et al. (2004) developed two MLE-based distances following a Poisson model for SAGE data. In Kim et al. (2007), motivated by analyzing an Arabidopsis dataset, a spectral clustering method was developed by separately modeling the shape and magnitude parameters of a gene expression profile and considering them in a new feature space. Smoothing spline clustering (Ma et al., 2006) is a method developed for time-course gene expression data, taking into account natural properties of gene expression over time, differences in gene expression within a cluster, and the effects of experimental measurement error and missing data. All of these methods have their unique advantages for certain data, yet also generalizable with modifications.

Easy interpretation and low computational costs are the main advantages of coexpression networks. However, it is well known

that empirical sample covariances or correlations have poor asymptotic behaviors under the high dimensional setting. Furthermore, it is unclear whether high coexpression necessarily implies relatedness in biological function. Although Wolfe et al. (2005) used gene ontologies to show coexpression leads to functional similarity, other works (Filkov et al., 2002; Gillis and Pavlidis, 2012) have shown it often yields high false positive and low prediction rates.

2.1.2. Information theoretic for measuring nonlinear relationships

Another class of methods used to investigate dependencies between genes is mutual information (MI). The MI between expression vectors of gene i and j is defined as

$$MI_{ij} = H_i + H_j - H_{ij}, \quad (1)$$

where

$$H_i = - \sum_{k=1}^{n_i} f_i(k) \log(f_i(k))$$

$$H_{ij} = - \sum_{k=1, l=1}^{n_i, n_j} f_{ij}(k, l) \log(f_{ij}(k, l)) \quad (2)$$

are the marginal and joint entropies of the expression variables, respectively. Since expression data are continuous, the expression range can be partitioned into discrete bins and $f_i(k)$ represents the frequency of expression measurements for gene i falling into the k th bin. Similarly, $f_{ij}(k, l)$ is the joint frequency for gene i and j with respect to bins k and l (Butte and Kohane, 2000). Alternatively, each observation in the bin can be weighted using a smoothing kernel (Daub et al., 2004; Steuer et al., 2002; Basso et al., 2005; Margolin et al., 2006) to reduce the influence of noise at the edges of the bins. Post-processing to arrive at the final network includes hard thresholding based on significance values obtained from permutation tests (Butte and Kohane, 2000; Daub et al., 2004) and edge pruning (Margolin et al., 2006).

MI is a more general way to measure gene relationships than Pearson correlation. The latter being zero does not imply statistical independence but the MI between two variables is zero only if they are independent. In practice, MI is shown to be able to capture nonlinear correlations between expression profiles (Daub et al., 2004), but may also yield almost identical results as Pearson correlation (Steuer et al., 2002). Whether one can directly interpret pairwise statistical dependence as functional similarity remains uncertain. Recently, Reshef et al. (2011) proposed a new measure named the maximal information coefficient (MIC) for detecting general associations between pairs of data. The measure is based on normalized estimates of MI over different ways of partitioning the data into grids. However, Kinney and Atwal (2014) argued that MIC has inferior power compared to MI and other correlation measures, and the claim that MIC is equitable for different types of dependence relationships was questioned.

Another relevant dependence measure is the Renyi Correlation, which is also known as the Maximal Correlation (Rényi, 1959). Let (X, Y) be a pair of dependent random variables. The Renyi Correlation, or Maximal Correlation, of X and Y is defined as

$$\rho(X, Y) = \max_{f(x), g(y)} \mathbb{E}[f(X)g(Y)],$$

where $f(x)$ and $g(y)$ are functions of x and y respectively such that $\mathbb{E}f(X) = \mathbb{E}g(Y) = 0$ and $\mathbb{E}f(X)^2 = \mathbb{E}g(Y)^2 = 1$. The Renyi correlation has the property that $\rho(X, Y) = 0$ if and only if X and Y are independent. If there is a linear correlation between the variables, then the Renyi correlation coincides with the Pearson correlation. Renyi correlation is grossly under-explored for its potential in biological applications.

2.1.3. Partial correlation/Gaussian graphical models

Both coexpression and MI only consider pairwise relationships between genes. However, in a real biological pathway, a gene may interact with a group of genes but not possess a strong marginal relationship with any individual member of the group. Such higher-level interactions can be potentially missing in the networks constructed by pairwise measures. In this sense, Gaussian graphical models (GGM) offer a more realistic way to represent complex gene networks due to its interpretation in terms of conditional correlations. Assuming a multivariate normal distribution for the expression vectors for a set of genes W , the GGM uses $(\Sigma)^{-1}$, the inverse of the gene covariance matrix (or precision matrix), as a measure for gene association patterns. This approach is closely related with the concept of partial correlations, noting that the partial correlation between genes i and j can be expressed as

$$\rho_{ij} = \text{cor}(i, j | W \setminus \{i, j\}) = \begin{cases} -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}, & i \neq j \\ 1, & i = j, \end{cases} \quad (3)$$

where ω_{ij} are elements in the precision matrix. Therefore genes i and j being conditionally independent is equivalent to the corresponding partial correlation and element in the precision matrix being zero. And nonzero entries in the precision matrix correspond to the presence of direct interaction between two genes having controlled for the effect of the other genes.

