BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 15 2012, pages 2029–2036
doi:10.1093/bioinformatics/bts312

Systems biology Advance Access publication June 8, 2012

Empirical Bayes conditional independence graphs for regulatory network recovery

Rami Mahdi[1,*], Abishek S. Madduri[1], Guoqing Wang[1], Yael Strulovici-Barel[1], Jacqueline Salit[1], Neil R. Hackett[1], Ronald G. Crystal[1] and Jason G. Mezey[1,2,*]

[1]Department of Genetic Medicine, Weill Cornell Medical College, New York, NY 10065, USA and [2]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Computational inference methods that make use of graphical models to extract regulatory networks from gene expression data can have difficulty reconstructing dense regions of a network, a consequence of both computational complexity and unreliable parameter estimation when sample size is small. As a result, identification of hub genes is of special difficulty for these methods.

Methods: We present a new algorithm, Empirical Light Mutual Min (ELMM), for large network reconstruction that has properties well suited for recovery of graphs with high-degree nodes. ELMM reconstructs the undirected graph of a regulatory network using empirical Bayes conditional independence testing with a heuristic relaxation of independence constraints in dense areas of the graph. This relaxation allows only one gene of a pair with a putative relation to be aware of the network connection, an approach that is aimed at easing multiple testing problems associated with recovering densely connected structures.

Results: Using *in silico* data, we show that ELMM has better performance than commonly used network inference algorithms including GeneNet, ARACNE, FOCI, GENIE3 and GLASSO. We also apply ELMM to reconstruct a network among 5492 genes expressed in human lung airway epithelium of healthy non-smokers, healthy smokers and individuals with chronic obstructive pulmonary disease assayed using microarrays. The analysis identifies dense sub-networks that are consistent with known regulatory relationships in the lung airway and also suggests novel hub regulatory relationships among a number of genes that play roles in oxidative stress and secretion.

Availability and implementation: Software for running ELMM is made available at http://mezeylab.cb.bscb.cornell.edu/Software.aspx.

Contact: ramimahdi@yahoo.com or jgm45@cornell.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 31, 2011; revised on May 1, 2012; accepted on May 22, 2012

1 INTRODUCTION

A combination of microarray and next-generation sequencing are producing data on the gene expression profiles that characterize the genomic and regulatory states of cell populations (Blencowe *et al.*, 2009; Shendure and Ji, 2008). These data contain information that can be mined to identify regulatory relationships responsible for genomic profiles and that are important for understanding processes underlying disease development (Sotiriou and Pusztai, 2009; Volinia *et al.*, 2006). A variety of algorithms making use of probabilistic graphical models, both directed and undirected, have been used to recover putative network structure from such data (Bansal *et al.*, 2007; Marbach *et al.*, 2010; Markowetz and Spang, 2007; Schlitt and Brazma, 2007). These approaches have been successful in recovering networks with broad-scale topology which are consistent with gene and pathway annotations, where some of the predicted relationships have been verified by experiments (Akavia *et al.*, 2010; Faith *et al.*, 2007). To deal with the high dimensionality and small sample size of training data that is typical of gene expression experiments, researchers have used various strategies, including regularized estimation of parameters (Friedman *et al.*, 2008; Li and Li, 2008), learning sparse networks (Aliferis *et al.*, 2003; Kalisch *et al.*, 2010), empirical estimation of testing statistics (Schäfer and Strimmer, 2005) and mutual information methods (Margolin *et al.*, 2006). These approaches were shown to improve the accuracy of the network inference especially when the recovered network is sparse. In spite of these advances, the goal of highly accurate recovery of regulation networks, which are not necessarily sparse, remains a difficult problem (Baralla *et al.*, 2009; Marbach *et al.*, 2010).

In this article, we present a new algorithm, Empirical Light Mutual Min (ELMM), for reconstructing undirected networks from gene expression data, which has properties well suited for recovering dense network regions when the size of the observed samples is small. ELMM is a variant of Bayesian network (BN) inference approaches that include reconstruction of an undirected graph, the skeleton of the BN, using conditional independence (CI) testing (Spirtes *et al.*, 2001). ELMM uses an empirical Bayes approach to learn the independence/dependence parameters from gene expression data, an approach that has been found to have good properties for dealing with small samples and noisy data (Schäfer and Strimmer, 2005). In addition, ELMM uses a heuristic relaxation of independence testing that aims at easing the effects of the multiple testing problem when recovering densely connected sub-graphs.

We assessed the performance of ELMM using 10 *in silico* networks simulated based on known regulatory mechanisms and network topologies in yeast and *Escherichia coli*. The analysis shows that ELMM has a higher accuracy compared with popular network recovery algorithms including GeneNet (Schäfer *et al.*,

*To whom correspondence should be addressed.

2006), ARACNE (Margolin *et al.*, 2006), GLASSO (Friedman *et al.*, 2008), FOCI (Magwene *et al.*, 2004) and GENIE3 (Huynh-Thu *et al.*, 2010) in terms of both overall network inference and identification of hub genes. We also present results from using ELMM on the network inference challenge of DREAM5 with a comparison to the best performing methods (see Supplementary materials). In addition, to demonstrate the utility of this method for real-world applications, we applied ELMM to model interactions among 5492 genes expressed in the human lung airway measured in a sample that included healthy non-smokers, healthy smokers and individuals with chronic obstructive pulmonary disease (COPD). Analysis of the recovered network based on the literature shows an overall enrichment for connecting functionally similar genes four times higher than expected at random. In addition, further analysis of dense sub-networks shows that they are consistent with known regulatory relationships and pathways. A number of suggestive functional annotations and relationships among genes in the recovered network are discussed, including novel regulatory relationships among a number of genes modulated by NRF2 that operate in oxidative stress response and a suggestive hub role of CREB3L1 which operates in mucus secretion in the lung.

