BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 22 2011, pages 3166–3172
doi:10.1093/bioinformatics/btr533

Systems biology

Advance Access publication September 28, 2011

Unsupervised detection of genes of influence in lung cancer using biological networks

Anna Goldenberg[1,*], Sara Mostafavi[3], Gerald Quon[3], Paul C. Boutros[2] and Quaid D. Morris[1,3]

[1]Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5S 3E1, [2]Informatics and Biocomputing Platform, Ontario Institute for Cancer Research, Toronto, ON M5G 0A3 and [3]Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Lung cancer is often discovered long after its onset, making identifying genes important in its initiation and progression a challenge. By the time the tumors are discovered, we only observe the final sum of changes of the few genes that initiated cancer and thousands of genes that they have influenced. Gene interactions and heterogeneity of samples make it difficult to identify genes consistent between different cohorts. Using gene and gene–product interaction networks, we propose a principled approach to identify a small subset of genes whose network neighbors exhibit consistently high expression change (in cancerous tissue versus normal) regardless of their own expression. We hypothesize that these genes can shed light on the larger scale perturbations in the overall landscape of expression levels.

Results: We benchmark our method on simulated data, and show that we can recover a true gene list in noisy measurement data. We then apply our method to four non-small cell lung cancer and two pancreatic cancer cohorts, finding several genes that are consistent within all cohorts of the same cancer type.

Conclusion: Our model is flexible, robust and identifies gene sets that are more consistent across cohorts than several other approaches. Additionally, our method can be applied on a per-patient basis not requiring large cohorts of patients to find genes of influence. Our approach is generally applicable to gene expression studies where the goal is to identify a small set of influential genes that may in turn explain the much larger set of genome-wide expression changes.

Availability: The code is available at http://morrislab.med.utoronto.ca/~anna/cannet.zip

Contact: anna.goldenberg@utoronto.ca

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

Received on October 29, 2010; revised on September 13, 2011; accepted on September 14, 2011

1 INTRODUCTION

The reality of cancer diagnosis is harsh. Despite continued advances in surgical, pharmacological and radiological treatments, mortality remains high, with lung cancer as the leading cause of cancer-related deaths in the United States (Jemal *et al.*, 2008). One of the primary challenges of lung cancer is the detection of small, early-stage tumors. Most cancers are identified and the tumor is excised for molecular analysis at a late stage. By then, we are only able to observe the final sum of changes to cell structure and function reflected in differential expression of a few genes that initiated cancer and thousands of genes affected through gene interactions.

Because of the cascading effect, it is difficult to identify genes that are directly responsible for tumorigenesis from gene expression data alone. Recently, a large number of studies have investigated the use of known gene–gene relationships, such as physical interaction of their proteins, to characterize cancer genes (Jonsson and Bates, 2006; Milenkovic and Przulj, 2008; Taylor *et al.*, 2009) and to identify more robust signatures of diseases including cancer (Chuang *et al.*, 2007; Fortney *et al.*, 2010; Hwang and Park, 2009). For instance, it has been shown that genes whose expression has changed are more likely to be surrounded by genes whose expression has also changed (Efroni *et al.*, 2007; Subramanian *et al.*, 2005; Vogelstein and Kinzler, 2006; Watters and Roberts, 2004; Yan and Sun, 2008). Initial attempts have relied on pre-defined pathways or gene sets, limiting these analyses to well-studied genes or known pathway memberships. A more natural approach is to study the change in expression of a gene directly in the context of the interaction network of gene products, for instance, by identifying *active subnetworks* that are predictive of the disease (Chuang *et al.*, 2007; Fortney *et al.*, 2010; Hwang and Park, 2009).

Development of accurate models for identifying important genes in cancer still face a number of key challenges. First, many methods depend on the knowledge of 'seed genes' known to be involved in a particular disease [e.g. Lage *et al.* (2007); Navlakha and Kingsford (2010); Vandin *et al.* (2011)]. Second, the procedure for estimating parameters in these models sometimes combine *ad hoc* heuristics with underdetermined systems of equations (Ergün *et al.*, 2007), which may not be robust to expression or network noise. Third, most methods require availability of a sufficiently large number of samples in order to reach conclusions about important genes.

In this article, we start to address these challenges within a framework for detecting genes of influence given a set of gene expression profiles from tumor and healthy samples, and a network of gene–gene interactions. Our method does not require seed genes; does not have to simultaneously infer the network, circumventing the second challenge outlined above; and can be applied on the

*To whom correspondence should be addressed.

individual patient basis achieving great resolution and flexibility. We take a more direct approach for identifying potentially important genes by asking: which genes can explain away the majority of the observed gene expression changes of their neighbors? We use an interaction network that includes both regulatory and physical protein–protein interactions. Using simulated expression profiles, we show that our model can recover genes of influence with high precision and recall under varying levels of noise. We then apply our procedure to four different cohorts of lung cancer patients, and identify candidate genes of influence for each patient in each dataset independently. We find that genes of influence found by our method are more consistent across patients and datasets compared with other approaches, indicating that despite increasing the resolution of our analysis to a per-patient level, our results are not overfitting.