The major difficulty of estimating the precision matrix arises from the high dimensional nature of gene expression data. Various regularized estimation methods have been proposed to address this “curse of dimensionality”. Edwards (2000) proposed a backward selection scheme to remove weak edges in the estimated Σ^{-1} . Schäfer and Strimmer (2005) chose to estimate Σ^{-1} directly using the Moore–Penrose pseudo-inverse (Penrose, 1955) and using the bagged average of all bootstrap estimates. Since gene networks are believed to be inherently sparse, Li and Gui (2006) introduced in-built sparsity in their estimated Σ^{-1} by a threshold gradient descent algorithm. Noting that regressing the expression vector X_i for gene i on the other expression vectors X_j ,

$$X_i = \sum_{j \neq i} \beta_{ij} X_j + \epsilon_i, \quad (4)$$

where the coefficients $\beta_{ij} = \rho_{ij} \sqrt{\omega_{jj}/\omega_{ii}}$, sparsity can be more naturally incorporated in a penalized regression setting (Meinshausen and Bühlmann, 2006; Peng et al., 2009). A rich wealth of the literature exists on the problem of estimating sparse precision matrix in high dimensional GGMs (Zhou et al., 2011; Yuan and Lin, 2007; Friedman et al., 2007).

The above partial correlation based approaches have attractive theoretical properties and their asymptotic behaviors are extensively studied. However, the kind of biological inference they are capable of achieving is still limited. In the current literature, partial correlation is usually calculated conditioned on either all of the available genes or a more or less arbitrary subset of them that may contain noisy (biologically unrelated) genes. As pointed out in De La Fuente et al. (2004) and Kim et al. (2012), the selection of a proper set of genes on which the correlation in (3) is conditioned is critical. The inclusion of noisy genes in the set $W \setminus \{i, j\}$ may introduce spurious dependencies and consequently false edges in the estimated network. There are also efforts on using lower order partial correlations (De La Fuente et al., 2004; Magwene and Kim, 2004; Wille et al., 2004; Wille and Bühlmann, 2006) which condition on one or two other genes. Li (2002) considered how the first-order partial correlation changes depending on the expression level of the conditional gene, which acts as a surrogate variable for varying cellular state. This measure termed liquid association was used to identify candidate genes involved in urea

cycle (Li, 2002) and multiple sclerosis (Li et al., 2007). These methods, however, lose sensitivity for inferring higher level gene associations and cannot guarantee to eliminate the effect of noisy genes. Kim et al. (2012) proposed to minimize the impact of noisy genes by conditioning on a small set (3–5 genes) of known pathway genes, or “seed genes”. When such prior biological information is unavailable, Wang et al. (2014) introduced a new method of estimating the strength of gene group interactions using sparse canonical correlation analysis (SCCA) coupled with repeated random partition and subsampling of the expression dataset. By separating the genes into two groups, SCCA searches for meaningful linear group relationships which, reframed in a similar regression setting as (4), gives estimates proportional to partial correlations conditioned on different sets of signal genes (with noisy genes eliminated through sparsity). Subsampling allows for the discovery of multiple interacting groups simultaneously by stepping through subsets of the genes with varying signal strengths. The final edge weight matrix averages the results from all the random partitions and subsamples to obtain an aggregated measure of partial correlations of different orders. Performance comparison with popular coexpression measures on both simulated and real data show the new method leads to better accuracy and more biologically meaningful results.

2.2. Inferring Bayesian (directed) networks

From marginal dependencies in coexpression measure to conditional dependencies in partial correlation based approaches, the methods discussed above attempt to capture gene relationships using probabilistic dependencies of different kinds. However, they all lead to the construction of undirected graphs and hence unable to represent causal relationships between genes. Bayesian networks (BNs) for gene regulatory networks, pioneered by Friedman et al. (2000), are directed acyclic graphs (DAGs) that characterize the joint distribution of nodes (genes) as a series of local probability distributions. Denoting gene i as X_i , the joint distribution of all nodes is given by

$$\mathbb{P}(X_1, \dots, X_p) = \prod_{i=1}^p \mathbb{P}(X_i | Pa^G(X_i)), \quad (5)$$

where $Pa^G(X_i)$ are all the parent nodes of X_i in the DAG G . The joint distribution can be factorized this way because of the Markov assumption of BNs: given its parents, each node is independent of its non-descendants. In this sense, each directed edge can be interpreted as a causal link. A BN implies both a set of conditional dependencies and conditional independences. Two different DAGs can encode the same set of conditional independences (Pearl and Verma, 1991), and the goal of BN inference algorithms is to infer these equivalent classes of DAGs.

Reconstructing a BN based on expression data D involves finding the best DAG G that describes D , and each G is evaluated using a Bayesian score which is the posterior probability of G ,

$$S(G : D) = \log \mathbb{P}(G|D) = \log \mathbb{P}(D|G) + \log \mathbb{P}(G) + \text{const}. \quad (6)$$

The computation of the posterior probability is two-fold: (i) learning the graph G given observed data; (ii) learning the local conditional probabilities given G .

The second problem amounts to parameter estimation, which can be accomplished via a number of algorithms such as sum-product, MLE, MAP and EM depending on the form of the conditional probabilities (discrete, continuous or mixture distribution (Friedman et al., 2000)) and whether any node has missing information. Prior information concerning the distributions of parameters and graphs is also incorporated in the final computation of the scoring function. It is important that the scoring function chosen should be decomposable to the local scores from

each node for computational efficiency. The function should also contain features that guard against overfitting. Popular schemes to achieve this goal include using the BIC criterion and Bayesian Dirichlet equivalent (BDe) (Cooper and Herskovits, 1992; Yoo et al., 2002). A comparison of different scoring schemes can be found in Hartemink et al. (2001) and Yu et al. (2002).

The first problem, however, is a lot more challenging as theoretically it requires us to consider all possible topologies of DAGs, which is super-exponential in search space dimension. Furthermore, the high dimensional nature of expression data lead to many DAGs that score equally well. A number of heuristic algorithms have been developed to walk through the space of possible DAGs, including greedy hill-climbing, simulated annealing and genetic algorithms (Yu et al., 2002). Often the algorithms explore the neighborhood of a topology by adding, deleting or reversing the direction of an edge to make incremental changes at a time. To further reduce the search space, biological assumptions and priors can be employed to limit the number of parents a child node is allowed to have, and coexpression clustering can be applied to arrive at a set of most likely parent/child nodes. Rather than choosing a single optimal G , a number of DAGs scoring comparably can be compared for the selection of consistent topological features. Summaries of how to infer BNs can be found in e.g. Heckerman (1996) and Needham et al. (2007).

