Gene expression

Prestige centrality-based functional outlier detection in gene expression analysis

Ali Torkamani* and Nicholas J. Schork

The Scripps Translational Science Institute and Scripps Genomic Medicine, Scripps Health and The Scripps Research Institute, 3344 North Torrey Pines Court, Room 306 La Jolla, CA 92037 USA

Received on March 13, 2009; revised on June 11, 2009; accepted on June 19, 2009

Advance Access publication June 23, 2009

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Traditional gene expression analysis techniques capture an average gene expression state across sample replicates. However, the average signal across replicates will not capture activated gene networks in different states across replicates. For example, if a particular gene expression network is activated within a subset or all sample replicates, yet the activation state across the sample replicates differs by the specific genes activated in each replicate, the activation of this network will be washed out by averaging across replicates. This situation is likely to occur in single cell gene expression experiments or in noisy experimental settings where a small sub-population of cells contributes to the gene expression signature of interest.

Results and Implementation: In this light, we developed a novel network-based approach which considers gene expression within each replicate across its entire gene expression profile, and identifies outliers across replicates. The power of this method is demonstrated by its ability to enrich for distant metastasis related genes derived from noisy expression data of CD44+/CD24-/low tumor initiating cells.

Contact: atorkama@scripps.edu; atorkama@scrippshealth.org

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Current genome-wide array-based gene expression techniques capture a snapshot of gene expression within a specific population of cells. Gene expression in individual cells is known to be stochastic in nature, with gene and protein expression occurring in short bursts (Cai et al., 2006; Kaern et al., 2005; Kaufmann and van Oudenaarden, 2007; Volfson et al., 2006). Since the expression profile of each individual cell is captured at a discrete point in time, the cell-to-cell variability of gene expression will vary greatly, and the overall snapshot represents the average gene expression state across the population of cells. In well-defined, homogeneous, populations of cells the average expression state is likely to accurately reflect gene expression networks activated in the population. However, gene expression signatures derived from single cell expression experiments clearly do not benefit from this averaging. Additionally, in expression analyses of heterogeneous cell populations, where the gene expression signature of interest is contained within a minority of cells, the stochastic nature of gene expression is likely to play a more significant role in the resultant gene expression signature. In this case, only the most dominant gene expression states within the sub-population of interest will be apparent, and averaging across multiple replicates may wash out the signal of interest if the dominant gene expression state differs significantly across replicates.

An example of a heterogeneous cell population with a minority of cells of biological interest is CD44+/CD24-/low putative breast cancer stem cells (Al-Hajj et al., 2003). Although there is a great deal of controversy surrounding whether or not this cell population represents true cancer stem cells (Lewis, 2008), this cell population has undeniable tumor initiating properties and prevalence of CD44+/CD24-/low cells is strongly associated with distant, overt, hematogenous metastases (Abraham et al., 2005). However, only a small proportion of cancer cells sorted for the CD44+/CD24-/low markers are capable of initiating tumors. Approximately 200 sorted cells must be injected to give rise to tumors in NOD/SCID mice (Al-Hajj et al., 2003). A gene signature derived from tumorigenic CD44+/CD24-/low breast cancer cells was demonstrated to be associated with both overall and metastasis-free survival in patients with breast cancer (Liu et al., 2007). However, the signature, described by Liu et al., was derived from comparing tumorigenic cancer cells to normal breast epithelial cells rather than comparison of tumorigenic cancer cells to non-tumorigenic (CD24+) cancer cells.

A comparison of gene expression patterns in tumorigenic breast cancer cells and non-tumorigenic breast cancer cells to normal breast epithelium resulted in gene lists that overlapped significantly (60% overlap) suggesting the gene signature may represent oncogenically transformed cells rather than invasiveness (Massagué, 2007). If gene expression noise from the majority of CD44+/CD24-/low cells which are unable to initiate tumors partially masks the signature of the true tumor initiating cells in the population, averaging across replicates of CD44+/CD24-/low cell populations may further obscure the invasiveness signature, thus we propose an alternative network-based gene expression analysis technique.

