IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 29, 2019, accepted October 17, 2019, date of publication October 24, 2019, date of current version November 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949415

Recognizing Novel Tumor Suppressor Genes Using a Network Machine Learning Strategy

RAN ZHAO[1], LEI CHEN[1,2], BO ZHOU[3], ZI-HAN GUO[1], SHUAIQUN WANG[1], AND AORIGELE[4]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
[2]Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai 200241, China
[3]School of Basic Medicine, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China
[4]Faculty of Engineering, University of Toyama, Toyama 930-8555, Japan

Corresponding authors: Lei Chen (chen_lei1@163.com) and Bo Zhou (zhoub@sumhs.edu.cn)

This work was supported in part by the Natural Science Foundation of Shanghai under Grant 17ZR1412500, and in part by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 18dz2271000.

ABSTRACT Extensive research on tumor suppressor genes (TSGs) is helpful to understand the pathogenesis of cancer and design effective treatments. However, using traditional experiments to identify TSGs is of high costs and time-consuming. It is an alternative way to design effective computational methods for screening out latent TSGs. Up to now, some computational methods have been proposed to predict new TSGs. However, these methods did not contain a learning procedure to extract essential properties of validated TSGs, reducing their efficiencies. In this study, a novel computational method was proposed to identify latent TSGs. To this end, we downloaded the validated TSGs from the TSGene database (Version 1.0). These TSGs together with other genes were represented by features that were extracted from protein-protein interaction networks in STRING via a powerful network embedding method, Mashup. Then, thirty random forest models were constructed and used to predict latent TSGs. 135 inferred TSGs were obtained, where 28 genes have been included in the TSGene database (Version 2.0). Our method had better performance than some previous methods according to the validated TSGs in the TSGene database (Version 2.0). For the rest 107 inferred TSGs, some of them can be confirmed to be TSGs with solid literature support. Finally, our method can overcome the defects that only genes with strong associations to validated TSGs can be identified because we obtained several inferred TSGs that had weak associations to validated TSGs and can be novel TSGs with high probabilities.

INDEX TERMS Tumor suppressor gene, network embedding method, mashup, machine learning, random forest.

I. INTRODUCTION

Cancer is one of the leading causes of human death in the world. Based on the information reported by World Health Organization (WHO), more than 8.8 million people directly die from cancer all over the world, which takes approximate 1/6 of all global deaths. The costs on cancer prevention, diagnosis, and treatment reach up to 1.16 trillion dollars [1]. Thus, in the past years, lots of investigators devoted themselves to study the main pathogeneses of cancers. Up to now, part of them has been uncovered. However, we are still on a long way to understand the mechanism of cancers. According to current knowledge, genetic background and environmental factors are two main causes for forming cancers [2]. For different types of cancers, some related genes have been discovered. However, there still exist hidden genes waiting for us to be discovered.

The cancer-related genes can be simply divided into two types: oncogenes and tumor suppressor genes (TSGs) [3]. In general, oncogenes can promote the tumor initiation, while TSGs can protect cells from malignant alterations [4]. Based on the two-hit hypothesis, which explains the genetic contribution on tumor initiation [5], [6], it is more difficult for us to discover TSGs than oncogenes. Identification of TSGs via traditional experiments is of high costs and time-consuming. With the development of computer science, a plenty of advanced computational methods have been proposed, which provide new ways to identify novel TSGs. In recent years,

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

some studies investigated TSGs via designing computational methods. In 2014, Yang et al. investigated TSGs based on gene ontology (GO) terms and biological pathways [7]. Several key GO terms and pathways were extracted via machine learning algorithms, which were deemed to be important factors for identifying TSGs. Later, Chen et al. proposed a shortest path (SP)-based method to mine novel TSGs in a protein-protein interaction (PPI) network based on current confirmed TSGs [8]. 205 novel TSGs were proposed in their study. Recently, Chen et al. further proposed two other network-based methods for predicting novel TSGs [9]. These two methods adopted Laplacian heat diffusion (LHD) and random walk with restart (RWR) algorithms to search novel TSGs in a PPI network, respectively. Obtained genes were further screened by three tests. 140 genes and 41 genes were discovered by these two methods, which were deemed to be novel TSGs.

The previous methods proposed in [8] and [9] were all network-based methods. However, they did not contain a learning procedure to capture essential properties of validated TSGs, reducing their efficiencies. Furthermore, methods in [9] employed screening tests to improve the prediction quality. However, they also excluded hidden TSGs with weak associations to validated ones, that is, only genes with strong associations to validated TSGs can be identified. In this work, a novel computational method was proposed to identify novel TSGs, which can partly overcome above-mentioned shortcomings. Our method still employed PPI networks reported in STRING [10]. A network embedding algorithm, Mashup [11], was applied on these networks to access the feature vector representations of genes. Then, several random forest (RF) [12] models were built to learn the differences between validated TSGs, retrieved from TSGene database (Version 1.0) [13], and other genes, thereby predicting novel TSGs. 135 inferred TSGs were identified by our method, in which several of them were included in the updated TSGene database (Version 2.0). For some other inferred TSGs, an extensive analysis was performed to prove that they can be novel TSGs with high probabilities.

II. MATERIALS AND METHODS

A. MATERIALS

The purpose of the study was to find out latent TSGs according to validated TSGs. Thus, we retrieved validated TSGs from TSGene database (Version 1.0, https://bioinfo.uth.edu/TSGene1.0/) [13]. 716 human TSGs were obtained. These genes were also investigated in previous studies [7]–[9]. Because we employed the PPI networks reported in STRING, in which proteins are represented by Ensembl IDs. Thus, we extracted the proteins of these 716 genes and mapped them onto their Ensembl IDs. After removing Ensembl IDs that did not occur in the PPI networks, 631 Ensembl IDs, representing proteins of TSGs, were accessed. For convenience, we denoted the set comprising these 631 Emsembl IDs as S_1.

In addition, to validate the performance of our method for inferring novel TSGs, we further employed the human TSGs reported in TSGene database (Version 2.0, https://bioinfo.uth.edu/TSGene) [14]. 1,217 TSGs were obtained. After similar process described in the above paragraph, 1,011 Ensembl IDs were extracted and they constituted the set S_2. Compared with the Ensemble IDs in S_1, there were 449 Ensembl IDs in $S_2 - S_1$, that is, 449 TSGs were added into the TSGene database (Version 2.0).

B. PROTEIN-PROTEIN INTERACTION AND NETWORK CONSTRUCTION

Similar to previous studies [8], [9], we also employed the PPI networks reported in STRING (https://string-db.org/, Version 10.0) [10] to construct our method. However, all PPI networks were used in this work. In fact, the PPI information in STRING reports the associations of proteins from several aspects of proteins. Thus, for each protein aspect, a PPI network can be constructed. In detail, from the downloaded file '9606.protein.links.detailed.v10.txt.gz', which contains 4,274,001 PPIs involving 19,247 human proteins, each PPI contains two Ensembl IDs and eight scores, titled by 'Neighborhood', 'Fusion', 'Cooccurence', 'Coexpression', 'Experiment', 'Database', 'Textmining', 'Combined_score'. All scores are between 1 and 999. The first seven scores measure the associations of proteins from seven different aspects of proteins, while the last score 'Combined_score' integrates above seven scores with a naive Bayesian fashion [15]. Because we did not know whether the integration scheme for seven scores was optimal for building our method, the last 'Combined_score' was not adopted in our study. For other seven scores, they were denoted by $S_N(p_1, p_2)$, $S_F(p_1, p_2)$, $S_{CO}(p_1, p_2)$, $S_{CE}(p_1, p_2)$, $S_E(p_1, p_2)$, $S_D(p_1, p_2)$ and S_T (p_1, p_2), respectively.

Based on above-mentioned PPIs, we built seven PPI networks as follows. Since each network was generated in a similar way, we only gave the description of the network using 'Neighborhood' score. The network with 'Neighborhood' score defined the 19,247 human proteins as nodes and two nodes were adjacent if and only if the 'Neighborhood' score between them was greater than zero. For convenience, such network was denoted by N_N. With same procedures, other six networks were constructed, which were denoted by N_F, N_{CO}, N_{CE}, N_E, N_D, and N_T, respectively. The sizes of each network (i.e., number of edges) are shown in **Table 1**.

In previous studies [8], [9], the methods were executed on the PPI network with 'Combined_score'. Although this network can widely measure the associations of proteins, it inevitably ignores some special aspects of proteins. Thus, in this work, we directly used the individual seven PPI networks to design the method, which may contain more information of protein, thereby providing more opportunities to discover hidden TSGs.

TABLE 1. Detailed information of seven networks.

Network	Source score	Size
N_N	Neighborhood	76214
N_F	Fusion	2060
N_{CO}	Cooccurence	23739
N_{CE}	Coexpression	768962
N_E	Experiment	1736931
N_D	Database	212430
N_T	Textmining	3816497

C. NETWORK EMBEDDING METHOD

Seven PPI networks were constructed as mentioned in Section II.B, which is one difference from previous studies. On the other hand, previous studies did not contain a procedure to learn the essential properties of validated TSGs, reducing the efficiencies of their methods. Thus, in this work, we tried to employ a learning procedure on validated TSGs. Because majority machine learning algorithms require the input sample to be a vector, a powerful network embedding method, Mashup [11], was employed to extract features of proteins from above-constructed seven PPI networks. Mashup is a compact network embedding method [11]. It first extracts the basic features of each node from each network and then fuses several feature vectors, which are derived from different networks, for the same node into one compact vector. The features obtained by Mashup can abstract the protein associations in a system level, which may include informative essential properties of proteins. To date, it has been applied to tackle several biological problems [16]–[23]. A brief description for extracting features via Mashup was as below.

