Identification of Novel Cancer-Related Genes with a Prognostic Role Using Gene Expression and Protein-Protein Interaction Network Data

Peng Li1,2, Bo Sun1 and Maozu Guo2

1Beijing Normal University, Beijing, China
2Beijing University of Civil Engineering and Architecture, Beijing, China

Early cancer diagnosis and prognosis prediction are necessary for cancer patients. Effective identification of cancer-related genes and biomarkers and survival prediction for cancer patients would facilitate personalized treatment of cancer patients. This study aimed to investigate a method for integrating data regarding gene expression and protein-protein interaction networks to identify cancer-related prognostic genes via random walk with restart algorithm and survival analysis. Known cancer-related genes in protein-protein interaction networks were considered seed genes, and the random walk algorithm was used to identify candidate cancer-related genes. Thereafter, using the univariate Cox regression model, gene expression data were screened to identify survival-related genes. Furthermore, candidate genes and survival-related genes were screened to identify cancer-related prognostic genes. Finally, the effectiveness of the method was verified through gene function analysis and survival prediction. The results indicate that the cancer-related genes can be considered prognostic cancer biomarkers and provide a basis for cancer diagnosis.


Keywords: cancer genes, random walk algorithm, PPIN, survival analysis, biomarkers

1. Introduction

Cancer is associated with various factors but is essentially caused by gene mutations [1, 2]. Therefore, identification of cancer-related genes and biomarkers would potentially reveal the causes and therapeutic targets of cancer and help predict the prognosis of cancer patients, thus enabling personalized and precise therapies for cancer patients. Rather than a single genetic abnormality, cancer is caused by the interaction of genes with regulatory factors in complex biological networks [3]. These factors affect gene expression, thus influencing cancer pathogenesis and progression.

Systems biology, which emerged in the 1990s, has provided novel insights into complex diseases such as cancer [4, 5]. The study focus has shifted from single genes or pathways to multiple gene modules and interrelationships within pathways. Through integrated analyses of multi-level, heterogeneous data, factors influencing tumorigenesis can be fully considered. Accordingly, systems biology is not only a novel strategy for basic research but it also provides a direction to current studies on cancer. Systems bioinformatics was recently proposed, which linked systems biology and classical bioinformatics [6]. Systems bioinformatics harnesses methods involved in network science to integrate and extract information across different data sources. Networks provide
a robust scaffold wherein different biological data can be integrated [7]. Using quantitative
descriptions of various biological networks, such as protein-protein interaction networks (PPIN), gene regulatory networks, or metabol-
ic interaction networks, numerous biological computing problems, e.g., the identification of
relevant genes of a particular disease [8, 9], network-based biomarker discovery [10, 11], and network-based drug-discovery [12, 13] can be
solved using network-based methods.

Human gene regulation is complex and occurs through various regulatory factors. During tu-
morigenesis and cancer progression, mutations in key genes destabilize biological networks,
leading to an imbalance in biological systems, driving the entire system towards tumorigene-
sis [14, 15]. With the accumulation of key variants, the oncogenic state is further aggravated.
However, among all mutations, few genes drive tumorigenesis. Hence, numerous studies have
attempted to integrate biological data to effec-
tively identify cancer-related genes.

High-throughput sequencing technologies have
yielded extensive biological data for different biological systems [16]. These data are collect-
ively referred to as multi-omics data and in-
clude gene expression, DNA copy number vari-
ation, DNA methylation, miRNA, and lncRNA data. Various omics methods have provided
insights into disease pathogenesis and patho-
physiology from various perspectives [17, 18].
Although the use of single-omics data to iden-
tify cancer-related genes has yielded numer-
ous valuable results, a single data source does
not provide complete information regarding a
gene, and the results are significantly affected
by noise. Therefore, numerous studies have
investigated methods to integrate multi-omics
data to effectively identify cancer-related genes
and biomarkers [19, 20].

Currently, multi-omics data are widely used to
identify molecular biomarkers, with albeit low
reproducibility [21]. To integrate multi-omics
data, it is often difficult to identify the associa-
tions among multi-omics data. Furthermore,
inclusion of multi-omics data is time-consum-
ing and increases the computational complex-
ity. We propose a novel approach to identify cancer-related genes using only PPIN and gene
expression data.

This study aimed to identify novel cancer-relat-
ed genes with prognostic role through random
walk with restart algorithm (RWR) and sur-
ival analysis methods. Cancer-related genes
were first considered seed genes in the PPIN
and the RWR algorithm was used to identify
candidate cancer-related genes. Thereafter, us-
ing the univariate Cox regression model, gene
expression data were screened to identify sur-
vival-related genes. Cancer-related prognostic
genes were screened on the basis of candidate
genes and survival-related genes. This method
was applied for lung squamous cell carcinoma
(LUSC). The method was validated through
gene function analysis and survival prediction.