2 METHODS

The ultimate goal of our work is to find a small set of genes that can explain some of the high changes in expression levels observed across a larger set of genes. To tackle this task, we use the information encoded in known gene and gene product interactions, usually represented as a network, which has an edge if there is a direct interaction between gene products A and B, and does not have an edge if there is no known interaction. If all neighbors of A have a high expression change, we hypothesize that node A might have influenced its neighbors to significantly change in cancer compared with normal tissue. We conclude then, that the contribution *influence factor* of node A should be high if the change in expression levels between cancerous and healthy tissue of its surrounding neighbors are high. Note, that we derived the influence factor of node A without referring to its observed value. This setup allows us to capture changes in target expression due to post-translational modifications, where the change of expression levels of the target genes is the only indication of the activity of the gene itself (its own mRNA level does not change). Transcription factors (TFs) are typically expressed at lower levels compared with non-TF genes (Vaquerizas *et al.*, 2009). The idea of summarizing the information about the gene using its network neighborhood has already appeared in Pradines *et al.* (2004), where the authors found that using network neighborhood to summarize a gene's impact is a way to smooth over different types of gene regulation. Similar ideas have been used to infer the activity levels of transcription factors from their known target genes (Gao *et al.*, 2004). In what follows, we propose an objective function that allows us to formalize the search for the influence factors for all genes simultaneously.

2.1 Notation

For the purpose of describing our model, we use the terms gene and protein interchangeably. Let a binary square matrix $W \in \{0,1\}^{d\times d}$ represent an interaction network, where $W_{ij}=1$ if genes i and j are connected in the network, and $W_{ij}=0$ if they are not. d is the total number of genes in the network. W represents direct (first order) connectivity in the network, connecting genes that have direct interactions. $W^2 = W\times W$, the second power of W, is the set of second-order interactions and, generally, the n-th power of W: W^n, is the n-th order, i.e. each row (or column, the matrix is symmetric) of W^n contains information about the number of paths of length n between all pairs of genes in the network. In this work, we are only interested in the existence of a path, rather than the number of them, thus we binarize the W^n matrix. Since we are not interested in how many paths start and finish in the same gene, we set the diagonal of the matrix to zero. Finally, we assume that for two genes i and j, the more direct path is more likely to be of influence among all paths of different orders. For example, if i and j have a direct and a second-order connection, we assume that the influence would propagate across the direct connection, and we would need to set the second-order connection to 0. In formal terms, this condition can be stated as follows: $[W^n]_{ij}=0$ if $\exists k:[W^k]_{ij}>0, k=1\ldots n-1$. The modified version of W^n employed here is W^{n*}:

$$W^{n*} = \begin{cases} 1 & \text{if } [W^n]_{ij}>0,\ i\neq j \\ & \text{and } \forall k=1\ldots n-1:[W^k]_{ij}=0. \\ 0 & \text{otherwise} \end{cases}$$

Let $\vec{z}\in\Re^{d\times 1}$ represent the log of the ratio in expression of our d genes between tumor and normal tissue. The change of expression is measured between the current expression level of a gene (for e.g. in a lung cancer patient) and the 'normal' expression level (as measured in healthy patients). We describe the exact measurement procedure and the distribution of $\vec{z}$ in Section 3.

2.2 Formulation

Our objective function formalizes our assumption that each gene's expression change can potentially be explained by the influence of its neighbors. In the simplest case, the i-th gene's expression change, given by z_i, can be explained by the added effects of its *direct* neighbors: $z_i \approx \sum_j \alpha_j W_{ij}$ where α_j is an unknown factor of influence of gene j on its neighbors that we have to estimate. To extend this formulation to indirect neighbors, we note that influence of n-th order neighborhood is $\sum_j \alpha_{jn}[W^{n*}]_{ij}$ where α_{jn} can, for example, be constrained to decrease with increasing n. Thus, the additive effect of all neighbors up to order n on gene i is $\sum_n\sum_j \alpha_{jn}[W^{n*}]_{ij}$. We can then formulate our hypothesis of influence in the most general way as $z_i \approx \sum_n\sum_j \alpha_{jn}[W^{n*}]_{ij}$. Our goal is to find the unknown influence factors $\vec{\alpha}$ that allow us to explain the expression change of all d genes. We can achieve this goal by finding $\vec{\alpha}$ that minimizes the following squared error: $\arg\min_{\vec{\alpha}} \sum_{i=1}^{d}\left[\sum_{j=1}^{d}\sum_n \alpha_{jn}[W^{n*}]_{ij} - z_i\right]^2$.

