doi: 10.20532/cit.2019.1004856

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

Peng Li[1,2], Bo Sun[1] and Maozu Guo[2]

[1]Beijing Normal University, Beijing, China
[2]Beijing 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 univariant 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.

ACM CCS (2012) Classification: Applied computing → Life and medical sciences → Systems biology

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 metabolic 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 tumorigenesis and cancer progression, mutations in key genes destabilize biological networks, leading to an imbalance in biological systems, driving the entire system towards tumorigenesis [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 effectively identify cancer-related genes.

High-throughput sequencing technologies have yielded extensive biological data for different biological systems [16]. These data are collectively referred to as multi-omics data and include gene expression, DNA copy number variation, DNA methylation, miRNA, and lncRNA data. Various omics methods have provided insights into disease pathogenesis and pathophysiology from various perspectives [17, 18]. Although the use of single-omics data to identify cancer-related genes has yielded numerous 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 associations among multi-omics data. Furthermore, integration of multi-omics data is time-consuming and increases the computational complexity. We propose a novel approach to identify cancer-related genes using only PPIN and gene expression data.

This study aimed to identify novel cancer-related genes with prognostic role through random walk with restart algorithm (RWR) and survival 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, using the univariant Cox regression model, gene expression data were screened to identify survival-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 widely used mathematical models for analyzing biological data [22]. Networks are a simple and efficient abstraction of biological systems. In a biological network, a node represents a biomolecule, 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 representational model of a series of biological networks 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 cancer-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 directly associated in a biological network. The method 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 pathogenic genes for related diseases [25]. However, 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 disease-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 of 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.

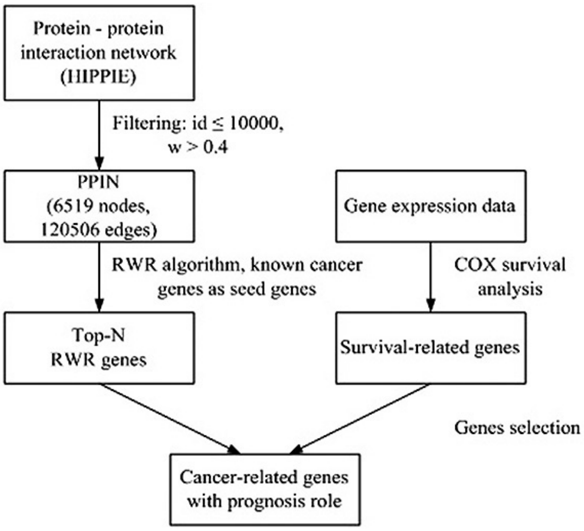
Protein - protein interaction network (HIPPIE)
Filtering: id ≤ 10000, w > 0.4
PPIN (6519 nodes, 120506 edges)
Gene expression data
RWR algorithm, known cancer genes as seed genes
COX survival analysis
Top-N RWR genes
Survival-related genes
Genes selection
Cancer-related genes with prognosis role

Figure 1. Flowchart for identifying cancer-related genes with prognostic role.

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]:

$$p^{t+1} = \left(1-r\right)WP^{t} + r\frac{l}{n} \tag{1}$$

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]:

$$p^{t+1} = \left(1-r\right)WP^{t} + rP^{0} \tag{2}$$

Where the initial value P^0 is an n × 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 expression 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*, *RABGGTB*, *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 1. Key gene descriptions and chromosomal information.