2 BACKGROUND: NETWORK INFERENCE WITH CONDITIONAL INDEPENDENCE GRAPHS

As with related algorithms that use graphical models to recover regulatory networks (Friedman *et al.*, 1999; Schäfer and Strimmer, 2005), we assume that regulatory relationships among variables can be divided into direct and indirect relations. Assuming a case where relevant network variables have been measured, conditional independence testing provides a theoretically sound approach for disentangling these relations. For example, consider the hypothetical directed regulatory relationships represented by the solid arrows in Figure 1. When measurements of gene expressions generated by such networks are analyzed, a correlation can be observed among pairs of directly interacting genes, as well as the indirectly related genes. Using tests of conditional independence, the dependence between indirectly related genes can be negated by conditioning on a set of variables along the paths through which the indirect relation is propagated (Neapolitan, 2004; Spirtes *et al.*, 2001). For example, genes *B* and *C* in Figure 1a are expected to be correlated because they are co-regulated by gene *A*, but conditioning on *A* can reveal that *B* and *C* are independent. Similarly, in Figure 1b, genes *D* and *G* become independent when conditioning on either *E* or *F*. When the indirect effect is propagated through multiple paths, such as the case of the genes *K* and *L* in Figure 1c, a higher order conditional independence test is needed where *K* and *L* can only be found independent when conditioning on the genes *H* and *M* simultaneously. In contrast, directly interacting genes cannot be found conditionally independent when conditioning on any set of other genes.

Using the properties of direct/indirect relations, the graph of direct interactions can, in theory, be recovered from observational data by recovering a graph, i.e. consistent with the dependencies and conditional independence entailed by the observed data. This type of graphs is typically referred to as a CI graphs. CI inference algorithms typically construct the graph by connecting all correlated pairs of variables, unless they can be found conditionally independent when conditioning on any set of other variables (Spirtes *et al.*, 2001).

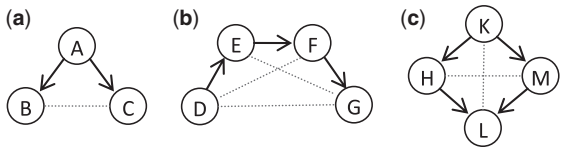

Fig. 1. Examples of three (a–c) regulatory network relationships among measured genes (circles). A solid arrow represents a direct regulatory relationship between genes. Dotted lines represent indirect relations between genes, where the corresponding genes are expected to be correlated

Usually, the graph is constructed iteratively and the search for possible conditional independence is restricted to conditioning on locally neighboring nodes in the graph.

Under the faithfulness assumption (Neapolitan, 2004; Spirtes *et al.*, 2001) and when using the appropriate independence testing method, CI inference is proven to recover the correct graph in the asymptotic limit. However, in spite of this theoretical soundness, there are known challenges for the application of such methods for reconstructing regulation networks from gene expression data:

1. Due to non-ideal sampling, such as non-independence among samples and batch effects during the measurement process, standard parametric conditional independence testing that ignores these contributions can be unreliable for constructing accurate graphs from gene expression data (Schäfer and Strimmer, 2005).

2. Edges in dense areas of a graph are more likely to be incorrectly rejected due to multiple testing ((Tsamardinos and Brown, 2008); R. Mahdi and J. Mezey, submitted for publication). For example, to correctly connect the parent node *A* in Figure 2a to all of its child nodes 1–7, all independence tests for each edge, when conditioning on all subsets of all other neighbors, must be correctly rejected. This conservative approach causes the probability of rejecting a true edge to increase with the number of neighboring variables.

3. For highly connected genes, conditioning on all subsets of all neighboring genes can be computationally impractical.

3 METHODS: EMPIRICAL LIGHT MUTUAL MIN ALGORITHM

ELMM reconstructs an undirected graph representing regulatory network relations using an iterative constraint-based approach. The algorithm incorporates two new interconnected components aiming at improving the inference of the graph. First, a mutual dependence criterion is used to measure the unexplainable dependence between pairs of variables. The proposed criterion is an extension of the empirical Bayes approach previously proposed for graphical Gaussian model (GGM) inference (Schäfer and Strimmer, 2005), where the goal is to assess the mutual dependence relationships among variables in a manner that empirically adjusts for non-ideal sampling conditions. Second, to ease the multiple testing problem in dense areas of the graph and the associated high computational complexity, we use a heuristic procedure for relaxing independence constraints by selectively allowing only one gene of a pair with a putative relation to use the other gene in independence testing. This relaxation also results in a fast learning that can scale for large network recovery (see Supplementary material for complexity analysis).

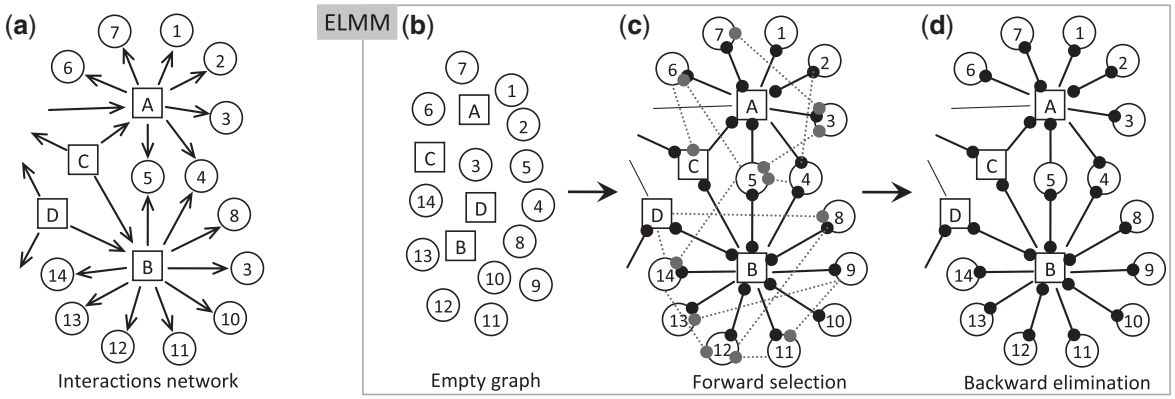

Fig. 2. Example of a regulatory network (**a**), where arrows indicate direct regulation. ELMM starts with empty graph (**b**) and performs forward selection (**c**) followed by backward elimination (**d**) based on edge ranking. Dotted lines (**c**) represent candidate false discovery edges that can be eliminated in the backward elimination. The undirected graph (**d**) is one possible output of ELMM where a solid circle at the end of the edge indicates the node is aware of the neighbor at the other end of the edge