Because we want to find a small number of genes that explain the changes network-wide and because the majority of the genes might not have significant detectable influence on their neighbors, we must regularize our likelihood to produce a sparse solution. We propose to use a combination of $\ell_1 = \|\vec{\alpha}\|_1$ and $\ell_2 = \|\vec{\alpha}\|_2$ regularization factors known as elastic net penalty (Friedman *et al.*, 2010; Zou and Hastie, 2005): $P_\xi(\alpha) = \frac{1}{2}(1-\xi)\|\vec{\alpha}\|_2^2 + \xi\|\vec{\alpha}\|_1$, where ξ is the trade-off parameter between sparsity (ℓ_1-norm) and shrinkage (ℓ_2-norm), $\xi=1$ corresponding to the Lasso penalty. The motivation for the elastic net penalty comes from the fact that when there are two sufficiently correlated covariates that are predictive of the outcome, the ℓ_1 penalty makes an arbitrary selection between the two, whereas the elastic net penalty selects both of them but shrinks the importance of each of them based on the ξ parameter. In our case, the elastic net penalty allows to recover all genes that can explain observed expression change rather than just one of them.

Our full objective function can then be generally written as the sum of the error term and a parameter regularization term corresponding to sparseness as described above:

$$\arg\min_{\vec{\alpha}} \sum_{i=1}^{d}\Big[\sum_{j=1}^{d}\sum_n \alpha_{jn}[W^{n*}]_{ij} - z_i\Big]^2 + \lambda P_\xi(\vec{\alpha}_n), \qquad (1)$$

where $\vec{\alpha}_n = \{\alpha_{1n}, \alpha_{2n}, \ldots, \alpha_{dn}\}$ are the influence factors up to order n to be estimated and λ is the regularization parameter.

The order of adjacency, n, in practice is considered to be no greater than 2: $n\in\{1,2\}$, restricting the influence only to the first- and second-order neighbors in the network. The reason for that is 2-fold; on one hand, for most nodes third-degree neighborhood spans the extent of almost the entire network, making neighborhoods of order beyond 3 not very informative of individual genes, and on the other, currently there is very little information about the extent of physical influence of a gene on the expression level of its neighbor that is 3 hops away.

In its simplest form, Equation (1) can be reduced to only consider first-degree interactions as in the First-Degree model (FD) [Equation (2)].

$$\arg\min_{\vec{\alpha}} \sum_{i=1}^{d} \left[\sum_{j=1}^{d} \alpha_j W_{ij} - z_i \right]^2 + \lambda P_\xi(\vec{\alpha}). \quad (2)$$

A second model that we call Expression-change Gene network of second Order (EGO) considers second-order interactions in addition to first-degree neighbors. An independent parameter β indicates influence from the second-order neighbors [$n=2$ in Equation (1)].

$$\arg\min_{\vec{\alpha},\vec{\beta}} \sum_{i=1}^{d} \left[\sum_{j=1}^{d} \left(\alpha_j W_{ij} + \beta_j [W^{2*}]_{ij}\right) - z_i \right]^2 + \lambda_1 P_{\xi_1}(\vec{\alpha}) + \lambda_2 P_{\xi_2}(\vec{\beta}). \quad (3)$$

While EGO model requires $2d$ parameters, twice as many as FD, it is able to consider second-order influence of genes on each other.

2.3 Optimization and model selection

To solve for the parameters of EGO and FD, we use a coordinate descent procedure (Friedman *et al.*, 2010), which solves for the entire regularization path available as part of the glmnet package. To do so, we write Equations (2) and (3) as the solution to a linear regression problem with the elastic net penalty. Let $\bar{W} = [W\ W^{2*}]$ and $\vec{\gamma} = [\vec{\alpha}^T\ \vec{\beta}^T]^T$. Then we can obtain $\vec{\gamma}$ (which is composed of $\vec{\alpha}$ and $\vec{\beta}$) by solving:

$$\arg\min_{\vec{\gamma}} (\bar{W}\vec{\gamma} - z)^T (\bar{W}\vec{\gamma} - z) + \lambda P_\xi(\vec{\gamma}). \quad (4)$$

In order to find suitable settings for the regularization parameter λ, we use the Bayesian Information Criterion (BIC). In particular, we compute the BIC score for various settings of λ: $\mathrm{BIC}(\lambda) = d\log(\hat{\sigma_\lambda}^2) + N_\lambda \log(d)$ where N_λ is the number of non-zero elements in $\vec{\gamma}$ for a particular setting of λ and $\hat{\sigma_\lambda}$ is the error variance. We then choose a setting of λ that results in the lowest BIC score. The use of BIC score is common in the elastic-net regularized regression setting [e.g. Zou and Zhang (2009)]. We choose ξ in the range of $\xi \in [0.1\ \ 1]$ that gives us the maximum number of non-zero entries in γ.