2. Methods to Identify Cancer-Related
Genes Based on Network Models

Networks are currently one of the most wide-
ly used mathematical models for analyzing bi-
ological data [22]. Networks are a simple and
efficient abstraction of biological systems. In a
biological network, a node represents a bio-
molecule, such as a gene, protein, or metabolite,
and an edge between nodes represents physical
or functional interactions, including transcriptional binding, protein interaction relationship,
genetic interaction, or biochemical reaction.
Most network-based methods generate a repre-
sentational model of a series of biological net-
works by integrating different types of data, and
then different network-based analysis methods
can be used to identify cancer-related genes
[23,24]. Network-based identification of can-
cer-related genes is primarily based on direct
network neighbors, network structure-based
methods, and machine learning-based methods.
Methods based on direct network neighbors are
used to determine whether two genes are direct-
ly associated in a biological network. The meth-
od assumes that if two genes are associated in
a network, they are functionally related. Thus,
network neighbors of known pathogenic genes
can be screened to identify candidate patho-
genic genes for related diseases [25]. Howev-
er, such a simple approach is highly inaccurate
because it would yield false genes associated
with diseases via irrelevant edges; furthermore,
genes not directly interacting with known dis-
ease-related genes would be missed.
The network structure-based approach considers the topology of the entire network, not only direct network neighbors. This method primarily includes the shortest path method and the random walk method. The shortest path method investigates the functional similarity between a disease-related gene and candidate genes in the network. This method has been successfully applied to predict pathogenic genes associated with Alzheimer’s disease [26]. The shortest path method is a local network structure method, while the random walk method completely considers global information regarding the entire network to determine the similarity between a candidate gene and a known pathogenic gene, and ranks genes on the basis of similarity. Sebastian et al. [27] first proposed the random walk algorithm to predict disease-related genes and reported that the performance of this method is better than that of the direct network neighbors, the shortest path, and diffusion kernel methods. Zhu et al. [28] considered known human papillomavirus (HPV) genes from HPVbase as seed genes, used the random walk with restart algorithm to identify candidate genes in the PPIN, and then filtered candidate genes using permutation and association tests to identify HPV-related genes. Li et al. [29] used known epigenetic factors as seed nodes to identify potential epigenetic factors using the random walk algorithm in the PPIN. Vanunu et al. [30] first proposed the random walk method to predict disease-related gene modules. Luo et al. [31] integrated PPIN, gene-disease associations, and disease-similarity network data to construct heterogeneous networks and identified candidate disease genes using the double random walk algorithm.

Machine learning-based methods primarily include classification and clustering methods. The classification method falls under the supervised learning method. A classifier is trained through different characteristics of the known oncogenes and non-oncogenes in biological networks and is used to predict candidate gene function. Xu et al. [32] predicted oncogenes on the basis of the topological characteristics of the PPIN. Ying Cui et al. [33] analyzed the topological properties of weighted PPIN and used random forest classifiers to predict disease-related genes. Clustering methods fall under the unsupervised learning method, and most of them perform clustering on the basis of the associations among the nodes of the biological networks. Bader et al. [34] proposed an MCODE algorithm on the basis of k-core to identify the most widely used network module, which is often used as the underlying method for other methods. Nepusz et al. [35] proposed the ClusterONE method, which can mine protein complexes in protein-protein interaction networks. Because the classification-based method relies on a large amount of tagged data, its application is limited. Using the traditional clustering algorithm, sub-network modules are identified as cancer-related gene modules. However, traditional methods for mining cancer-related genes are primarily focused on the operation of the network itself, and less consideration is given to real regulatory relationships among biomolecules.

Biological network analysis has revealed that genes causing similar or identical diseases potentially interact directly or indirectly in the network [36]. Often in studies on disease-related genes, a group of genes is usually associated with a disease, and we wish to infer new genes that may be associated with that disease. The shortest path method and direct network neighbors do not provide global information regarding the network structure and cannot detect complex relationships among network nodes; hence, the random walk method is widely applicable to predict disease-related genes.

3. Methods

Because genes encode proteins, a PPIN is considered a genetic relationship network [37, 38]. For a weighted PPIN, the weight on the edge represents the confidence or intensity of an interaction between two genes. A PPIN $G = (V, E, W)$, where $V$ is the set of nodes and each node represents a protein molecule. $E$ is the set of interactions between the nodes in $V$. If $u \in V$ and $v \in V$, then $e_{uv} \in E$. $W$ represents the weight set of edges. If there is an edge between $u$ and $v$, then $w_{uv} \in W$.