3.1 Mutual dependence criterion

In order to rank candidate edges, we use an estimate of the joint conditional posterior probability that the dependence between the corresponding two variables cannot be explained by conditioning on any of their neighbors, which is equivalent to the posterior probability that the two variables are connected in the correct graph. This criterion is defined as follows:

$$P(E_{ij}|CP_i,CP_j,D)\cong P(E_{ij}|CP_i,D)\times P(E_{ij}|CP_j,D), \quad (1)$$

where E_{ij} is an edge between nodes i and j, CP_i and CP_j are the sets of neighboring variables of i and j, respectively, and D is the observed data. In equation (1), $P(E_{ij}|CP_i,D)$ is a one-sided conditional posterior probability that the variables i and j are not conditionally independent when conditioning on any subset of CP_i. The factorization of the joint conditional posterior in equation (1) is approximate because it ignores a possible correlation between the results of independence tests when conditioning on subsets of CP_i or subsets of CP_j. We also refer to $P(E_{ij}|CP_i,CP_j,D)$ as a conditional posterior because we condition on a certain graph structure that implies CP_i and CP_j are the neighbor sets of i and j, respectively. For the rest of the article, we refer to $P(E_{ij}|CP_i,D)$ as the one-sided conditional posterior probability of dependence (CPPD) of i on j and we refer to $P(E_{ij}|CP_i,CP_j,D)$ as the joint CPPD between i and j.

A critical part of the approximation in equation (1) is the need to estimate the one-sided conditional posterior probability $P(E_{ij}|CP_i,D)$ that a variable i is not conditionally independent from a variable j when CP_i is thought to be the set containing the parent and child variables of i (and similarly for variable j). In this work, without loss of generality, the partial correlation method is used for conditional independence testing, an approach that is appropriate for the continuous measurements typical of gene expression data. Partial correlations can also be computed efficiently, which in turn makes large-scale network inference more practical. The partial correlation ρ_{ij} is a measure of the statistical dependence between two variables i and j, while the conditional partial correlation $\rho_{ij|S}$ is a measure of the dependence between i and j that cannot be explained by the set of variables S. Since the true partial correlations cannot be computed exactly from limited data, we will have to rely on independence tests based on the sample partial correlations (i.e $\hat{\rho}_{ij}$, $\hat{\rho}_{ij|S}$) (see Supplementary material).

In order to compute $P(E_{ij}|CP_i,D)$ from the sample partial correlations, we need to compute the probability that none of the true partial correlations is insignificant when conditioning on all subsets of CP_i. Although there does not seem to be simple and unique formulation of such a conditional posterior, it is intuitive that it should have an inverse relation with the evidence that one or more partial correlations are zero. For example, one way to approximate $P(E_{ij}|CP_i,D)$ is to use the probability that the least significant conditional partial correlation is actually significant as follows:

$$P^{(1)}(E_{ij}|CP_i,D)=\min_{S\subseteq CP_i} P(\rho_{ij|S}\neq 0|\hat{\rho}_{ij|S}). \quad (2)$$

Using equation (2) to approximate $P(E_{ij}|CP_i,D)$ is equivalent to the approach used by regular constraint-based methods (Kalisch and Bühlmann, 2007; Tsamardinos *et al.*, 2006), where i and j are inferred to be not directly related if at least one conditional independence test is not rejected. However, this approach can be dependent on the size of CP_i, where $P^{(1)}(E_{ij}|CP_i,D)$ tends to decrease as CP_i becomes larger due to the increasing probability of finding small sample conditional partial correlations by chance due to the large number of tests. To ease this issue, we use a smoother dependence function that combines multiple partial correlations instead of the overall minimum as follows:

$$P^{(2)}(E_{ij}|CP_i,D)=\left(\prod_{z=0}^{Z} P_z(E_{ij}|CP_i,D)^{w_z}\right)^{1/\sum w_z}, \quad (3)$$

where

$$P_z(E_{ij}|CP_i,D)=\min_{S\subseteq CP_i,|S|\leq z} P(\rho_{ij|S}\neq 0|\hat{\rho}_{ij|S})$$

and

$$w_z=\begin{cases} 1 & \text{if } z=0 \text{ or } P_z(E_{ij}|CP_i)<P_{z-1}(E_{ij}|CP_i) \\ 0 & \text{otherwise} \end{cases}.$$

What is important to recognize about either approximation (2) or (3) is that when given a large enough sample, both of them will approach zero if and only if CP_i contains a subset of variables that fully explains the dependence between i and j ($\exists S\subseteq CP_i\, s.t.\, i \perp\!\!\!\perp j|S$), whereas, they will go to 1 if such a subset does not exist. In other words, both of these approximations are sufficient to recover the correct graph in the asymptotic limit (R. Mahdi and J. Mezey, submitted for publication). However, when given a small sample data, the approximation in equation (2) is expected to be sensitive to multiple testing, while the approximation in equation (3) eases this issue by taking the geometric average over the least significant partial correlations resulting from considering different sizes of the conditioning sets. This provides a smoothing effect that eases the problem that different nodes in the graph have different numbers of neighbors.