The BN has a number of advantages as a modeling framework. The probabilistic setup offers a natural way to incorporate latent variables, prior knowledge and the possibility that gene expression levels are stochastic with noise. Some missing data can also be handled. However, in order to infer all this additional information, more parameters need to be estimated and hence more high quality data is required. For this reason, the application of BNs has been centered around yeast data, and the success in higher organisms and larger networks is still limited. Conceptually, feedback loops, which is a common feature in many pathways, cannot be modeled under this framework since all BNs are acyclic. Although the linkages can be potentially causal, they are still qualitative and do not indicate whether a regulation is activation or repression. These problems can be analyzed using extensions of BNs such as dynamic BNs (Yu et al., 2004; Murphy, 2002) and BNs on perturbation data (Pe'er et al., 2001). Perturbation gene expression data, obtained from perturbation experiments (by knockout or RNA interference), offers an important source of information for estimating directed relationships and networks (Markowitz et al., 2007; Tresch and Markowitz, 2008; Shojaie et al., 2013). One line of approach has been based on incorporating the nested structure of the observed perturbation effects (Markowitz et al., 2007).

2.3. Identifying groups of genes with dense interactions

When a reconstructed gene network has topological structures reflecting real gene interactions, the problem of identifying functional modules can be reframed in the context of pattern recognition or clustering. As we expect these modules to have higher within-group homogeneity, the problem corresponds to finding highly connected subunits within a network, which can be considered as candidate genes acting in individual regulatory systems.

Clustering has been a popular and well studied pattern recognition tool in numerous fields. General reviews of various clustering techniques can be found in Kaufman and Rousseeuw (2005); Theodoridis and Koutroumbas (2005); Jain et al. (1999), with more specific focus on gene expression data in D'haeseleer et al. (2000); Jiang et al. (2004); Kerr et al. (2008). We first note that it is sometimes meaningful to cluster both the genes and samples, especially when groups of samples correspond to distinct

phenotypes or experimental conditions. Biclustering techniques (Cheng and Church, 2000; Madeira and Oliveira, 2004) are required to resolve this case. Clustering can also be applied to expression vectors directly using heuristic algorithms (Self Organizing Maps (Tamayo et al., 1999)), genetic algorithms (Gesù et al., 2005) or model based approaches (Expectation Maximization (Yeung et al., 2001; Muro et al., 2003); variational Bayes (Teschendorff et al., 2005)). We will focus on gene-based clustering utilizing information from a reconstructed gene network.

Most methods discussed in Section 2.1 give rise to a similarity matrix, which can be converted into a distance metric for K-means and hierarchical clustering (either agglomerative or divisive), both widely used in gene expression studies (Tavazoie et al., 1999; Eisen et al., 1998; Alon et al., 1999). K-means assigns each gene to exactly one cluster and requires a pre-defined cluster number. The lack of robustness and the greedy nature of most implemented algorithms are some of its drawbacks. To account for the situation where one gene participates in several pathways, fuzzy versions of K-means have been developed (Dembélé and Kastner, 2003; Fu and Medico, 2007) to associate each gene with multiple clusters. Biologically, one may also argue that functional modules are hierarchical in nature (Barabási and Oltvai, 2004) which provides a natural setting for hierarchical clustering. Since Eisen's work, hierarchical clustering has remained the most widely used tool in analyzing gene expression. The main ambiguity lies in the interpretation of the tree structure and where to define a cut-off to produce the final clusters. Efforts to address this can be found in e.g. Langfelder et al. (2008). Variants of spectral clustering techniques are also widely explored for detecting communities/blocks in sparse networks (Ramesh et al., 2010).

Given gene relationships are now represented by a graph, another natural approach is to consider functional modules as tightly connected subgraphs. Ben-Dor et al. (1999) developed CAST, an algorithm that constructs one cluster at a time by adding and dropping genes iteratively according to a similarity measure. CLICK, proposed by Sharan et al. (2003), assumes edge weights between all pairs of genes follow a mixture normal distribution with a higher mean for within-cluster edges. The parameters and cluster memberships are estimated using EM methods. This idea that nodes have different connectivities depending on their cluster memberships is adopted in a more general probabilistic graph model known as Stochastic Block Model (SBM). The SBM, formally introduced by Holland et al. (1983), generalizes the Erdős-Rényi model and defines a family of probability distributions for a graph with node set $\{1, 2, \dots, p\}$ and K node blocks as follows.

- Let $\mathbf{C} = (C_1, C_2, \dots, C_p)$ denote the set of labels such that $C_i = k$ if the node i belongs to block k .

$$\mathbf{C} \stackrel{i.i.d}{\sim} \text{Multinomial}(\boldsymbol{\gamma}),$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_K)$ is the vector of proportions.

- Let $\boldsymbol{\pi} = (\pi_{lk})_{1 \leq l, k \leq K}$ be a symmetric matrix of block dependent edge probability matrix and \mathbf{A} be the adjacency matrix. Conditioned on the block labels \mathbf{C} , (\mathbf{A}_{ij}) for $i < j$ are independent, and $P(\mathbf{A}_{ij} | \mathbf{C}) = P(\mathbf{A}_{ij} = 1 | C_i = l, C_j = k) = \pi_{lk}$.