The flowchart of the method used to identify cancer-related genes with prognostic role is shown in Figure 1.
3.1. Identification of Cancer-Related Genes Using the RWR Algorithm

The concept underlying the random walk algorithm is to begin with specific pathogenic genes and random walk particles spread along the interlinks in the PPIN. After the steady state, the candidate disease genes are ranked in accordance with the gene score. The random walk model has been previously discussed [39]. The RWR algorithm used herein includes parameter \( r \) to adjust for the probability of random particles remaining at the original node based on the basic random walk algorithm.

The Pagerank algorithm was the first to adopt the restart idea [40]:

\[
p^{t+1} = (1-r)WP^t + r \frac{l}{n}
\]

(1)

Where \( W \) represents the adjacency matrix normalized by the network column. \( r \) is the probability of restart. \( l \) is an all-1 column vector of length \( n \) (\( n \) is the number of nodes). \( P^t \) is the probability vector at which the starting nodes jump to all nodes in the network at time step \( t \).

RWR is an improved Pagerank algorithm [41]:

\[
p^{t+1} = (1-r)WP^t + rP^0
\]

(2)

Where the initial value \( P^0 \) is an \( n \times 1 \) probability vector and seed nodes of the known function in the vector are set to equal probability values, with the sum of probabilities being equal to 1, and the remaining nodes were set to 0. The difference between the Pagerank algorithm and RWR algorithm is the setting of the initial vector. In the Pagerank algorithm, the values of the initial vectors were set to 1, indicating that the calculation can be initiated from any node with uniform probability. In contrast, the RWR algorithm can only be initiated from specified seed nodes.

First, the transition probability matrix \( W \) is obtained in accordance with the structure of the PPIN. Second, we constructed an initial vector \( P^0 \) based on prior knowledge of the specific cancer. The known cancer gene nodes in \( P^0 \) were set to be equal probability values, with the sum of the probabilities equal to 1 and initial values of other nodes were set to 0. The iterative operation was performed in accordance with Equation 2 until convergence. The operation is generally considered to have converged when the difference between the vector \( P^{t+1} \) and vector \( P^t \) was less than \( 10^{-6} \). Parameter \( r \) was set to 0.8. Finally, the probabilities of candidate tumorigenic genes were assessed on the basis of the calculated \( P^t \) value. For convenience, the ranked top-N genes identified via the random walk algorithm were called RWR genes.

3.2. Identification of Survival-Related Genes

Herein, gene expression data were used to screen for survival-related genes. Potential abnormalities include missing data values. For each gene, missing values were first examined, retaining only genes with expression values greater than 0 in more than 80% of samples and greater than 1 in more than 10% of samples. The mean value was then used to complement the missing value.

For gene expression data, the following processing was performed: \( x' = \log_2 (x + 1) \), where \( x \) represents the value of gene expression data. The data were then normalized to the range [0, 1] using min-max normalization. In order to better process the data, a 4-digit significant number was reserved for standardized data.

Using a univariate Cox proportional hazards regression model (CPH), genes in the gene ex-
pression dataset were screened to identify survival-related genes. If the *p*-value of a gene's likelihood ratio test was less than 0.05, the gene was considered to be associated with survival.

### 3.3. Screening for Cancer-Related Prognostic Genes

The intersection of RWR genes of preceding ranked top-N and survival-related genes was considered to detect cancer-related prognostic genes. To determine whether the genes are cancer-related, gene function and pathway analysis were performed. Through survival prediction analysis, we confirmed whether the identified genes can distinguish high- and low-risk groups and whether they are of high prognostic significance.

For this study, the CPH and the Kaplan-Meier survival analysis were implemented using R package survival analysis. All data pre-processing and performance measurements were implemented using Python and Scikit-learn 0.19.2. Hierarchical clustered heatmaps were implemented using R package *pheatmap*.

### 4. Materials and Experimental Results

#### 4.1. Materials

Lung cancer is the most common cause of cancer-related mortality worldwide, of which lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) are the most common subtypes. Herein, we used LUSC data to evaluate the identification method of cancer-related prognostic genes. The PPIN data used herein were obtained from HIPPIE [42]. For each PPI, HIPPIE specifies a value indicating the credibility and strength of the interaction. We selected interactive data with an interaction value greater than 0.4, representing moderate confidence or intensity. To simplify the calculation, we selected nodes with Ensembl IDs of less than 10,000. The resulting PPIN included 6519 nodes and 120,506 edges. Gene expression data were obtained from Firebrowse (http://firebrowse.org/?cohort=LUSC&download_dialog=true), including 552 LUSC samples, among which 501 were tumor samples and 51 were normal samples.