Gene name	Gene ID	Description	Chromosomal information
CALM2	805	Calmodulin 2	hs2 (47160082, 47176936)
RGL2	5863	Ral Guanine Nucleotide Dissociation Stimulator Like 2	hs6 (33291654, 33299388)
HSPA1A	3303	Heat Shock Protein Family A (Hsp70) Member 1A	hs6 (31815543, 31817942)
HSPA4	3308	Heat Shock Protein Family A (Hsp70) Member 4	hs5 (133052013, 133106449)
JUN	3725	Jun Proto-Oncogene, AP-1 Transcription Factor Subunit	hs1 (58780791, 58784047)
ARAF	369	A-Raf Proto-Oncogene, Serine/Threonine Kinase	hsX (47561100, 47571921)
PLCG1	5335	Phospholipase C Gamma 1	hs20 (41137519, 41177626)
SRC	6714	SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase	hs20 (37344685, 37405432)
ZNF189	7743	Zinc Finger Protein 189	hs9 (101398830, 101410660)
HSPD1	3329	Heat Shock Protein Family D (Hsp60) Member 1	hs2 (197486584, 197500274)
CDKN1A	1026	Cyclin Dependent Kinase Inhibitor 1A	hs6 (36676250, 36687339)
PPP2CA	5515	Protein Phosphatase 2 Catalytic Subunit Alpha	hs5 (134194332, 134226073)
HSPG2	3339	Heparan Sulfate Proteoglycan 2	hs1 (21822244, 21937310)
IRS1	3667	Insulin Receptor Substrate 1	hs2 (226731317, 226798790)
RABGGTB	5876	Rab Geranylgeranyltransferase Subunit Beta	hs1 (75786194, 75795090)
BCL2L1	598	BCL2 Like 1	hs20 (31664452, 31723963)
AXIN1	8312	Axin 1	hs16 (287440, 355226)
UBE2E1	7324	Ubiquitin Conjugating Enzyme E2 E1	hs3 (23805955, 23891640)
GRIN2D	2906	Glutamate Ionotropic Receptor NMDA Type Subunit 2D	hs19 (48394875, 48444937)

Table 2. Function ID and function name.

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

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 3. The comparison of performance measures between five classifiers.

Metric	SVM	Logistic regression	Naive Bayes	Decision tree	Random forest
Accuracy	0.975	0.976	0.908	0.958	0.975
Precision	0.976	0.980	0.908	0.980	0.979
Recall	0.996	0.994	1.000	0.974	0.994
F1 score	0.986	0.987	0.951	0.977	0.986

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.

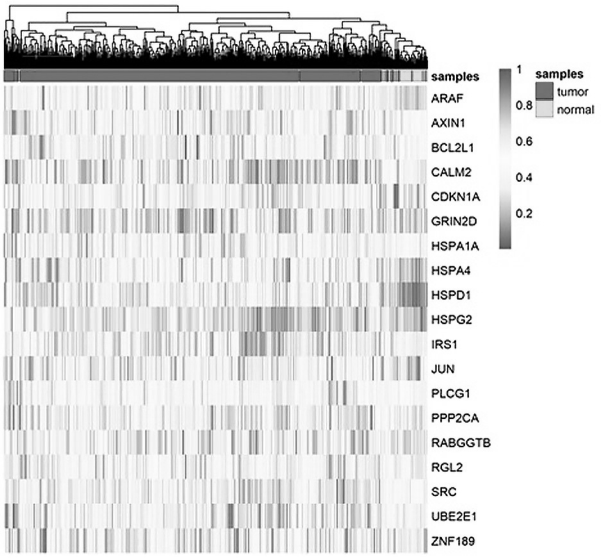

Figure 2. Distinguishing between normal samples and tumor samples. Hierarchical clustering was performed for the gene expression data containing the 19 key genes. To the first row, the dark gray represents tumor samples and the light gray represents normal samples. Each row of other parts represents 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*, *YWHAB*, *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.

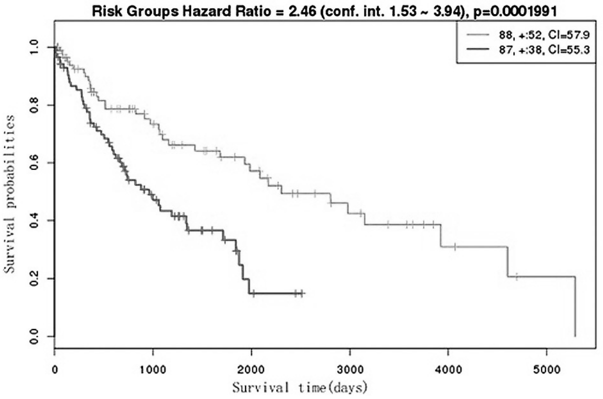

Figure 3. Survival analysis for cancer-related genes with prognostic role. The *x*-axis represents time (unit: day), and the *y*-axis represents the global survival ratio. The dark gray line represents the high-risk group, and the light gray represents the low-risk group. In the upper right corner, the 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).