3.2 Semi-parametric independence testing

To compute $P(\rho_{ij|S}\neq 0|\hat{\rho}_{ij|S})$ which is required in equation (3), we use a variant of the empirical Bayes approach proposed by Schäfer and Strimmer (2005), where the partial correlations are assumed to be generated from a

mixture of a null and an alternative distribution ($H_0 : \rho = 0$ and $H_A : \rho \neq 0$). Using Bayes rule of posterior probability, $P(\rho_{ij|S} \neq 0 | \hat{\rho}_{ij|S})$ is computed as follows:

$$P(\rho_{ij|S} \neq 0 | \hat{\rho}_{ij|S}) = \frac{\pi_A \times f_A(\hat{\rho}_{ij|S})}{\pi_A \times f_A(\hat{\rho}_{ij|S}) + \pi_0 \times f_0(\hat{\rho}_{ij|S}, \hat{\kappa}_{i,h})}, \quad (4)$$

where $\hat{\rho}_{ij|S}$ is the sample partial correlation, f_0 and f_A are the distribution of the sample partial correlations under the null and the alternative hypothesis, respectively, while π_0 and π_A are the corresponding prior probabilities. Since the sample partial correlation between truly related variables is likely to be dependent on the true structure of the network, which is typically unknown, for the alternative hypothesis ($H_A : \rho_{ij|S} \neq 0$), we make use of a simplifying assumption that the sample partial correlations follow a uniform distribution in the range $[-1, 1]$ ($f_A(\hat{\rho}_{ij|S}) = 0.5, \forall \rho$). On the other hand, from Hotelling (1953), when the true partial correlation is zero ($H_0 : \rho_{ij} = 0$), the sample partial correlation $\hat{\rho}_{ij}$ has the following distribution:

$$f_0(\hat{\rho}_{ij}) = (1 - \hat{\rho}_{ij}^2)^{(\kappa-3)/2} \frac{\Gamma(\kappa/2)}{\pi^{1/2}\Gamma[(\kappa-1)/2]}, \quad (5)$$

where Γ is the gamma function and κ is the distribution degree of freedom. Under the assumption of ideal sampling conditions, the degree of freedom becomes the number of observations ($\kappa = n$) and when i and j are conditionally independent upon knowing the set $S_{\setminus i,j}$ ($H_0 : \rho_{ij|S} = 0$), the sample conditional partial correlation $\rho_{ij|S}$ also follows the same distribution but with a degree of freedom ($N - |S| + 1$). However, since ideal sampling conditions do not hold for gene expression data, the actual degree of freedom κ of the null distribution is unknown and could vary considerably from the expected value. To estimate κ and the unknown mixture proportions of the null and alternative hypotheses π_0 and π_A, as in Schäfer and Strimmer (2005), we use likelihood maximization of observing the zero-order sample partial correlations (conditioning on the empty set:ϕ) generated from the mixture model as follows:

$$(\hat{\kappa}, \hat{\pi}_0, \hat{\pi}_A) = arg \max_{\kappa, \pi_0, \pi_A} \prod_{i \neq j} f(\hat{\rho}_{ij} | \kappa, \pi_0, \pi_A), \; s.t. \; \pi_0 + \pi_A = 1, \quad (6)$$

where

$$f(\hat{\rho}_{ij} | \kappa, \pi_0, \pi_A) = \pi_0 f_0(\hat{\rho}_{ij}, \kappa) + \pi_A f_A(\hat{\rho}_{ij}) \quad (7)$$

is the likelihood of observing the partial correlation $\hat{\rho}_{ij}$ generated by the mixture model. In this work, we take this approach a step further and empirically estimate the degree of freedom, for every single variable separately, for both partial correlations and conditional partial correlations separately as follows:

$$\hat{\kappa}_{i,h} = arg \max_{\kappa} \prod_{i \neq j, |S| = h} f(\hat{\rho}_{ij|S} | \kappa, \hat{\pi}_0, \hat{\pi}_A), \quad (8)$$

where $\hat{\kappa}_{i,h}$ is the degree of freedom of the null distribution of the conditional partial correlations of the variable i with other variables when conditioning on sets of size h. This approach has the added benefit of adjusting the degree of freedom when the conditioning sets vary in size. In addition, it is anticipated that, as with gene expression measurements data, the effect of latent variables and noise in measurements is likely to vary between genes. Therefore, the actual degree of freedom is expected to be different for different genes and sizes of conditioning sets. Note that the likelihood estimation in equation (8) is performed over the sample conditional partial correlations between a variable i and all other variables when conditioning on sets of variables found to be the most correlated with i. To maximize equations (6) and (8), we use a grid search to maximize the log likelihood.

3.3 Heuristic relaxation of independence constraints

Even with the smoothing approximation of equation (3), when a node in the graph becomes highly connected, the chance that at least one conditioning independence test will be accepted by accident increases due to the large number of tests. As a result, the approximation of the one-sided conditional probability $P(E_{ij}|CP_i, D)$ in equation (3) will tend to decrease as the number of neighboring variables of i increases. To control for this issue, we incorporate a heuristic relaxation of independence constraints that aims at limiting the decrease in the one-sided conditional posterior that occurs as the number of neighbors of a given node increases by adaptively limiting the number of neighbors used for conditional independence testing.

In general, since every node in any graph has its own neighbor set, it is anticipated that most pairs of variables will have asymmetric one-sided dependence (i.e. $P(E_{ij}|CP_i, D) \neq P(E_{ij}|CP_j, D)$), where the node with the lower one-sided CPPD is more likely to be the node with the larger neighbor set. ELMM algorithm takes advantage of this asymmetry to identify node with the lower one-sided CPPD as the node that might be suffering from multiple testing when assessing edges connecting to this node. Those edges, in turn, become candidates for constraint relaxation. To do this, the algorithm relaxes independence constraints as the graph $\hat{G}$ is being constructed by allowing only the node with the higher one-sided CPPD to become aware of a newly added edge E_{ij} as follows:

if $P(E_{ij}|CP_i, D) \geq P(E_{ij}|CP_j, D)$ **then**
  $CP_i \leftarrow CP_i \cup j$
end if
if $P(E_{ij}|CP_j, D) \geq P(E_{ij}|CP_i, D)$ **then**
  $CP_j \leftarrow CP_j \cup i$
end if