Given a PPI network N_j ($j \in \{N, F, CO, CE, E, D, T\}$), Mashup uses the random walk with restart (RWR) algorithm [9], [24], [25] to access the raw features of proteins. In detail, for a node in the PPI network N_j, the RWR algorithm is executed on such network with this node as the seed node. After RWR algorithm stops, each node in the network receives a probability. Obtained probabilities are aligned in a vector to generate the feature vector of the given node. For formulation, the feature vector of p_i on N_j was denoted by V_j^i. Clearly, seven feature vectors can be accessed from seven PPI networks. It is necessary to fuse them into one vector. On the other hand, the raw feature vector always has a high dimension. A dimensionality reduction procedure can be done simultaneously. Suppose X^i represents the final feature vector of protein p_i and W_j^i denotes the context feature vector of p_i in the network N_j. The following procedures are to determine the optimal components in X^i and W_j^i. Let $\tilde{V}_j^i$ be a vector for protein p_i in network N_j. Its components can be determined by X^i and W_j^i as follows:

$$\tilde{V}_{jk}^{i} = \frac{\exp((X^i)^T W_j^k)}{\sum_{k'} \exp((X^i)^T W_j^{k'})} \quad k = 1, 2, \ldots, n \tag{1}$$

where n represents the total number of different proteins in all PPI networks. Clearly, the vector $\tilde{V}_j^i$ should be approximate to V_j^i as much as possible, thereby determining the optimal values in X^i and W_j^i. Thus, mashup solves the following optimization problem

$$\underset{X^i, W_j^i}{\text{minimize}} \ \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} D_{KL}(V_j^i || \tilde{V}_j^i) \tag{2}$$

where m is the total number of networks and $D_{KL}(\bullet)$ is the function of KL-divergence (relative entropy).

In this study, the Mashup program was retrieved from http://cb.csail.mit.edu/cb/mashup/. Default parameters were used and we set the dimension of output vector to be 500.

D. RANDOM FOREST

RF [12] is a classic machine learning algorithm, which consists of several decision trees. Although decision tree is a relatively weak classifier. However, RF has been deemed to be an excellent and strong classifier [26]. In bioinformatics and computational biology, it is always an important candidate for building different models [17], [27]–[32]. When constructing a decision tree in RF, two random selection procedures are adopted. The first one is used to generate the dataset. In detail, randomly select samples, with replacement, which are as many as those in the original dataset, to construct a dataset. Based on samples in the above-mentioned dataset, the second one is used to randomly select features for extending the tree. After the predefined number of decision trees have been constructed, a RF classifier is built. For a query sample, each decision tree yields a predicted result. The RF integrates these results via majority voting.

In this study, a tool 'RandomForest' in Weka was directly employed, which implements the RF. Default parameters were used. The main parameter, number of decision tree, was ten. The Weka can be downloaded at https://www.cs.waikato.ac.nz/ml/weka/downloading.html.

E. THE PROPOSED METHOD

As mentioned in Section II.A, 631 proteins (Ensembl IDs) of human TSGs were accessed from TSGene database (Version 1.0). We tried to learn the essential properties of these proteins via RF models. To this end, these proteins were termed as positive samples, while the rest unlabeled proteins in the PPI networks were deemed as negative samples. However, there were 18,616 negative samples, which were much more than positive samples. The RF model directly constructed on all samples would have poor performance. Thus, 18,616 proteins were randomly and equally divided into thirty parts, where 16 contained 621 proteins and 14 comprised 620 proteins. Proteins in each part were combined with positive samples to constitute a dataset. Thirty datasets, denoted as $DS_1, DS_2, \ldots, DS_{30}$, were generated. A RF model was built on each dataset.

Above-constructed RF models can learn some essential differences between TSGs and other genes. Thus, they can be further used to identify hidden TSGs from unlabeled 18,616 genes in the PPI network. For each of these genes,

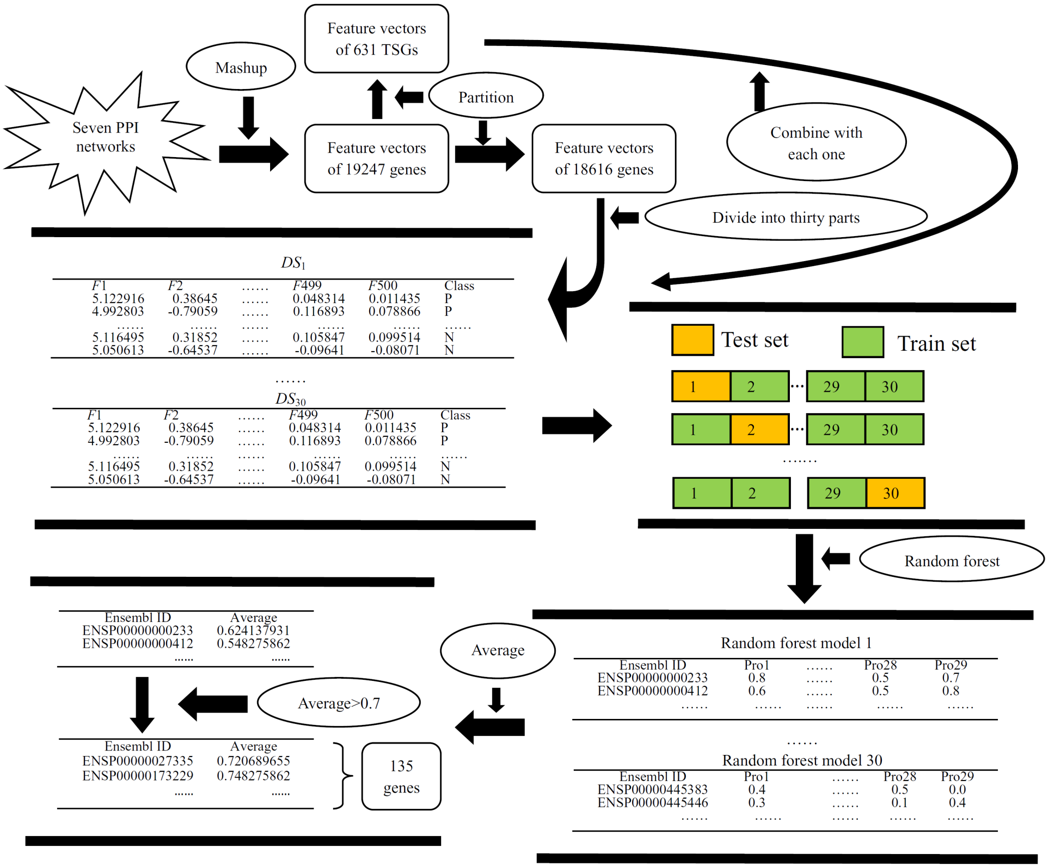

Seven PPI networks
Mashup
Feature vectors of 631 TSGs
Feature vectors of 19247 genes
Partition
Feature vectors of 18616 genes
Combine with each one
Divide into thirty parts
DS_1
F1 F2 …… F499 F500 Class
5.122916 0.38645 …… 0.048314 0.011435 P
4.992803 -0.79059 …… 0.116893 0.078866 P
…… …… …… …… …… ……
5.116495 0.31852 …… 0.105847 0.099514 N
5.050613 -0.64537 …… -0.09641 -0.08071 N
……
DS_{30}
F1 F2 …… F499 F500 Class
5.122916 0.38645 …… 0.048314 0.011435 P
4.992803 -0.79059 …… 0.116893 0.078866 P
…… …… …… …… …… ……
5.116495 0.31852 …… 0.105847 0.099514 N
5.050613 -0.64537 …… -0.09641 -0.08071 N
Test set
Train set
1 2 … 29 30
1 2 … 29 30
…….
1 2 29 30
Random forest
Random forest model 1
Ensembl ID Pro1 …… Pro28 Pro29
ENSP00000000233 0.8 …… 0.5 0.7
ENSP00000000412 0.6 …… 0.5 0.8
…… …… …… …… ……
……
Random forest model 30
Ensembl ID Pro1 …… Pro28 Pro29
ENSP00000445383 0.4 …… 0.5 0.0
ENSP00000445446 0.3 …… 0.1 0.4
…… …… …… …… ……
Average
Ensembl ID Average
ENSP00000000233 0.624137931
ENSP00000000412 0.548275862
…… ……
Average>0.7
Ensembl ID Average
ENSP00000027335 0.720689655
ENSP00000173229 0.748275862
…… ……
135 genes

FIGURE 1. Flow chart to illustrate the procedures of the proposed method. The Mashup was applied on seven PPI networks to encode 19247 genes, where 631 were validated tumor suppressor genes (TSGs). Rest unlabeled 18616 genes were divided into thirty parts and TSGs were poured into each part to constitute a dataset. A random forest model was built on each dataset and used to produce a probability of each gene not included in this model to be a novel TSG. An average probability was calculated for each unlabeled gene and those with such values larger than 0.7 were selected as inferred TSGs.

it was fed into 29 RF models (the RF model containing it as a negative sample was not used). 29 predicted results can be accessed. To improve the accuracy, we extracted the probability of each gene to be a TSG (positive sample) rather than the predicted class, that is, 29 probabilities of each gene can be obtained. The average value of these probabilities was calculated and picked up as the final measurement to indicate its likelihood to be novel TSGs. Evidently, a gene with a high average probability was more likely to be a novel TSG. To select reliable TSGs, we can set a high threshold for the average probability to screen out most possible genes. These genes were called inferred TSGs in the following text.

The procedures described above are illustrated in **Figure 1**.