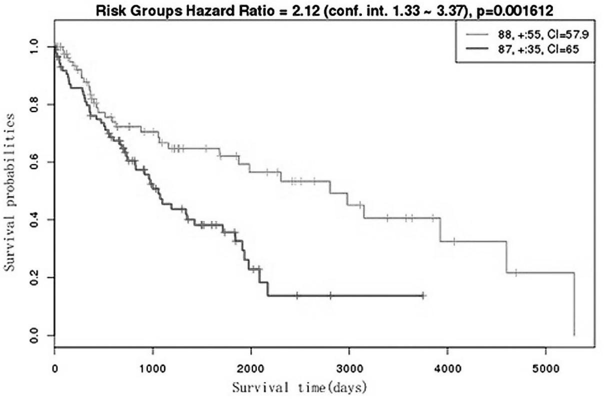

Figure 4. Survival analysis for top-19 RWR genes. The *x*-axis represents time (unit: day), and the *y*-axis represents the global survival ratio. The dark gray line represents the high-risk group, and the light gray line represents the low-risk group.

On comparing the two curves, the differences gradually increased with time, and the two groups of genes significantly differed between

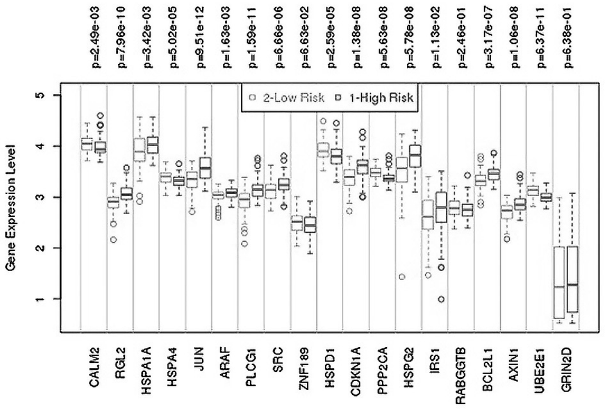

Figure 5. Gene expression levels for cancer-related genes with prognostic role. The *x*-axis represents genes, and the *y*-axis represents the expression levels of key genes in the high- and low-risk groups. The *p*-value is a statistical test variable, and the *p*-value is less than 0.05 for the significant difference.

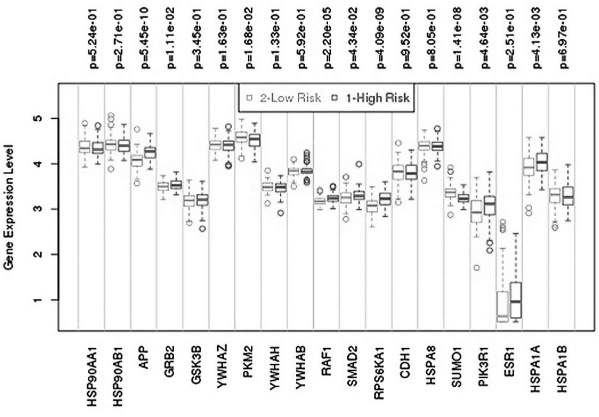

Figure 6. Gene expression levels for top-19 RWR genes. The *x*-axis represents genes, and the *y*-axis represents the expression levels of key genes in the high- and low-risk groups. The *p*-value is a statistical test variable, and the *p*-value is less than 0.05 for the significant difference.

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.

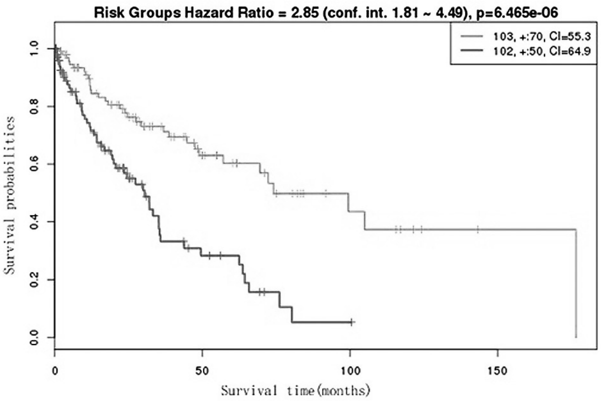

Figure 7. Survival analysis for cancer-related genes with prognosis role. The *x*-axis represents time (unit: month), and the *y*-axis represents the global survival ratio. The dark gray line represents the high-risk group, and the light gray line represents the low-risk group. In the upper right corner, 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).