We describe a method and analysis of gene expression patterns which builds on techniques developed for social network analysis, in which network perturbation ‘scores’ are assigned to individual genes on the basis of their contribution to gene expression patterns identified from more global genomic gene expression patterns across all genes, rather than from individual state-dependent gene expression values. The scores and measures we use are similar to
the 'prestige measure of centrality', where the status of an 'actor' (in our case a single gene's network perturbation score) is recursively related to the status of the 'actors' (i.e. other genes' expression values) to which it is connected (Wasserman and Faust, 1994). Network-based approaches have previously been shown to identify genes of functional or prognostic value in similar contexts (Barabasi and Oltavi, 2004; Carter et al., 2004; Horvath and Dong, 2008; Horvath et al., 2006).

Our method forms connections between genes based upon their co-expression with one another, and scores genes based upon the strength of those connections (e.g. the correlation or mutual information measure), the expression level of those connected genes, and the importance of those connected genes in the network as a whole, defined by the degree centrality of the connected gene. In addition, we adapt our gene expression relative to the overall state of each cell population, where gene scores are most strongly influenced by collections of other genes which are closely related to the gene being scored. This scoring controls for random gene expression fluctuations and amplifies coordinated gene expression fluctuations across closely connected gene networks. Prestige centrality scores determined in this manner result in highly correlated scores between replicates, and require a strong network-based perturbation, rather than the fluctuation of the expression of a single gene, in order to produce 'outlier' genes whose expression levels break from the correlation structure associated with a particular cellular state. Thus, genes of interest are identified by a coordinated activation of their genes of interest are identified by a coordinated activation of their (Lauss et al., 2008) and that outlier status enriches for genes, either metastases related or tumorigenesis-related genes, which are activated in primary bulk tumors that will ultimately result in distant metastases.

2 SYSTEM AND METHODS

In order to derive a network-based prestige centrality score relevant to breast cancer, we must first reconstruct the breast tissue/breast cancer co-expression network. This could be done via a number of different algorithms, including straightforward analysis of pair-wise gene expression correlations and contrasts; however, we chose to use the ARACNE algorithm because of its proven performance and computational tractability on a whole-genome scale (Margolin et al., 2006). Essentially, ARACNE was used to calculate a mutual information score for every gene pair represented on the Affymetrix U133 2.0 Plus platform (20 135 genes) using numerous datasets representing multiple perturbed states of normal and cancerous breast tissue (Bild et al., 2006; Bourdeau et al., 2007; Duss et al., 2007; Huang et al., 2007; Ince et al., 2007; Johnstone et al., 2006; Loi et al., 2007; Richardson et al., 2006; Schuetz et al., 2006; Turashvili et al., 2007) (466 samples in total, 386 tumor samples, 80 normal samples). We include both normal and cancerous breast tissue because we expect cancerous tissue to represent the 'prestige measure of centrality', where the status of an 'actor' (in our case a single gene's network perturbation score) is recursively related to the status of the 'actors' (i.e. other genes' expression values) to which it is connected (Wasserman and Faust, 1994). Network-based approaches have previously been shown to identify genes of functional or prognostic value in similar contexts (Barabasi and Oltavi, 2004; Carter et al., 2004; Horvath and Dong, 2008; Horvath et al., 2006).

The mutual information scores (I) were used to construct a symmetric, undirected, adjacency matrix (A), where self connections are not allowed (i.e. the diagonal is set to 0), and the connection strength between genes x and y is simply equivalent to the mutual information score. The elements, a_{xy} of the matrix are thus defined as: a_{xy}=I(x,y). This adjacency matrix provided the framework for the calculation of prestige centrality scores.