2.4 Normalization

The above model ignores the fact that hub genes (i.e. genes that have a large degree in the interaction network) have many more interactions than non-hub genes. Therefore, without normalizing the matrices W and W^{2*}, the coefficients assigned to the high-degree nodes will likely be much smaller than those assigned to low-degree nodes. To correct for this, we normalize the matrices W and W^{2*} using symmetric normalization. Let D be a diagonal matrix with diagonal entries equal to the node degrees: $D_{ii} = \sum_j W_{ij}$, then the symmetrically normalized matrix is given by $D^{-1/2} W D^{-1/2}$.

2.5 Network data

Our EGO and FD models aim at identifying genes that have significantly changed neighborhoods between the tumor and normal populations. To obtain the highest coverage of the genes in our study, we have constructed our network from a combination of three sources of interactions: physical protein interaction data [downloaded from the HPRD database (Mishra, 2006)], interactions that are derived from pathway membership [downloaded from Pathway Commons (Cerami, 2010)] and regulatory interactions as described in Ravasi (2010). The final network consists of 7708 genes and 49 433 interactions and contains genes that have at least one neighbor and are not part of stand-alone pairs.

3 RESULTS

To illustrate our methodology, we first study a simulated scenario: we impute influential genes in a real interaction network (as described in Section 2.5) and then recover them with our two proposed models. We then apply our method to four lung cancer datasets, using the same interaction network. Since in the real data scenario we do not have the information about which genes have most profound effect on their neighbors, we evaluate our performance in an unbiased way by examining consistency of the found genes among the four independent lung cancer datasets.

3.1 Simulating important genes

We determine how accurately we can recover true *influential* genes in a real network under the assumption of correctness of our methods. We simulate the expression changes as follows:

(1) Randomly sample $K \in \{10, 20, 30, 40, 50\}$ genes and set them to be the known genes of interest.
(2) For each of the K genes, sample $\vec{\alpha}$ uniformly at random on the interval $[0.2, 1]^d$ (the same for $\vec{\beta}$ in the EGO model).
(3) Use the EGO or FD model to propagate the changed expression values to their first- and second-order neighbors, where appropriate.
(4) Add the noise η to all the genes in the network.

The log of expression change (z) is known to be distributed approximately as $N(0, \sigma)$ (Chen *et al.*, 1997). To construct the expression-change noise model, we estimate σ from the housekeeping genes for which the expression level in the given data is not supposed to change. In our data, the basic noise model of the log of expression change is estimated to be $\eta \sim N(0, 0.0842)$.

Note that Step (2) above corresponds to selecting the 'true influential genes', which is then followed by propagating influence to the rest of the network. The range of simulated influence values $\vec{\alpha} : \forall i, \alpha_i \in [0.2, 1]$ is a scaled version of the realistic alpha values, selected for the ease of interpretation. Having simulated the expression change in all genes, our goal is to determine which genes were the initiators of the observed expression change. We use the area under precision recall curve (AUPR) to evaluate how well EGO and FD recover the true influential genes in each simulation scenario. Figure 1 shows the performance of EGO on EGO-simulated expression on five different gene set sizes ($K \in \{10, 20, 30, 40, 50\}$). As shown, EGO can recover true influential gene sets with high precision. EGO performs better on smaller gene sets. Such bias can be explained by the fact that the more genes are affecting their neighbors, the more confounding the effect will be with some of the weaker signal getting harder to recover.

We also investigated how well each of our models can identify the true influential set if the data were generated based on the assumptions of the other. Figure 2 shows the performance of EGO and FD, where the simulated expression is generated under the opposing model. As expected, EGO can accurately recover the true influential genes when the expression is simulated through the FD model, finding second-order interactions to be irrelevant. However, FD fails to recover the second-order effect and thus results in much lower accuracy, when the expression levels were propagated to first- and second-order neighbors according to the EGO model.

In summary, our simulation results show that FD and EGO, can accurately identify the set of true influential genes when the simulated expression coincides with the appropriate model

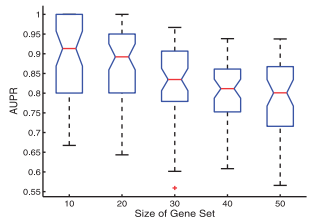

Fig. 1. Performance of EGO in recovering the true influential gene set. The x-axis represent the gene set size (K), the y-axis shows AUPR. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme datapoints. Red crosses indicate sample min/max/outliers. We performed 100 experiments for each gene set size.

