Por Sae-Seaw
Final Paper – Bioc218
June 6, 2010

A review of graphical models for gene regulatory network inference using microarray data

Introduction

An important problem in systems biology is the inference of gene regulatory networks within complex living organisms. The availability of genome-wide gene expression technologies has enabled researchers to make considerable progress towards achieving this goal. With the description of complete genome sequences, DNA microarray technology allows scientists to simultaneously investigate the whole transcriptome on a single chip. However, the challenging task is to efficiently extract useful information and use the data to infer gene regulatory networks.

One way to tackle the problem of reverse engineering gene networks from DNA microarray data is by using multivariate probabilistic models, known as graphical models [1]. Graphical models are promising tools for the analysis of gene networks because they allow the stochastic description of net-like association and dependence structures in complex high-dimensional data, such as microarray data. In addition, they also offer an advanced statistical framework for inference [2]. Among various graphical models, Bayesian networks (BNs) and graphical Gaussian models (GGMs) have been used widely to infer gene regulatory networks from gene expression data. In this review, the two graphical models, BNs and GGMs, will be introduced, followed by a critical review of relevant computational methods for reverse engineering gene regulatory networks. These are summarized in Table 1.

Bayesian Networks

A Bayesian network is a graphical model for probabilistic relationships among a set of random variables, and the relationships are represented by a directed acyclic graph. In modeling gene networks, each node represents one gene, and the relationships between the genes are described by a joint probability distribution that captures properties of conditional independence between the genes. The genes on which the probability is conditioned are called the parent genes, and these parent genes regulate the child gene. However, BNs only model probabilistic dependencies among variables and not causality. Thus, the parents of a gene are not necessarily also the direct causes of its behavior [3]. BNs offer several advantages in inferring gene networks from microarray data. It can describe arbitrary combinatorial control of gene expression and thus is not limited to pairwise interactions between genes [4]. Moreover, due to their probabilistic nature, BN algorithms are capable of handling noisy data, which are often found in biological experiments [5].

BN algorithms, which are used to analyze steady-state data, are unable to infer networks involving cycles, such as feedback loops. This is the principal limitation of the BN models. However, a dynamic Bayesian network (DBN), which is an extension of BN, can be used to infer cyclic phenomena that are prevalent in biological systems. In addition, DBN algorithms can also infer direction of causality because they incorporate temporal information [6].

To find the network that best describes probabilistic relationships between variables, the score of each graph is calculated to find the graph with the maximum score. The Bayesian Information Criteria (BIC) and Bayesian Dirichlet equivalence (BDe) are the most popular scoring matrices, and both scores incorporate a penalty for complexity to guard against overfitting of data [7]. Because, ideally, all possible sets of directed acyclic graphs linking the genes must be assessed, learning BNs is computationally expensive as the number of graphs is super-exponential in the number of genes. Therefore, various heuristic search algorithms have been used instead in an attempt to optimize some scoring function. However, the problem with heuristic searches is that they often find local maxima and do not converge to the globally optimal solution [8].

Graphical Gaussian Models

In order to elucidate functional associations and infer gene networks from genome expression data, a popular and simple strategy is to compute the standard Pearson correlation between any two genes. An edge is drawn between two genes if the absolute pairwise correlation coefficient exceeds a prespecified threshold [9]. The resulting graph is called a relevance network where missing edges denote marginal independence. Although the advantages of the relevance network are its straight forward approach and low computational cost, this approach is only of limited use for understanding gene interaction. A high correlation coefficient between two genes could be due to direct interaction, indirect interaction, or regulation by a common gene. Therefore, relevance networks are powerful tools for determining “independence” between gene pairs (suggested by the absence of correlation), but not for elucidating the dependence network [10].