Herbst *et al.* [43] reviewed the recent advancements in lung cancer and its treatment and discussed the most common genetic alterations in LUSC and LUAD. We selected all 28 known LUSC genes (EGFR, ERBB2, ERBB3, FGFR1, FGFR2, FGFR3, Pten, PIK3CA, KRAS, NRAS, RASA, HRAS, NF1, STK11, AKT1, AKT2, AKT3, BRAF, TSC1, TSC2, MAP2K1, CDKN2A, KEAP1, CUL3, Mtor, NFE2L2, TP53, and RB1) and mapped them onto the PPIN, wherein proteins were represented by Ensembl IDs. These 28 Ensembl IDs were considered the seed nodes to identify novel LUSC cancer-related genes by using the RWR algorithm.

#### 4.2. Analysis of Experimental Results

We selected the top 200 RWR genes and all survival-related genes, and their intersection included 19 genes. We considered the 19 cancer-related prognostic genes.

#### 4.2.1. Cancer-Related Genes

The 19 cancer-related prognostic genes were as follows: CALM2, RGL2, HSPA1A, HSPA4, JUN, ARAF, PLCG1, SRC, ZNF189, HSPD1, CDKN1A, PPP2CA, HSPG2, IRS1, RABGTA, BCL2L1, AXIN1, UBE2E1, and GRIN2D. Among these, ARAF, CDKN1A, and PLCG1 are associated with LUSC. The ID, description, and chromosomal information regarding each key gene (including the chromosome number of the gene and the start site information of the chromosome to which the gene belongs) are shown in Table 1. We searched GeneCards (https://www.genecards.org) for the function of each gene and found that the genes were associated with cancer or complex diseases.

#### 4.2.2 Analysis of Gene Function and Pathways

The DAVID (The Database for Annotation, Visualization and Integrated Discovery, https://david.ncifcrf.gov) database provides a complete set of functional annotation tools...
available to assess the biological implications of the list of genes. To further confirm the role of these 19 key genes, we used the DAVID tool to perform Gene Ontology and KEGG pathway analysis. The specific names of the key functions and path ID are shown in Table 2. The function ID of the gene ontology is prefixed with GO, and the path ID of KEGG is prefixed with hsa. As shown in Table 2, these functions and pathways significantly influence cancer pathogenesis and progression. For example, the dysregulation of GO: 0010634, GO: 0010941, GO: 0043392, and GO: 0033160 would aggravate the division and proliferation of cells, with extremely important effects. Hsa05205 represents cancer proteoglycans. Hsa05200 represents the cancer pathway. Hsa05219 represents bladder cancer. Hsa05210 represents colorectal cancer. Hsa05203 represents viral carcinogenesis. Therefore, the functions and pathways regulated by these 19 genes are closely associated with the evolution of LUSC.

### 4.2.3. Differences Between Normal Samples and Tumor Samples

To verify the effect of these genes on development of LUSC, gene expression data were used to distinguish between normal samples and tumor samples. The performance of the proposed method was measured on the basis of the classification performance in accordance with the gene expression data of these 19 key genes. The number of classes of the omics data, such as gene expression data, DNA methylation data, is usually unbalanced, and the number of tumor samples (positive) is often greater than that of normal samples (negative). First, we accounted for the 19 key genes from the dataset and determined four performance measures, i.e., the accuracy, precision, recall, and F1 score with five classifiers, i.e., support vector machine (SVM), logistic regression, naïve bayes, decision tree, and random forest. Furthermore, we performed hierarchical clustering for the expression data