The inference problems for SBMs involve both node classification and parameter estimation, and a block with a high internal edge probability can be considered as a potential functional module. Due to the intricacy of its graph structures, how to fit a SBM has remained an active area of research and a number of inference algorithms have been proposed, from methods including MLE, EM (Snijders and Nowicki, 1997), Gibbs sampling (Nowicki and Snijders, 2001), and graph modularity (Newman and Girvan, 2004) to methods that scale

up to larger networks (variational methods (Daudin et al., 2008; Latouche et al., 2012), belief propagation (Decelle et al., 2011), spectral clustering (Rohe et al., 2011)), and pseudo-likelihood (Amini et al., 2013)). More asymptotic analysis of some of the estimation procedures can be found in Bickel and Chen (2009); Celisse et al. (2012); Bickel et al. (2013). Another technical difficulty in fitting a SBM is related to specifying the number of blocks K . A few data-driven approaches based on Integrated Classification Likelihood (Daudin et al., 2008), node degree gaps (Channarond et al., 2012), variational (Hofman and Wiggins, 2008) and spectral clustering (Fishkind et al., 2013) methods, but extensive testings of these procedures using real data remain to be performed.

SBMs and the concept of communities as modularities have been applied in Guimerà and Amaral (2005); Daudin et al. (2008); Airoldi et al. (2008) to recover community structures in biochemical networks. However, wide applicability of SBMs to large-sized gene networks has yet to be verified partly due to the scalability of the current algorithms and the intrinsic sparsity of gene networks, which causes identifiability problems for parameter estimation. Conceptually, SBMs are also too simplistic to account for real network features such as degree variation within blocks and overlapping blocks. These can be addressed to some extent using a degree-corrected SBM (Karrer and Newman, 2011) and mixed membership SBM (Airoldi et al., 2008). As active theoretical research in probabilistic graph models continues, we have reasons to believe this will propel more development in the application of these methods to gene expression data in the future.

3. Dynamic gene networks

The types of gene networks discussed so far have all been static, describing only the network topologies and qualitative features of gene relationships. They do not capture the dynamic nature of real networks and cannot yield quantitative predictions of gene behaviors.

Boolean networks is one of the earliest dynamic models proposed (Kauffman, 1969) that simplifies regulation dynamics as a directed graph, where each node is a binary variable and its change of state between consecutive time points is regulated by a Boolean function of its parent nodes. Because the states are finite, all trajectories of the system are periodic and the attractors can be used to explain stable states in cell cycles. Empirical estimations of the Boolean functions require imposing constraint on network topology or using coexpression or information theoretic approaches to reduce the search space (Liang et al., 1998; Ideker et al., 2000).

As mentioned in Section 2.1, BNs can be extended to capture temporal relationships between the variables. In a dynamic BN, the joint probability factorizes into local probabilities of each node associated with every time point, where the parents of a node can include nodes from previous time points. Although dynamic BNs possess rich statistical properties, parameter estimation is computationally expensive and often discretization of gene expression levels or simplifying graph topology based on prior knowledge is necessary to make parameter estimation feasible (Kim et al., 2004; Zou and Conzen, 2005; Zhu et al., 2010).

Another popular class of dynamic models is based on differential equations (DEs), which models the rate of change in the expression level of a gene as a function of the expression of other genes (often including external perturbations). DE approaches mainly differ in the functional form used, ranging from linear functions (Yeung et al., 2002; Gardner et al., 2003; di Bernardo et al., 2005; Bonneau et al., 2006), power law models (Savageau, 1991) to complex nonlinear functions (Mazur et al., 2009). While solutions for linear systems can be found using linear algebra and

regression techniques, solving more complex systems often require evolutionary algorithms as search strategies (Spieth et al., 2004; Kimura et al., 2005). Since DEs are deterministic in nature, extensions to stochastic DEs have also been proposed (Chen et al., 2005).

Despite having appealing conceptual features, the main drawbacks of these methods lie in the nature of the data. The amount of extensive inference required implies more sample measurements are needed (Akutsu et al., 1999; Yeung et al., 2002). One way to alleviate the dimensionality problem is to combine multiple time-course data and perform inference on the integrated data. Along this line, (Wang et al., 2006) proposed a method that solves a set of DEs for each dataset and finally pools the results into a consistent, sparse network. Furthermore, in order for the data to capture the underlying dynamics of regulatory systems, expression measurements need to be taken on a slowly changing system or finely spaced in time. A recent survey on modeling dynamic biochemical systems can be found in Bar-Joseph et al. (2012).

4. Network reconstruction beyond a single data type

Decades of genomic research have fueled the development of numerous experimental and computational techniques and led to the curation of a large number of databases, such as TRANSFAC (Wingender et al., 2001), KEGG (Okuda et al., 2008), DAVID (Dennis et al., 2003), Cytoscape (Kohl et al., 2011) and NCBI GEO (Barrett et al., 2009). These databases have compiled large amounts of information on gene expression profiles, TF binding motifs, SNP data, PPIs and other biochemical interactions. Performing inference of gene networks based on this type of known knowledge leads to semi-supervised or supervised approaches which typically outperform unsupervised ones. Combining gene expression and other genetic information has also proven fruitful in genome-wide association studies (Xiong et al., 2012).

