Data-driven Networking Reveals 5-Genes Signature for Early Detection of Lung Cancer

Vladimir Kuznetsov
Bioinformatics Institute,
Singapore
vladimirk@bii.a-star.edu.sg

Sterling Thomas
Virginia Commonwealth University
sthomas@vcu.edu

Danail Bonchev
Virginia Commonwealth University
dgbonchev@vcu.edu

Abstract

A new strategy is developed for a gene signature search proceeding from a biological basis in analyzing microarray databases. The procedure involves a combination of known and original methods for correlation, statistical, and network analysis. The application of the strategy to lung adenocarcinoma resulted in a 5-gene signature, which included four genes not associated earlier with lung adenocarcinoma. 93-96% accuracy of classification of cancer vs. normal was achieved. The final stage of our procedure included expanding of the gene signature network to a 43 gene/protein network, which showed that the five genes are in the cross-talk of 24 pathways, providing thus information for mechanistic analysis.

1. Introduction

Lung cancer is characterized by the highest rate of cancer mortality in the United States, a rate much higher than that of prostate, colorectal and breast cancer – 86% vs. 7%, 15% and 37%, respectively [1]. The lack of reliable methods for early diagnosis results in lung cancer detection at an advanced stage, when it is too late for an effective treatment. The recent advance in technology made possible large scale screenings of genes and proteins, and produced the first markers for early lung cancer screening and survival prognosis [2-8].

While producing encouraging results, these gene signature studies have also certain limitations. The large groups of genes that are offered as potential markers could hardly be directly applicable to clinical purposes [9], which would be ideally satisfied by a single highly discriminating gene. Gene signature classification and prognostic studies are frequently biased by the previous findings in the field.

In our study we proceeded from a different strategy for searching gene signature for early detection of lung adenocarcinoma. While also using microarray databases analysis from more than one microarray database, we search for gene ontology categories that would produce the best basis for identifying small size and high sensitivity gene signatures. The genes from the selected categories are then subjected to correlation analysis, which reduces strongly the pool of potential candidates proceeding from a selected cut-off value for correlation significance. A consensus network is then built from the intersecting genes and significant gene-gene statistical correlations from the databases used. The set of genes thus selected is analyzed to identify the subset of genes that provides the highest accuracy of cancer vs. normal patients’ classification. The final step of the procedure includes a search for a physical and biological equivalent of the network built. A network is constructed from the genes and proteins identified in the previous step, and the information found in public databases for their interactions. The network thus built necessarily includes other proteins, which cross-talk to those preliminary selected. This provides an important gene ontology feedback on the pathways and biological processes involved in the complex integrated carcinogenesis process.

2. Methods

2.1. Microarray and Gene Data

mRNA expression data from individuals diagnosed with adenocarcinoma of the lung were acquired from Oncomine [10] and from Genome Expression Omnibus (GEO, 2007) [11]. From Oncomine we acquired U95A platform expression data (12,600 probes) from a classification study of patients with lung malignancies [2] including 62 adenocarcinomas and 17 normal samples. The 62 adenocarcinomas were selected based on agreement between assessments of two independent pathologists. Samples where one report did not indicate pure adenocarcinoma were excluded, and the same was done with the data for patients with secondary metastasis of a different morphology. This produced a dataset of pure adenocarcinomas with no metastasis, tumor sizes of 1 to 8 cm, and all stages. The Bhattacharjee et al. expression data [2] were thus partitioned into five subsets, four of which containing 15 or 16 tumor samples each randomly distributed.
from the pool of 62 samples, and one set of expression data with 17 normal samples. From GEO we acquired expression data from 27 adenocarcinoma patients with accompanying U133A platform expression (22,284 probes) from adjacent normal tissue (Su et al. 2007[12]). Expression data was thus presented by two datasets of 27 samples each (normal and adenocarcinoma tissues).

2.2. Gene List Creation

We compiled gene lists from a novel strategy that proceeds from certain hypothesis as to which biological processes and molecular functions are of importance in carcinogenesis. Recent studies on the mechanisms of lung cancer pay a considerable attention on the chromatin structure changes, and histone as the chief protein component of chromatin, during carcinogenesis [13,14]. Thus, one of the specific targets of the study was to identify proteins that are strongly affected in the different stages of lung adenocarcinoma. Other classes of proteins related to DNA modification, MAP Kinase activity, and transcription were also considered to be of importance. Two lists were created. The first one with 98 genes was based on Gene Ontology Database [GO] [15] classifications provided by Affymetrix Inc. The second list with 43 genes was a merge of Panther [16] and David [17] ontology and GO terms, of the four categories described above.

2.3. Correlation Analysis

We began analysis by randomly organizing our patient pool and calculating the Spearman non-linear correlation coefficients $\rho$ between pairs of expressed genes. A network map was assembled from each dataset linking genes with coefficients higher than 0.7 and less than $-0.7$ ($p<0.01$).