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Gene ID</th>
<th>Description</th>
<th>Chromosomal information</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALM2</td>
<td>805</td>
<td>Calmodulin 2</td>
<td>hs2 (47160082, 47176936)</td>
</tr>
<tr>
<td>RGL2</td>
<td>5863</td>
<td>Ral Guanine Nucleotide Dissociation Stimulator Like 2</td>
<td>hs6 (33291654, 33299388)</td>
</tr>
<tr>
<td>HSPA1A</td>
<td>3303</td>
<td>Heat Shock Protein Family A (Hsp70) Member 1A</td>
<td>hs6 (31815543, 31817942)</td>
</tr>
<tr>
<td>HSPA4</td>
<td>3308</td>
<td>Heat Shock Protein Family A (Hsp70) Member 4</td>
<td>hs5 (133052013, 133106449)</td>
</tr>
<tr>
<td>JUN</td>
<td>3725</td>
<td>Jun Proto-Oncogene, AP-1 Transcription Factor Subunit</td>
<td>hs1 (58780791, 58784047)</td>
</tr>
<tr>
<td>ARAF</td>
<td>369</td>
<td>A-Raf Proto-Oncogene, Serine/Threonine Kinase</td>
<td>hsX (47561100, 47571921)</td>
</tr>
<tr>
<td>PLCG1</td>
<td>5335</td>
<td>Phospholipase C Gamma 1</td>
<td>hs20 (41137519, 41177626)</td>
</tr>
<tr>
<td>SRC</td>
<td>6714</td>
<td>SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase</td>
<td>hs20 (37344685, 37405432)</td>
</tr>
<tr>
<td>ZNF189</td>
<td>7743</td>
<td>Zinc Finger Protein 189</td>
<td>hs9 (101398830, 101410660)</td>
</tr>
<tr>
<td>HSDP1</td>
<td>3329</td>
<td>Heat Shock Protein Family D (Hsp60) Member 1</td>
<td>hs2 (197486584, 197500274)</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>1026</td>
<td>Cyclin Dependent Kinase Inhibitor 1A</td>
<td>hs6 (36676250, 36687339)</td>
</tr>
<tr>
<td>PPP2CA</td>
<td>5515</td>
<td>Protein Phosphatase 2 Catalytic Subunit Alpha</td>
<td>hs5 (134194332, 134226073)</td>
</tr>
<tr>
<td>HSPG2</td>
<td>3339</td>
<td>Heparan Sulfate Proteoglycan 2</td>
<td>hs1 (21822244, 21937310)</td>
</tr>
<tr>
<td>IRS1</td>
<td>3667</td>
<td>Insulin Receptor Substrate 1</td>
<td>hs2 (226731317, 226798790)</td>
</tr>
<tr>
<td>RABGGTB</td>
<td>5876</td>
<td>Rab Geranylgeranyltransferase Subunit Beta</td>
<td>hs1 (75786194, 75795090)</td>
</tr>
<tr>
<td>BCL2L1</td>
<td>598</td>
<td>BCL2 Like 1</td>
<td>hs20 (31664452, 31723963)</td>
</tr>
<tr>
<td>AXIN1</td>
<td>8312</td>
<td>Axin 1</td>
<td>hs16 (287440, 355226)</td>
</tr>
<tr>
<td>UBE2E1</td>
<td>7324</td>
<td>Ubiquitin Conjugating Enzyme E2 E1</td>
<td>hs3 (23805955, 23891640)</td>
</tr>
<tr>
<td>GRIN2D</td>
<td>2906</td>
<td>Glutamate Ionotropic Receptor NMDA Type Subunit 2D</td>
<td>hs19 (48394875, 48444937)</td>
</tr>
</tbody>
</table>
Identification of Novel Cancer-Related Genes with a Prognostic Role...

...for the 19 key genes, and used heatmaps to obtain the results of the classification.

A 10-fold cross-validation test was conducted for performance evaluation. We selected the 19 key genes from the gene expression gene dataset for the test. For classification problems, we used stratified sampling to ensure that the proportion of tumor samples and normal samples in the training set and test set was the same as in the original data set. We used five classifiers to determine four performance measures. From the measurement of accuracy, precision, recall, and F1 score in Table 3, it is shown that these five classifiers could produce a comparable performance for sample type prediction. So these 19 genes can distinguish between tumor samples and normal samples.

We performed hierarchical clustering for the gene expression data for the 19 key genes. Euclidean distance was used as distance metric.

<table>
<thead>
<tr>
<th>Function ID</th>
<th>Function term</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0043547</td>
<td>positive regulation of GTPase activity</td>
</tr>
<tr>
<td>GO:0010634</td>
<td>positive regulation of epithelial cell migration</td>
</tr>
<tr>
<td>GO:0071902</td>
<td>positive regulation of protein serine/threonine kinase activity</td>
</tr>
<tr>
<td>hsa05205</td>
<td>Proteoglycans in cancer</td>
</tr>
<tr>
<td>GO:0043066</td>
<td>negative regulation of apoptotic process</td>
</tr>
<tr>
<td>hsa05200</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td>hsa05219</td>
<td>Bladder cancer</td>
</tr>
<tr>
<td>hsa05210</td>
<td>Colorectal cancer</td>
</tr>
<tr>
<td>GO:0043524</td>
<td>negative regulation of neuron apoptotic process</td>
</tr>
<tr>
<td>GO:0010907</td>
<td>positive regulation of glucose metabolic process</td>
</tr>
<tr>
<td>GO:0050821</td>
<td>protein stabilization</td>
</tr>
<tr>
<td>hsa05120</td>
<td>Epithelial cell signaling in Helicobacter pylori infection</td>
</tr>
<tr>
<td>GO:0010941</td>
<td>regulation of cell death</td>
</tr>
<tr>
<td>GO:0046628</td>
<td>positive regulation of insulin receptor signaling pathway</td>
</tr>
<tr>
<td>GO:2000811</td>
<td>negative regulation of anoikis</td>
</tr>
<tr>
<td>GO:0031954</td>
<td>positive regulation of protein autophosphorylation</td>
</tr>
<tr>
<td>GO:0043392</td>
<td>negative regulation of DNA binding</td>
</tr>
<tr>
<td>GO:0043065</td>
<td>positive regulation of apoptotic process</td>
</tr>
<tr>
<td>hsa05203</td>
<td>Viral carcinogenesis</td>
</tr>
<tr>
<td>GO:0033138</td>
<td>positive regulation of peptidyl-serine phosphorylation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>SVM</th>
<th>Logistic regression</th>
<th>Naive Bayes</th>
<th>Decision tree</th>
<th>Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.975</td>
<td>0.976</td>
<td>0.908</td>
<td>0.958</td>
<td>0.975</td>
</tr>
<tr>
<td>Precision</td>
<td>0.976</td>
<td>0.980</td>
<td>0.908</td>
<td>0.980</td>
<td>0.979</td>
</tr>
<tr>
<td>Recall</td>
<td>0.996</td>
<td>0.994</td>
<td>1.000</td>
<td>0.974</td>
<td>0.994</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.986</td>
<td>0.987</td>
<td>0.951</td>
<td>0.977</td>
<td>0.986</td>
</tr>
</tbody>
</table>
The clustered heatmap is shown in Figure 2. According to Figure 2, it can be also found that these 19 key genes can distinguish obviously between tumor samples and normal samples.