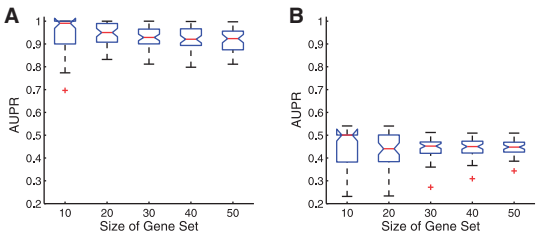

Fig. 2. Performance of EGO and FD in recovering the true influential gene set, when the gene expression change is simulated in contrast to each model's assumption. (**A**) EGO tested on FD simulated expression. (**B**) FD when tested on EGO simulated expression. The x-axis represent the gene set size (K), and the y-axis shows AUPR. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme datapoints. Red crosses indicate sample min/max/outliers. We performed 100 experiments for each gene set size.

assumptions. In addition, the EGO model can also recover the true gene set when the simulated expression conforms to the FD model.

3.2 Robustness to noise in the network

Since to identify genes of interest our model relies on the network connectivity, here we inspect the stability of our performance with respect to the noise in the network. To do so, we have examined scenarios with 0, 1, 10, 25 and 50% of missing edges and added edges (independently). Each network perturbation was performed by removing (or adding) edges uniformly at random. We have found that our list is robust to small perturbations in the network, at least up to 1% of either addition or deletion of edges. At 10% of noise, the median performance drops from 0.9 to 0.75 AUPR for removed edges and to 0.5 for randomly added edges. This confirms our hypothesis that the methodology is much more sensitive to the addition of false edges rather than to missing true ones (at random). The results for the full range of noise levels are shown on Figure 3.

3.3 Experiments on lung cancer dataset

Motivated by our results on the simulated data, here we investigate gene importance in lung cancer using gene expression across four independent cohorts of patients available as part of the Director's Challenge Study [Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma *et al*. (2008)]. Our goal is to show that our EGO model identifies genes across

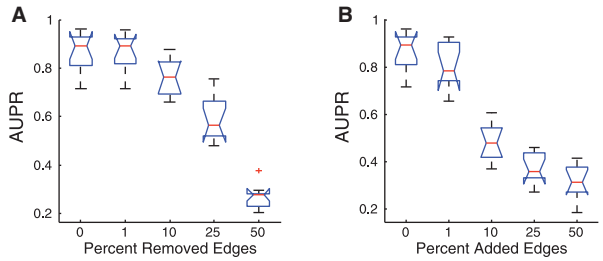

Fig. 3. (**A**) AUPR score (y-axis) for EGO method corresponding to recovering the original gene list in the presence of noise (missing edges) in the network. The amount of perturbation is indicated in percent on the x-axis. (**B**) AUPR of recovering the original gene list in the presence of false positive (added) edges. The edges are added or deleted uniformly at random.

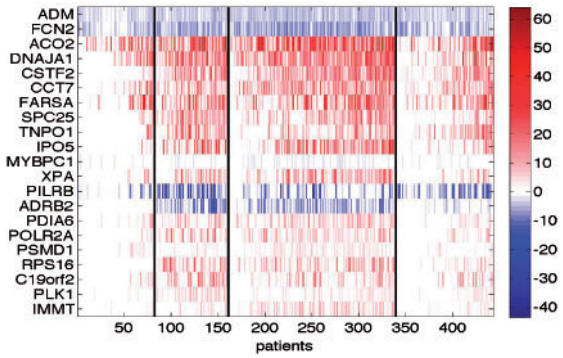

Fig. 4. The heat map shows 21 genes (rows) with non-zero influence factors (non-zero α or β) in at least 50% of patients as found by EGO in one or more of the four Director cohorts (separated by black vertical lines). The colors indicate the value of the influence factor.

these four datasets more consistently than alternative approaches. The four cohorts that we have examined contained 82, 79, 178 and 104 patients and 49 controls as described in Director's Challenge Study. To account for batch effects, we normalized each cohort using RMA (Rafael, 2003). We used SAM (Tusher *et al.*, 2001) to obtain normalized differential expression. For a given dataset, we ran EGO to find influential genes for each patient independently. We then evaluated consistency at two levels: (i) across patients within the same dataset and (ii) across different cohorts. In the following section, we examine the consistency of our predictions across the four datasets, and compare EGO results to (i) FD, (ii) a subnetwork discovery method Chuang *et al.* (2007) and (iii) independent t-testing. It took 1 h to run EGO on all the 443 patients in parallel on a cluster of 400 nodes with <5 GB of RAM per run.

3.3.1 Consistency among lung cancer cohorts To investigate consistency of our method, we have applied our approach to each patient in the four cohorts independently. Figure 4 shows the coefficients α or β, as found by EGO (genes influential in both first- and second-order neighborhoods would appear twice). The genes are ordered by consistency across patients: genes found influential in more patients are at the top, and by increase in explanation power

in a single patient; patients whose differential profile yielded higher number of influential genes are further to the right.