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 ($p = 6.465\text{e-}06$) is better than that shown in Figure 8 ($p = 7.836\text{e-}06$). So, it is also indicated that our proposed method can better identify prognostic cancer-related genes than the random walk algorithm alone.

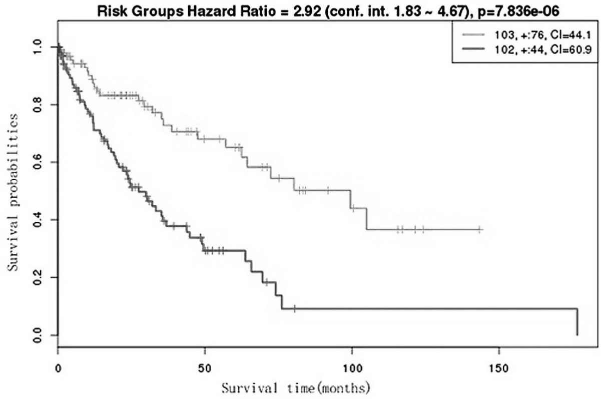

Figure 8. Survival analysis for top-19 RWR genes. The x-axis represents time (unit: month), and the y-axis represents the global survival ratio. The dark gray line represents the high-risk group, and the light gray line represents the low-risk group.

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.

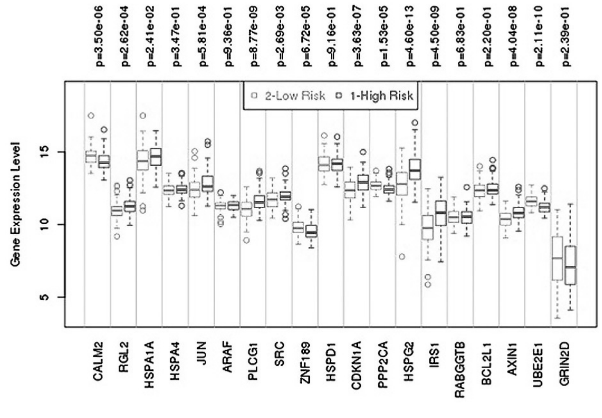

Figure 9. Gene expression levels for cancer-related genes with prognostic role. The x-axis represents genes, and the y-axis represents the expression levels of key genes in the high- and low-risk groups. The p-value is a statistical test variable, and the p-value is less than 0.05 for the significant difference.

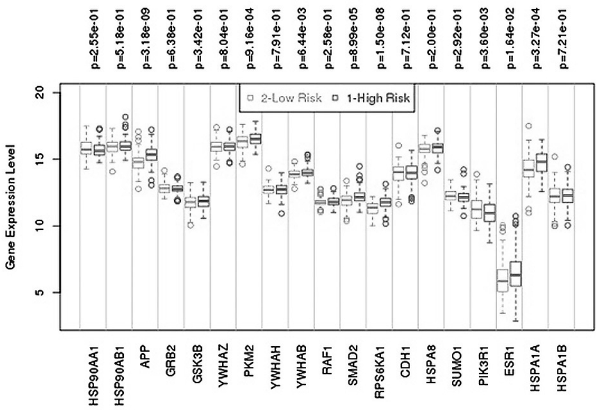

Figure 10. Gene expression levels for top-19 RWR genes. The x-axis represents genes, and the y-axis represents the expression levels of key genes in the high- and low-risk groups. The p-value is a statistical test variable, and the p-value is less than 0.05 for the significant difference.