Figure 2. Distinguishing between normal samples and tumor samples. Hierarchical clustering was performed for the gene expression data containing the 19 key genes. To the first row, the dark gray represents tumor samples and the light gray represents normal samples. Each row of other parts represents a gene, and each column represents a sample, indicating expression value of a gene under a specific sample. The color depth indicates the size of the gene expression value.

4.2.4. Survival Analysis

To determine the key role of potential cancer biomarkers in the development and progression of LUSC, we utilized SurvExpress biomarker validation tool for survival analysis [44]. The SurvExpress online tool contains two LUSC datasets from The Cancer Genome Atlas database (TCGA). For comparison with the proposed method, we only used the random walk algorithm to identify cancer-related genes. We selected the top 19 genes as key genes including the following: HSP90AA1, HSP90AB1, APP, GRB2, GSK3B, YWHAZ, PKM, YWHAH, YWHAZ, RAF1, SMAD2, RPS6KA1, CDH1, HSPA8, SUMO1, PIK3R1, ESR1, HSPA1A and HSPA1B. We performed survival analysis using the two datasets for these two groups of genes.

The LUSC dataset one comprises 175 LUSC samples including 88 low-risk group samples and 87 high-risk group samples. We performed survival analysis using the Cox proportional hazards model. The Kaplan-Meier curves of the results are shown in Figures 3–4.

Figure 3. Survival analysis for cancer-related genes with prognostic role. The x-axis represents time (unit: day), and the y-axis represents the global survival ratio. The dark gray line represents the high-risk group, and the light gray represents the low-risk group. In the upper right corner, the number on the left represents the number of people in each group. The number with the '+' sign in the middle represents the number of lost visitors, and the number on the right represents the Concordance Index (c-index).

Figure 4. Survival analysis for top-19 RWR genes. The x-axis represents time (unit: day), and the y-axis represents the global survival ratio. The dark gray line represents the high-risk group, and the light gray line represents the low-risk group.

On comparing the two curves, the differences gradually increased with time, and the two groups of genes significantly differed between
the two groups of patients \((p < 0.05)\) with respect to patient survival. The classification effect shown in Figure 3 \((p = 0.0001991)\) is better than that shown in Figure 4 \((p = 0.001612)\), indicating that our proposed method could better identify prognostic cancer-related genes than the random walk algorithm alone.

The experiment results showed that the genes identified by the proposed method were differentially expressed in the high- and low-risk groups. In the present study, we analyzed the expression levels of key genes in the tumor samples. The \(t\)-test was used for each gene between the high- and low-risk groups. Figures 5-6 show the differences in the expression levels of key genes identified using the two methods in the dataset. If the \(p\)-value of the \(t\)-test of a gene was less than 0.05, the gene was considered as differentially expressed gene. In Figure 5, the \(p\)-value of 16 genes is less than 0.05, but in Figure 6, only the \(p\)-value of 9 genes is less than 0.05. So, the genes identified using the currently proposed method displayed marked differences in their expression levels in the high- and low-risk groups.

The LUSC dataset two comprises 205 LUSC samples including 103 low-risk group samples and 102 high-risk group samples. We performed survival analysis using the Cox proportional hazards model. The Kaplan-Meier curves of the results are shown in Figures 7-8.