Given a partially known network, network inference can be considered as a supervised classification problem where the object to be classified is a pair of nodes and often a feature vector is defined for each pair by transforming features available for each node. Various classification algorithms including support vector machines, logistic regression can be utilized to learn pairwise connections. Alternatively, one can predict whether there is an edge between a newly added node and any existing node by learning individually a subnetwork associated with each node of interest. Supervised network inference has been applied to a number of metabolic, PPI, and gene regulatory networks (Bleakley et al., 2007; Yip and Gerstein, 2009; Cerulo et al., 2010). For gene networks in particular, SEREND (Ernst et al., 2008) is the first semi-supervised learning method that integrated information from verified TF binding motifs and a compendium of gene expression data to reconstruct transcriptional regulatory networks in *E. coli*. Another frequently used supervised learning method is SIRENE (Mordelet and Vert, 2008), which uses coexpression behavior of known target genes of TFs to predict binding targets of new TFs. Wang et al. (2009) designed a Bayesian network framework to predict TF cooperativity by integrating 15 genomic features. In particular, they narrowed down on the prediction of a subnetwork of TFs and were able to achieve accurate results. Other studies, including unsupervised methods, that incorporate ChIP-chip, motif, PPIs and phenotypic data have been performed (Bar-Joseph et al., 2003; Tanay et al., 2004; Lemmens et al., 2006; Sabatti and James, 2006; Wang et al., 2009). Recent surveys and comparisons with unsupervised methods are provided in De Smet and Marchal (2010) and Maetschke et al. (2014).

Integrated analyses of genomic data have also found numerous applications beyond inferring gene networks. For metabolic networks, various constrained-based modeling methods (Haggart et al., 2011; Orth et al., 2010; Price et al., 2004) together with efforts to integrate high throughput transcriptomic, proteomic and metabolomic data (Becker and Palsson, 2008; Shlomi et al., 2008; Yizhak et al., 2010) have led to the reconstruction and curation of a large number of organism-specific genome-scale metabolic networks (Feist et al., 2009) capable of predicting reaction fluxes and quantifying metabolic activities. In PPI networks, predicting PPI in silico can be achieved using phylogenetic profiling, sequence homology, structural information or Bayesian framework integrating various genomic features (Pellegrini et al., 1999; Aloy and Russell, 2003; Jansen et al., 2003; Jensen et al., 2009).

Another important application of network reconstruction and integrative data analysis is to identify biomarkers relevant to disease or biological processes under investigation. Using gene expression data, typical methods for finding disease biomarkers rank genes based on their discriminative capacities in relation to different physiological classes, such as disease versus health states. Network-based biomarker discovery approaches, combined with integrating different types of “omic” data, are also used to detect the changes in the “activity” or “behavior” of the reconstructed or known networks across different disease states, from which a more comprehensive and complete picture of disease biomarker activities can be gleaned. Azuaje (2010) provides a nice survey on studies related to disease biomarkers and biological interaction networks.

5. Conclusion

Statistical methods for network reconstruction were reviewed with the main focus on those applicable to gene expression data. When inferring an undirected network, key issues involved include: (i) the selection of an appropriate coexpression measure and (ii) the selection of a community detection method for identifying gene functional groups. As discussed in Section 2.1, choosing an effective coexpression measure depends on the nature of the gene interactions one wants to capture. The latter issue is related to the assumed/expected structure of the target network. Node degree distribution, network conductivity and assortativity are example factors need to be considered. For a directed network such as a gene regulatory work, the inference of edge direction is a fundamental issue. Usually time-course data or perturbation data are needed for determining the causal or driving factors, as presented in Section 2.3. The inference of edge direction is also often aided by integrating other types of data with gene expression data. Dynamic network models discussed in Section 3 allow the reconstructed networks to vary over time, thus more truthfully reflecting the behaviors of real networks and enabling quantitative predictions. However, this comes at a cost of requiring more samples taken at a fine time resolution. As more and more rich, large-scale genomic data are generated via high-throughput technologies, data integration has become a key theme in many studies listed in Section 4 to improve inference beyond what a single type of data can. Despite having proven successful in many individual cases, finding a unified framework for integrating diverse genomic data remains a fertile ground with many uncharted territories. All in all, in order to choose an appropriate method for performing a network analysis, a deep understanding of the biological nature of the target network and the statistical properties of the data are indispensable. It is our hope that this paper has provided a high-level overview of the statistical issues related to gene networks and will serve as a guide for choosing different methods to model them.

Acknowledgement

This work is partly supported by an NIH grant U01HG007031 and an NSF grant DMS-1160319.