Graphical Gaussian models (GGMs), however, offer appropriate statistical strategies to construct gene association networks where only direct interactions among genes are depicted by edges. The key idea behind GGMs is to use partial correlations as a measure of conditional independence between any two genes. That is, the correlation between two genes is measured after the common effects of all other genes are removed. If the partial correlation is different from the standard correlation and approaches zero, it can be inferred that the original correlation is spurious, as the control genes might be either common anteceding cause, or intervening genes [11]. Thus, in these models an edge between two genes represents a direct interaction, and a path connecting two genes represents an indirect interaction mediated by other genes in the path [12]. GGMs are more powerful than relevance networks in describing gene networks, as non-zero correlated gene pairs would not be joined by an edge when they influence each other only indirectly through other genes. In contrast to BNs, GGMs are undirected graphical models, hence they are conceptually more simple and also do not suffer from a restriction inherent in BNs, which cannot contain feedback loops such as directed cycles. However, one disadvantage of the undirected edges in GGMs is that the resulting networks cannot describe directionality or causality [3].

For a microarray dataset, an observed expression data matrix contains *n* rows, corresponding to the samples from *n* different experimental conditions, and *p* columns, corresponding to the genes being probed. Under the GGM approach, the covariance matrix is calculated, and the partial correlation matrix is computed from the inverse of the covariance matrix. The GGM is then constructed based on the rule that no edge is included in the graph if the absolute value of partial correlation coefficient is less than some prespecified threshold [11]. Unfortunately, this standard GMM theory can only be applied when the sample size *n* is larger than the number of genes *p*. Otherwise, the sample covariance and correlation matrices are not positive definite and cannot be inverted, which in turn prevents the direct computation of partial correlation coefficients [13].

Although today's high-throughput facilities allow us to investigate experimentally a greatly increased number of features, the number of samples cannot be similarly increased. In a typical microarray dataset, the number of genes *p* is usually in the order of tens of thousands, but the number of observations *n* is in the order of tens. This poses a serious challenge to any statistical inference procedure. To cope with this "small *n*, large *p*" problem in GGMs, two main strategies have been proposed in the literature: computation of limited-order partial correlations, and use of shrinkage estimators of the covariance matrix to infer GGMs. Examples of methods that apply these strategies will be described in the next section.

Table 1: A list of computational methods for gene regulatory network inference from DNA microarray data presented in this paper.

Approach	Reference	Graphical Model	Key feature
Banjo	Yu *et al.* 2004	BN	Applies a novel influence score to estimate both the regulatory sign and relative magnitudes of the interaction.
Seeded Bayesian Networks	Djebbari & Quackenbush 2008	BN	Incorporates prior information about gene-gene interactions to seed the BN analysis.
Limited-order partial correlations	Wille *et al.* 2004	GGM	Employs first-order conditional independence instead of computing full-order partial calculations as in a full GGM.
Shrinkage covariance estimators	Schäfer & Strimmer 2005b	GGM	Uses a shrinkage approach to obtain reliable covariance estimators
GeneNet	Opgen-Rhein & Strimmer 2007	GGM	Introduces standardized partial variance to convert undirected GGMs into partially directed graphs

Banjo: Yu *et al.* 2004

Banjo is a gene network inference software that has been developed by Yu *et al.* (2004). It is based on BN algorithms and implements both BN and DBN. Therefore, it can analyze both steady-state and time-series data. Heuristic approaches are used to search the network space to find the graph with the best score, which is computed using the BDe metric.

To recover more meaningful, more interpretable, and more accurate networks, Yu *et al.* developed a novel *influence score* for BN interactions that attempts to estimate both the regulatory sign (activation (+) or repression (-)) and relative magnitudes of the interactions between a child variable and each of its parents. This score is computed from the conditional probability that a child is in one expression state given that its parent set is in another state. Intuitively, a parent is presumably an activator if there is a high probability that a child is in a high expression state when the parent is highly expressed and that the child is in a low expression state when the parent is low expressed. Conversely, a parent is presumably a repressor if there is a high probability that a child is in a low expression state when the parent is highly expressed and that the child is in a high expression state when the parent is low expressed. When an influence score is zero or close to zero, it means either that the sign of regulation is difficult to determine, or the regulation strength is very weak. Thus, this is useful in eliminating low-scoring false positive links, although the number of true positives is not increased [5].