As a result of this procedure, during network recovery, a node connected to many neighbors will only be aware of a fraction of them, while a node with a small number of connections will tend to be aware of most or all of its neighboring nodes (see supplement for the complete pseudo-code). For example, when recovering an undirected graph of the network in Figure 2a where a hub such as B regulates many genes (5–14), B will become connected to many nodes, but it will only use a fraction of them for independence testing. In contrast, all the child nodes will be aware of the hub B and will use it for testing of conditional independence. The undirected graph in Figure 2d is one possible output of ELMM where a solid circle at the end of the edge indicates that the node is aware of the neighbor at the other end of the edge. Since most of the child nodes will be aware of their parent nodes (the hub genes), dependencies among child nodes will be negated due to the common parent being used in conditional independence tests and false edges among child nodes can be correctly rejected. An additional benefit of this relaxation is that every node will condition on a smaller number of neighboring subsets. This property allows ELMM to perform fast network reconstruction and to scale well for high dimensional networks (see Supplementary material for complexity analysis).

In the proposed approach, the graph is constructed iteratively, where in every iteration, edges are ranked using equation (1). The algorithm starts with an empty graph and edges with the highest joint conditional posterior are iteratively added until a maximum number of edges is reached. Afterward, edges with the least joint conditional posterior are successively deleted, and the order at which edges are deleted is then used to rank all candidate edges. The complete pseudo-code is presented in the Supplementary material.

4 SIMULATIONS AND DATA

4.1 *In silico* simulated networks

Ten *in silico* networks were simulated using the GeneNetWeaver (GNW) simulation tool version 3.0 (Marbach *et al.*, 2009; Schaffter *et al.*, 2011). The first network was part of the Dream5 challenge for assessing network recovery algorithms (Marbach *et al.*, 2010). We additionally simulated nine more networks based on transcription regulatory networks of *E.coli* (bacteria) and *Saccharomyces cerevisiae* (yeast). The simulation is based on a kinetic model of gene regulation and translation activities with

both independent and synergistic gene regulation. In addition, the simulation incorporates models of noise at both levels, network dynamics and expression level measurements. The 10 network were further split into two sets based on the number of genes used in the simulation. The first set contains six network with sizes ranging from 1400 to 1656 genes each, while the second set contains four networks of sizes ranging from 2000 to 2500 genes each. A detailed description of the simulation is made available in the supplementary material and the reader is referred to Marbach *et al.* (2009) and Schaffter *et al.* (2011) for further details. Table S2 of the Supplementary material provides summary statistics of the 10 networks.

4.2 Lung airway dataset

The gene expression dataset used in this study is composed of 272 samples obtained by bronchial brushing from two different locations: large airway and small airway (Raman *et al.*, 2009). All samples were analyzed using Affymetrix Human Genome U133 Plus 2.0 microarrays. To avoid the problem of probe sets mapping to multiple genes, we used the custom mapping provided by Dai *et al.* (2005) to convert chip probes expression measurements into a single expression measurements for genes with unique Entrez-numbers. The participating individuals were characterized as non-smokers or smokers and were further labeled as healthy or having COPD. The Supplementary material provides complete details of the demographic information of the samples.

The analyzed microarray data include data on ∼23,000 protein-coding genes, an unknown subset of which are operating in the regulation and response behaviors of the pulmonary environment. Even with the scaling properties of ELMM, simultaneous genome-wide network inference is not computationally feasible. We therefore used the following selection procedure to limit the number of genes considered for network inference:

1. Genes that were not present (expressed) in at least 40% of the microarray samples were filtered out leaving 10,900 genes.
2. Hierarchical clustering was performed among the 10,900 remaining genes and a subset of linked clusters containing a total of 5492 genes was selected. The hierarchical clustering cutoff was selected to group approximately half of the genes.

The first of these filtering steps removed lowly expressed (or unexpressed) genes and the second identified a subset of genes that are expected to be enriched for intra-regulatory interactions. We additionally added extra variables to represent the following known covariates: age, gender, smoking and COPD status (see Supplementary meterial). Additionally, we applied ELMM in two phases. First, unconstrained network inference was applied among all the 5492 genes. Afterward, all genes were ranked based on the number of their interactions. In the second phase, the 550 most connected genes were selected and ELMM network inference was reapplied such that every selected edge must connect to at least one of the predetermined 550 hub candidate genes. This strategy was adopted to leverage the known hub-like topologies in regulation networks and to provide a network that is easier to analyze.

Assessing the performance of network inference when analyzing gene expression data of a human tissue is difficult, given the lack of large number of verified interactions and the many unknown tissue specific gene regulation. We therefore relied on gene annotation databases and published literature to assess the recovered network. Specifically, we used the semantic similarity tool provided by Wang *et al.* (2007) which uses the Gene Ontology database of gene annotations. We also performed further analysis of the most connected 20 hub genes recovered using the Gene-Card database and the DAVID analysis tool. Finally, we used known lung regulatory relationships reported in the primary literature to assess individual hubs.

5 RESULTS

5.1 Analysis of *in silico* data

The 10 *in silico* networks were used to assess the performance of ELMM for network reconstruction compared with five popular network inference methods: GeneNet (Schäfer *et al.*, 2006) based on empirical graphical Gaussian model (GGM), GLASSO (Friedman *et al.*, 2008) based on sparse GGM, ARACNE (Margolin *et al.*, 2006) based on mutual information, FOCI (Magwene *et al.*, 2004) based on low-order partial correlations and GENIE3 (Huynh-Thu *et al.*, 2010) based on decision trees for feature selection.

Figure 3 shows the average precision–recall (PR) curves for each of the methods for recovering the two sets of network. In our evaluation, the correct undirected graph was considered to be the graph where there is an edge between the two genes if and only if there was a direct regulatory relationship between the two genes.