3 ALGORITHM

First, the adjacency matrix, A, is used to calculate a 'score' for the importance or relative contribution of each gene to the breast tissue/breast cancer network. This score is equivalent to the 'degree of centrality' (C(D)) computed from the weighted adjacency matrix as the sum of all the mutual information scores for each gene. If we define n as the number of genes, the degree centrality score is defined as:

\[ C_D(x) = \sum_{y=1}^{n} I(x,y). \]  

These degree centralities scores are then standardized to range between 0 and 1 (standardization is only useful for ease of interpretation of the degree centrality scores, it is not required). We define the 'prestige centrality score' (P) of each gene \( x \) in each sample \( i \) as the sum of the following product for all genes \( y \) connected to gene \( x \)—degree centrality of gene \( y \), \( C_D(y) \), multiplied by the log fold change in expression of the connected gene \( y \) in sample \( i \), \( E_i(y) \), multiplied by the connection strength between gene \( x \) and gene \( y \) (mutual information, \( I(x,y) \)). Formally:

\[ P(x) = \sum_{y=1}^{n} I(x,y) \cdot C_D(y) \cdot E_i(y). \]  

Note two unique features of this method: (i) the score of each gene depends upon the expression level of all other genes, and the expression level of the gene itself has no bearing upon its score; (ii) given the adjacency matrix and degree centralities for any given gene, its prestige centrality score can be calculated from a dataset even if that gene is not represented on the expression chip since the score for that gene can be imputed from the expression level of all genes represented on the expression chip. In a sense, gene scores are adjusted across the entire gene expression profile of the sample, penalizing random fluctuations in individual genes while amplifying coordinated fluctuations across closely connected networks. Additionally, the relevance of a gene expression fluctuation is weighted by its connectivity to the network as a whole, by the degree centrality measure.

\[ \text{Functional outlier detection in gene expression analysis} \]
As can be observed in Figure 1, even completely uncorrelated data changes across its nearby network neighborhood. Where $P_i(x)$ represents a matrix containing the prestige centrality scores of gene $x$ over samples $i$, $P_{avg}(x)$ is a matrix of the means of $P_i(x)$ not including the $i$-th sample, and $S$ is the estimated covariance matrix of the entire dataset not including gene $x$. Thus, the final jackknifed outlier distances of the prestige centrality score for gene $x$ identifies genes whose prestige centrality score, in either one or a subset of samples, lies outside of the overall correlation structure, presumably through a concentration of outlier gene expression fold changes across its nearby network neighborhood.

4 IMPLEMENTATION

4.1 Correlation structure of the prestige centrality score

Multiple replicates within a dataset, whether raw gene expression data are correlated or not, are forced into a highly correlated structure via the prestige centrality score. To explore the relationship between the original expression data’s correlation structure, and the correlation structure derived from the prestige centrality score, we generated $100$ random ‘cancer’ sample expression patterns with $100$ genes per pattern. Each simulated gene expression value was randomly assigned a value between zero and one. Next, we generated a random adjacency matrix by randomly generating $100 \times 100$ mutual information scores ranging between $0$ and $1$. The degree centrality scores were then calculated, as previously described, as the sum of all mutual information scores for a gene. Using these randomly generated expression and mutual information values, we calculated all pair-wise correlations for the raw expression data and derived prestige scores. After each round, the expression levels of a single gene were set equal to one another across all ‘cancer’ samples to increase the correlation coefficient between all samples. This was done recursively until all expression sets were equivalent. As can be observed in Figure 1, even completely uncorrelated data can be forced into a strong correlation structure with a correlation coefficient $\geq 0.76$ (Fig. 1). This property of the prestige centrality score is an important one, because, ultimately, we aim to identify genes forced out of the strong correlation structure induced by the prestige centrality transformation of raw gene expression data. Outliers from this highly correlated structure will be induced by gene expression fluctuations across a gene’s network neighborhood, thus identifying the gene as central to a perturbed gene expression network.

4.2 Application to CD44+CD24-/low

In order to verify the biological relevance of outlier genes identified by the prestige centrality transformation, we determined whether or not prestige centrality outliers calculated from tumorigenic CD44+CD24-/low expression data do, in fact, represent genes important in mediating recurrence by determining whether the consensus gene expression signature for breast cancer recurrence (Breast-CGLS) (Lauss et al., 2008) is overrepresented among prestige centrality outliers. Metastasis is intimately related to recurrence, although not equivalent. For example, drug efflux transporters would be expected to mediate recurrence but not metastasis.