The use of the influence score offers a particular strength to Banjo as identifying activation and repression interactions is very important in the study of biological systems. Because Banjo employs heuristic methods, one weakness of Banjo is that the search could be trapped in local maxima [5]. As a consequence of the DBN algorithms and the influence score, Banjo is very accurate, but it has low sensitivity. Banjo is a probabilistic algorithm that requires the estimation of probability density distributions, therefore its performance depends on large number of data points. It works relatively well in recovering the true network with large quantities of data. As the quantity of sampled data is decreased, influence score representation is also decreased, but the accuracy of the designated sign is not affected [7]. Due to the limitation of statistical inference methods, more data will be required to learn more complex networks accurately. Using other types of gene association data together with microarray expression data can potentially enhance the ability of BNs to accurately recover gene network structures [14].

Seeded Bayesian Networks: Djebbari and Quackenbush 2008

To cope with the limitations of learning BNs from the imperfect microarray datasets that usually provide too few data points to constrain potential models and from the local maxima problem of heuristic algorithms, Djebbari and Quackenbush (2008) proposed a new BN approach to construct genetic networks from microarray data. They employed prior knowledge of preliminary network topologies to provide a useful bias serving as soft constraints and to seed the search for a network graph with the best topology.

Djebbari and Quackenbush applied the co-occurence method described in Jenssen *et al.* (2001) to infer potential functional associations between genes [15]. For instance, if two and only two genes are described in a single article indexed in PubMed, then an interaction between them is assumed. This method assigns a co-occurrence edge weight, which counts the number of times an interaction appears in the literature relative to the total number of manuscripts surveyed, as prior probabilities of interactions between the two genes. In addition, by limiting networks to papers containing two and only two genes, they can remove publications that include whole-genome studies and generate network topologies exhibiting a scale-free behavior. Although limiting to two genes is conservative and some interactions might be missed, it allows a prior network to have the highest possible confidence without resorting to more ambitious text-mining approaches [8].

In addition to the literature, prior network structures are also deduced from high-throughput yeast two hybrid protein-protein interaction screens. These datasets represent an unbiased screen for interactions, and thus the protein-protein interaction networks have a uniform distribution for the prior probabilities for all edges. Because a BN is a directed acyclic graph, the initial network used to seed the search must be directed graphs as well. Edges in the undirected literature and protein-protein interaction networks are subject to direction assignments using a depth-first search algorithm, which is commonly used for cycle detection [16]. Instead of using heuristic search, the authors perform model averaging through non-parametric bootstrapping (resampling with replacement), which allows them to assign confidence values to the individual interactions [8].

To compare the performance of BN analyses with and without network seeds, Djebbari and Quackenbush compared the resulting networks to known pathways based on the KEGG pathway database and evaluated the ability to reproduce known interactions between genes. The use of prior network seeds improves the ability of BN analyses to learn known interactions between genes relative to a standard, unseeded BN analysis. By varying the bootstrap confidence threshold, the authors could show a tradeoff between sensitivity and specificity in detecting interactions. Using high confidence thresholds yields high specificity but low sensitivity; many interactions are missed, including potential novel interactions. Although this seeded BN approach outperforms a standard BN analysis in recovering known interactions and can at least extract network graphs from a gene list and refine the graph using expression data, its ability in discovering new interactions and build testable networks is still questionable. This will depend on one's ability to manage the tradeoff between specificity and sensitivity and to validate the resulting networks, as lowering the confidence threshold not only increases sensitivity, but also leads to the identification of many spurious edges [8].

Limited-order partial correlations: Wille *et al.* 2004

Wille *et al.* (2004) used a modified GGM approach based on limited-order partial correlations to avoid the "small *n*, large *p*" problem. To explore dependencies between two genes, Wille *et al.* do not jointly condition on all remaining genes at a time. Instead, all pairwise partial correlations are considered separately. The conditioning set is restricted to single variables. Thus, this method is computed for so called *first-order partial correlation coefficients*. The sample Pearson's correlation coefficient between two genes A and B is computed to measure coexpression. Then, for all triples of genes A, B, C, effects of the other gene C on the correlation coefficient are examined by computing the partial correlation coefficient conditioned on C and not on all other *p*-1 genes. Similarly to the full GGM approach, if the expression level of C is independent of A and B, the partial correlation coefficient would not differ from the standard correlation coefficient. However, if gene C coregulates both genes, the partial correlation coefficient is expected to be close to zero. To combine these into a network and identify direct coregulation between genes, an edge between two genes will be drawn when their pairwise correlation is not the effect of a third gene. However, if there is at least one C that makes the partial correlation coefficient equal or close to zero, no edge will be drawn between the two genes [17].