Based on the PR curve evaluation, on average ELMM outperformed the other five methods in recovering the two sets of networks. Although ARACNE was competitive to ELMM in recovering the second set of networks, its performance seems to drop steeply if someone is considering higher than 15% recall. On the other hand, GENIE3 that was the winner of the Dream5 network inference challenge performed poorly in this evaluation. We believe this is due to the fact that the set of transcription factors was not known to any of the methods as in the Dream5 challenge. In sum, when considering the unconstrained case of network inference, regardless as to whether the goal is minimizing the percentage of false positives or for correctly recovering a greater proportion of the true network, ELMM had better overall performance.

The better performance of ELMM was also clear when considering the ability of each method to identify hub genes (Fig. 4). For each network, we determined a hub gene to be any gene that directly regulates more than six other genes, where the choice of six was made mainly to focus evaluation on the most connected genes. By this criterion, the number of hub genes in the first set of networks was 123, 94, 90, 88, 90 and 88, while in the second set of networks it was 135, 84, 79 and 78.

To assess performance for hub gene identification, we used the output of each algorithm to rank genes based on the number of other genes to which they are connected (see Supplementary material for details). Figure 4 presents a plot of the percentage of correctly identified hubs (Y-axis = recall) versus the number of retrieved hubs (X-axis) normalized by the actual number of all true hubs from each network. Based on these results, ELMM showed a superior accuracy in retrieving hub genes. For example, if we are to retrieve $1X$ (X = true number of hubs) of the most significant candidate hub genes, we would be able, on average, to retrieve 24% of the actual hub

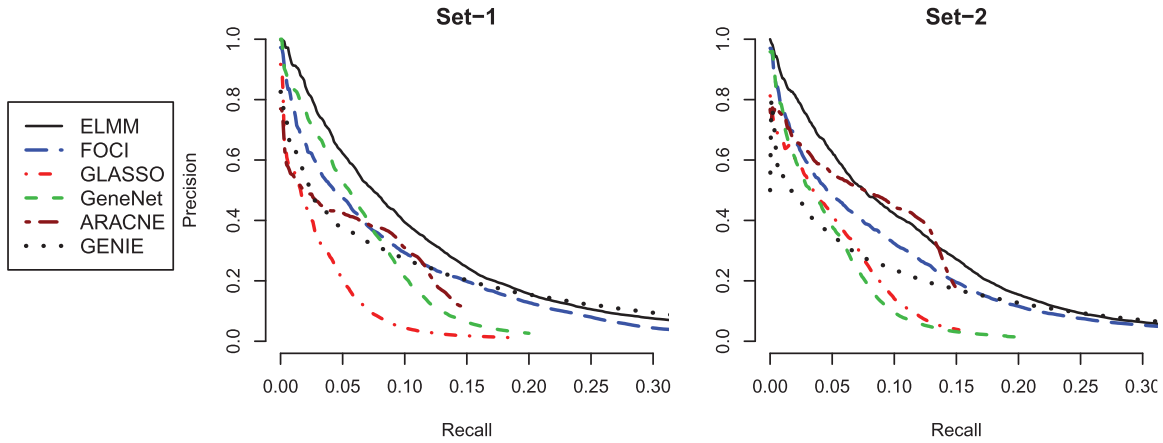

Fig. 3. Average PR curves for network recovery when ELMM, FOCI, GLASSO, GeneNet, ARACNE and GENIE were applied to recover two sets of *in silico* networks (Set-1 and Set-2). Precision = TP/(TP + FP) is plotted on the *Y*-axis versus recall = TP/(TP + FN) on the *X*-axis where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives

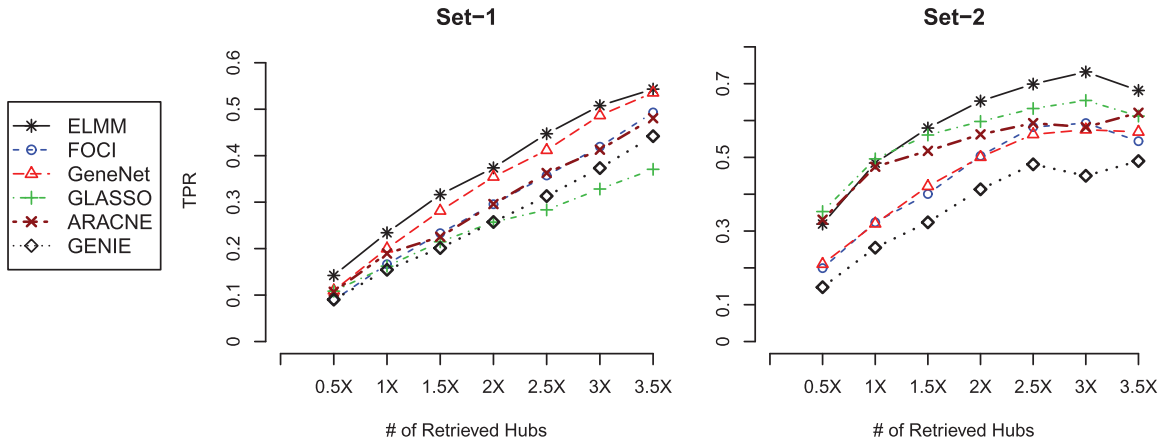

Fig. 4. Average recall of the correct hub genes (*Y*-axis = true positive rate (TPR)) versus the number of identified hub genes (*X*-axis = number of top candidate hub genes being considered) when using ELMM, FOCI, GLASSO, GeneNet, ARACNE and GENIE to recover the *in silico* networks of the two sets (Set-1 and Set-2). *X*-axis is normalized by the number of all true hubs from every network

genes in the first set of networks and ∼44% of the hub genes in the second set of networks.

5.2 Analysis of the human lung airway epithelium

5.2.1 Lung airway epithelium network and gene ontology analysis Figure S2 of the Supplementary material shows a representative sub-network of 1500 genes of the inferred network. The graph also highlights the genes affected by smoking and shows a high co-localization among those genes. This topology was expected, since genes up- or down-regulated as a group due to smoking will result in induced dependencies between the expression levels of those genes. However, genes that respond to the effect of smoking are likely to share common regulation control mechanisms (Hübner *et al.*, 2009), and the topology of the recovered network suggests regulatory relationships and possible hub genes that may be involved in evoking the smoking response.