Data from Liu et al. (2007), including nine tumorigenic cancer cell replicates, three non-tumorigenic cancer cell replicates and three replicates of normal breast epithelium was subject to the prestige centrality transformation. Specifically, we calculated a log base 2-fold change in expression per gene per tumorigenic cancer cell sample versus the average expression per gene in normal breast tissue and applied the prestige centrality algorithm followed by multivariate outlier analysis (see Supplementary Table 1 for ranked list). The expression network associated with the identified outlier genes must be sufficiently perturbed within any subset of replicates so as to place its prestige centrality score outside of the correlation structure obtained over all genes, suggesting the presence of a true concerted, functional signal involving the oulying genes.

For 357 Breast-CGLS genes represented in the Liu et al., tumorigenic breast cancer cell dataset, the mean outlier distance was higher than the average outlier distance (CGLS mean $= 3.16 \pm 0.05$, others $2.24 \pm 0.007$) and highly statistically significant by Rank Sums test (one way Chi square approximation $\chi^2 = 135.7$, $p = 2.4 \cdot 10^{-31}$) or ANOVA ($F$-ratio = 331, $p = 2.3 \cdot 10^{-73}$). Of the top 100 outliers (out of a total of 20,135 genes), about one fourth (23) belonged to the Breast-CGLS (cumulative probability of 23 or more outliers $p = 1.8 \cdot 10^{-19}$ based on the hypergeometric distribution. Therefore, the prestige centrality transformation is useful in predicting biologically functional genes.

Additionally, we investigated whether inclusion of normal breast tissue in the network reconstruction altered the results significantly. The 80 normal tissue samples were removed from the original expression data and network reconstruction was performed again. We found that mutual information scores were only changed by 5% on average and that gene ranks determined by analysis of the Liu et al. tumorigenic breast cancer cell dataset where unchanged for the top outliers. Inclusion of a small number of normal tissue samples
4.3 Distant metastases

It has been reported that the prevalence of CD44+CD24-/low breast cancer cells is associated with distant metastases (Abraham et al., 2005). To demonstrate the prestige centrality data analysis method enriches for metastases related genes, we evaluated the significance of the outlier distance in enriching for genes involved in distant metastasis in an independent test set and compared the enrichment with standard analyses (Wang et al., 2005). Our independent test set involves gene expression profiles of primary clinical lymph-node-negative breast cancer samples with 179 relapse free patients and 107 patients who developed distant metastases.

First, we determined whether the prestige centrality outlier distances calculated from Liu et al. (2007) more accurately captured a metastasis gene signature as compared with standard differential expression or outlier analysis of the raw Liu et al. expression data. We calculated a t-test statistic for genes differentially expressed in patients who developed distant metastases versus those who did not from Wang et al. (2005), and compared the strength of the correlation between these t-test statistics versus the prestige centrality outlier distances, direct calculation of outliers from the raw expression data or standard differential expression analysis p-values of tumorigenic cells versus normal breast epithelium from Liu et al. (2007) (Table 1). Clearly, prestige centrality outlier scores more accurately capture genes differentially expressed in tumors which are destined to undergo metastases as compared with simple differential expression in tumorigenic versus normal breast tissue, or the identification of outlier genes based upon raw expression of tumorigenic cancer cells. This suggests that stochastic fluctuations in gene expression in the heterogeneous CD44+/CD24-/low breast cancer cell populations creates noise which can be regularized by the prestige centrality algorithm. Other centrality measures do not have this advantage. For example, we compared utilizing degree centrality as a weight for expression values and found the results were nearly identical to utilizing raw expression values (correlation of results = 0.97). This is likely due to the fact that utilizing only degree centrality does not induce regularization of gene expression patterns, which is the primary strength of the prestige centrality approach.