Due to the simplification in modeling small subnetworks of three genes, this approach offers two advantages. First, it can avoid the dimensionality problem that occurs when trying to estimate very high-order conditional interactions. Thus, this approach can be applied to datasets with moderate sample sizes. Second, because the running time required to calculate conditional correlations increases at least exponentially as the order of interactions increases, restricting to first-order conditional interactions reduces the computational cost [17]. However, from a statistical point of view, the resulting network constitutes something inbetween a full GGM (full-order) and a relevance network (zero-order) model based on standard correlations. Therefore, missing edges could indicate either conditional or marginal independence [18].

Shrinkage covariance estimators: Schäfer and Strimmer 2005b

When the number of genes *p* by far exceeds the number of available samples *n*, standard GGM methods cannot be used to obtain partial correlation coefficients due to statistical unreliability of small samples. A simple approach frequently used to reduce variance is bootstrap aggregation or bagging. Schäfer and Strimmer (2005a) applied this strategy to improve the accuracy of estimates of the correlation and covariance matrices. Once regularized estimates of partial correlation are identified, they employed an heuristic based on empirical Bayes multiple testing in order to find an optimal network. However, the bootstrap variance reduction is computationally expensive, especially when dealing with several thousands of genes in the genomic settings [2].

Due to this drawback, Schäfer and Strimmer (2005b) have proposed an alternative method to obtain reliable covariance estimators by using a shrinkage estimator, also known as a biased estimator [10]. This method was first introduced by Ledoit and Wolf (2003) [19]. The estimator shrinks the sample covariance matrix towards a low-dimensional (biased) estimator of the covariance matrix. The biased estimator is constrained and contains fewer parameters than the unconstrained, unbiased one. Thus, the constrained estimator will exhibit considerable bias and a lower variance than its unconstrained counterpart. Both estimators are combined to generate a linear shrinkage estimator, instead of choosing between one of the two extremes.

Therefore, the shrinkage estimator can be used to minimize the mean squared error by finding the best trade-off between error due to bias and error due to variance [10].

The ability of this method on inferring gene networks was illustrated by applying to a real microarray dataset to reverse engineer an *E. coli* subnetwork. Schäfer and Strimmer showed that the covariance shrinkage estimator provides large overall gains in the accuracy and in the power to recover the true network structure suggested by the hub topology compared with their previous approach described in Schäfer and Strimmer (2005a) [2]. Moreover, this method is much less computationally expensive.

GeneNet: Opgen-Rhein and Strimmer 2007

GGM is a correlation network, which is an undirected graph that does not describe directionality. Moreover, correlations not only confound direct and indirect interactions, but also do not distinguish between cause and effect. Thus, it is only of limited use in representing the causal processes such as gene regulatory networks. Therefore, causal analysis usually requires different algorithms, such as Bayesian networks, which describe causal relations by a directed acyclic graph (DAG). However, these methods generally work well only when dealing with small numbers of variables and with large sample size, which is usually not the case for microarray data [20]. Thus, Opgen-Rhein and Strimmer (2007) have proposed a new approach implemented in GeneNet, which convert correlation (GGM) into causation networks.

GeneNet applied an algorithm which is an extension of the GGM inference approach proposed in Schäfer and Strimmer (2005b). In the first step, the correlation network is transformed into a partial correlation network using the covariance shrinkage estimator to uncover topology of the network. Secondly, the undirected GGM is converted into a partially directed graph. Edges are removed from the independence graph to obtain the underlying DAG. An undirected edge between genes A and B in a partial correlation graph can be interpreted as a bidirected edge, in the sense that gene A influences gene B and vice versa. Thus, a directed edge can be implied by removing one of these two directions [20].