The results also indicate that the genes obtained by the two methods can significantly distinguish between the two groups of patients and were closely related to patient survival \((p < 0.05)\). The classification effect shown in Figure 7 \((p = 6.465e-06)\) is better than that shown in Figure 8 \((p = 7.836e-06)\). So, it is also indicated that our proposed method can better identify prognostic cancer-related genes than the random walk algorithm alone.
The experiment results showed that the genes identified by the proposed method were differentially expressed in the high- and low-risk groups. The $t$-test was used for each gene between the high- and low-risk groups. Figures 9–10 show the differences in the expression levels of key genes identified using the two methods in the dataset. In Figure 9, the $p$-value of 13 genes was less than 0.05, but in Figure 10, only the $p$-value of 8 genes was less than 0.05. So, it is indicated that our proposed method can identify more differentially expressed genes than the random walk algorithm alone. The key genes as potential cancer-related genes play crucial role in cell proliferation and differentiation.

Compared with direct network neighbors and the shortest path method, the random walk method provided more precise results considering all possible edges among genes. The application of random walk to gene ranking can thus identify the key genes strongly associated with known cancer genes. Using univariate survival analysis of gene expression data, cancer-related genes with prognostic roles could be screened. Hence, the genes identified via random walk and survival analysis had a more prognostic role than those identified using only the random walk method. These genes cannot only remarkably distinguish between tumor samples and normal samples, but also between high- and low-risk groups, thus being applicable as a cancer biomarker with practical significance for the diagnosis, treatment, and prognosis of cancer.

5. Discussion

In this paper, a new identification method of cancer-related genes with prognosis role based on PPIN and gene expression data is proposed. The method was applied for LUSC and 19 key genes were identified. The roles of the genes were validated through gene function analysis and survival prediction. The results indicated that the cancer-related genes could be considered prognostic cancer biomarkers and provided the basis for cancer diagnosis and treatment.
In addition to experimental verification, we confirmed their involvement in cancer pathogenesis, according to recent literature. CALM2 was an important gene affected by B-Myb knockdown in human cancers, including NSCLC [45]. Silencing of RGL2 inhibited cell population growth in anchorage dependent and independent conditions, and was important in human NSCLC [46]. Tracz-Gaszewska et al. [47] employed lung cancer cell lines that constitutively overexpressed heat shock proteins and have shown that HSPA1A facilitated the binding of mutated p53 to the TAp73α protein. HSPA4 knockdown would be effective in therapies in cancer with head and neck squamous cell carcinoma [48]. ShanJu et al. [49] constructed genetic and epigenetic networks (GENs) with LUAD and LUSC data, and identified essential biomarkers including JUN in each progression stage of LUAD and LUSC. Mutant ARAF was an oncogenic driver in LUAD and an indicator of sorafenib response [50]. PLCG1 might be involved in smoking-induced lung cancer by interacting with SRC, and might be responsible for the development of smoking-induced lung cancer [51]. The inhibition of apoptosis by fibronectin was enhanced by SRC overexpression and reversed by SRC knockdown in lung cancer cells [52]. Paul et al. [53] used a set of mutations from 22 cancers, and detected 151 putative cancer drivers including ZNF189. HSPD1, as a key marker of mitochondrial biogenesis, had the highest predictive value and was effective in predicting tumor progression in NSCLC patients [54]. Germline mutations in CDKN1A associated with idiopathic pulmonary fibrosis risk were detected in most lung cancer samples [55]. Tan et al. [56] analyzed microarray databases to screen the key genes related to NSCLC, and identified PPP2CA as differentially expressed genes in the down-regulated co-pathways. Ilhan-Mutlu et al. [57] studied the expression pattern of angiogenesis-related genes in brain metastases of lung cancer and melanoma, and identified HSPG2 more than 50-fold up-regulated in all diagnosis groups compared to control. Function mutation of IRS1 could be related to development of lung cancer, and played a role in the phenotype of lung cancer [58]. RABGGTB was overexpressed in the refractory de novo diffuse large B-cell lymphoma (DLBCL), and was a potential target for drug intervention [59]. The BET (bromodomain and extraterminal) proteins that bind acetylated histones and recruit protein complexes regulated expression of BCL2L1 in oncogenesis [60]. Yang et al. [61] studied the mechanisms of X-ray irradiation in increasing Axin expression, and analyzed AXIN1 down-regulated in many cases of lung cancer. UBE2E1 was the top-ranked genes across the NSCLC pathologic stages by the MFSector method [62]. Sun et al. [63] identified differentially expressed genes between LUAD tissues and nontumor tissues from RNA sequencing data, and distinguished GBIN2D related to LUAD.

