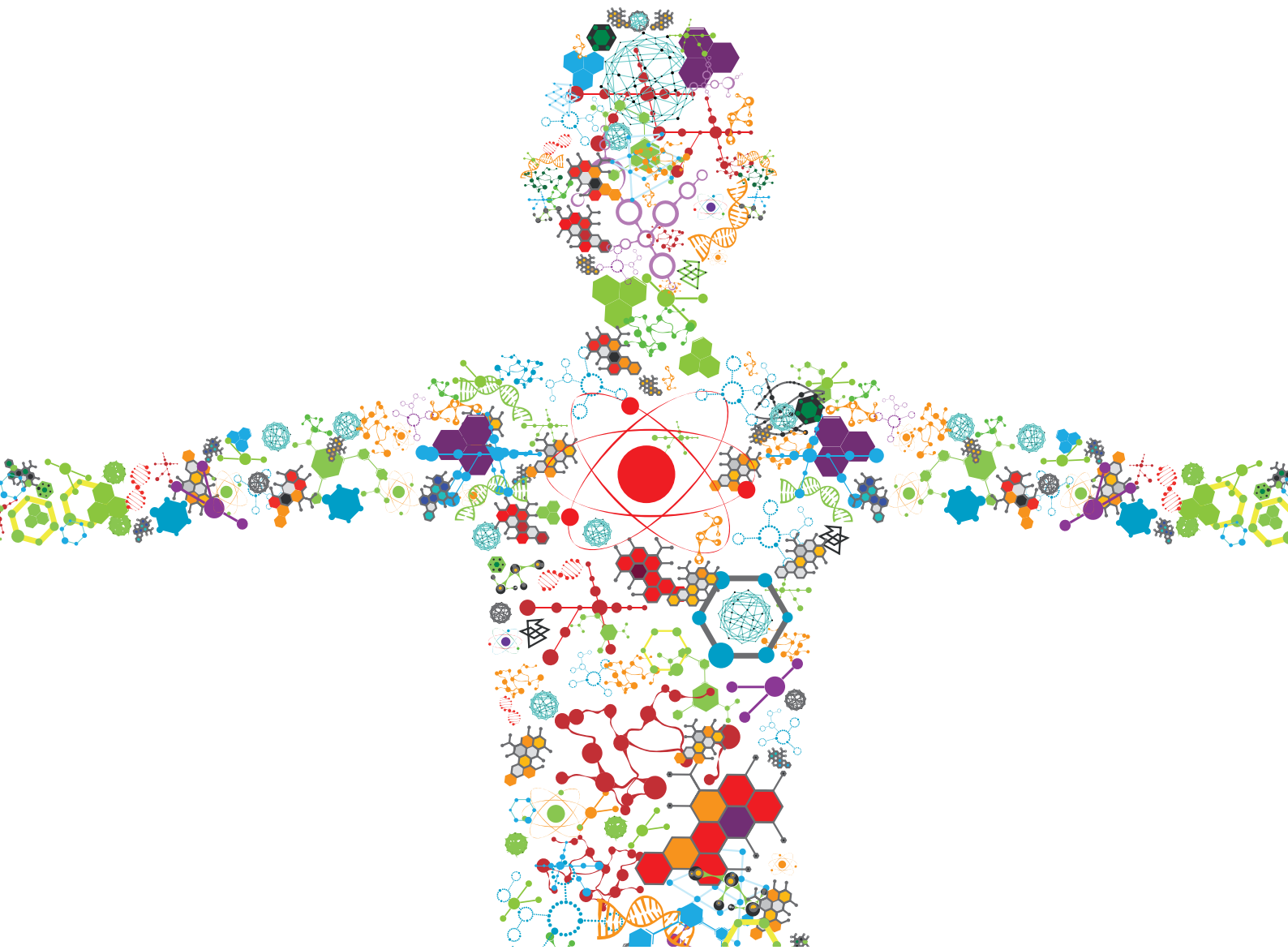
EXPLAINABLE INTELLIGENT PROCESSING OF BIOLOGICAL RESOURCES INTEGRATING DATA, INFORMATION, KNOWLEDGE, AND WISDOM