To visually demonstrate the degree of enrichment for genes differentially expressed in tumors which are destined to undergo metastases by the prestige centrality method versus the other aforementioned methods, genes were rank ordered by their outlier distance in the case of prestige centrality, by their outlier distance calculated on the basis of the raw expression data, or by their t-test statistic in the case of standard differential gene expression analysis. To visually depict the level of enrichment, we moved up the gene rankings, from the lowest ranking to the highest ranking gene, for each analysis type, and calculated the average absolute t-test statistic for all genes ranked above the current gene, using the t-test statistics of Wang et al. (2005) (Fig. 2). In other words, the first data point corresponds to the overall average t-test statistic, the next data point represents the average t-test statistic for all genes except for the lowest ranking gene, and so on until the last data point represents only the t-test statistic for the highest ranking gene, i.e. each data point represents a progressively larger threshold outlier distance.

<table>
<thead>
<tr>
<th>Method</th>
<th>Prestige centrality</th>
<th>Expression outlier</th>
<th>Differential expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation with metastases genes</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td>Comparison with prestige centrality</td>
<td>p = 1.00</td>
<td>p = 1.11 x 10^{-8}</td>
<td>p = 9.3 x 10^{-13}</td>
</tr>
</tbody>
</table>

Table 1. Correlation with metastases-related genes

For visual purposes, Figure 2 displays the gene order for all analysis types on the prestige centrality outlier scale, preserving the gene order for all analyses, and allowing for comparison to prestige centrality. Figure 2 displays enrichment for metastasis related genes by the prestige centrality method (triangles), by determining outliers based upon differential expression of tumorigenic versus normal breast epithelium (circles). The prestige centrality method clearly enriches for genes with progressively higher absolute t-test statistics. The average test statistic increases exponentially suggesting that outlier genes with increasing outlier distance are involved in distant metastases.

for calculating the average t-test statistic for all genes above this threshold.

For visual purposes, Figure 2 displays the gene order for all analysis types on the prestige centrality outlier scale, preserving the gene order for all analyses, and allowing for comparison to prestige centrality. Figure 2 displays enrichment for metastasis related genes by the prestige centrality method (triangles), by determining outliers based upon differential expression of tumorigenic versus normal breast epithelium (circles). Figure 2 clearly demonstrates a progressive enrichment of metastasis-related genes by the prestige centrality method. Figure 3 displays this enrichment analysis if genes are ordered randomly. Similarly, genes from Wang et al. (2005) were randomly assigned to outlier distances and the average t-test statistic for genes above each threshold outlier distance was calculated. This process was repeated 10 times and the average of the 10 iterations is plotted above. As expected, when genes are ordered randomly, the average t-test statistic does not deviate from the overall mean in any systematic way (Fig. 3).
4.4 Top genes predict metastasis-free survival

Finally, we demonstrate that the prestige centrality method is capable of identifying genes predictive of metastasis-free survival, with equivalent to slightly improved predictive power over the invasiveness gene signature (IGS) (Liu et al., 2007). One hundred and five genes of the IGS match genes within the lymph-node-negative breast cancer dataset investigating distant metastases (Wang et al., 2005). Thus, we compared the predictive power of the top 105 genes identified by the prestige centrality method with the 105 genes of the IGS (Fig. 4). There is absolutely no overlap between the IGS and prestige centrality genes. We chose to perform a simple logistic regression to minimize any bias that may be introduced by more sophisticated prediction algorithms. As can be observed in Figure 4, the prestige centrality genes (area under the curve (AUC) = 0.83 ± 0.02) predict metastases free survival slightly better than the IGS (AUC = 0.78 ± 0.03) (p < 0.0001). AUC and p-values were calculated empirically (Lasko et al., 2005). Note this analyses is neither meant to propose a new predictive signature nor to claim a significant improvement in predictability, but is meant to merely demonstrate the relevance of genes identified by the prestige centrality method to metastases. The comparison of the two predictive signatures is certainly affected by the number of genes within the IGS not present in our test dataset. Additionally, many of the genes identified by the prestige centrality method may mediate metastasis, but may not be differentially expressed in tumors prone to metastases. Genes identified by the prestige centrality method may simply be closely connected to differentially expressed genes important in mediating metastasis, and thus would rank highly by the prestige centrality method, but do not necessarily display differential expression themselves. Therefore, although the prestige centrality method does identify genes differentially expressed in tumors which ultimately undergo metastases, the purpose of the method is to identify genes participating in networks which are functionally important to the process of metastases regardless of individual gene level expression differences.

