Empirical Bayes conditional independence graphs for regulatory network recovery

Rami Mahdi1,*, Abishek S. Madduri1, Guoqing Wang1, Yael Strulovici-Barel1, Jacqueline Salti1, Neil R. Hackett1, Ronald G. Crystal1 and Jason G. Mezey1,2,*

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

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 epithelial cells of healthy non-smokers, healthy smokers and individuals with chronic obstructive pulmonary disease (COPD). We compare our results to ARACNE, FOCI, GENIE3 and GLASSO in the context of network topology, number of nodes and edges, and number of true positives and false positives.


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. 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. 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. 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 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.

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. 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. 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 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 reconstruction algorithms including GeneNet, ARACNE, FOCI, GENIE3 and GLASSO.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
As with related algorithms that use graphical models to recover regulatory networks, 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 1a. When measuring a number of genes expressed in 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 conditioning on one or more variables. This conservative approach causes the probability of rejecting a true edge to increase with the number of neighboring variables.

2 BACKGROUND: NETWORK INFERENCE WITH CONDITIONAL INDEPENDENCE GRAPHS

2.1 ALGORITHM

As with related algorithms that use graphical models to recover regulatory networks (Friedman et al., 2000; Meinshausen and Bühlmann, 2007; 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 1a. When measuring a number of genes expressed in 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 one or more variables. This conservative approach causes the probability of rejecting a true edge to increase with the number of neighboring variables.

2.2 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 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 inference of the graph.
Conditional Independence Graphs for Network Recovery

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:

$$P(E_0|CP_i, CP_j, D) \geq P(E_0|CP_i, D) \times P(E_0|CP_j, D),$$  \hspace{1cm} (1)

where $E_k$ 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 $P(E_0|CP_i, CP_j, D)$ 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 $P(E_0|CP_i, CP_j, D)$ 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 $P(E_0|CP_i, CP_j, D)$ as a conditional posterior because we condition on a certain graph structure that implies $CP_i$ and $CP_j$ are neighbor sets of $i$ and $j$, respectively. For the rest of the article, we refer to $P(E_0|CP_i, CP_j, D)$ as the one-sided conditional posterior probability of dependence (CPDD) of $i$ on $j$ and we refer to $P(E_0|CP_i, CP_j, D)$ as the joint CPDD between $i$ and $j$.

A critical part of the approximation in equation $P(E_0|CP_i, CP_j, D)$ is the need to estimate the one-sided conditional posterior probability $P(E_0|CP_i, CP_j, D)$ 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 $\rho_{ij}$ is a measure of the statistical dependence between two variables $i$ and $j$, while the partial conditional correlation $\rho_{ij|kl}$ 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 have to rely on independence tests based on the sample partial correlations (i.e., $\rho_{ij}$, $\rho_{ij|kl}$) (see Supplementary material).

In order to compute $P(E_0|CP_i, CP_j, D)$ 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 $P(E_0|CP_i, CP_j, D)$ is to use the probability that the least significant conditional partial correlation is actually significant as follows:

$$P(E_0|CP_i, CP_j, D) = \min_{S \subseteq CP_i} \max_{S \subseteq CP_j} P(\rho_{ij} \neq 0 | \rho_{ij}^S = 0).$$  \hspace{1cm} (2)

Using equation $P(E_0|CP_i, CP_j, D)$ is equivalent to the approach used by regular constraint-based methods (Kalisch and Bühlmann, 2005), where the partial correlations are assumed to be generated from a variant of the empirical Bayes approach proposed by Schäfer and Strimmer (2005), 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 $P(E_0|CP_i, CP_j, D)$ 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:

$$P(E_0|CP_i, CP_j, D) = \left( \sum_{S \subseteq CP_i} P(\rho_{ij}^S \neq 0 | \rho_{ij}^S) \right)^{-1/2},$$  \hspace{1cm} (3)

where

$$P(\rho_{ij}^S \neq 0 | \rho_{ij}^S) = \min_{S \subseteq CP_i} \max_{S \subseteq CP_j} P(\rho_{ij} \neq 0 | \rho_{ij}^S)$$

and

$$w_j = \begin{cases} 1 & \text{if } z = 0 \text{ or } P(E_0|CP_i, CP_j) < P(E_0|CP_i) \leq \rho_{ij}^S \text{ or } z = 1, \hfill \\ 0 & \text{otherwise}. \end{cases}$$

What is important to recognize about either approximation $P(E_0|CP_i, CP_j, D)$ or $P(E_0|CP_i, CP_j, D)$ 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 $j$ (i.e., $S \subseteq CP_i$), 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. Meizel, submitted for publication). However, when given a small sample data, the approximation in equation $P(E_0|CP_i, CP_j, D)$ is expected to be sensitive to multiple testing, while the approximation in equation $P(E_0|CP_i, CP_j, D)$ cases 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 cases the problem that different nodes in the graph have different numbers of neighbors.