F. PERFORMANCE MEASUREMENTS

In this study, we first constructed thirty RF models. In these models, proteins of TSGs were deemed as positive samples and others were termed as negative samples. Thus, these models were binary classification models. Tenfold cross-validation [33], [34] was adopted to test their performance. For the predicted results in binary classification, we generally count four values: true positive (TP), false negative (FN), false positive (FP) and true negative (TN). Accordingly, two measurements: Prediction accuracy (ACC) and Matthews correlation coefficient (MCC) [17], [27], [35]–[39] can be computed. Their definitions are as follows:

$$\begin{cases} ACC = \dfrac{TP+TN}{TP+FN+FP+TN} \\ MCC = \dfrac{TP\times TN - FP\times FN}{\sqrt{(TN+FN)\times(TN+FP)\times(TP+FN)\times(TP+FP)}} \end{cases} \tag{3}$$

Besides, we also integrated 30 RF models to infer novel TSGs. Based on the TSGs reported in TSGene database (Version 2.0) and not included in TSGene database (Version 1.0), that is, members in $S_2 - S_1$, we can evaluate the performance of our method with Precision (P), Recall (R) and F1-measure [17], [27], [40], which can be calculated by

$$\begin{cases} R = \dfrac{TP}{TP+FN} \\ P = \dfrac{TP}{TP+FP} \\ F1-measure = \dfrac{2\times P\times R}{P+R} \end{cases} \tag{4}$$

TABLE 2. Performance of three methods by comparing the inferred genes with new included genes in TSGene database.

Method	Number of inferred TSGs	Number of inferred TSGs included in TSGene database (Version 2.0)	Precision	Recall	F1-measure
The proposed method	135	28	**0.207**	**0.062**	**0.096**
LHD-based method	41	2	0.049	0.004	0.008
RWR-based method	140	20	0.143	0.045	0.068

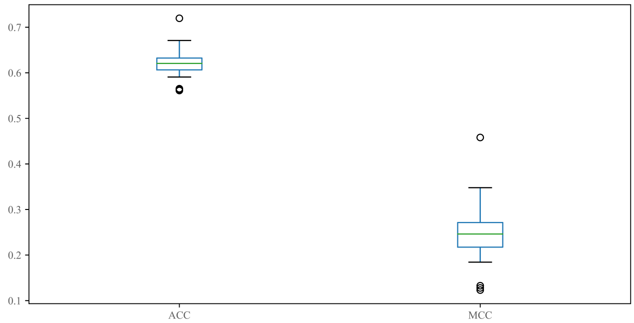


FIGURE 2. Box plot to show the performance of thirty RF models.

III. RESULTS

A. PERFORMANCE OF THIRTY RF MODELS

The unlabeled genes in the PPI networks were deemed as negative samples when constructing RF models, inducing that negative samples were much more than positive samples. A strategy described in Section II.E was proposed, resulting in 30 datasets, on each of which an RF model was built. Ten-fold cross-validation was adopted to evaluate these models. The predicted results were counted as ACC and MCC, which were shown in **Figure 2**. It can be observed that majority ACCs were between 0.6 and 0.7, while most MCCs varied between 0.2 and 0.35. The performance of these models was not very high because some unlabeled genes may be hidden TSGs. Anyway, they can still capture more or less essential properties of TSGs, thereby helping infer novel TSGs.

B. INFERRED TUMOR SUPPRESSOR GENES YIELDED BY OUR METHOD

As mentioned above, 30 RF models were built. For each unlabeled gene in the PPI network, it was feed into 29 RF models as a testing sample because it took part in the construction of one RF model. These 29 models yielded 29 probabilities to indicate its likelihood to be novel TSGs. Clearly, the average probability can fully represent such likelihood. The average probabilities of 18,616 unlabeled genes are provided in Table S1. By setting the threshold of average probability to be 0.7, 135 inferred TSGs were picked up.

Among 135 inferred genes, 28 genes were reported in TSGene database (Version 2.0), that is, they were in the set $S_2 - S_1$. If $S_2 - S_1$ was set as the benchmark testing dataset, we can calculate P, R and F1-measure as mentioned in Section II.F, which are listed in **Table 2**. They were 0.207, 0.062 and 0.096, respectively. Although they look low, they were higher than those yielded by previous methods, which would be analyzed in Section III.C.

C. COMPARISON WITH PREVIOUS METHODS

There were some previous studies focusing on identifying TSGs with computational methods. In [8], a SP-based method was proposed for inferring new TSGs. However, their method used an old version of PPI network (Version 9.0) in STRING, which was quite different from the PPI networks (Version 10.0) used in this study. Thus, it is not fair to compare their results with ours. Here, we employed the results in [9] to make comparisons, which adopted the same version of PPI networks. There were two methods, namely LHD-based and RWR-based methods, respectively, in such study. 140 inferred genes were yielded by the RWR-based method, while LHD-based method produced 41 inferred genes.

Among the 140 inferred genes yielded by the RWR-based method, 20 genes have been included in TSGene database (Version 2.0). For the 41 genes produced by the LHD-based method, only two genes were in TSGene database (Version 2.0). Likewise, we calculated the P, R and F1-measure for these two methods, which are also listed in **Table 2**. They were 0.143, 0.045 and 0.068 for RWR-based method, respectively, while they were 0.049, 0.004 and 0.008, respectively, for LHD-based method. These measurements were all lower than those of our method, indicating the utility of our method. In detail, the F1-measure of our method was about 0.03 and 0.09 higher than those of other two methods.

In addition, a Venn diagram was plotted in **Figure 3(A)** to illustrate the numbers of common genes and different genes in three inferred TSG sets. It can be observed that among 135 inferred TSGs yielded by our method, two genes were also predicted by other two methods, three genes were only predicted by LHD-based method, 24 genes were only identified by RWR-based method, and 106 genes were not recognized by other two methods. Interestingly, the inferred genes of our method were much similar to those of RWR-based method. It is reasonable because we used the features of proteins that were refined from the raw features derived by the RWR algorithm. The detailed predicted results of 135 inferred TSGs are listed in **Table S2**.

The LHD-based and RWR-based methods contained three screening tests. The association test excluded the candidate genes without highest confidence associations ('combined_score' no less than 900) to validated TSGs. In another

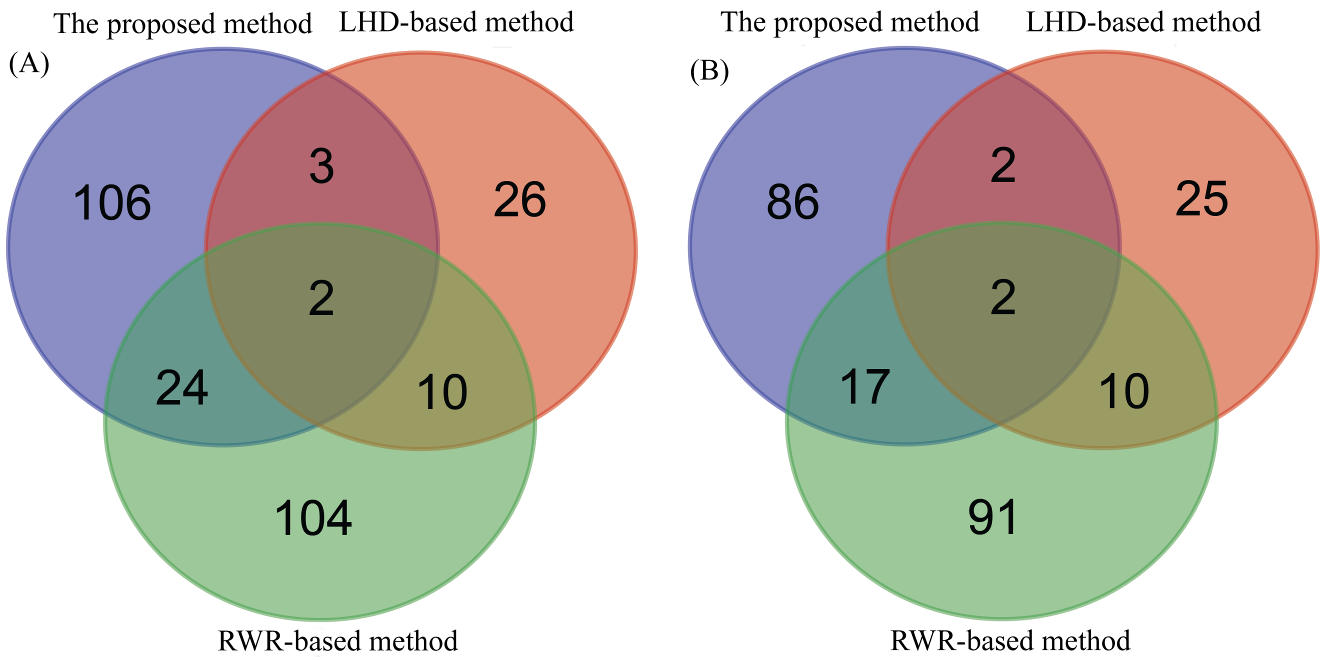


FIGURE 3. Venn Diagrams to show inferred TSGs yielded by three methods using validated TSGs reported in TSGene database (Version 1.0). (A) Venn Diagram to show all inferred TSGs yielded by three methods; (B) Venn Diagram to show inferred TSGs yielded by three methods, which were not included in TSGene database (Version 2.0).

word, only candidate genes with highest confidence associations to validated TSGs can be selected by these two previous methods. However, this is not the case. To indicate this fact, we picked up genes in $S_2 - S_1$ and extract all PPIs between them and TSGs in S_1. According to cutoff values reported in STRING, we counted the numbers of PPIs with 'Combined_score' in [0, 399], [400, 699], [700, 899] and [900, 999] for each gene in $S_2 - S_1$, which are listed in Table S3. Most genes in $S_2 - S_1$ had few PPIs with highest confidence scores. The number of PPIs with highest confidence scores ('combined_score' no less than 900) for each gene in $S_2 - S_1$ is counted in **Figure 4**. It can be seen that there were 199 genes without PPIs with highest confidence scores, meaning that these genes were impossible predicted by LHD-based or RWR-based methods. For the 135 inferred TSGs yielded by our method, 26 genes had no PPIs with highest confidence associations to validated TSGs, in which one was included in TSGene database (Version 2.0). It is indicated that our method can deeply mine novel TSGs based on validated TSGs because it can infer candidates with weak associations to validated ones.