EDITED BY: Yucong Duan and Yungang Xu
PUBLISHED IN: Frontiers in Bioengineering and Biotechnology,
Frontiers in Genetics and Frontiers in Plant Science



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-136-6
DOI 10.3389/978-2-88974-136-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

EXPLAINABLE INTELLIGENT PROCESSING OF BIOLOGICAL RESOURCES INTEGRATING DATA, INFORMATION, KNOWLEDGE, AND WISDOM

Topic Editors:
Yucong Duan, Hainan University, China
Yungang Xu, Xi'an Jiaotong University, China

Citation: Duan, Y., Xu, Y., eds. (2022). Explainable Intelligent Processing of Biological Resources Integrating Data, Information, Knowledge, and Wisdom. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-136-6

Table of Contents

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 24 March 2020
doi: 10.3389/fbioe.2020.00183

Predicting Bacteriophage Enzymes and Hydrolases by Using Combined Features

*Hong-Fei Li[1,2], Xian-Fang Wang[2] and Hua Tang[1]**

[1] Department of Pathophysiology, Key Laboratory of Medical Electrophysiology, Ministry of Education, Southwest Medical University, Luzhou, China, [2] School of Computer and Information Engineering, Henan Normal University, Henan, China

Bacteriophage is a type of virus that could infect the host bacteria. They have been applied in the treatment of pathogenic bacterial infection. Phage enzymes and hydrolases play the most important role in the destruction of bacterial cells. Correctly identifying the hydrolases coded by phage is not only beneficial to their function study, but also conducive to antibacteria drug discovery. Thus, this work aims to recognize the enzymes and hydrolases in phage. A combination of different features was used to represent samples of phage and hydrolase. A feature selection technique called analysis of variance was developed to optimize features. The classification was performed by using support vector machine (SVM). The prediction process includes two steps. The first step is to identify phage enzymes. The second step is to determine whether a phage enzyme is hydrolase or not. The jackknife cross-validated results showed that our method could produce overall accuracies of 85.1 and 94.3%, respectively, for the two predictions, demonstrating that the proposed method is promising.

Keywords: bacteriophage enzymes, hydrolase, analysis of variance, sequence feature, classification

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Balachandran Manavalan,
Ajou University, South Korea
Yongchun Zuo,
Inner Mongolia University, China

***Correspondence:**
Hua Tang
huatang@swmu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *02 February 2020*
Accepted: *24 February 2020*
Published: *24 March 2020*

Citation:
Li H-F, Wang X-F and Tang H (2020) Predicting Bacteriophage Enzymes and Hydrolases by Using Combined Features. Front. Bioeng. Biotechnol. 8:183. doi: 10.3389/fbioe.2020.00183

INTRODUCTION

Bacteriophage, as safe agent, can lyse and infect specific bacteria without destroying natural beneficial microflora (Parmar et al., 2018). Hydrolytic enzymes encoded by phages are key ingredients of lysis, which is helpful to fighting bacterial pathogens, especially those that cannot be killed by antibiotics and chemicals. In fact, in some countries, they have been used therapeutically to treat bacterial infections that do not respond to antibiotics (Thiel, 2004; Parfitt, 2005; Keen, 2012). They have also been used as a food safety tool to reduce bacterial contamination (Pirisi, 2000). Hence, rapid detection of bacteriophage and hydrolase responsible for antibacterial drugs is a growing necessity for public health.

Because of abuse of antibiotics, certain resistant viruses cannot be effectively controlled. This problem can be resolved by therapy of phage hydrolytic that disintegrates host viruses during releasing progeny phage. Therefore, the identification of hydrolases encoded by phages has become an important research topic. It not only has been studied in chemistry and physics through

Abbreviations: CTD, composition transition and distribution; SVM, support vector machine; RF, random forest; MLP, multilayer perceptron; KNN, k-nearest neighbors; Sn, sensitivity; Sp, specificity; Ac, accuracy; MCC, matthew correlation coefficient; PseAAC, pseudo–amino acid composition; GTPC, grouped tripeptide composition; ROC, receiver operating characteristic; AUC, area under receiver operating characteristic (ROC) curve; GGDC, g-gap dipeptide composition; ANOVA, analysis of variance; RBF, Radial Basis Function; ORFs, Open Reading Frames.