Thus, all 19 genes are associated with the development and progression of cancer, of which 17 genes are related to lung cancer. The results illustrate to some extent the effectiveness of our method. Through the functions and pathways analysis of these 19 genes, we found that they were highly correlated with cancer. For example, the disorders of GO: 0043392, GO: 0010634, hsa05203, hsa05205 can exacerbate the division and proliferation of cancer cells.

Using RWR algorithm and CPH model, we identified potential LUSC cancer-related genes. These 19 cancer-related genes together with the known 28 cancer genes can be considered biomarkers of cancer prognosis. The number of biomarkers obtained in the study is very small, which provides great convenience for detection. The identified genes can be applied as cancer biomarkers with practical significance for the diagnosis, treatment, and prognosis of cancer.

6. Conclusion

Many complex diseases are caused by dysfunctions in related regulatory networks, not only mutations in individual molecules. Mining for key cancer-related genes or modules from biological networks has recently gained increasing attention in studies on cancer. Identification of genes and modules would help design cancer treatments and enable an early cancer diagnosis. Herein, we propose a method for integrating gene expression and PPIN data to identify prognostic cancer-related genes through random walk with restart algorithm and survival analysis. The effectiveness of the method was verified through gene function analysis and survival analysis.
This study has some limitations regarding the use of the random walk algorithm to predict cancer-related genes. For example, only network topology was considered, while topology attributes and the weight of the nodes in the PPIN were not. Therefore, screening of candidate genes in a PPIN tends to result in a higher error rate. Furthermore, in the iterative operation, the matrix multiplication operation is repeated, which requires a large memory space and a long operation time. In subsequent studies, we will optimize the random walk algorithm to account for weight information of the edges and improve the efficiency of the algorithm.

Acknowledgement

This research was supported by the national natural science foundation of China [grant number 61532014].

References


[60] L. T. Lam et al., "Vulnerability of Small-Cell Lung Cancer to Apoptosis Induced by the Combination of BET Bromodomain Proteins and BCL2 Inhibitors", Molecular Cancer Therapeutics, vol. 16, no. 8, pp. 1511–1520, 2017. http://dx.doi.org/10.1158/1535-7163.MCT-16-0459
http://dx.doi.org/10.1186/1471-2407-13-368

http://dx.doi.org/10.1186/s12859-019-2775-8

http://dx.doi.org/10.2147/OTT.S97357

---

**Contact addresses:**

Peng Li  
Beijing Normal University  
Beijing  
e-mail: lipeng1@bucea.edu.cn

Bo Sun*  
Beijing Normal University  
Beijing  
e-mail: tosunbo@bnu.edu.cn  
*Corresponding author

Maozu Guo  
Beijing University of Civil Engineering and Architecture  
Beijing  
e-mail: guomaozu@bucea.edu.cn

---

**Received:** August 2019  
**Revised:** January 2020  
**Accepted:** February 2020

**PENG LI** was born in Dongying, Shandong, China, in 1975. He received his BSc degree in computer science from Qingdao University, Shandong, China, in 1997, and his MSc degree in computer science from Shandong University of Technology, China, in 2000. Since 2003, he has been a faculty member as a Lecturer in the School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture (BUCEA), China. He is currently working toward the PhD degree at the School of Artificial Intelligence, Beijing Normal University (BNU). His research focuses on machine learning, bioinformatics and data mining.

**BO SUN** was born in Jiangsu, China, in 1966. He received the BSc degree in computer science from Beihang University (BUAA), Beijing, China, in 1988, the MSc and PhD degrees separately in natural language processing and computer-aided education from Beijing Normal University (BNU), Beijing, in 1991 and 2003, respectively. From 1999 to 2004, he was an Associate Professor with the Computer Science Department at BNU. Since 2004, he has been a Professor at the College of Information Science and Technology, BNU. His research interests include machine learning, deep neural networks, pattern recognition and natural language processing. Dr. Sun is now a Professor at the School of Artificial Intelligence, the Director of Intelligent Computing and Software Research Center, at BNU, PhD supervisor, a senior member of ACM, China Computer Federation and China Society of Image and Graphics.

**MAOZU GUO** was born in Shandong, China, in 1966. He received the BSc degree in computer science from Harbin Engineering University, China, in 1988, the MSc and PhD degrees separately in computer science from Harbin Institute of Technology (HIT), in 1991 and 1997, respectively. From 1998 to 2002, he was an Associate Professor at the College of Information Science and Technology at HIT. From 2002 to 2016, he was a Professor at the College of Information Science and Technology, HIT. Since 2016, he has been a Professor at the Beijing University of Civil Engineering and Architecture (BUCEA), Beijing. His research interests include bioinformatics, machine learning, and data mining. Dr. Guo is now the dean of the School of Electrical and Information Engineering at BUCEA, Professor, PhD supervisor and member of CCF.