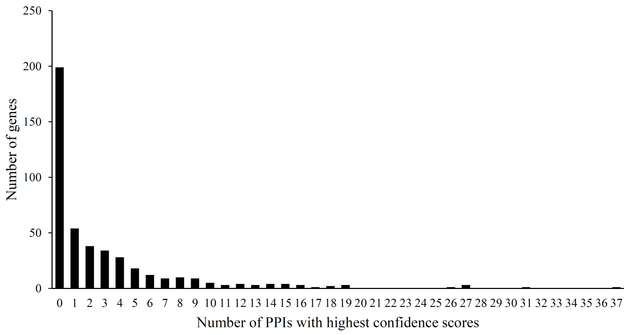


FIGURE 4. Distribution of TSGs in TSGene database (Version 2.0) but not in TSGene database (Version 1.0) based on their PPIs with highest confidence associations to TSGs in TSGene database (Version 1.0).

IV. DISCUSSION

41 inferred genes was obtained by the LHD-based method and RWR-based method yielded 140 inferred genes [9]. In this study, we designed a novel method with a learning procedure for identifying novel TSGs. 135 inferred TSGs were obtained. Among these genes, 28 genes have been included in TSGene database (Version 2.0), indicating they were actual TSGs. Of the remaining 107 genes, 86 genes were exclusively predicted by our method (**Figure 3(B)**), that is, they were not identified by LHD-based or RWR-based methods. These inferred TSGs can be a useful complement to the inferred TSGs reported in previous studies. Here, we selected some of them to make analyses, which are listed in **Table 3**. Related experimental studies confirmed that they can act as tumor suppressors, validating the reliability of our results. According to the published studies support, we divided these inferred TSGs into two groups: (I) Inferred TSGs with solid literature support; and (II) Inferred TSGs only with a considerable degree of functional relevance, which require more intensive studies. The detailed analyses for each gene are presented below.

A. INFERRED TSGS WITH SOLID LITERATURE SUPPORT

RERG encodes a Ras superfamily small GTP binding and hydrolyzing protein (GTPase). *RERG* has been predicted to

TABLE 3. Detailed analysis of important inferred TSGs yielded by our method.

Ensembl ID	Gene symbol	Full name	Probability	Category	References
ENSP00000256953	*RERG*	Ras-like estrogen-regulated growth inhibitor	0.7414	Solid literature	[41-47]
ENSP00000229002	*RERGL*	RAS-like and estrogen-regulated growth inhibitor like	0.7724	Solid literature	[48-51]
ENSP00000359866	*BMP5*	bone morphogenetic protein 5	0.7448	Solid literature	[52-59]
ENSP00000225688	*RASD1*	Ras Related Dexamethasone Induced 1	0.7345	Solid literature	[60-65]
ENSP00000430333	*SLIT3*	Slit Guidance Ligand 3	0.7207	Solid literature	[66-76]
ENSP00000432743	*TP53AIP1*	tumor protein p53 regulated apoptosis inducing protein 1	0.7172	Solid literature	[77-82]
ENSP00000216807	*BRMS1L*	breast cancer metastasis-suppressor 1-like	0.7034	Solid literature	[83-93]
ENSP00000435412	*RPS6KA1*	Ribosomal Protein S6 Kinase A1	0.8000	Functional relevance	[94]
ENSP00000368884	*RPS6KA3*	Ribosomal Protein S6 Kinase A3	0.8448	Functional relevance	[95-97]
ENSP00000202556	*PPP1R13B*	Protein Phosphatase 1 Regulatory Subunit 13B	0.7414	Functional relevance	[98-100]
ENSP00000341268	*TRADD*	Tumour necrosis factor receptor (TNFR)- associated via death domain	0.7138	Functional relevance	[80, 101-103]

be closely connected with the inhibition of breast cancer cell proliferation and tumor formation [41], and its expression has correlated inversely with patient survival and the development of distant metastases [42]. Significantly hypermethylated *RERG* was also observed in colorectal adenocarcinoma [43], [44], breast cancer [45], and nasopharyngeal carcinoma [46]. Recent publications have shown that *RERG* participates in tumor suppression by regulating extracellular signal-regulated kinase (ERK) and nuclear factor (NF)-κB signaling pathway in breast cancer [45], nasopharyngeal carcinoma [46] and prostatic carcinoma [47]. So far, the function of *RERGL* is unknown, however, protein alignment suggests that the protein encoded by *RERGL* shares the majority of conserved regions and GTP-binding regions with the protein encoded by *RERG*. Depletion of *RERGL* was observed in colorectal cancer patients [48]; and the expression of *RERGL* was significantly correlated with the overall survival time with colorectal cancer patients [49]. The possibility of tumor related characteristics of *RERGL* was also highlighted by two reports, which found five mutations in the colorectal cancer tumor from five colorectal cancer cases [50], [51]. Therefore, it is reasonable to forecast that *RERG* and *RERGL* may function as TSGs.

BMP5 encodes a member of the bone morphogenetic protein family, which is part of the transforming growth factor-beta (TGF-β) superfamily. Although BMPs have been originally identified by their ability to induce bone and cartilage development [52], it has been shown to affect tumorigenesis in a variety of tumors [53]. For example, *BMP5* is down-regulated in various human cancers, including melanoma [54], adrenocortical carcinoma [55], breast cancer [56], and pancreatic cancer [57]. Two recent studies uncovered the distinctive role of *BMP5* in sporadic colorectal cancer (CRC). One is a genomic and transcriptomic profiling based study, which identified the common alterations in *BMP5* and its effect on cell growth and migration, implying its potential tumor suppressor function in human CRC [58]. The other study in CRC determined *BMP5* is the direct target of miR-32, an oncogene, and loss of tumor suppressor *BMP5* may partially due to the miR-32 dysregulation. Therefore, *BMP5* may definitely be a potential TSG [59].

RASD1 is a member of Ras superfamily of small GTPases coding gene and is induced by dexamethasone. *RASD1* is localized to human chromosome 17p11.2, a region associated with a high incidence of heterozygous deletions and deletions in cancer [60], [61]. Since*RASD1* has ~35% similarity to each major RAS subfamilies [62], the available evidence suggests that unlike other RAS family members, *RASD1* may play different roles in various cancer cells. Overexpression of *RASD1* results in inhibition of the growth of breast cancer, renal cell carcinoma and lung adenocarcinoma cell lines [63], [64]. A recent study showed that overexpressing *RASD1* had no significant influence on the proliferation of glioma cells, but inhibited glioma cell migration and invasion [65]. The study also found that high levels of *RASD1* predict good survival in patients with astrocytoma [65]. Therefore, *RASD1* may be a potential tumor suppressor.

Slit is a family of secreted extracellular matrix proteins, which regulate neuronal orientation and branching during nervous system development. In mammals, three *SLIT* genes, *SLIT1*, *2* and *3*, have been characterized to date. The role of *SLIT1* as a tumor suppressor gene has been discussed in our previous study [9]. *SLIT2* is frequently inactivated in various cancers [104]–[108] and its tumor suppressive role has been well-studied. It has been reported that hypermethylation and subsequent down-regulation of *SLIT3* are found in a variety of cancers including thyroid cancer [66], pancreatic ductal adenocarcinoma [67], gastric cancer [68], colorectal cancer [69], nasopharyngeal cancer [70], cervical cancer [71], ovarian cancer [72], lung carcinoma [73] and hepatocellular carcinoma [74]. In addition, *SLIT3* has been shown to inhibit tumor growth in mouse models [75] and impair cancer cell invasion and migration [66], [73], [76], suggesting that *SLIT3* acts as a tumor suppressor in a variety of cancers by repressing the tumor growth and progression.

TP53AIP1 encodes p53AIP1 protein, a mitochondrial membrane protein and the downstream target of p53 tumor suppressor that plays an important role in the tumor suppressor p53 dependent apoptotic signaling [78]. *TP53AIP1* is activated by UV exposure, causing its protein to accumulate in mitochondria and is significantly increased during UV-induced apoptosis [77], [78]. Some studies have shown that *TP53AIP1* plays a role in the progression of different cancer types. For example, *TP53AIP1* has been shown to be a prognostic factor in primary non-small cell lung cancer [79]. *TP53AIP1* was detected to be significantly reduced in breast cancer tissues. Moreover, the survival rate of breast cancer patients with low *TP53AIP1* levels is lower than those with high *TP53AIP1* levels [80]. Truncating *TP53AIP1* mutations have also been suggested to increase the risk of prostate cancer [81]. Overexpression of *TP53AIP1* up-regulates p53 levels in liver hepatocellular carcinoma cells, thereby inducing apoptosis and cell cycle arrest [82]. Therefore, *TP53AIP1* may definitely be a potential TSG.

BRMS1L, as another predicted TSG, shares 57% identical amino acid sequence with *BRMS1*, raising the possibility that they have similar functions. It is well known that *BRMS1* is a tumor suppressor that is involved in the ability of many tumors [83]–[86]. However, functional studies of *BRMS1L* have only recently been reported. It has been reported that *BRMS1L* is a novel downstream target of the p53 family and may be an inhibitor of cancer cell invasion and migration [87] through regulation of Wnt signaling pathway [88], [89], including breast cancer [88], adenoid cystic carcinoma [90], lung cancer [91] and ovarian cancer [89], [92]. Recent publications also suggested that reduced *BRMS1L* expression is associated with poor prognosis in breast cancer [88], ovarian cancer [89], glioma [93]. Therefore, *BRMS1L* might act as a putative cancer metastasis suppressor and a candidate for a clinical prognostic marker.