There are 3, 16, 16 and 7 genes that appeared in >50% of the patients in Director 1, 2, 3 and 4 datasets, respectively. The overlap between those four sets of genes is shown in Table 1. Three genes present in >50% of the population across all four cohorts are the top three genes in Figure 4: ADM, FCN2 and ACO2. ADM was found to explain its first-order neighbors (non-zero α coefficient) and FCN2 and ACO2 were found to significantly impact their second-order neighbors (non-zero β). Functionally, ADM has been previously associated with tumor growth (Miller *et al.*, 1996), while FCN2 is part of a module associated with immune response. In our network, FCN2 and its neighborhood are functionally enriched in the activation of immune response ($q = 1.37e\text{-}13$) and immune effector process ($q = 1.06e\text{-}16$, where q-value is FDR corrected P-value). FCN2 itself is a known part of the innate immune system (Garred *et al.*, 2009). ADM and FCN2 and their neighborhoods are shown on Figure 5. While these genes are currently not associated with lung cancer directly, each have been implicated in relevant processes: ADM's inhibition, for example, decreases lung angiogenesis (Vadivel *et al.*, 2010); while lower levels of FCN2 have been associated with recurrent respiratory infections in children (Cedzynski *et al.*, 2009); ACO2 is involved in the Kreb's cycle, which is part of the respiratory system.

We provide the network and dataset characteristics of these three genes in Table 2. We also list the statistics for TP53 and EFGR genes widely known to be associated with cancer. From Table 2, it is evident that the known genes can be recovered by simply studying the degree distribution. Our method instead focuses on genes that could not be revealed by the simple network characteristics, nor by the differential expression levels, but by considering network and differential expression together in principled fashion.

Table 1. Number of genes overlapping between pairs of datasets based on genes found influential by EGO, PinnacleZ and *t*-testing, respectively

EGO/PinnacleZ/*t*-test	D2(16,5)	D3(16,11)	D4(7,33)
D1 (3,10)	3* / 0 / 0	3* / 0 / 0	3* / 0 / 1*
D2 (16, 5)		10* / 3* / 0	7* / 0 / 0
D3 (16,11)			7* / 0 / 0

D1, D2, D3, D4 stand for Director 1,2,3 and 4 datasets. The numbers in parentheses next to the dataset name indicate the total number of genes found influential by EGO in >50% of the patients in each of these datasets and PinnacleZ (we use the same number of genes for *t*-testing as found in EGO).
*Indicates whether the overlap is significant according to the hypergeometric test (the total number of genes is 7708).

3.3.2 Comparison with other methods We have tested the consistency performance of our EGO method against FD, the subnetwork method by Chuang *et al.* (2007) and *t*-tests. We found that there was no overlap between the genes found by FD. Briefly, Chuang *et al.* (2007) finds subnetworks of genes that result in best performance on a classification task (e.g. adenocarcinoma versus normal). It requires a gene network and expression profiles for a set of patients. In the original work, the approach was used to predict metastasis in breast cancer, i.e. the found subnetworks were optimized to predict a binary class indicating whether the breast cancer in the given patient has metastasized or not. We have used the code available as Cytoscape plugin named PinnacleZ as was suggested by the authors using 100 random trials, ST1 and ST2

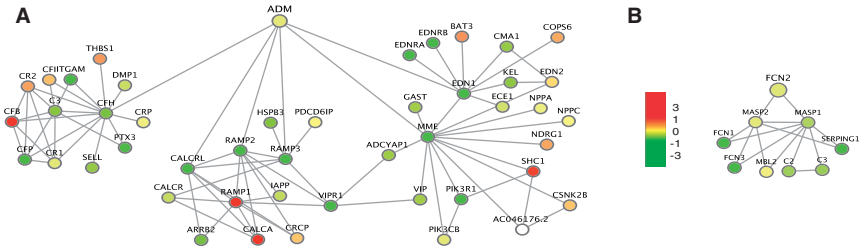

Fig. 5. (**A**) ADM gene and its first- and second-order neighborhoods in our network. The colors represent change in expression levels. ADM was found to be associated with downregulation in its first-order neighborhood. (**B**) FCN2 gene and its first- and second-order neighborhoods in our network. The colors represent change in expression levels. FCN2 was found to be associated with downregulation in its second-order neighborhood.

Table 2. A table of network and expression statistics for the (top) three genes found by EGO to be consistent for over 50% of the patients in all four cohorts considered; (bottom) two known oncogenes for reference

Gene name	$\log_2$(EC)	First-degree median [$\log_2$(EC)]	Second-degree median [$\log_2$(EC)]	Mean(α)	Mean(β)	Function
ADM	−0.11	6 (−1.49)	40 (−0.24)	−3.32	0	Lung angiogenesis; tumor growth
FCN2	−0.02	2 (−0.14)	6 (−0.60)	0	−4.87	respiratory tract infection; innate immunity
ACO2	+0.50	11 (+0.14)	166 (+0.40)	0	+11.10	Kreb's cycle (respiratory system)
EGFR	−0.22	157 (+0.03)	2320 (+0.05)	0	0	Known oncogene
TP53	+0.72	255 (+0.18)	2833 (+0.10)	0	0	Known oncogene

EC stands for the mean expression change ratio across all patients.