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 MFSelector 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

[1] R. M. Ricke and J. M. van Deursen, "Aneuploidy in Health, Disease, and Aging", *The Journal of Cell Biology*, vol. 201, no. 1, pp. 11–21, 2013. https://doi.org/10.1083/jcb.201301061

[2] B. Vogelstein *et al.*, "Cancer Genome Landscapes", *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013. https://doi.org/10.1126/science.1235122

[3] A. L. Barabási and Z. N. Oltvai, "Network Biology: Understanding the Cell's Functional Organization", *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004. https://doi.org/10.1038/nrg1272

[4] A. L. Barabási *et al.*, "Network Medicine: A Network-Based Approach to Human Disease", *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011. https://doi.org/10.1038/nrg2918

[5] M. Vidal *et al.*, "Interactome Networks and Human Disease", *Cell*, vol. 144, no. 6, pp. 986–998, 2011. https://doi.org/10.1016/j.cell.2011.02.016

[6] A. Oulas *et al.*, "Systems Bioinformatics: Increasing Precision of Computational Diagnostics and Therapeutics Through Network-Based Approaches", *Briefings in Bioinformatics*, vol. 20, no. 3, pp. 806–824, 2019. http://dx.doi.org/10.1093/bib/bbx151

[7] J. L. Robinson and J. Nielsen, "Integrative Analysis of Human Omics Data Using Biomolecular Networks", *Molecular BioSystems*, vol. 12, no. 10, pp. 2953–2964, 2016. http://dx.doi.org/10.1039/c6mb00476h

[8] T. Gui *et al.*, "Identification of Hepatocellular Carcinoma-Related Genes with a Machine Learning and Network Analysis", *Journal of Computational Biology*, vol. 22, no. 1, pp. 63–71, 2015. https://doi.org/10.1089/cmb.2014.0122

[9] L. Chen *et al.*, "OPMSP: A Computational Method Integrating Protein Interaction and Sequence Information for the Identification of Novel Putative Oncogenes", *Protein & Peptide Letters*, vol. 23, no. 12, pp. 1081–1094, 2016. https://doi.org/10.2174/0929866523666161021165506

[10] X. Ma *et al.*, "Multiple Network Algorithm for Epigenetic Modules via the Integration of Genome-Wide DNA Methylation and Gene Expression Data", *BMC Bioinformatics*, vol. 18, 2017. https://doi.org/10.1186/s12859-017-1490-6

[11] S. Jamal *et al.*, "Integrating Network, Sequence and Functional Features Using Machine Learning Approaches Towards Identification of Novel Alzheimer Genes", *BMC Genomics*, vol. 17, 2016. https://doi.org/10.1186/s12864-016-3108-1

[12] J. C. Siavelis *et al.*, "Bioinformatics Methods in Drug Repurposing for Alzheimer's Disease", *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 322–335, 2016. https://doi.org/10.1093/bib/bbv048

[13] M. M. Bourdakou *et al.*, "Discovering Gene Re-Ranking Efficiency and Conserved Gene-Gene Relationships Derived from Gene Co-Expression Network Analysis on Breast Cancer Data", *Scientific Reports*, vol. 6, 2016. https://doi.org/10.1038/srep20518

[14] P. F. Jonsson and P. A. Bates, "Global Topological Features of Cancer Proteins in the Human Interactome", *Bioinformatics*, vol. 22, no. 18, pp. 2291–2297, 2006. https://doi.org/10.1093/bioinformatics/btl390

[15] Y. Q. Qiu *et al.*, "Detecting Disease Associated Modules and Prioritizing Active Genes Based on High Throughput Data", *BMC Bioinformatics*, vol. 11, p. 26, 2010. https://dx.doi.org/10.1186/1471-2105-11-26

[16] J. A. Reuter *et al.*, "High-Throughput Sequencing Technologies", *Molecular Cell*, vol. 58, no. 4, pp. 586–597, 2015. http://dx.doi.org/10.1016/j.molcel.2015.05.004

[17] B. M. Shivakumar *et al.*, "Copy Number Variations are Progressively Associated with the Pathogenesis of Colorectal Cancer in Ulcerative Colitis", *World Journal of Gastroenterology*, vol. 21, no. 2, pp. 616–622, 2015. https://dx.doi.org/10.3748/wjg.v21.i2.616

[18] S. Zhao *et al.*, "Epigenome-Wide Tumor DNA Methylation Profiling Identifies Novel Prognostic Biomarkers of Metastatic-Lethal Progression in Men Diagnosed with Clinically Localized Prostate Cancer", *Clinical Cancer Research*, vol. 23, no. 1, pp. 311–319, 2017.
https://doi.org/10.1158/1078-0432.CCR-16-0549