References

- Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P., 2008. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* 9, 1981–2014.
- Akutsu, T., Miyano, S., Kuhara, S., 1999. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: *Pacific Symposium on Biocomputing*, pp. 17–28.
- Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.* 96, 6745–6750.
- Aloy, P., Russell, R.B., 2003. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* 19, 161–162.
- Amini, A.A., Chen, A., Bickel, P.J., Levina, E., 2013. Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Stat.* 41, 2097–2122.
- Azuaje, F., 2010. *Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine*. John Wiley & Sons, West Sussex, UK.
- Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A., Gifford, D.K., 2003. Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* 21, 1337–1342.
- Bar-Joseph, Z., Gitter, A., Simon, I., 2012. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genetics* 13, 552–564.
- Barabási, A.L., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genetics* 5, 101–113.
- Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muetter, R.N., Edgar, R., 2009. NCBI GEO: archive for high-throughput functional genomic data. *Nucl. Acids Res.* 37, D885–D890.
- Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A., 2005. Reverse engineering of regulatory networks in human b cells. *Nat. Genetics* 37, 382–390.
- Becker, S.A., Palsson, B.O., 2008. Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* 4, e1000082.
- Ben-Dor, A., Shamir, R., Yakhini, Z., 1999. Clustering gene expression patterns. *J. Comput. Biol.* 6, 281–297.
- Bickel, D.R., 2005. Probabilities of spurious connections in gene networks: application to expression time series. *Bioinformatics* 21, 1121–1128.
- Bickel, P., Chen, A., 2009. A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci.* 106, 21068–21073.
- Bickel, P.J., Choi, D., Chang, X., Zhang, H., 2013. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Stat.* 41, 1922–1943.
- Bleakley, K., Biau, G., Vert, J.P., 2007. Supervised reconstruction of biological networks with local models. *Bioinformatics* 23, i57–i65.
- Bonneau, R., Reiss, D.J., Shannon, P., Facciotti, M., Hood, L., Baliga, N.S., Thorsson, V., 2006. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* 7, R36.
- Butte, A.J., Kohane, I.S., 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: *Pacific Symposium on Biocomputing*, pp. 418–429.
- Cai, L., Huang, H., Blackshaw, S., Liu, J.S., Cepko, C., Wong, W.H., 2004. Clustering analysis of SAGE data using a Poisson approach. *Genome Biol.* 5, R51.
- Carter, S.L., Brechbühler, C.M., Griffin, M., Bond, A.T., 2004. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 20, 2242–2250.
- Celisse, A., Daudin, J.J., Pierre, L., 2012. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* 6, 1847–1899.
- Cerulo, L., Elkan, C., Ceccarelli, M., 2010. Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinform.* 11, 228.
- Channarond, A., Daudin, J.-J., Robin, S., 2012. Classification and estimation in the stochastic block model based on the empirical degrees. *Electron. J. Stat.* 6, 2574–2601.
- Chen, K.C., Wang, T.Y., Tseng, H.H., Huang, C.Y., Kao, C.Y., 2005. A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. *Bioinformatics* 21, 2883–2890.
- Chen, T., Filkov, V., Skiena, S., 2001. Identifying gene regulatory networks from experimental data. *Parallel Comput.* 27, 141–162.
- Cheng, Y., Church, G.M., 2000. Biclustering of expression data. *Int. Conf. Intell. Syst. Mol. Biol.* 8, 93–103.
- Cooper, G.F., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* 9, 309–347.
- Daub, C.O., Steuer, R., Selbig, J., Kloska, S., 2004. Estimating mutual information using B-spline functions — an improved similarity measure for analysing gene expression data. *BMC Bioinform.* 5, 118.
- Daudin, J.J., Picard, F., Robin, S., 2008. A mixture model for random graphs. *Stat. Comput.* 18, 173–183.
- De La Fuente, A., Bing, N., Hoeschele, I., Mendes, P., 2004. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20, 3565–3574.
- De Smet, R., Marchal, K., 2010. Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8, 717–729.
- Decelle, A., Krzakala, F., Moore, C., Zdeborová, L., 2011. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* 84, 066106.
- Dembéle, D., Kastner, P., 2003. Fuzzy C-means method for clustering microarray data. *Bioinformatics* 19, 973–980.
- Dennis, G.J., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A., 2003. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* 4, P3.
- D'haeseleer, P., Liang, S., Somogyi, R., 2000. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726.
- di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E., Collins, J.J., 2005. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* 23, 377–383.
- Edwards, D.I., 2000. *Introduction to Graphical Modelling*, 2nd ed Springer, New York, USA.
- Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* 95, 14863–14868.
- Ernst, J., Beg, Q.K., Kay, K.A., Balázs, G., Oltvai, Z.N., Bar-Joseph, Z., 2008. A semi-supervised method for predicting transcription factor–gene interactions in *Escherichia coli*. *PLoS Comput. Biol.* 4, e1000044.
- Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L., Palsson, B.O., 2009. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7, 129–143.
- Filkov, V., Skiena, S., Zhi, J., 2002. Analysis techniques for microarray time-series data. *J. Comput. Biol.* 9, 317–330.
- Fishkind, D., Sussman, D., Tang, M., Vogelstein, J., Priebe, C., 2013. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* 34, 23–29.
- Friedman, J., Hastie, T., Tibshirani, R., 2007. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* 9, 432–441.
- Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620.
- Fu, L., Medico, E., 2007. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinform.* 8, 3.
- Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J., 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301, 102–105.
- Gesù, V.D., Giancarlo, R., Bosco, G.L., Raimondi, A., Scaturro, D., 2005. GenClust: a genetic algorithm for clustering gene expression data. *BMC Bioinform.* 6, 289.
- Gillis, J., Pavlidis, P., 2012. “Guilt by association” is the exception rather than the rule in gene networks. *PLoS Comput. Biol.* 8, e1002444.
- Guimerà, R., Amaral, L.A.N., 2005. Functional cartography of complex metabolic networks. *Nature* 433, 895–900.
- Haggart, C.R., Bartell, J.A., Saucerman, J.J., Papin, J.A., 2011. Whole-genome metabolic network reconstruction and constraint-based modeling. *Meth. Enzymol.* 500, 411–433.
- Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., Young, R.A., 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In: *Pacific Symposium on Biocomputing*, pp. 422–433.
- Heckerman, D., 1996. *A Tutorial on Learning with Bayesian Networks*. Technical Report, Learning in Graphical Models.
- Hofman, J.M., Wiggins, C.H., 2008. A Bayesian approach to network modularity. *Phys. Rev. Lett.* 100, 258701.
- Holland, P.W., Laskey, K.B., Leinhardt, S., 1983. Stochastic blockmodels: first steps. *Soc. Netw.* 5, 109–137.
- Horvath, S., Dong, J., 2008. Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* 4, e1000117.
- Ideker, T.E., Thorsson, V., Karp, R.M., 2000. Discovery of regulatory interactions through perturbation: inference and experimental design. In: *Pacific Symposium on Biocomputing*, pp. 305–316.
- Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. *ACM Comput. Surv.* 31, 264–323.
- Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M., 2003. A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* 302, 449–453.
- Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C., 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucl. Acids Res.* 37, D412–D416.
- Jiang, D., Tang, C., Zhang, A., 2004. Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.* 16, 1370–1386.
- Karrer, B., Newman, M.E.J., 2011. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* 83, 016107.
- Kauffman, S.A., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theoret. Biol.* 22, 437–467.
- Kaufman, L., Rousseeuw, P.J., 2005. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley and Sons, New York.