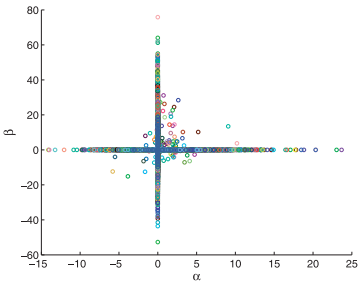

Fig. 6. Each of the 7708 genes influence on its first (x-axis) versus second order (y-axis) neighbors across all four Director cohorts.

P-value cutoffs of 0.05 and 20 000 ST3 trials. Additionally, max node degree was set to 200, min improvement to 0.05, max module size to 20 and max radius to 2. We have experimented with other parameters as well, but not all of them returned a result set. We found that the parameter settings above gave us the most consistent outcome for the four datasets in this study. PinnacleZ resulted in an overlap of three genes between Directors 2 and 3 (Table 1).

Independent t-tests for each of the genes in the four cohorts did not result in consistency across cohorts either. Out of 7708 genes in our network, over 6600 were significantly different between tumor and normal patients in each of the cohorts. We have taken the top (3, 16, 16 and 7) genes according to EGO gene list sizes and ordered genes by decreasing t-statistic in the case of a tie in P-value. There was no overlap between t-test gene lists across the four lung cancer cohorts on a full set of 11 911 genes. Once we restricted the list to the 7708 from our network, there was an overlap of one gene between datasets D1 and D4 as shown in Table 1.

In contrast, EGO finds three genes that overlap among all four cohorts: ADM, FCN2 and ACO2. We note that it may be beneficial to look at both approaches: one that reveals consistency in change in gene neighborhoods and one that reveals consistency across differentially expressed genes since the two approaches compliment each other. Full lists of genes for all methods for each of the Directors datasets are available in the Supplementary Materials.

3.3.3 First- versus second-order influence factors In addition to more consistency in our results, we can glean further information about genes by inspecting our influence factors α and β. Figure 6 shows the scatter plot of influence on the first- versus second-order neighbors for all genes across all patients. The majority of the mass on that plot is at (0,0). Those genes that have non-zero influence exhibit either first- or second-order influence but not both. This is interesting because there are no constraints that would not allow selection of the same gene twice. Since in our networks there are no second-order paths that have a corresponding first-order path between two genes (see definition of W^{n*} in Section 2.1), we are able to reveal genes that exhibit independent second-order influence and are not simply part of tightly knit clusters.

We have then examined each individual patient for genes that had both non-zero first- and second-order effects. We found that out of 443 patients, 74 had one gene with effect on both first- and second-order neighbors, and in four patients there were two genes that had effect on both. Further analysis showed that there were only three genes implicated in this joint influence for all patients considered. These were S100A12 (0,6,32,19), LGALS1 (2,0,8,4), PDIA6 (2,2,4,2), where the numbers in parenthesis indicate the number of patients in which these genes had non-zero α and β coefficients simultaneously. While LGALS1 and PDIA6 do not have a strong signal, the pro-inflammatory protein S100A12 is present in almost 20% of patients in Director cohorts 3 and 4. In addition, S100A12 is a known marker for chronic and acute lung diseases (Lorenz *et al.*, 2008). This fonding is in line with the other lung inflammation factors that we found to be consistently influential.

Our results lead us to believe that different genes exhibit different patterns of influence. It is thus very important to go beyond immediate neighborhoods to understand each gene's role and associations. We also note that most of the genes we found to be consistent among patients are those with second-rather than first-order influences. Indeed, from the 21 genes on Figure 4, only 4 (ADM, MYBPC1, PLK1 and PSMD1) are first-order genes, the remaining 17 genes were influential in second-order neighborhoods.

Finally, we have performed additional experiments on two completely independent cohorts of pancreatic cancer patients. We have again found much higher consistency by EGO as opposed to the alternatives considered. In particular, of the total of 107 genes found by EGO to be consistent among patients within each dataset alone, 26 genes overlap between the two datasets ($P=0$). PinnacleZ finds a total of 373 genes of which 17 genes overlap between the two datasets ($P=0.5\times10^{-5}$) (Supplementary Materials).

4 DISCUSSION

Interestingly, many of the genes identified as influential and consistent between cohorts were genes with significant second-order influence, alluding to the importance of post-translational modifiers. We believe that allowing for independent recovery of such genes (de-coupling of α and β parameters in our model) is one of the reasons for our superior performance. One can think of our method as a projection of many genes onto a lower dimensional set of influential genes. Our results reveal that these projections tend to be more stable. Additionally, all the genes that we have recovered while not being directly implicated in lung cancer, are associated with acute or chronic respiratory inflammation.