[19] S. Huang *et al.*, "More Is Better: Recent Progress in Multi-Omics Data Integration Methods", *Frontiers in Genetics*, vol. 8, 2017.
https://doi.org/10.3389/fgene.2017.00084

[20] Y. Li *et al.*, "A Review on Machine Learning Principles for Multi-View Biological Data Integration", *Briefings in Bioinformatics*, vol. 19, no. 2, pp. 325–340, 2018.
https://doi.org/10.1093/bib/bbw113

[21] Z. Z. Hu *et al.*, "Omics Based Molecular Target and Biomarker Identification", Bioinformatics for Omics Data: Methods and Protocols, Humana Press, pp. 547–571, 2011.
http://dx.doi.org/10.1007/978-1-61779-027-0_26

[22] M. Jalili *et al.*, "Unveiling Network-Based Functional Features Through Integration of Gene Expression into Protein Networks", *BBA-Molecular Basis of Disease*, vol. 1864, no. 6, pp. 2349–2359, 2018.
https://doi.org/10.1016/j.bbadis.2018.02.010

[23] P. Jia *et al.*, "dmGWAS: Dense Module Searching for Genome-Wide Association Studies in Protein–Protein Interaction Networks", *Bioinformatics*, vol. 27, no. 1, pp. 95–102, 2011.
https://doi.org/10.1093/bioinformatics/btq615

[24] Y. Li and J. Li, "Disease Gene Identification by Random Walk on Multigraphs Merging Heterogeneous Genomic and Phenotype Data", *BMC Genomics*, vol. 13, 2012.
https://doi.org/10.1186/1471-2164-13-S7-S27

[25] M. Oti *et al.*, "Predicting Disease Genes Using Protein-Protein Interactions", *Journal of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2006.
http://dx.doi.org/10.1136/jmg.2006.041376

[26] M. Krauthammer *et al.*, "Molecular Triangulation: Bridging Linkage and Molecular-Network Information for Identifying Candidate Genes in Alzheimer's Disease", *Proceedings of the National Academy of Sciences*, vol. 101, no. 42, pp. 15148–15153, 2004.
http://dx.doi.org/10.1073/pnas.0404315101

[27] S. Kohler *et al.*, "Walking the Interactome for Prioritization of Candidate Disease Genes", *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.
http://dx.doi.org/10.1016/j.ajhg.2008.02.013

[28] L. Zhu *et al.*, "Network-Based Method for Mining Novel HPV Infection Related Genes Using Random Walk With Restart Algorithm", *BBA-Molecular Basis of Disease*, vol. 1864, no. 6, pp. 2376–2383, 2018.
http://dx.doi.org/10.1016/j.bbadis.2017.11.021

[29] J. Li *et al.*, "A Computational Method Using the Random Walk with Restart Algorithm for Identifying Novel Epigenetic Factors", *Molecular Genetics and Genomics*, vol. 293, no. 1, pp. 293–301, 2018.
http://dx.doi.org/10.1007/s00438-017-1374-5

[30] O. Vanunu *et al.*, "Associating Genes and Protein Complexes with Disease via Network Propagation", *PLoS Computational Biology*, vol. 6, no. 1, 2010.
http://dx.doi.org/10.1371/journal.pcbi.1000641

[31] J. Luo and S. Liang, "Prioritization of Potential Candidate Disease Genes by Topological Similarity of Protein-Protein Interaction Network and Phenotype Data", *Journal of Biomedical Informatics*, vol. 53, pp. 229–236, 2015.
http://dx.doi.org/10.1016/j.jbi.2014.11.004

[32] J. Xu and Y. Li, "Discovering Disease-Genes by Topological Features in Human Protein-Protein Interaction Network", *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.
http://dx.doi.org/10.1093/bioinformatics/btl467

[33] Y. Cui *et al.*, "Discovering Disease-Associated Genes in Weighted Protein-Protein Interaction Networks", *Physica A: Statistical Mechanics and its Applications*, vol. 496, pp. 53–61, 2018.
http://dx.doi.org/10.1016/j.physa.2017.12.080