experimental methods, but also achieved good results in theory through recently popular machine learning algorithms. Some experiments have been performed to study the function of phage hydrolase (Kimura and Itoh, 2003; Rodriguez-Rubio et al., 2013). In addition, in the study of host cell lysis by hydrolytic enzyme activation, Kovalenko et al. (2019) found that the calcium could regulate phage-induced bacterial lysis. Although those biochemical-based methods can accurately recognize phage hydrolases and clearly elucidate the functional mechanism of the enzyme, it is time-consuming and expensive. Additionally, biochemical experiments always need rigorous experimental conditions, which will prevent most of scholars from doing more in-depth studies. Computational methods provide another chance to study phage hydrolase without the disadvantage of biochemical-based methods. Phylogenetic analysis or similarity search could find relative conservation of motifs among related species (Lin and Li, 2011; Liu et al., 2019). However, it is extremely diverse for phage Open Reading Frames (ORFs), of which more than 70% of them cannot find out similar genes with annotated functions in GenBank (Seguritan et al., 2012). Moreover, it is also time-consuming.

With the accumulation of more and more postgenomic data, some computational methods have been proposed to study the function of phage proteins. Riede and his colleagues (Riede et al., 1987) have proposed a model to predict tail-fiber proteins' three-dimensional structure of T-even–type phages. The results are consistent with electron microscopic data. Subsequently, a computer program was developed to identify DNA-binding regulatory proteins in bacteriophage T7 (White, 1987; Song et al., 2014; Zou et al., 2016a; Qu et al., 2019). Recently, the virion proteins encoded by phages were studied by using naive Bayes combined with primary sequence information (Feng et al., 2013). The proposed model could yield the overall accuracy (Ac) of 79.15%. By using feature selection technique, the overall Ac was improved to 85.02% (Ding et al., 2014). A free webserver called PVPred (Ding et al., 2014) was constructed for predicting phage virion proteins.

The success of previous works on the prediction of phage functional proteins (Feng et al., 2013; Ding et al., 2014) and enzyme prediction (Zuo et al., 2014; Ding H. et al., 2016) provided good strategy to discriminate hydrolases encoded by phages by transforming protein sequences into digital features and further establishing machine learning-based models. Thus, this work aims to develop a powerful computational model to recognize phage hydrolase by combining feature selection and expression of multiple features. The entire experiment was divided into two steps. First is to discriminate phage enzymes from phage nonenzymes and then to identify phage hydrolases from phage enzymes. In this model, the support vector machine (SVM) was applied as the algorithm to perform the classification. Different features were proposed to formulate protein samples and then inputted into SVM. The best features that can achieve the maximum accuracies were discovered by using analysis of variance (ANOVA). The model's performance was estimated by using jackknife cross-validation.

MATERIALS AND METHODS

Benchmark Dataset

Constructing a reliable benchmark dataset could guarantee the reliability of the proposed computational model (Ma et al., 2014; Liang et al., 2017; Yang et al., 2017; Wang et al., 2018; Cheng et al., 2019; Hu et al., 2019; Zheng et al., 2019). In this work, samples were gained from Ding H. et al. (2016), which were rigorously screened through the following three steps: (1) phage proteins have been annotated by standard operating procedure for UniProt manual curation (Swiss-Prot); (2) protein sequences samples containing illegal characters were deleted; (3) sequence identity in the dataset must be less than 30%, which was implemented by CD-HIT (Fu et al., 2012) software. Consequently, the definitive benchmark dataset contains 255 phage proteins, of which 124 proteins belong to phage enzymes (positive samples of set 1), and the remaining 131 are phage nonenzymes (negative samples of set 1). Furthermore, 124 phage enzymes are divided into 69 hydrolases (positive samples of set 2) and 55 nonhydrolases (negative samples of set 2), respectively. The following calculations are all based on these data.