B. INFERRED TSGS WITH A CONSIDERABLE DEGREE OF FUNCTIONAL RELEVANCE

RPS6KA1 and *RPS6KA3* with high probabilities (no less than 0.800) yielded by our method encode ribosomal S6 protein kinase (RSK) 1 and RSK2, respectively, which are widely expressed and respond to many growth factors, peptide hormones, and neurotransmitters. RSKs appear to have important roles in a variety of cellular processes, including gene transcription, cell proliferation, cell growth, and differentiation [109]. Intriguingly, our understanding of RSK function in carcinogenic process remains inconsistent and is also complicated by the fact that different RSK isoforms may manifest opposing functions. For example, in prostate cancer, expression of RSK1 and 2 proteins, analyzed by Western blot analysis, have been previously shown to increase when the cancer is localized in the primary site [110] and in bone metastases [111], which provide strong evidence that RSKs is an important driver in prostate cancer progression in bone. *In vivo* evidence of RSK function in head and neck squamous cell carcinoma (HNSCC) showed that higher RSK2 levels correlated with increased metastasis and knockdown of RSK2 in human HNSCC cells also reduced the metastasis of xenografts in mice, whereas RSK1 has no effect on HNSCC metastasis [94]. However, it has been reported that RSK2 plays an important role in the DNA damage pathway that maintains genomic stability by mediating cell cycle progression and DNA repair [95]. Another strong evidence suggests that RSK2 deficiency can result in dramatically decreased IFNγ secretion, leading to immunosuppression and accelerated colon cancer metastasis and growth [96]. Based on recurrent mutations in *RPS6KA3* (9.6%) observed in human liver tumors, some scholars believe that since RSK2 is a known inhibitor of the RAS/MAPK pathway, RSK2 can act as a tumor suppressor and its inactivation can lead to activation of the RAS pathway [97]. To fully unravel the role of *RPS6KA1* and *RPS6KA3* in the regulation of tumorigenesis, further experiments based on different tumor types and RSK isoforms are still needed to elucidate the underlying mechanisms.

PPP1R13B was predicted by our method as a potential TSG. Although there is no direct evidence to confirm such gene as an actual TSG, it is quite reasonable to speculate that *PPP1R13B* may be a latent TSG based on its significant relationship with tumorigenesis reported in recent publications. According to recent publications, *PPP1R13B* is a major member of the apoptosis-stimulating proteins of the p53 family (ASPPs) [98]. Previous studies have reported that the PPP1R13B-p53 complex can promote p53-induced apoptosis and regulate apoptosis by specifically enhancing the ability of p53 to bind to DNA and acting on the pro-apoptotic gene promoter [99], [100]. Considering that p53-associated genes, such as *PPP1R13B*, are strongly associated with the initiation and progression of tumor, *PPP1R13B* is a potential TSG.

The next identified gene, *TRADD* encodes tumor necrosis factor receptor type 1-associated DEATH domain protein that mediates both cell death and inflammatory signals. As for its role in the development of tumors, although there are not enough direct reports to confirm that this gene triggers the occurrence or development of tumors, there are some related clues. For example, *TRADD* is located within chromosome 16q22.1, a region that frequently loses heterozygosity in various tumor types [101], [102]. A recent *in vitro* experiments demonstrated that *TRADD* is a target protein required to mediate receptor-interacting protein kinase 3 (RIP3) independent apoptosis, because *TRADD* knockdown inhibits TNFα-induced RIP3 knockdown of caspase activation and apoptosis in L929 cells, and recovery of *TRADD* expression rescues L929 cells from TNFα induction sensitivity [80]. *In vivo* experiments indicate that *TRADD* deficiency in mice accelerates tumor formation in a chemically induced carcinogenesis model [103], implying its potential contribution as a tumor suppressor.

As discussed above, several inferred genes can be novel TSGs, indicating that our proposed method can really extract hidden TSGs. Furthermore, four above-discussed genes: *RERG*, *RERGL*, *RASD1* and *BRMS1L*, had weak associations with validated TSGs in S_1. It is impossible for LHD-based and RWR-based methods to predict them. However, they were identified by our method. For the rest inferred genes, we listed them in Table S2. It is believed that some of them can be actual TSGs.

V. CONCLUSION

In this study, we proposed a novel computational method with a learning procedure to identify novel TSGs. This method used the features derived from PPI networks via a powerful network embedding algorithm. A number of RF models were built, which can learn essential differences between TSGs and other genes. The final obtained 135 inferred TSGs were produced by these RF models. Many of them can be confirmed to be novel TSGs. Hopefully, the new findings in this study can provide new insights for investigating TSGs.

REFERENCES

[1] S. McGuire, "World cancer report 2014. Geneva, Switzerland: World health organization, international agency for research on cancer, WHO press, 2015," *Adv. Nutrition*, vol. 7, pp. 418–419, Mar. 2016.

[2] F. P. Perera, "Environment and cancer: Who are susceptible?" *Science*, vol. 278, pp. 1068–1073, Nov. 1997.

[3] C. Lobry, P. Oh, M. R. Mansour, A. T. Look, and I. Aifantis, "Notch signaling: Switching an oncogene to a tumor suppressor," *Blood*, vol. 123, pp. 2451–2459, Apr. 2014.

[4] M. Kavianpour, A. Ahmadzadeh, S. Shahrabi, and N. Saki, "Significance of oncogenes and tumor suppressor genes in AML prognosis," *Tumor Biol.*, vol. 37, pp. 10041–10052, Aug. 2016.

[5] O. Hino and T. Kobayashi, "Mourning Dr. Alfred G. Knudson: The two-hit hypothesis, tumor suppressor genes, and the tuberous sclerosis complex," *Cancer Sci.*, vol. 108, pp. 5–11, Jan. 2017.

[6] A. J. W. Paige, "Redefining tumour suppressor genes: Exceptions to the two-hit hypothesis," *Cellular Mol. Life Sci. CMLS*, vol. 60, pp. 2147–2163, Oct. 2003.

[7] J. Yang, L. Chen, X. Kong, T. Huang, and Y.-D. Cai, "Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway," *PLoS ONE*, vol. 9, no. 9, 2014, Art. no. e107202.

[8] L. Chen, J. Yang, T. Huang, X. Kong, L. Lu, and Y.-D. Cai, "Mining for novel tumor suppressor genes using a shortest path approach," *J. Biomol. Struct. Dyn.*, vol. 34, no. 3, pp. 664–675, 2016.

[9] L. Chen, Y.-H. Zhang, Z. Zhang, T. Huang, and Y.-D. Cai, "Inferring novel tumor suppressor genes with a protein-protein interaction network and network diffusion algorithms," *Mol. Therapy Methods Clin. Develop.*, vol. 10, pp. 57–67, Sep. 2018.

[10] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, S. Santos, K. P. Tsafou, and M. Kuhn, "STRING v10: Protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Res.*, vol. 43, pp. D447–D452, Oct. 2015.

[11] H. Cho, B. Berger, and J. Peng, "Compact integration of multi-network topology for functional analysis of genes," *Cell Syst.*, vol. 3, pp. 540–548, Dec. 2016.

[12] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[13] M. Zhao, J. Sun, and Z. Zhao, "TSGene: A Web resource for tumor suppressor genes," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D970–D976, 2013.

[14] M. Zhao, P. Kim, R. Mitra, J. Zhao, and Z. Zhao, "TSGene 2.0: An updated literature-based knowledgebase for tumor suppressor genes," *Nucleic Acids Res.*, vol. 44, pp. D1023–D1031, Jan. 2016.

[15] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork, "STRING: Known and predicted protein–protein associations, integrated and transferred across organisms," *Nucleic Acids Res.*, vol. 33, pp. D433–D437, Jan. 2005.

[16] Z.-H. Guo, L. Chen, and X. Zhao, "A network integration method for deciphering the types of metabolic pathway of chemicals with heterogeneous information," *Combinat. Chem. High Throughput Screening*, vol. 21, no. 9, pp. 670–680, 2018.

[17] X. Zhao, L. Chen, Z.-H. Guo, and T. Liu, "Predicting drug side effects with compact integration of heterogeneous networks," *Current Bioinf.*, to be published.

[18] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Commun.*, vol. 8, no. 1, p. 573, 2017.

[19] R. Wang, G. Liu, C. Wang, L. Su, and L. Sun, "Predicting overlapping protein complexes based on core-attachment and a local modularity structure," *BMC Bioinf.*, vol. 19, p. 305, Dec. 2018.

[20] G. W. Schwartz, J. Petrovic, Y. Zhou, and R. B. Faryabi, "Differential integration of transcriptome and proteome identifies pan-cancer prognostic biomarkers," *Frontiers Genet.*, vol. 9, p. 205, Jun. 2018.

[21] C.-Y. Ma, Y.-P. P. Chen, B. Berger, and C.-S. Liao, "Identification of protein complexes by integrating multiple alignment of protein interaction networks," *Bioinformatics*, vol. 33, pp. 1681–1688, Jun. 2017.

[22] J. Peng, H. Wang, J. Lu, W. Hui, Y. Wang, and X. Shang, "Identifying term relations cross different gene ontology categories," *BMC Bioinf.*, vol. 18, p. 573, Dec. 2017.

[23] X. Zhang, L. Chen, Z.-H. Guo, and H. Liang, "Identification of human membrane protein types by incorporating network embedding methods," *IEEE Access*, vol. 7, pp. 140794–140805, 2019.