[34] G. D. Bader *et al.*, "An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks", *BMC Bioinformatics*, vol. 4, 2003.
http://dx.doi.org/10.1186/1471-2105-4-2

[35] T. Nepusz *et al.*, "Detecting Overlapping Protein Complexes in Protein-Protein Interaction Networks", *Nature Methods*, vol. 9, no. 5, pp. 471–472, 2012.
http://dx.doi.org/10.1038/nmeth.1938

[36] J. Menche *et al.*, "Uncovering Disease-Disease Relationships Through the Incomplete Interactome", *Science*, vol. 347, no. 6224, pp. 841–850, 2015.
https://dx.doi.org/10.1126/science.1257601

[37] J. Chen *et al.*, "Disease Candidate Gene Identification and Prioritization Using Protein Interaction Networks", *BMC Bioinformatics*, vol. 10, 2009.
https://doi.org/10.1186/1471-2105-10-73

[38] S. Navlakha and C. Kingsford, "The Power of Protein Interaction Networks for Associating Genes with Diseases", *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, 2010.
https://doi.org/10.1093/bioinformatics/btq076

[39] E. A. Codling *et al.*, "Random Walk Models in Biology", *Journal of the Royal Society Interface*, vol. 5, no. 25, pp. 813–834, 2008.
http://dx.doi.org/10.1098/rsif.2008.0014

[40] L. Page *et al.*, "The Pagerank Citation Ranking: Bringing Order to the Web", Technical Report, Stanford InfoLab, 1999.

[41] J. Y. Pan *et al.*, "Automatic Multimedia Cross-Modal Correlation Discovery", in KDD, pp. 653–658, 2004. http://dx.doi.org/10.1145/1014052.1014135

[42] G. Alanis-Lobato *et al.*, "HIPPIE v2.0: Enhancing Meaningfulness and Reliability of Protein-Protein Interaction Networks", *Nucleic Acids Research*, vol. 45, no. D1, pp. D408–D414, 2017. http://dx.doi.org/10.1093/nar/gkw985

[43] R. S. Herbst *et al.*, "The Biology and Management of Non-Small Cell Lung Cancer", *Nature*, vol. 553, pp. 446–454, 2018. http://dx.doi.org/10.1038/nature25183

[44] R. Aguirre-Gamboa *et al.*, "SurvExpress: An Online Biomarker Validation Tool and Database for Cancer Gene Expression Data Using Survival Analysis", *PLoS ONE*, vol. 8, no. 9. https://doi.org/10.1371/journal.pone.0074250

[45] X. Fan *et al.*, "B-Myb Mediates Proliferation and Migration of Non-Small-Cell Lung Cancer via Suppressing IGFBP3", *International Journal of Molecular Sciences*, vol. 19, no. 5, pp. 1479–1498, 2018. http://dx.doi.org/10.3390/ijms19051479

[46] A. O. Santos *et al.*, "RalGPS2 is Essential for Survival and Cell Cycle Progression of Lung Cancer Cells Independently of Its Established Substrates Ral GTPases", *PLoS One*, vol. 11, no. 5, 2016. http://dx.doi.org/10.1371/journal.pone.0154840

[47] Z. Tracz-Gaszewska *et al.*, "Molecular Chaperones in the Acquisition of Cancer Cell Chemoresistance with Mutated TP53 and MDM2 Up-Regulation", *Oncotarget*, vol. 8, no. 47, pp. 82123–82143, 2017. http://dx.doi.org/10.18632/oncotarget.18899

[48] J. Ma *et al.*, "Systematic Analysis of Sex-Linked Molecular Alterations and Therapies in Cancer", *Scientific Reports*, vol. 6, no. 1, pp. 19119–19127, 2016. http://dx.doi.org/10.1038/srep19119

[49] S. J. Yeh *et al.*, "Comparing Progression Molecular Mechanisms Between Lung Adenocarcinoma and Lung Squamous Cell Carcinoma Based on Genetic and Epigenetic Networks: Big Data Mining and Genome-Wide Systems Identification", *Oncotarget*, vol. 10, no. 38, pp. 3760–3806, 2019. http://dx.doi.org/10.18632/oncotarget.26940