A second analysis was done by grouping genes using proprietary clustering algorithm [18] and calculating $\rho$ as a clustered group between normal expression and diseased expression. The clustering algorithm estimated node degree based on two-dimensional geometric clustering of expression data (i.e. gene X and gene Y represented as point $(X,Y)$). Cluster estimation was done using gap static [19], each gene produced an integer representing the estimated number of clusters. Each gene pair with cluster estimation > 1 was recorded producing a list of genes; the count of this gene list represents an unsupervised node degree estimate. Supervised optimization is done using $\rho$ to compare node degrees resulting from clustering when $\rho$ reaches statistical significance ($p<0.001$). The count of genes on this optimized list represents the final node degree. All optimized gene lists were combined to produce a group-wise network map representing a network of genes with significant group-wise correlation. The results of the group-wise and pair-wise mapping were then compared to produce a consensus network map.

Natural language searching (NLS) methods were used to identify literature support for potential relationships (i.e. regulatory and binding interactions) identified in previous steps. This analytical approach was combined with Pathways Studio [20] software to reveal potential biological implications of the newly discovered networks. Each network was analyzed to identify topological parameters and patterns, i.e., network diameter, node and link degree, and complexity [21]. The maximum number of potential interactions is thus determined.

2.4. Marker Identification

Markers were identified by partitioning datasets into normal and diseased and plotting expression levels. Overlap was then evaluated and calculated as a percentage of all records. This was completed for both datasets and for all genes from each list. Genes with expression overlap of less than 20% where reported as potential diagnostic markers (overlap represents the potential error in ability of gene(s) to discriminate between normal a diseased samples). The SAM method (Statistical Analysis of Microarrays) [22] was also used, in conjunction with PAM (Predictive Analysis of Microarrays) [23] analysis. PAM was used along with the statistically weighted voting classification method [24]. We used modified gene voting procedure including selection of informative (discriminative) genes and calculation of the statistically weighted votes of expression levels of individual genes in selected contribution (signature) of the genes. The weights are then summed over all signature genes. A Classification Decision Rule is thus formulated: If the sum of the weighted votes is positive, the sample is classified as lung adenocarcinoma; if the sum is negative, the sample is normal.

3. Results

3.1. Group-wise analysis & Information Flow

Spearman correlation analysis of both datasets produced a group of 9 genes: NCL, CDK7, MYST3, AR, BRCA1, PAX5, ELK1, ETV3, and HIRIP3 with high correlation in both datasets from a group of five
clusters (AR, HIRIP3, BRCA1, ELK1, CDK7; U95A mean ρ = 0.848, U133A mean ρ = 0.91). These correlations suggested connected networks of nine genes with 18 links for Su et al. data (U133A [12]) and 20 links for Bhattacharjee data (U95A [2]). Node connectivity ranges from 1 to 5 (NCL, BRCA1, AR).

![Fig.1. Nine nodes networks constructed from U95A [2] and U133A [12] microarray platforms built proceeding from correlation analysis (See Methods). The nine nodes overlap entirely, as do most (but not all) of their links.](image)

Within the network of nine nodes and 18 and 20 edges a sub-network of high correlation was identified consisting of Elk1, NCL, AR, ETV3 and PAX5. This network had mean ρ of 0.94 and a max of 0.98. Three more genes (E2F1, EHMT2, MAP3K11) were identified as lying within the network of these nine genes. Although they do not increase ρ, being connected only to the most highly correlated nine network nodes, these three genes were also considered of potential importance for early detection of lung adenocarcinoma. Five of the twelve genes thus selected are transcription factors.

Each node in the network of 12 genes provided a Gaussian distribution of ρ frequencies. When cut-off was applied to each node, an overall change of +1.88 fold was shown with a maximum of +7.75 (E2F1) and a minimum of -1.29 (CDK7).

### 3.2. Gene Signature Identification

The statistical significance of the selected 12 genes found to be promising for the early detection of lung adenocarcinoma was evaluated by using the SAM. Seven of the genes (BRCA1, CDK7, E2F1, EHMT2, ELK1, MYST3, and PAX5) were found to be highly significant, while AR, ETV3, HIRIP3, MAP3K11, and NCL were not statistically significant in context of differential expression, and were eliminated from the signature search. These genes are also poorly informative in context of individual classification of the patients. (The results with the details of the PAM analysis are available at request).

The modified PAM analysis allowed reducing further the set of genes highly sensitive to lung adenocarcinoma to a 5-genes signature: MYST3, CDK7, EHMT2, BRCA1, and E2F1. This set of genes provided 93% correct classification of the cancer and normal data. It is to be noted that only one of our final set of proteins (E2F1) has so far been related to lung cancer in a recently published set of 10 survival predictor genes [8]. The other four genes are proposed for the first time here as lung adenocarcinoma identifiers. The best results are manifested by MYST3, which failed to correctly identify only two cases. In addition, we validated our 5-gene signature using the U133A dataset of Su et al. [12], and achieved 96% accuracy with that data.