For every inferred relationship (pair of connected genes), the semantic similarity tool provided by Wang *et al.* (2007) was used to assess the similarity of annotations in the gene ontology database between the two genes for three criteria: molecular functionality, biological processes and cellular location. This analysis showed that among the top 10,000 inferred interactions, 1063 pairs of connected genes were found to have the same annotation (semantic similarity of 1.0) along at least one of the three criteria. When compared with 10,000 randomly selected connections among the 5492 genes, only 244 pairs were found to be functionally similar. Although it is unclear how such annotation similarities map to the actual regulatory network, we expect there to be some correspondence, such that a higher annotation similarity of the recovered network compared to random is consistent with the recovered network providing useful information concerning the true underlying network regulatory relationships responsible for the observed gene expression covariation.

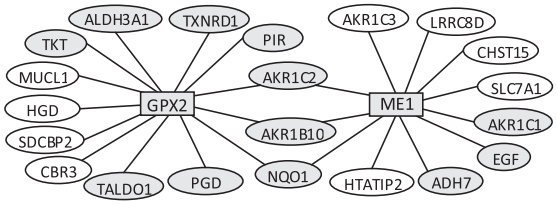

Fig. 5. NRF2 recovered sub-network where shaded genes are known to be regulated by NRF2. Only the most significant edges are displayed

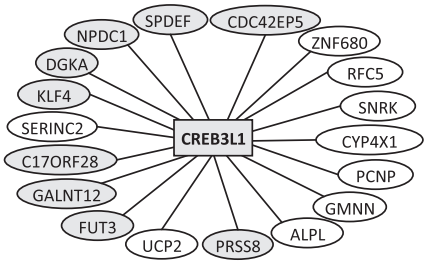

Fig. 6. CREB3L1 recovered sub-network, where shaded genes are involved in airway goblet cell differentiation, mucus production or cell secretion. Only the most 19 significant edges are displayed

Analysis of the most connected 20 hub genes recovered also indicated the network was consistent with broad-scale annotation. For example, the most connected gene was *CEACAM5*, which is known to play a role in intracellular signaling (Safran *et al.*, 2010). *CEACAM5* is also considered one of the best markers of cancer and is believed to activate integrin pathways (Chan and Stanners, 2007). The second most connected gene was *GPX2*, which plays a role in protection from toxicity. *GPX2* was connected to many genes regulated by the transcription factor NRF2 (Fig. 5). Being itself regulated by NRF2 that becomes active only after phosphorylation, the expression of *GPX2* is likely to be serving as a mirror of the regulatory effects of the phosphorylated NRF2. Other genes in the top 10 most connected genes were *AKR1B10*, *SNAI2*, *FOS* and *IFIT3*. In Yan *et al.* (2007), *AKR1B10* was shown to play a role in cell growth and the silencing of its expression was shown to inhibit the growth of colorectal cancer cells. Both *SNAI2* and *FOS* are known to encode transcription factors (Safran *et al.*, 2010), while *IFIT3* was previously reported to be crucial for the phosphorylation of the IRF3 transcription factor which in turns triggers anti-viral response (Liu *et al.*, 2011).

5.2.2 Analysis of hubs and sub-networks For a more detailed analysis of the most dense 20 sub-networks composed of hub-like structures, we performed a literature search for known relations between these genes that have been identified in the lung airway epithelium. A number of these sub-networks were found to be consistent with known pathways or regulatory relationships operating in the lung. In this section, we present two examples.

Figure 5 shows a sub-network of 23 genes, where 14 of these genes are known to be regulated by the NRF2 transcription factor in the lung airway (Hübner *et al.*, 2009). NRF2 is an oxidant-responsive transcription factor known to induce detoxifying and antioxidant genes. Although the network inference did not predict that NRF2 is directly connected to these genes, NRF2 only becomes functional when phosphorylated, so it is not necessarily surprising that NRF2 transcription is not reflective of the regulatory properties of this gene. Given this high enrichment for known NRF2-regulated genes in the recovered sub-network, the other nine genes, including *MUCL1*, *HGD*, *SDCBP2*, *CBR3*, *HTATIP2*, *SLC7A1*, *CHST15*, *LRRC8D* and *AKR1C3* seem to be candidates for being regulated by NRF2 or by one of the genes regulated by NRF.

Figure 6 shows a hub sub-network including 18 genes that are connected to the transcription factor CREB3L1, whose *Drosophila* homolog CREB3LA is the major regulator of secretory capacity (Fox *et al.*, 2010). Nine of these genes were reported to be associated with airway goblet cell differentiation, mucus production or cell secretion (Chen *et al.*, 2009; Davis and Dickey, 2008; Fahy and Dickey, 2010; Katz *et al.*, 2002; Safran *et al.*, 2010). This high enrichment also suggests that the other nine genes, including *SERINC2*, *UCP2*, *ALPL*, *GMNN*, *PCNP*, *CYP4X1*, *SNRK*, *RFC5* and *Znf467* might also play a role in secretion pathways.

6 CONCLUSIONS

ELMM is a scalable network recovery algorithm that has properties well suited for both edge recovery and identification of hub genes when dealing with gene expression data. The comparative accuracy in the identification of hub genes is of particular value for at least two major reasons. First, there are a number of indirect and direct lines of evidence that hub-rich topologies may be common in regulatory networks (Duarte and Zeng, 2011; Jordan *et al.*, 2004; Lukashin *et al.*, 2003). Second, identification of hubs may provide information on genes that tightly associated, even if the direct mode of interaction is hidden. Although the analysis of real data cannot determine properties of the candidate hubs beyond their putative position at the center of a set of genes that are co-regulated by the same control mechanisms, such highly connected hub genes are candidates for being a causal gene that affects the expression of other genes or as indicators of an underlying causal mechanism that co-regulates the set of genes connected to the same hub. In general, accurate and scalable methods for network inference such as ELMM will be increasingly important with the increasing availability and the improved quality of gene expression data.