One of the limitations of this work is modeling the change in expression among neighbors as strictly additive. We do not directly model the case where one neighbor is acting to increase and another to decrease the expression level of a gene, unless they do so consistently in their own neighborhoods resulting in jointly positive (negative) impact. Though capturing those relations would be ideal, it would require more data than are typically available and an increased number of parameters. Otherwise, the model may be unidentifiable [e.g. Ergün *et al.* (2007)].

In summary, we propose an unsupervised approach to combing through thousands of genes to identify a few of importance. We define genes of importance as having influence on expression change of their neighbors. Applied to a set of real cancer datasets, our method finds a set of genes consistent across four cohorts of adenocarcinoma lung patients considered. The influential gene set is three orders of magnitude smaller than the total number of genes. All the found genes have already been associated with respiratory inflammation processes and might warrant further investigation.

Funding: Canadian Institute of Health Research (MOP-93671 to Q.D.M.); Ontario Institute for Cancer Research through funding provided by the Government of Ontario (to P.C.B.).

Conflict of Interest: none declared.

REFERENCES

Cedzynski,M. *et al.* (2009) L-ficolin (ficolin-2) insufficiency is associated with combined allergic and infectious respiratory disease in children. *Mol. Immunol.*, **47**, 415–419.

Cerami,E. *et al.* (2010) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.

Chen,Y. *et al.* (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.*, **2**, 364–374.

Chuang,H. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**.

Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma. *et al*. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med*, **14**, 822–827.

Efroni,C. *et al.* (2007) Identification of key processes underlying phenotypes using biological pathway analysis. *PLoS One*, **2**, e425.

Ergün,A. *et al.* (2007) A network biology approach to prostate cancer. *Mol. Syst. Biol.*, **3**.

Fortney,K. *et al.* (2010) Inferring the functions of longevity genes with modular subnetwork biomarkers of caenorhabditis elegans aging. *Genome Biol.*, **11**, R13.

Friedman,J. *et al.* (2010) Regularized paths for generalized linear models via coordinate descent. *J. Stat. Softwr.*, **33**.

Gao,F. *et al.* (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Garred,P. *et al.* (2009) The genetics of ficolins. *J. Innate Immun.*, **2**, 3–16.

Hwang,T. and Park,T. (2009) Identification of differentially expressed subnetworks based on multivariate ANOVA. *BMC Bioinformatics*, **10**.

Jemal,A. *et al.* (2008) Cancer statistics, 2008. *CA Cancer J. Clin*., **58**, 71–96.

Jonsson,P. and Bates,P. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics*, **22**, 2291–2297.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Lorenz,E. *et al.* (2008) Different expression ratio of s100a8/a9 and s100a12 in acute and chronic lung diseases. *Respir. Med.*, **102**, 567–573.

Milenkovic,T. and Przulj,N. (2008) Uncovering biological network function via graphlet degree signatures.*Cancer Informat.*, **4**, 257–273.

Miller,M.J. *et al.* (1996) Adrenomedullin expression in human tumor cell lines. its potential role as an autocrine growth factor. *J. Biol. Chem.*, **271**, 23345–23351.

Mishra,G. *et al.* (2006) Human protein reference database – 2006 update. *Nucleic Acids Res.*, **34**, D411–D414.

Navlakha,S. and Kingsford,C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, **26**, 1057–1063.

Pradines,J. *et al.* (2004) Detection of activity centers in cellular pathways using transcript profiling. *J. Biopharm. Stat.*, **14**, 701–721.

Rafael,A. *et al*. (2003) Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.*, **31**, e15.

Ravasi,T. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, D411–D414.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Taylor,I. *et al.* (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, **27**, 199–204.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Vadivel,A. *et al.*(2010) Adrenomedullin promotes lung angiogenesis, alveolar development, and repair. *Am. J. Respir. Cell Mol. Biol.*, **43**, 152–160.

Vandin,E. *et al*. (2011) Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.*, **18**, 507–522.

Vaquerizas,J.M. *et al.* (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.

Vogelstein,B. and Kinzler,K. (2006) Cancer genes and the pathways they control. *Nat. Methods*, **10**, 789–799.

Watters,J. and Roberts,C. (2004) Developing gene expression signatures of pathway deregulation in tumors. *Mol. Cancer Ther.*, **5**, 2444–2449.

Yan,X. and Sun,F. (2008) Testing gene set enrichment for subset of genes: Sub-gse. *BMC Bioinformatics*, **9**, 362.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc.*, **67**, 301–320.

Zou,H. and Zhang,H.H. (2009) On the adaptive elastic-net with a diverging number of parameters. *Ann. Stat.*, **37**, 1733–1751.