- Kerr, G., Ruskin, H.J., Crane, M., Doolan, P., 2008. Techniques for clustering gene expression data. *Comput. Biol. Med.* 38, 283–293.
- Kim, K., Jiang, K., Teng, S.M., Feldman, L.J., Huang, H., 2012. Using biologically interrelated experiments to identify pathway genes in Arabidopsis. *Bioinformatics* 28, 815–822.
- Kim, K., Zhang, S., Jiang, K., Cai, L., Lee, I.B., Feldman, L.J., Huang, H., 2007. Measuring similarities between gene expression profiles through new data transformations. *BMC Bioinform.* 8, 29.
- Kim, S., Imoto, S., Miyano, S., 2004. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems* 75, 57–65.
- Kimura, S., Ide, K., Kashiwara, A., Kano, M., Hatakeyama, M., Masui, R., Nakagawa, N., Yokoyama, S., Kuramitsu, S., Konagaya, A., 2005. Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics* 21, 1154–1163.
- Kinney, J.B., Atwal, G.S., 2014. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci.* 111, 3354–3359.
- Kohl, M., Wiese, S., Warscheid, B., 2011. Cytoscape: software for visualization and analysis of biological networks. *Meth. Mol. Biol.* 696, 291–303.
- Kumari, S., Nie, J., Chen, H.S., Ma, H., Stewart, R., Li, X., Lu, M.Z., Taylor, W.M., Wei, H., 2012. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS ONE* 7, e50411.
- Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9, 559.
- Langfelder, P., Zhang, B., Horvath, S., 2008. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* 24, 719–720.
- Latouche, P., Birmelé, E., Ambroise, C., 2012. Variational Bayesian inference and complexity control for stochastic block models. *Stat. Modell.* 12, 93–115.
- Lemmens, K., Dhollander, T., De Bie, T., Monsieus, P., Engelen, K., Smets, B., Winderickx, J., De Moor, B., Marchal, K., 2006. Inferring transcriptional modules from chip-chip, motif and microarray data. *Genome Biol.* 7, R37.
- Li, H., Gui, J., 2006. Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* 7, 302–317.
- Li, K.-C., 2002. Genome-wide coexpression dynamics: theory and application. *Proc. Natl. Acad. Sci.* 99, 16875–16880.
- Li, K.-C., Palotie, A., Yuan, S., Bronnikov, D., Chen, D., Wei, X., Choi, O.-W., Saarela, J., Peltonen, L., 2007. Finding disease candidate genes by liquid association. *Genome Biol.* 8, R205.
- Liang, S., Fuhrman, S., Somogyi, R., 1998. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: *Pacific Symposium on Biocomputing*, pp. 18–29.
- Lim, J., Hao, T., Shaw, C., Patel, A.J., Szabó, G., Rual, J.F., Fisk, C.J., Li, N., Smolyar, A., Hill, D.E., Barabási, A.L., Vidal, M., Zoghbi, H.Y., 2006. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 125, 801–814.
- Ma, P., Castillo-Davis, C.I., Zhong, W., Liu, J.S., 2006. A data-driven clustering method for time course gene expression data. *Nucl. Acids Res.* 34, 1261–1269.
- Madeira, S.C., Oliveira, A.L., 2004. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1, 24–45.
- Maetschke, S.R., Madhamshettiwar, P.B., Davis, M.J., Ragan, M.A., 2014. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Brief. Bioinform.* 15, 195–211.
- Magwene, P., Kim, J., 2004. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.* 5, R100.
- Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla-Favera, R., Califano, A., 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform.* 7, S7.
- Markowitz, F., Kostka, D., Troyanskaya, O.G., Spang, R., 2007. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics* 23, i305–i312.
- Mazur, J., Ritter, D., Reinelt, G., Kaderali, L., 2009. Reconstructing nonlinear dynamic models of gene regulation using stochastic sampling. *BMC Bioinform.* 10, 448.
- Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* 34, 1049–1079.
- Mordelet, F., Vert, J.P., 2008. SIRENE: supervised inference of regulatory networks. *Bioinformatics* 24, i76–i82.
- Muro, S., Takemasa, I., Oba, S., Matoba, R., Ueno, N., Maruyama, C., Yamashita, R., Sekimoto, M., Yamamoto, H., Nakamori, S., Monden, M., Ishii, S., Kato, K., 2003. Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. *Genome Biol.* 4, R21.
- Murphy, K., 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning* (Ph.D. thesis). UC Berkeley, Computer Science Division.
- Needham, C.J., Bradford, J.R., Bulpitt, A.J., Westhead, D.R., 2007. A primer on learning in Bayesian networks for computational biology. *PLoS Comput. Biol.* 3, e129.
- Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. *Phys. Rev. E* 69, 026113.
- Nowicki, K., Snijders, T.A.B., 2001. Estimation and prediction for stochastic block-structures. *J. Am. Stat. Assoc.* 96, 1077–1087.
- Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., Kanehisa, M., 2008. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucl. Acids Res.* 36, W423–W426.
- Orth, J.D., Thiele, I., Palsson, B.O., 2010. What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248.
- Pearl, J., Verma, T., 1991. A theory of inferred causation. in: (KR 1991), pp. 441–452.
- Pe'er, D., Regev, A., Elidan, G., N., F., 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17, S215–S224.
- Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* 96, 4285–4288.
- Peng, J., Wang, P., Zhou, N., Zhu, J., 2009. Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Assoc.* 104, 736–746.
- Penrose, R., 1955. A generalized inverse for matrices. *Math. Proc. Cambridge Philos. Soc.* 51, 406–413.
- Price, N.D., Reed, J.L., Palsson, B.O., 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* 2, 886–897.
- Ramesh, A., Trevino, R., VON Hoff, D.D., Kim, S., 2010. Clustering context-specific gene regulatory networks. In: *Pacific Symposium on Biocomputing*, pp. 444–455.
- Rényi, A., 1959. On measure of dependence. *Acta Math. Academiae Scientiarum Hungarica* 10, 441–451.
- 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., 2011. Detecting novel associations in large data sets. *Science* 334, 1518–1524.
- Rohe, K., Chatterjee, S., Yu, B., 2011. Spectral clustering and the high-dimensional stochastic block model. *Ann. Stat.* 39, 1878–1915.
- Sabatti, C., James, G.M., 2006. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics* 22, 739–746.
- Savageau, M.A., 1991. Biochemical systems theory: operational differences among variant representations and their significance. *J. Theor. Biol.* 151, 509–530.
- Schäfer, J., Strimmer, K., 2005. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21 (6), 754–764.
- Sharan, R., Maron-Katz, A., Shamir, R., 2003. CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics* 19, 1787–1799.
- Shlomi, T., Cabili, M.N., Herrgård, M.J., Palsson, B.O., Ruppin, E., 2008. Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.* 26, 1003–1010.
- Shojaie, A., Jauhiainen, A., Kallitsis, M., Michailidis, G., 2013. Inferring regulatory networks by combining perturbation screens and steady state gene expression profiles. *arXiv:1312.0335*.
- Snijders, T.A.B., Nowicki, K., 1997. Estimation and prediction for stochastic block-models for graphs with latent block structure. *J. Classif.* 14, 75–100.
- Soler-López, M., Zanzoni, A., Lluís, R., Stelzl, U., Aloy, P., 2011. Interactome mapping suggests new mechanistic details underlying Alzheimer's disease. *Genome Res.* 21, 364–376.
- Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P. O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297.
- Spieth, C., Streichert, F., Speer, N., Zell, A., 2004. A memetic inference method for gene regulatory networks based on S-Systems. In: *Proceedings of the IEEE Congress on Evolutionary Computation*, pp. 152–157.
- Steuer, R., Kurths, J., Daub, C.O., Weise, J., Selbig, J., 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 18, S231–S240.
- Stuart, J.M., Segal, E., Koller, D., Kim, S.K., 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* 96, 2907–2912.
- Tanay, A., Sharan, R., Kupiec, M., Shamir, R., 2004. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci.* 101, 2981–2986.
- Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., 1999. Systematic determination of genetic network architecture. *Nat. Genetics* 22, 281–285.
- Teng, S.L., Huang, H., 2009. A statistical framework to infer functional gene relationships from biologically interrelated microarray experiments. *J. Am. Stat. Assoc.* 104 (486), 465–473.
- Teschendorff, A.E., Wang, Y., Barbosa-Morais, N.L., Brenton, J.D., Caldas, C., 2005. A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics* 21, 3025–3033.
- Theodoridis, S., Koutroumbas, K., 2005. *Pattern Recognition*, 4th edition Academic Press, San Diego, USA.
- Tresch, A., Markowitz, F., 2008. Structure learning in nested effects models. *Stat. Appl. Genetics Mol. Biol.* 7, 9.
- Wang, Y., Joshi, T., Zhang, X.S., Xu, D., Chen, L., 2006. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics* 22, 2413–2420.
- Wang, Y., Zhang, X.-S., Chen, L., 2009. A network biology study on circadian rhythm by integrating various omics data. *OMICS: J. Integr. Biol.* 13, 313–324.
- Wang, Y., Zhang, X.-S., Xia, Y., 2009. Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. *Nucl. Acids Res.* 37, 5943–5958.
- Wang, Y.X.R., Jiang, K., Feldman, L.J., Bickel, P.J., Huang, H., 2014. Inferring gene association networks using sparse canonical correlation analysis. *arXiv:1401.6504*.
- Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., Somogyi, R., 1998. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci.* 95, 334–339.
- Wille, A., Bühlmann, P., 2006. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genetics Mol. Biol.* 5, 1.

- Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., Bühlmann, P., 2004. Sparse graphical Gaussian modeling of the isoprenoid gene network in *arabidopsis thaliana*. *Genome Biol.* 5, 1–13.
- Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüss, M., Schacherer, F., Thiele, S., Urbach, S., 2001. The TRANSFAC system on gene expression regulation. *Nucl. Acids Res.* 29, 281–283.
- Wolfe, C.J., Kohane, I.S., Butte, A.J., 2005. Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. *BMC Bioinform.* 6, 227.
- Xiong, Q., Ancona, N., Hauser, E.R., Mukherjee, S., Furey, T.S., 2012. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Res.* 22, 386–397.
- Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L., 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977–987.
- Yeung, M.K., Tegnér, J., Collins, J.J., 2002. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci.* 99, 6163–6168.
- Yip, K.Y., Gerstein, M., 2009. Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics* 25, 243–250.
- Yizhak, K., Benyamini, T., Liebermeister, W., Rupp, E., Shlomi, T., 2010. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* 26, i255–i260.
- Yoo, C., Thorsson, V., Cooper, G., 2002. Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. In: *Pacific Symposium on Biocomputing*, pp. 498–509.
- Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D., 2002. Using Bayesian network inference algorithms to recover molecular genetic regulatory networks. In: *International Conference on Systems Biology*.
- Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D., 2004. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 20, 3594–3603.
- Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. *Biometrika* 94, 19–35.
- Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genetics Mol. Biol.* 4, 17.
- Zhou, S., Rütimann, P., Xu, M., Bühlmann, P., 2011. High-dimensional covariance estimation based on Gaussian graphical models. *J. Mach. Learn. Res.* 12, 2975–3026.
- Zhu, J., Chen, Y., Leonardson, A.S., Wang, K., Lamb, J.R., Emilsson, V., Schadt, E.E., 2010. Characterizing dynamic changes in the human blood transcriptional network. *PLoS Comput. Biol.* 6, e1000671.
- Zou, M., Conzen, S.D., 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* 21, 71–79.