### 3.3. Building a network from literature data for gene and protein interactions

The twelve genes identified above as candidate markers were imported into the ResNet 5.0 database of the Pathway Studio software package [20]. The database includes different types of gene-gene, gene-protein and protein-protein interactions, such as physical binding between proteins, promoter binding, regulation, protein posttranslational modification (phosphorylation, acetylation, etc.), and even molecular transport, providing thus an integrated view on biological process of interest. As seen in Fig. 2, the twelve proteins/genes are well connected with each other, either directly or via other proteins, and the resulting network might be considered as capturing essential part of the mechanism of the lung adenocarcinoma. The presence in the network of “connecting” proteins, which are well known for their role in cancer, such as TP53, MYC, EP300, RUNX1 and others, together with seven histone or histone-related proteins (HIST2H3C, H2BFQ, H4F8, HDAC1, HDAC3, HIRA, and HIRIP3) and nine mitogen-activated protein kinases, might be regarded in support of our hypothesis for the importance of histones in the process of carcinogenesis. Adding also the presence of the APP receptor and CCL3 ligand, one may conclude that the network shown in Fig. 2 seems to contain most of the elements needed for advancing detailed hypotheses for the mechanism of the lung adenocarcinoma.

![Fig. 2 reveals some details of the regulatory interactions. One may single out the potential importance of DAXX, which regulates negatively major players like TP53, AR, PAX5 and RARB. Regulatory hubs besides TP53 and MAPK1 include also PAX5 with the record number of 20 interactions. A specific hub is the receptor protein APP which is regulated by three of our selected 12 proteins (NCL, E2F1 and MAP3K11), and on its turn activates MYST3 and down regulates NOTCH1.](image)
Additional information on the complex regulatory mechanisms involving some of the selected proteins was inferred from Pathway Studio 5.0 software. It shows external regulatory proteins acting on eight proteins from our list of twelve. Very high degree of external regulation is shown for AR, E2F1 and BRCA1, whereas no such regulation was found for ELK1, MYST3, EHTM2 and HIRIP3.

3.4. Gene Ontology Analysis

The extension of the set of twelve potential marker genes to an integrated set of 43 genes/proteins provided important information for the pathways and biological functions the five signature genes take part in. We used in parallel David [17] and Panther [16] gene ontology. Fig. 3 summarizes some results from the Panther pathway analysis. The 43 genes have shown a statistically significant involvement in 24 different pathways. (The detailed information is provided by request from the authors). The leading role, with the highest number of genes incorporated, is played by RAS, PDGF, TGF-beta, Toll Receptor, B-Cell Activation, EGF receptor, Interferon, Interleukin, and other pathways. The information obtained, along with the integrated network shown in Fig. 3, provides a basis for mechanistic analysis.

4. Discussion

The search for predictive and prognostic sets of genes proceeds most often from a purely statistical analysis of microarray data. In this study, aiming at identification of markers for early detection of lung adenocarcinoma, we proceeded from a different strategy. We selected several gene ontology categories of human genes, assumed to be of potential importance for carcinogenesis. Using microarray databases we searched for genes from these categories, the expression of which is significantly changed in carcinoma vs normal lung tissue. Applying both well-known and proprietary techniques for correlation analysis, statistical evaluation, and network analysis we arrived to a set of twelve genes, from which a set of five genes highly sensitive to lung adenocarcinoma emerged. The novelty of our approach resulted in identifying genes like MYST3, CDK7, BRCA1, and EHTM2, which have not been offered as potential diagnostics markers of lung adenocarcinoma before. In addition, the 96% accuracy of classification achieved with the U133 database of Su et al. [12], is comparable to the best results reported in the literature (see [6],[8] for references), only one of which (E2F1) appeared in our set. It is important to note that two of the five genes in our final set (EHTM2 and E2F1) were predicted to be potentially very sensitive to lung cancer proceeding from purely topological properties of the mini-network build from the correlation analysis of the microarray data. The validation of the proposed 5-gene signature with a set of U133A chips [12], very different in design, patients, and even race, produced similar discrimination of normal and adenocarcinoma lung tissues. The subsequent network analysis by expanding the initial set of twelve candidate genes to an integrated network of 43 genes/proteins provided important feedback on the intracellular environment within which the signature genes act. This environment included well-known players like TP53, MYC, EP300,
and others, and proved to be in the cross-talk of 24 different pathways. All this provided ample information for future mechanistic analysis. In conclusion, it may be noted that the novel strategy for search of gene signatures proceeding from a biological basis seems promising, and will be tested in other types of disease, including different types of cancer.

5. References