In GeneNet, Opgen-Rhein and Strimmer introduced a new term called *standardized partial variance* (SPV), which measures the proportion of variance that cannot be explained by other variables. In another word, SPV equals the ratio of the partial variance to the variance, and the value ranges from 0 to 1. Thus, when there is no correlation between two genes and the partial variance equals the variance, SPV = 1. If gene A is the cause of gene B and A is parentless (A → B), B would tend to be better explained than A, which implies that B has a lower SPV. Based on this intuition, they can impose directionality from the less well explained variable (large SPV, independent variable) to the more well explained one (lower SPV, dependent variable) [20].

They evaluated the algorithm for discovering causal structure by analyzing a large *Arabidopsis thaliana* expression dataset. The resulting graph is a partially directed network containing both directed and undirected nodes. This demonstrates a distinct advantage of this approach because it does not force directions onto the edges when directions cannot be determined due to complex interactions among the nodes. Since this approach is approximate and non-iterative, it is computationally inexpensive and allows screening large-scale dataset for causal structure. However, it lacks iterative refinements in the algorithm to remove spurious edges, which can be very time consuming [20].

Conclusion

One of the main problems for all statistical inference algorithms when applied to typical microarray data is the curse of dimensionality when the number of genes by far exceeds the number of observations. Proposed BN and GGM approaches in the literature employ different strategies to cope with this problem. Currently, there is not one approach fits all. Different situations may require different reverse engineering strategies. The value of each approach discussed here will be judged based on its utility. For example, one may use GeneNet first to screen large-scale dataset for causal structure to predict a testable model as this approach is not computationally expensive. Subsequently, a more computationally expensive approach, such as full BN modeling, can be applied to refine the network. Since the curse of dimensionality is still an intrinsic problem in genetic network inference from microarray data, novel statistical methods remain to be discovered in order to overcome this limitation in systems biology.

References

1. Whittaker, J. (1990). Graphical models in applied multivariate statistics. New York: Wiley.

2. Schäfer, J., and Strimmer K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21(6): 754-764.

3. Werhli, A.V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics* 22(20): 2523-2531.

4. Hartemink, A.J., Gifford, D., Jaakkola, T., and Young, R. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* 422-433.

5. Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., and Jarvis, E.D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 20(18): 3594-3603.

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

7. Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Sys. Biol.* 3:78.

8. Djebbari, A., and Quackenbush , J. (2008). Seeded Bayesian networks: Constructing genetic networks from microarray data. *BMC Sys. Biol.* 2:57.

9. Butte, A.J., and Kohane, I.S. (2000). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac. Sym. Biocomput.* 5:415-426.

10. Schäfer, J., and Strimmer K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. App. in Gen. & Mol. Biol.* 4(1): 32.

11. Wu, X., Ye, Y., and Subramanian, K.R. (2003). Interactive analysis of gene interactions using graphical Gaussian model. *ACM SIGKDD Workshop on Data Mining in Bioinformatics* 3:63-69.

12. Jones, B., and West, M. (2005). Covariance decomposition in undirected Gaussian graphical models. *Biometrika* 92(4): 779-786.

13. Hastie, T., Tibshirani, R., and Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.

14. Hartemink, A.J., Gifford, D., Jaakkola, T., and Young, R. (2002). Combining location and expression data for principled discovery of genetic regulatory network. *Pac. Symp. Biocomput.* 7:436-449.

15. Jenssen, T.K., Laegreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28(1): 21-28.

16. Cormen, T.H., Leiserson, C.E., and Rivest, R.L. (1990). Introduction to Algorithms. Cambridge: MIT Press.

17. Wille, A., Zimmerman, P., Vránova, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelić, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P., Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.* 5: R92.

18. Schäfer, J., and Strimmer K. (2005c). Learning large-scale graphical Gaussian models from genomic data. *AIP Conference Proceedings*. 776: 263-276.

19. Ledoit, O., and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance*, 10: 603-621.

20. Opgen-Rhein, R., and Strimmer, K. (2007). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Sys. Biol*. 1:37