[24] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *Amer. J. Hum. Genet.*, vol. 82, pp. 949–958, Apr. 2008.

[25] L. Chen, T. Liu, and X. Zhao, "Inferring anatomical therapeutic chemical (ATC) class of drugs using shortest path and random walk with restart algorithms," *BBA Mol. Basis Disease*, vol. 1864, pp. 2228–2240, Jun. 2018.

[26] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, Oct. 2014.

[27] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Math. Biosci.*, vol. 306, pp. 136–144, Dec. 2018.

[28] L. Chen, Y.-H. Zhang, M. Zheng, T. Huang, and Y.-D. Cai, "Identification of compound–protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds," *Mol. Genet. Genomics*, vol. 291, pp. 2065–2079, Dec. 2016.

[29] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, "Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features," *J. Proteome Res.*, vol. 9, pp. 4992–5001, Oct. 2010.

[30] L. Wei, P. Xing, R. Su, G. Shi, Z. S. Ma, and Q. Zou, "CPPred-RF: A sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency," *J. Proteome Res.*, vol. 16, pp. 2044–2053, May 2017.

[31] L. Wei, P. Xing, J. Tang, and Q. Zou, "PhosPred-RF: A novel sequence-based predictor for phosphorylation sites using sequential information only," *IEEE Trans. Nanobiosci.*, vol. 16, no. 4, pp. 240–247, Jun. 2017.

[32] Q. Zhang, X. Sun, K. Feng, S. Wang, Y. H. Zhang, and S. Wang, "Predicting citrullination sites in protein sequences using mRMR method and random forest algorithm," *Combinat. Chem. High Throughput Screening*, vol. 20, pp. 164–173, Dec. 2017.

[33] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 1995, pp. 1137–1145.

[34] J.-P. Zhou, L. Chen, and Z.-H. Guo, "iATC-NRAKEL: An efficient multi-label classifier for recognizing anatomical therapeutic chemical (ATC) classes of drugs," *Bioinformatics*, to be published.

[35] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[36] H. Cui and L. Chen, "A binary classifier for the prediction of EC numbers of enzymes," *Current Proteomics*, vol. 16, no. 5, pp. 381–389, 2019.

[37] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, and J. Song, "*iFeature*: A Python package and Web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, pp. 2499–2502, Jul. 2018.

[38] J. Song, Y. Wang, F. Li, T. Akutsu, N. D. Rawlings, G. I. Webb, and K.-C. Chou, "iProt-sub: A comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites," *Briefings Bioinf.*, vol. 20, no. 2, pp. 638–658, Mar. 2019.

[39] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb, and K. C. Chou, "iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings Bioinf.*, to be published.

[40] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[41] B. S. Finlin, C.-L. Gau, G. A. Murphy, H. Shao, T. Kimel, R. S. Seitz, Y.-F. Chiu, D. Botstein, P. O. Brown, and C. J. Der, "RERG is a novel RAS-related, estrogen-regulated and growth-inhibitory gene in breast cancer," *J. Biol. Chem.*, vol. 276, pp. 42259–42267, Nov. 2001.

[42] H. O. Habashy, D. G. Powe, E. Glaab, G. Ball, I. Spiteri, N. Krasnogor, J. M. Garibaldi, E. A. Rakha, A. R. Green, C. Caldas, and I. O. Ellis, "RERG (Ras-like, oestrogen-regulated, growth-inhibitor) expression in breast cancer: A marker of ER-positive luminal-like subtype," *Breast Cancer Res. Treat.*, vol. 128, pp. 315–326, Jul. 2011.

[43] B. Øster, K. Thorsen, P. Lamy, T. K. Wojdacz, L. L. Hansen, K. Birkenkamp-Demtröder, S. Laurberg, T. F. Ørntoft, and C. L. Andersen, "Identification and validation of highly frequent CpG island hypermethylation in colorectal adenomas and carcinomas," *Int. J. Cancer*, vol. 129, pp. 2855–2866, Dec. 2011.

[44] X. Luo, R. Huang, H. Sun, Y. Liu, H. Bi, J. Li, H. Yu, J. Sun, S. Lin, B. Cui, and Y. Zhao, "Methylation of a panel of genes in peripheral blood leukocytes is associated with colorectal cancer," *Sci. Rep.*, vol. 6, Jul. 2016, Art. no. 29922.

[45] J.-Y. Ho, R.-J. Hsu, J.-M. Liu, S.-C. Chen, G.-S. Liao, H.-W. Gao, and C.-P. Yu, "MicroRNA-382-5p aggravates breast cancer progression by regulating the RERG/Ras/ERK signaling axis," *Oncotarget*, vol. 8, pp. 22443–22459, Apr. 2017.

[46] W. Zhao, N. Ma, S. Wang, Y. Mo, Z. Zhang, and G. Huang, "RERG suppresses cell proliferation, migration and angiogenesis through ERK/NF-κB signaling pathway in nasopharyngeal carcinoma," *J. Exp. Clin. Cancer Res.*, vol. 36, p. 88, Jun. 2017.

[47] Y. Xiong, H. Huang, S. Chen, H. Dai, and L. Zhang, "ERK5-regulated RERG expression promotes cancer progression in prostatic carcinoma," *Oncol. Rep.*, vol. 41, pp. 1160–1168, Feb. 2019.

[48] R. Yang, B. Chen, K. Pfütze, S. Buch, V. Steinke, E. Holinski-Feder, S. Stöcker, W. von Schönfels, T. Becker, H. K. Schackert, and B. Royer-Pokora, "Genome-wide analysis associates familial colorectal cancer with increases in copy number variations and a rare structural variation at 12P12.3," *Carcinogenesis*, vol. 35, pp. 315–323, Feb. 2014.

[49] H. Y. Liu and C. J. Zhang, "Identification of differentially expressed genes and their upstream regulators in colorectal cancer," *Cancer Gene Therapy*, vol. 24, pp. 244–250, Jun. 2017.

[50] Cancer Genome Atlas Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, pp. 330–337, Jul. 2012.

[51] A. J. Bass *et al.*, "Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion," *Nature Genet.*, vol. 43, pp. 964–968, Sep. 2011.

[52] B. L. Hogan, "Bone morphogenetic proteins: Multifunctional regulators of vertebrate development," *Genes Develop.*, vol. 10, pp. 1580–1594, Jul. 1996.

[53] A. Singh and R. J. Morris, "The Yin and Yang of bone morphogenetic proteins in cancer," *Cytokine Growth Factor Rev.*, vol. 21, pp. 299–313, Aug. 2010.

[54] T. B. Ro, R. U. Holt, A.-T. Brenne, H. Hjorth-Hansen, A. Waage, O. Hjertner, A. Sundan, and M. Borset, "Bone morphogenetic protein-5, -6 and -7 inhibit growth and induce apoptosis in human myeloma cells," *Oncogene*, vol. 23, pp. 3024–3032, Apr. 2004.

[55] I. K. Johnsen, R. Kappler, C. J. Auernhammer, and F. Beuschlein, "Bone morphogenetic proteins 2 and 5 are down-regulated in adrenocortical carcinoma and modulate adrenal cell proliferation and steroidogenesis," *Cancer Res.*, vol. 69, pp. 5784–5792, Jul. 2009.

[56] M. Romagnoli, K. Belguise, Z. Yu, X. Wang, E. Landesman-Bollag, D. C. Seldin, D. Chalbos, S. Barillé-Nion, P. Jézéquel, M. L. Seldin and G. E. Sonenshein, "Epithelial-to-mesenchymal transition induced by TGF-β1 is mediated by blimp-1–dependent repression of BMP-5," *Cancer Res*, vol. 72, pp. 6268–6278, Dec. 2012.

[57] S. Virtanen, E.-L. Alarmo, S. Sandström, M. Ampuja, and A. Kallioniemi, "Bone morphogenetic protein -4 and-5 in pancreatic cancer–novel bidirectional players," *Exp. Cell Res.*, vol. 317, pp. 2136–2146, Sep. 2011.

[58] E. Chen *et al.*, "Alteration of tumor suppressor BMP5 in sporadic colorectal cancer: A genomic and transcriptomic profiling based study," *Mol. Cancer*, vol. 17, Dec. 2018, Art. no. 176.

[59] E. Chen, Q. Li, H. Wang, P. Zhang, X. Zhao, F. Yang, J. Yang, "MiR-32 promotes tumorigenesis of colorectal cancer by targeting BMP5," *Biomed. Pharmacotherapy*, vol. 106, pp. 1046–1051, Oct. 2018.

[60] T. Koga, H. Iwasaki, M. Ishiguro, A. Matsuzaki, and M. Kikuchi, "Losses in chromosomes 17, 19, and 22q in neurofibromatosis type 1 and sporadic neurofibromas: A comparative genomic hybridization analysis," *Cancer Genet. Cytogenetics*, vol. 136, pp. 113–120, Jul. 2002.

[61] M. W. Stacey, J. Wang, R. L. Byrd, J. M. Liu, and W. G. Kearns, "Nuclear receptor co-repressor gene localizes to 17p11.2, a frequently deleted band in malignant disorders," *Genes, Chromosomes Cancer*, vol. 25, pp. 191–193, Jun. 1999.

[62] M. J. Cismowski and S. M. Lanier, "Activation of heterotrimeric G-proteins independent of a G-protein coupled receptor and the implications for signal processing," in *Reviews of Physiology Biochemistry and Pharmacology*, vol. 155. Berlin, Germany: Springer, 2005, pp. 57–80.

[63] G. Vaidyanathan, M. J. Cismowski, G. Wang, T. S. Vincent, K. D. Brown, and S. M. Lanier, "The RAS-related protein AGS1/RASD1 suppresses cell growth," *Oncogene*, vol. 23, pp. 5858–5863, Jul. 2004.