[50] M. Imielinski *et al.*, "Oncogenic and Sorafenib-Sensitive ARAF Mutations in Lung Adenocarcinoma", *The Journal of Clinical Investigation*, vol. 124, no. 4, pp. 1582–1586, 2014. http://dx.doi.org/10.1172/JCI72763

[51] Z. Yang *et al.*, "Identification of Gene Markers in the Development of Smoking-Induced Lung Cancer", *Gene*, vol. 576, no. 1, pp. 451–457, 2016. http://dx.doi.org/10.1016/j.gene.2015.10.060

[52] S. Qin *et al.*, "Fibronectin Protects Lung Cancer Cells Against Docetaxel-Induced Apoptosis by Promoting Srcand Caspase-8 Phosphorylation", *Tumour Biology*, vol. 37, no. 10, pp. 13509–13520, 2016. https://doi.org/10.1007/s13277-016-5206-8

[53] P. Ashford *et al.*, "A CATH Domain Functional Family Based Approach to Identify Putative Cancer Driver Genes and Driver Mutations", *Scientific Reports*, vol. 9, 2019. http://dx.doi.org/10.1038/s41598-018-36401-4

[54] F. Sotgia and M. P. Lisanti, "Mitochondrial Markers Predict Survival and Progression in Non-Small Cell Lung Cancer (NSCLC) Patients: Use as Companion Diagnostics", *Oncotarget*, vol. 8, no. 40, pp. 68095–68107, 2017. http://dx.doi.org/10.18632/oncotarget.19677

[55] J. A. Hwang *et al.*, "Genomic Profiles of Lung Cancer Associated with Idiopathic Pulmonary Fibrosis", *The Journal of Pathology*, vol. 244, no. 1, pp. 25–35, 2018. http://dx.doi.org/10.1002/path.4978

[56] X. Tan and M. Chen, "MYLK and MYL9 Expression in Non-Small Cell Lung Cancer Identified by Bioinformatics Analysis of Public Expression Data", *Tumour Biology*, vol. 35, no. 12, pp. 12189–12200, 2014. http://dx.doi.org/10.1007/s13277-014-2527-3

[57] A. Ilhan-Mutlu *et al.*, "Expression Profiling of Angiogenesis-Related Genes in Brain Metastases of Lung Cancer and Melanoma", *Tumour Biology*, vol. 37, no. 1, pp. 1173–1182, 2016. http://dx.doi.org/10.1007/s13277-015-3790-7

[58] G. Gorgisen *et al.*, "Identification of Novel Mutations of Insulin Receptor Substrate 1 (IRS1) in Tumor Samples of Non-Small Cell Lung Cancer (NSCLC): Implications for Aberrant Insulin Signaling in Development of Cancer", *Genetics and Molecular Biology*, vol. 42, no. 1, pp. 15–25, 2019. http://dx.doi.org/10.1590/1678-4685-gmb-2017-0307

[59] J. Linderoth *et al.*, "Genes Associated with the Tumour Microenvironment are Differentially Expressed in Cured Versus Primary Chemotherapy-Refractory Diffuse Large B-Cell Lymphoma", *British Journal of Haematology*, vol. 141, no. 4, pp. 423–432, 2008. http://dx.doi.org/10.1111/j.1365-2141.2008.07037.x

[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

[61] L. Yang *et al.*, "Axin Gene Methylation Status Correlates with Radiosensitivity of Lung Cancer Cells", *BMC Cancer*, vol. 13, 2013. http://dx.doi.org/10.1186/1471-2407-13-368

[62] S. Tian, "Identification of Monotonically Differentially Expressed Genes for Non-Small Cell Lung Cancer", *BMC Bioinformatics*, vol. 20, 2019. http://dx.doi.org/10.1186/s12859-019-2775-8

[63] Y. Sun *et al.*, "Identifying Candidate Agents for Lung Adenocarcinoma by Walking the Human Interactome", *OncoTargets and Therapy*, vol. 9, pp. 5439–5450, 2016. http://dx.doi.org/10.2147/OTT.S97357

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

Contact addresses:

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

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

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

Peng Li was born in Dongying, Shandong, China, in 1975. He received his BSc degree in computer science from Qingdao University, Shangdong, 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 process 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.