5 DISCUSSION

Network reconstruction has been increasingly used to identify genetic networks of biological importance (Ideker and Sharan, 2008; Lee et al., 2008; Mani et al., 2008; Oldham et al., 2008; Zhang and Horvath, 2005). In this article, we leverage network reconstruction techniques and concepts from social networking analysis to draw multi-gene inferences from noisy gene expression data. Our novel network scoring approach allows genes to be scored based upon coordinated fluctuations within networks they associate with or contribute to in order to tease apart irrelevant contaminating gene expression noise from true coordinated expression events. Using this approach, we are able to identify important gene expression events within a subset of sample replicates which would be missed by averaging across other noisy replicates. This analysis technique should be useful in setting where a homogenous cell population is unavailable. We expect that other correlation measures, such as Pearson’s correlation, may lead to slightly different results, but suggest that mutual information captures non-linear relationships and thus, should provide the most biologically relevant results.

As noted, an extremely provocative field of study, where homogenous cell populations are unavailable, is the study of cancer stem cells or tumor initiating cells. To demonstrate the power of the prestige centrality approach in this context, we applied it to putative breast cancer stem cells. Cancer stem cells have been notoriously difficult to characterize and isolate, with no known markers exclusively expressed by cancer stem cells (Visvader and Lindeman, 2008). Thus, isolation of putative cancer stem cells by surrogate markers results in a mixed collection of tumor initiating cells and other cancer cells without invasive capabilities. By applying the prestige centrality method to CD44+/CD24−/low, we were able to demonstrate its ability to identify biologically important gene expression signals in the face of contaminating gene expression noise.

 Examination of the top-ranked genes by the prestige centrality method produces genes well known to play a role in cell motility, migration, filopodia formation, centrosome organization and metastases. Gene ontology enrichment analysis of the top...
200 genes reveals strong enrichment for mitotic genes, and genes involved in immune cell differentiation and proliferation (Supplementary Table 2) (Zhang et al., 2004). The excess of immune cell differentiation and proliferation genes likely represents the epithelial to mesenchymal transition required for metastasis (Thiery, 2003). Some clear examples of genes which may be directly involved in metastasis include GMFG (rank 14) and its role in cytoskeletal reorganization during cell migration (Ikeda et al., 2006), NCKAP1L and its role in filopodia formation (Weiner et al., 2007), HMMR (rank 37) and its role in metastasis (Wang et al., 1998), CCR7 (rank 60) and its role in cell migration (Till et al., 2002), and LSP1 (rank 77) and its role in transendothelial migration (Liu et al., 2005). The top-ranked gene MPEGI, a peroxin-like protein (Mahl et al., 2004) and a series of granzymes in our network, lead to the intriguing hypothesis that in invasive breast cancer cells may utilize the cytotoxic peroxin/granzyme cytolytic pathway to migrate through dense tissue or the primary tumor itself (Trapani and Smyth, 2002).

In conclusion, we show the utility of the prestige centrality method to the identification of functionally relevant subsets of genes and samples in a heterogeneous collection of expression studies, and demonstrate its performance by identifying metastasis related genes in CD4+CD24−low tumor initiating cell expression data.

ACKNOWLEDGEMENTS

The authors would like to thank Brinnhilde Felding-Habermann, Mihaela Loinger and Deindre O’Sullivan for their helpful discussion on breast cancer stem cells.

Funding: The National Institute on Aging Longevity Consortium (grant number U19 AG023122-01 to N.J.S. and his laboratory); The NIMH-funded Genetic Association Information Network Study of Bipolar Disorder National (grant number 1 R01 MH078151-01A to N.J.S. and his laboratory); National Institutes of Health grants (grant numbers N01 MH22005, U01 DA024417-01, P50 MH018755-01 to N.J.S. and his laboratory); Scripps Genomic Medicine and the Scripps Translational Sciences Institute Clinical Scholarship to A.T.

Conflict of Interest: none declared.

REFERENCES