[64] G. S. Dalgin, D. T. Holloway, L. S. Liou, and C. DeLisi, "Identification and characterization of renal cell carcinoma gene markers," *Cancer Inf.*, vol. 3, pp. 65–92, Feb. 2007.

[65] S. Gao, L. Jin, G. Liu, P. Wang, Z. Sun, and Y. Cao, "Overexpression of RASD1 inhibits glioma cell migration/invasion and inactivates the AKT/mTOR signaling pathway," *Sci. Rep.*, vol. 7, p. 3202, Jun. 2017.

[66] H. Guan, G. Wei, J. Wu, D. Fang, Z. Liao, and H. Xiao, "Down-regulation of miR-218-2 and its host gene SLIT3 cooperate to promote invasion and progression of thyroid cancer," *J. Clin. Endocrinol. Metabolism*, vol. 98, pp. E1334–E1344, Aug. 2013.

[67] K. Nones *et al.*, "Genome-wide DNA methylation patterns in pancreatic ductal adenocarcinoma reveal epigenetic deregulation of SLIT-ROBO, ITGA2 and MET signaling," *Int. J. Cancer*, vol. 135, pp. 1110–1118, Sep. 2014.

[68] J. Tie, Y. Pan, L. Zhao, K. Wu, J. Liu, and S. Sun, "MiR-218 inhibits invasion and metastasis of gastric cancer by targeting the Robo1 receptor," *PLoS Genet*, vol. 6, p. e1000879, Mar. 2010.

[69] H. Yu, G. Gao, L. Jiang, L. Guo, M. Lin, X. Jiao, W. Jia, and J. Huang, "Decreased expression of miR-218 is associated with poor prognosis in patients with colorectal cancer," *Int. J. Clin. Expression Pathol.*, vol. 6, no. 12, pp. 2904–2911, 2013.

[70] W. Shi, C. Bastianutto, A. Li, B. Perez-Ordonez, R. Ng, K. Y. Chow, W. Zhang, I. Jurisica, K. W. Lo, A. Bayley, J. Kim, B. O'Sullivan, L. Siu, E. Chen, and F. F. Liu, "Multiple dysregulated pathways in nasopharyngeal carcinoma revealed by gene expression profiling," *Int. J. Cancer*, vol. 119, pp. 2467–2475, Nov. 2006.

[71] G. Narayan, C. Goparaju, H. Arias-Pulido, A. M. Kaufmann, A. Schneider, M. Dürst, M. Mansukhani, B. Pothuri, and V. V. Murty, ''Promoter hypermethylation-mediated inactivation of multiple Slit-Robo pathway genes in cervical cancer progression,'' *Mol. Cancer*, vol. 5, p. 16, May 2006.

[72] R. E. Dickinson, K. S. Fegan, X. Ren, S. G. Hillier, and W. C. Duncan, ''Glucocorticoid regulation of SLIT/ROBO tumour suppressor genes in the ovarian surface epithelium and ovarian cancer cells,'' *PLoS ONE*, vol. 6, no. 11, p. e27792, 2011.

[73] C. Zhang, H. Guo, B. Li, C. Sui, Y. Zhang, X. Xia, Y. Qin, L. Ye, F. Xie, H. Wang, M. Yuan, L. Yuan, and J. Ye, ''Effects of slit3 silencing on the invasive ability of lung carcinoma A549 cells,'' *Oncol. Rep.*, vol. 34, pp. 952–960, Aug. 2015.

[74] L. Ng, A. K. M. Chow, J. H. W. Man, T. C. C. Yau, T. M. H. Wan, and D. N. Iyer, ''Suppression of Slit3 induces tumor proliferation and chemoresistance in hepatocellular carcinoma through activation of GSK3β/β-catenin pathway,'' *BMC Cancer*, vol. 18, p. 621, Jun. 2018.

[75] R. Marlow, P. Strickland, J. S. Lee, X. Wu, M. Pebenito, M. Binnewies, E. K. Le, A. Moran, H. Macias, R. D. Cardiff, S. Sukumar, L. Hinck, ''SLITs suppress tumor growth *in vivo* by silencing SDF1/CXCR4 within breast epithelium,'' *Cancer Res.*, vol. 68, pp. 7819–7827, Oct. 2008.

[76] A. E. Denk, S. Braig, T. Schubert, and A. K. Bosserhoff, ''Slit3 inhibits activator protein 1-mediated migration of malignant melanoma cells,'' *Int. J. Mol. Med.*, vol. 28, pp. 721–726, Nov. 2011.

[77] K. Matsuda, K. Yoshida, Y. Taya, K. Nakamura, Y. Nakamura, and H. Arakawa, ''p53AIP1 regulates the mitochondrial apoptotic pathway,'' *Cancer Res.*, vol. 62, pp. 2883–2889, May 2002.

[78] K. Oda, H. Arakawa, T. Tanaka, K. Matsuda, C. Tanikawa, T. Mori, M. Nishimori, K. Tamai, T. Tokino, Y. Nakamura, and Y. Taya, ''*p53AIP1*, a potential mediator of p53-dependent apoptosis, and its regulation by Ser-46-phosphorylated p53,'' *Cell*, vol. 102, no. 6, pp. 849–862, Sep. 2000.

[79] S. I. Yamashita, Y. Masuda, N. Yoshida, H. Matsuzaki, T. Kurizaki, Y. Haga, S. Ikei, M. Miyawaki, Y. Kawano, M. Chujyo, and K. Kawahara, ''p53AIP1 expression can be a prognostic marker in non-small cell lung cancer,'' *Clin. Oncol.*, vol. 20, pp. 148–151, Mar. 2008.

[80] Y. Liang, S. Wang, and J. Liu, ''Overexpression of tumor protein p53-regulated apoptosis-inducing protein 1 regulates proliferation and apoptosis of breast cancer cells through the PI3K/Akt pathway,'' *J Breast Cancer*, vol. 22, pp. 172–184, Jun. 2019.

[81] X. Wang, F. Wang, K. Taniguchi, R. S. Seelan, L. Wang, and K. E. Zarfas, ''Truncating variants in *p53AIP1* disrupting DNA damage–induced apoptosis are associated with prostate cancer risk,'' *Cancer Res.*, vol. 66, pp. 10302–10307, Nov. 2006.

[82] Y. Jiang, H. Chen, H. Jia, Y. Xu, G. Liu, Y. Wang, X. Yang, and Y. Lu, ''Adenovirus Ad-p53AIP1-mediated gene therapy and its regulation of p53-MDM2 interactions,'' *Expression Therapy Med.*, vol. 1, pp. 363–368, Mar. 2010.

[83] Y. Liu, M. W. Mayo, A. S. Nagji, E. H. Hall, L. S. Shock, and A. Xiao, ''BRMS1 suppresses lung cancer metastases through an E3 ligase function on histone acetyltransferase p300,'' *Cancer Res.*, vol. 73, pp. 1308–1317, Feb. 2013.

[84] J. Li, Y. Cheng, D. Tai, M. Martinka, D. R. Welch, and G. Li, ''Prognostic significance of BRMS1 expression in human melanoma and its role in tumor angiogenesis,'' *Oncogene*, vol. 30, pp. 896–906, Feb. 2011.

[85] R. X. Cui, N. Liu, Q. M. He, W. F. Li, B. J. Huang, and Y. Sun, ''Low BRMS1 expression promotes nasopharyngeal carcinoma metastasis *in vitro* and *in vivo* and is associated with poor patient survival,'' *BMC Cancer*, vol. 12, p. 376, Aug. 2012.

[86] P. A. Phadke, K. S. Vaidya, K. T. Nash, D. R. Hurst, and D. R. Welch, ''BRMS1 suppresses breast cancer experimental metastasis to multiple organs by inhibiting several steps of the metastatic process,'' *Amer. J. Pathol.*, vol. 172, pp. 809–817, Mar. 2008.

[87] R. Koyama, M. Tamura, T. Nakagaki, T. Ohashi, M. Idogawa, H. Suzuki, T. Tokino, and Y. Sasaki, ''Identification and characterization of a metastatic suppressor BRMS1L as a target gene of p53,'' *Cancer Sci.*, vol. 108, pp. 2413–2421, Dec. 2017.

[88] C. Gong, S. Qu, X. B. Lv, B. Liu, W. Tan, Y. Nie, F. Su, Q. Liu, H. Yao, and E. Song, ''BRMS1L suppresses breast cancer metastasis by inducing epigenetic silence of FZD10,'' *Nat. Commun.*, vol. 5, p. 5406, Nov. 2014.

[89] P. Cao, S. Zhao, Z. Sun, N. Jiang, Y. Shang, Y. Wang, J. Gu, and S. Li, ''BRMS1L suppresses ovarian cancer metastasis via inhibition of the β-catenin-wnt pathway,'' *Exp. Cell Res.*, vol. 371, pp. 214–221, Oct. 2018.

[90] J. Hao, X. Jin, Y. Shi, and H. Zhang, ''miR-93-5p enhance lacrimal gland adenoid cystic carcinoma cell tumorigenesis by targeting BRMS1L,'' *Cancer Cell Int.*, vol. 18, p. 72, May 2018.

[91] I. T. Chiang, W. S. Wang, H. C. Liu, S. T. Yang, N. Y. Tang, and J. G. Chung, ''Curcumin alters gene expression-associated DNA damage, cell cycle, cell survival and cell migration and invasion in NCI-H460 human lung cancer cells *in vitro*,'' *Oncol. Rep.*, vol. 34, pp. 1853–1874, Oct. 2015.