Funding: GW* was supported in part by National Institutes of Health T32 HL094284. This work was supported by National Institutes of Health P50 HL084936 and by National Science Foundation grants IOS1026555 and DEB0922432.

Conflict of Interest: none declared.

REFERENCES

Akavia,U. *et al.* (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.

Aliferis,C. *et al.* (2003) Causal Explorer:Causal Probabilistic Network Learning Toolkit for Biomedical Discovery. In *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Scienes, METMBS '03, June 23–26, 2003, Las Vegas, Nevada, USA*, pp. 371–376.

Bansal,M. *et al.* (2007) How to infer gene networks from expression profiles. *Mol. Sys. Biol.*, **3**.

Baralla,A. *et al.* (2009) Inferring gene networks: dream or nightmare? *Ann. NY. Acad. Sci.*, **1158**, 246–256.

Blencowe,B. *et al.* (2009) Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Develop.*, **23**, 1379.

Chan,C. and Stanners,C. (2007) Recent advances in the tumour biology of the gpi-anchored carcinoembryonic antigen family members ceacam5 and ceacam6. *Curr. Oncol.*, **14**, 70.

Chen,G. *et al.* (2009) Spdef is required for mouse pulmonary goblet cell differentiation and regulates a network of genes associated with mucus production. *J. Clin. Invest.*, **119**, 2914.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res.*, **33**, e175.

Davis,C. and Dickey,B. (2008) Regulated airway goblet cell mucin secretion. *Annu. Rev. Physiol.*, **70**, 487–512.

Duarte,C. and Zeng,Z. (2011) High-confidence discovery of genetic network regulators in expression quantitative trait loci data. *Genetics*, **187**, 955–964.

Fahy,J. and Dickey,B. (2010) Airway mucus function and dysfunction. *New England J. Medi.*, **363**, 2233–2247.

Faith,J. *et al.* (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Fox,R. *et al.* (2010) The creba/creb3-like transcription factors are major and direct regulators of secretory capacity. *J. Cell Biol.*, **191**, 479.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Friedman,N. *et al.* (1999) Learning Bayesian network structure from massive datasets: the parse candidate algorithm. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence UAI*, pp. 206–215.

Hotelling,H. (1953) New light on the correlation coefficient and its transforms. *J. R. Stat. Soc. B.*, **15**, 193–232.

Hübner,R. *et al.* (2009) Coordinate control of expression of nrf2-modulated genes in the human small airway epithelium is highly responsive to cigarette smoking. *Mol. Med.*, **15**, 203.

Huynh-Thu,V. *et al.* (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.

Jordan,I. *et al.* (2004) Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.*, **21**, 2058.

Kalisch,M. and Bühlmann,P. (2007) Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mac. Learn. Res.*, **8**, 613–636.

Kalisch,M. (2010) *PCAlg: Estimation of CPDAG/PAG and Causal Inference using the IDA Algorithm*. R package version 1.0-2.

Katz,J. *et al.* (2002) The zinc-finger transcription factor klf4 is required for terminal differentiation of goblet cells in the colon. *Development*, **129**, 2619.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175.

Liu,X. *et al.* (2011) Ifn-induced tpr protein ifit3 potentiates antiviral signaling by bridging mavs and tbk1. *J. Immunol.*, **187**, 2559–2568.

Lukashin,A. *et al.* (2003) Topology of gene expression networks as revealed by data mining and modeling. *Bioinformatics*, **19**, 1909.

Magwene,P. *et al.* (2004) Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.*, **5**, R100.

Marbach,D. *et al.* (2009) Generating realistic *in silico* gene networks for performance assessment of reverse engineering methods. *J. Comput. Biol.*, **16**, 229–239.

Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Nat. Acad. Sci.*, **107**, 6286.

Margolin,A. *et al.* (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.

Markowetz,F. and Spang,R. (2007) Inferring cellular networks—a review. *BMC Bioinformatics*, **8**, S5.

Neapolitan,R.E. (2004) *Learning Bayesian Networks*. Pearson Printice Hall, London, UK.

Peter,S. (2001) *Causation, Prediction, and Search, 2nd edn*, Vol 1. The MIT Press.

Raman,T. *et al.* (2009) Quality control in microarray assessment of gene expression in human airway epithelium. *BMC Genomics*, **10**, 493.

Safran,M. *et al.* (2010) Genecards version 3: the human gene integrator. *Database*, **2010**.

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

Schäfer,J. *et al.* (2006) Reverse engineering genetic networks using the genenet package. *Newsletter R Project*, **6**, 50.

Schaffter,T. *et al.* (2011) Genenetweaver: *in silico* benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**, 2263–2270.

Schlitt,T. and Brazma,A. (2007) Current approaches to gene regulatory network modeling. *BMC bioinformatics*, **8**, S9.

Shendure,J. and Ji,H. (2008) Next-generation dna sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Sotiriou,C. and Pusztai,L. (2009) Gene-expression signatures in breast cancer. *New England J. Medi.*, **360**, 790–800.

Tsamardinos,I. *et al.* (2006) The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learn.*, **65**, 31–78.

Tsamardinos,I. and Brown,L. (2008) Bounding the false discovery rate in local bayesian network learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence*, pp. 1100–1105.

Volinia,S. *et al.* (2006) A microrna expression signature of human solid tumors defines cancer gene targets. *Proc. Nat. Acad. Sci. USA*, **103**, 2257.

Wang,J. *et al.* (2007) A new method to measure the semantic similarity of go terms. *Bioinformatics*, **23**, 1274.

Yan,R. *et al.* (2007) Aldo–keto reductase family 1 b10 gene silencing results in growth inhibition of colorectal cancer cells: Implication for cancer intervention. *Int. J. Cancer*, **121**, 2301–2306.