3.2 Semi-parametric independence testing

To compute $P(\rho_{ij} \neq 0 | \rho_{ij}^S)$ which is required in equation $P(E_0|CP_i, CP_j, D)$, 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 $H_1: \rho \neq 0$). Using Bayes rule of posterior probability, $P(\rho \neq 0|\hat{\rho})$ is computed as follows:

$$P(\rho \neq 0|\hat{\rho}) = \frac{x_1 \times f_0(\hat{\rho}|\hat{\rho})}{x_1 \times f_0(\hat{\rho}|\hat{\rho}) + x_0 \times f_0(\hat{\rho})},$$

(4)

where $\hat{\rho}_{ij}$ is the sample partial correlation, $x_1$ and $x_0$ are the distribution of the sample partial correlations under the null and the alternative hypothesis, respectively, while $x_1$ and $x_0$ 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 ($H_1: \rho \neq 0$), we make use of a simplifying assumption that the sample partial correlations follow a uniform distribution in the range $[-1,1]$ (e.g., $x_0=0.5, x_1=1$). On the other hand, from Schäfer and Strimmer (2005), when the true partial correlation is zero ($H_0: \rho_{ij}=0$), the sample partial correlation $\hat{\rho}_{ij}$ has the following distribution:

$$f(\hat{\rho}_{ij}) = (1-\rho^2)^{-1/2} \Gamma(n/2) \Gamma((n-1)/2)/\Gamma(n/2) \Gamma((n-1)/2),$$

(5)

where $\Gamma$ is the gamma function and $\kappa$ is the 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 $S_{ij}$ ($H_0: \rho_{ij}=0$), the sample partial correlation $\hat{\rho}_{ij}$ 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 $\kappa$ of the null distribution is unknown and could vary considerably from the expected value. To estimate $\kappa$ and the unknown mixture proportions of the null and alternative hypotheses $\pi_n$ and $\pi_0$, as in Schäfer and Strimmer (2005), we use likelihood maximization of observing the zero-order sample partial correlations (conditioning on the empty set $\phi$) generated from the mixture model as follows:

$$\hat{\kappa}, \hat{\pi}_n, \hat{\pi}_0 = \text{arg max } \prod_{i,j \in \phi} f(\hat{\rho}_{ij}|\pi_n, x_1) \times \prod_{i,j \not\in \phi} f(\hat{\rho}_{ij}|\pi_0, x_0),$$

(6)

where

$$f(\hat{\rho}_{ij}|\pi_n, x_1) = \pi_n f_0(\hat{\rho}_{ij} \mid \kappa) + \pi_0 f_0(\hat{\rho}_{ij}),$$

(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 as follows:

$$\hat{\kappa}_i = \text{arg max } \prod_{j \in \phi} f(\hat{\rho}_{ij}|\pi_n, x_1),$$

(8)

3.3 Heuristic relaxation of independence constraints

Even with the smoothing approximation of equation (5) 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 $P(E_0|C_{P_i}, D)$ in equation (6) 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. $P(E_0|C_{P_i}, D) \neq P(E_0|C_{P_j}, D)$), 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 $G$ 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 $P(E_0|C_{P_i}, D) \geq P(E_0|C_{P_j}, D)$ then $C_{P_i}\leftarrow C_{P_j} \cup \{i\}$.
- If $P(E_0|C_{P_i}, D) \geq P(E_0|C_{P_j}, D)$ then $C_{P_j}\leftarrow C_{P_i} \cup \{j\}$.

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 B where a hub such as $H_0$ regulates many genes ($5-14$), it will become connected to many nodes, but it will only use a fraction of them for independence testing. In contrast, all child nodes will be aware of the hub if and will use it for testing of conditional independence. The undirected graph in Figure B 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 (6). 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. 2012). 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 Schäffer et al. (2013) 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 (Haman et al. 2007). 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. (2006) 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 ELM, 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 material). 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. (2008) 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 (Schäffer 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. 2006) based on low-order partial correlations and GENIE3 (Huynh-Thu et al. 2011) based on decision trees for feature selection.

Figure 4 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, ELM had better overall performance.

The better performance of ELMM was also clear when considering the ability of each method to identify hub genes (Fig 4B). 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 4B 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.

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.
The second most connected gene was AKR1C3, which seems to be candidates for being regulated by NRF2 or by protection from toxicity. Other genes in the top 10 most connected genes were regulatory effects of the phosphorylated NRF2. Other genes regulated by the transcription factor NRF2 (Fig. 5). Being itself regulated by NRF2 that becomes active only after phosphorylation, the expression of the IRF3 transcription factor which in turns triggers anti-viral response is also considered one of the best markers of cancer and is likely to be serving as a mirror of the regulatory effects of the phosphorylated NRF2. Other genes that tightly associated, even if the direct mode of interaction is not necessarily surprising that NRF2 becomes active only after phosphorylation, the expression of GPX2 is likely to be connected to the same hub in airway goblet cell differentiation, mucus production or cell secretion. Only the most 19 significant edges are displayed also suggests that the other nine genes, including SERINC2, UCP2, ALPL, GMNN, PCNP, CYP4X1, SNRK, RFC5 and Znf467 might 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. Figure 6 shows a hub sub-network including 18 genes that are connected to the transcription factor CREB3L1, where shaded genes are involved in airway goblet cell differentiation, mucus production or cell secretion. Only the most 19 significant edges are displayed

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

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 6 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. These genes are either directly connected to these genes, NRF2 only becomes functional when phosphorylated, or they are not necessarily surprising that NRF2 does not predict that NRF2 is directly connected to these genes, NRF2 only becomes functional when phosphorylated, or they are not necessarily surprising that NRF2 does not predict that NRF2 has been previously 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 SNAIL2 and FOXD1 are known to encode transcription factors (Safran et al. 2010), while PHF1 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).

REFERENCES