[92] Y. Hu, Q. Zhang, J. Cui, Z. J. Liao, M. Jiao, Y. B. Zhang, Y.-H. Guo, and Y.-M. Gao, ''Oncogene miR-934 promotes ovarian cancer cell proliferation and inhibits cell apoptosis through targeting BRMS1L,'' *Eur. Rev. Med. Pharmacol. Sci.*, vol. 23, pp. 5595–5602, Jul. 2019.

[93] J. Lv, H. Yang, X. Wang, R. He, L. Ding, and X. Sun, ''Decreased BRMS1L expression is correlated with glioma grade and predicts poor survival in glioblastoma via an invasive phenotype,'' *Cancer Biomark*, vol. 22, no. 2, pp. 311–316, 2018.

[94] S. Kang, S. Elf, K. Lythgoe, T. Hitosugi, J. Taunton, W. Zhou, L. Xiong, D. Wang, S. Muller, S. Fan, S.-Y. Sun, A. I. Marcus, T.-L. Gu, R. D. Polakiewicz, Z. Chen, F. R. Khuri, D. M. Shin, and J. Chen, ''p90 ribosomal S6 kinase 2 promotes invasion and metastasis of human head and neck squamous cell carcinoma cells,'' *J. Clin. Invest*, vol. 120, no. 4, pp. 1165–1177, Apr. 2010.

[95] H. C. Lim, L. Xie, W. Zhang, R. Li, Z. C. Chen, and G. Z. Wu, ''Ribosomal S6 Kinase 2 (RSK2) maintains genomic stability by activating the Atm/p53-dependent DNA damage pathway,'' *PLoS ONE*, vol. 8, no. 9, p. e74334, 2013.

[96] K. Yao, C. Peng, Y. Zhang, T. A. Zykova, M. H. Lee, and S. Y. Lee, ''RSK2 phosphorylates T-bet to attenuate colon cancer metastasis and growth,'' *Proc. Nat. Acad. Sci. USA*, vol. 114, pp. 12791–12796, Nov. 2017.

[97] G. Amaddeo, C. Guichard, S. Imbeaud, and J. Zucman-Rossi, ''Next-generation sequencing identified new oncogenes and tumor suppressor genes in human hepatic tumors,'' *Oncoimmunology*, vol. 1, pp. 1612–1613, Dec. 2012.

[98] M. Yamashita, E. Nitta, and T. Suda, ''Regulation of hematopoietic stem cell integrity through p53 and its related factors,'' *Ann. NY Acad. Sci.*, vol. 1370, no. 1, pp. 45–54, Apr. 2016.

[99] M. Yamashita, E. Nitta, and T. Suda, ''[Maintenance of hematopoietic stem cell integrity and regulation of leukemogenesis by p53 and its coactivator Aspp1],'' *Rinsho Ketsueki*, vol. 56, pp. 2426–2433, Dec. 2015.

[100] J. Zak and X. Lu, ''ASPP1: A guardian of hematopoietic stem cell integrity,'' *Cell Stem Cell*, vol. 17, no. 1, pp. 3–5, Jul. 2015.

[101] G. Wang, C. H. Huang, Y. Zhao, L. Cai, Y. Wang, S. J. Xiu, Z. W. Jiang, S. Yang, X. T. Zhao, W. Huang, and J. R. Gu, ''Genetic aberration in primary hepatocellular carcinoma: Correlation between p53 gene mutation and loss-of-hetero- zygosity on chromosome 16q21-q23 and 9p21-p23,'' *Cell Res.*, vol. 10, pp. 311–323, Dec. 2000.

[102] P. J. Hainsworth, K. L. Raphael, R. G. Stillwell, R. C. Bennett, and O. M. Garson, ''Cytogenetic features of twenty-six primary breast cancers,'' *Cancer Genet. Cytogenetic*, vol. 53, no. 2, pp. 205–218, Jun. 1991.

[103] I. I. C. Chio, M. Sasaki, D. Ghazarian, J. Moreno, S. Done, T. Ueda, S. Inoue, Y.-L. Chang, N. J. Chen, and T. W. Mak, ''TRADD contributes to tumour suppression by regulating ULF-dependent p19^{Arf} ubiquitylation,'' *Nat. Cell Biol.*, vol. 14, pp. 625–633, May 2012.

[104] R. C. Tseng, S. H. Lee, H. S. Hsu, B. H. Chen, W. C. Tsai, and C. Tzao, ''SLIT2 attenuation during lung cancer progression deregulates beta-catenin and E-cadherin and associates with poor prognosis,'' *Cancer Res.*, vol. 70, pp. 543–551, Jan. 2010.

[105] C. Alvarez, T. Tapia, V. Cornejo, W. Fernandez, A. Munoz, and M. Camus, ''Silencing of tumor suppressor genes *RASSF1A*, *SLIT2*, and *WIF1* by promoter hypermethylation in hereditary breast cancer,'' *Mol. Carcinog*, vol. 52, pp. 475–487, Jun. 2013.

[106] A. Dallol, D. Morton, E. R. Maher, and F. Latif, ''SLIT2 axon guidance molecule is frequently inactivated in colorectal cancer and suppresses growth of colorectal carcinoma cells,'' *Cancer Res.*, vol. 63, pp. 1054–1058, Mar. 2003.

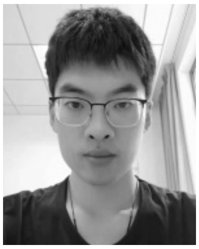
[107] R. Dong, J. Yu, H. Pu, Z. Zhang, and X. Xu, ''Frequent *SLIT2* promoter methylation in the serum of patients with ovarian cancer,'' *J. Int. Med. Res.*, vol. 40, no. 2, pp. 681–686, 2012.

[108] J. Jin, H. You, B. Yu, Y. Deng, N. Tang, and G. Yao, "Epigenetic inactivation of *SLIT2* in human hepatocellular carcinomas," *Biochem. Biophys. Res. Commun.*, vol. 379, no. 1, pp. 86–91, Jan. 2009.

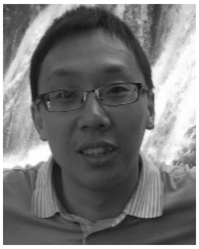
[109] F. J. Sulzmaier and J. W. Ramos, "RSK isoforms in cancer cell invasion and metastasis," *Cancer Res.*, vol. 73, pp. 6099–6105, Oct. 2013.

[110] D. E. Clark, T. M. Errington, J. A. Smith, H. F. Frierson, M. J. Weber, and D. A. Lannigan, "The serine/threonine protein kinase, p90 ribosomal S6 kinase, is an important regulator of prostate cancer cell proliferation," *Cancer Res.*, vol. 65, pp. 3108–3116, Apr. 2005.

[111] G. Yu, Y. C. Lee, C. J. Cheng, C. F. Wu, J. H. Song, G. E. Gallick, L.-Y. Yu-Lee, J. Kuang, and S.-H. Lin, "RSK promotes prostate cancer progression in bone through ING3, CKAP2, and PTK6-mediated cell survival," *Mol. Cancer Res.*, vol. 13, pp. 348–357, Feb. 2015.



RAN ZHAO received the B.S. degree in software engineering from Tianjin Polytechnic University, Tianjin, China, in 2017. He is currently pursuing the M.S. degree with Shanghai Maritime University. In 2018, he moved to Shanghai Maritime University to study computer application technology. His current interests include bioinformatics, computational biology, and computer vision.



LEI CHEN received the B.S. and M.S. degrees in operational researches from East China Normal University, in 2004 and 2007, respectively, and the Ph.D. degree in system analysis and integration from East China Normal University, in 2010. He studied mathematics with East China Normal University, from 2000 to 2007. In 2007, he moved to the Software Engineering Institute, East China Normal University, to study computer science.

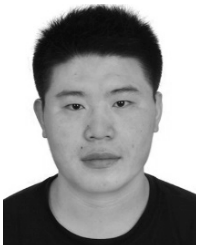
In 2010, he joined the College of Information Engineering, Shanghai Maritime University, where he was an Associate Professor. His interests include bioinformatics, computational biology, graph theory, and algorithm design.

Dr. Chen is a member of the China Computer Federation and the Chinese Association for Artificial Intelligence. He is the Editorial Board Member of *Current Bioinformatics* and *Current Proteomics* and the Section Editor of *Combinatorial Chemistry & High Throughput Screening*.

BO ZHOU received the B.S. degree in clinical medicine and the M.S. degree in genetics from the Huazhong University of Science and Technology, in 2001 and 2004, respectively, and the Ph.D. degree in oncology from the Second Military Medical University, in 2012. She studied at the Huazhong University of Science and Technology, from 1996 to 2004. In 2016, she joined Shanghai University of Medicine and Health Sciences as a Lecturer. Her interests include bioinformatics and clinical trials of new targeted drugs in colon cancer.

ZI-HAN GUO received the B.S. degree from Tianjin Polytechnic University, in 2017. He is currently pursuing the M.S. degree with Shanghai Maritime University. He studied computer science and technology at Tianjin Polytechnic University, from 2013 to 2017. In 2017, he moved to Shanghai Maritime University to study computer application technology. His interests include bioinformatics, computer vision, and natural language processing.

SHUAIQUN WANG received the B.S. degree from Inner Mongolia Normal University, Huhehaote, China, in 2007, the M.S. degree from Shanghai Maritime University, Shanghai, China, in 2010, and the D.E. degree from Tongji University, Shanghai, China, in 2015. She is currently with Shanghai Maritime University. Her current research interests include bioinformatics, intelligent optimization, and algorithm design.



AORIGELE received the B.S. degree in computer science and technology from Inner Mongolia Normal University, Huhehaote, China, in 2007, and the D.E. degree from Toyama University, Toyama, Japan, in 2016. His interests include bioinformatics and optimization problems.