Protein Feature Extraction

The perfect expression of protein sequences by digital features can dramatically increase the Ac and robust of computing models (Wang et al., 2008, 2010; Song et al., 2010, 2018; Zuo et al., 2017; Basith et al., 2018; Chen W. et al., 2018; Wei et al., 2018b; Boopathi et al., 2019; Ding et al., 2019; Manavalan et al., 2019b; Shen et al., 2019; Tan et al., 2019; Zhang and Liu, 2019; Zhu et al., 2019). The specific order of residues in the peptide sequence dictates the protein to fold up into a special three-dimensional structure. Thus, the interaction between two residues in a protein is a main factor to characterize the protein. In the past 20 years, scholars have developed dipeptide composition to formulate peptide samples (Tang et al., 2016). However, the feature can only describe the short-range interaction between two residues. In fact, there are lots of long-range interaction for a protein in three-dimensional space. For example, the secondary structures (α helix and β sheet) were formed by the interaction of two nonadjoining residues. Hence, it will be more reasonable to investigate the performance of other kinds of correlations.

Based on the above analysis and other peer works (Ding and Li, 2015), in this work, the g-gap dipeptide composition (GGDC), which is extended from general dipeptide composition, is used as the main feature to denote the residues' correlation in the original peptide sequence. For the perfect expression of the sample, the combination of GGDC, pseudo–amino acid composition (PseAAC), grouped tripeptide composition (GTPC), and composition transition and distribution (CTD) is used as the final feature vector. Pseudo–amino acid composition provides the correlation of physical and chemical properties between two residues (Chen et al., 2016; Yang et al., 2016). Grouped tripeptide composition provides tripeptide information (Tan et al., 2019). CTD provides distribution patterns of a specific structural property for residues (Cheng et al., 2018) and indirectly

contains information about 20 amino acid residues, so PseAAC, in our work, does not contain amino acid information.

G-Gap Dipeptide Composition

The GGDC proposed by Ding et al. (2014) is the extension of the proximate dipeptide composition, because proteins contain deep correlation of residues relating with hydrogen bonding in secondary structure. For different g, the protein sequence P with L residues is expressed by a 400-dimensional GGDC as follows:

$$P = [f_1^g, f_2^g, ..., f_\varepsilon^g, ..., f_{400}^g]^T \tag{1}$$

where T is called the transposing operator, the f_ε^g can be calculated by:

$$f_\varepsilon^g = n_\varepsilon^g/(L - g - 1) \tag{2}$$

where the n_ε^g denotes the absolute occurrence number of the GGDC in a protein. Since previous studies (Ding H. et al., 2016) have shown that $g = 2$ has the best prediction effect, only 2-gap was used in our experiments.

Pseudo–Amino Acid Composition

Hydrophobicity, hydrophilicity, and other physicochemical properties are important characteristics of amino acids. In order to incorporate these properties with amino acid composition, two types of PseAAC were used. In our work, motived by PseAAC, the protein sample, can be expressed as follows:

$$\begin{cases} \tau_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{k,k+1}^1 \\ \tau_2 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{k,k+1}^2 \\ \cdots \\ \tau_n = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{k,k+1}^n \\ \tau_{n+1} = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{k,k+2}^1, (l < L) \\ \tau_{n+2} = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{k,k+2}^2 \\ \cdots \\ \tau_{\lambda n} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{k,k+\lambda}^n \end{cases} \tag{3}$$

$H_{k,k+\lambda}^n$ th residue and the $(k+\lambda)$-th residue; L is length of sample. After experimental comparison, we selected 10 physical and chemical properties containing hydrophobicity, hydrophilicity, amino acid side chain group mass, -COOH group dissociation constant, -NH3 group dissociation constant, isoelectric point at 25°C, rigidity, flexibility, irreplaceability, and polarity. We used $\lambda = 15$.

GTPC and CTD

iFeature is a comprehensive Python-based toolkit that contains four major functions: feature representation, dimensionality reduction algorithms, feature selection algorithms, and feature clustering algorithms (Chen Z. et al., 2018). In our study, we have used GTPC and CTD provided by iFeature (Chen Z. et al., 2018) to extract numerical descriptors from samples. Grouped tripeptide composition converts protein sequences into 125-dimensional digital features expressed as follows:

$$f(r, s, t) = \frac{N_{rst}}{N-1}, \quad r, s \in \{g1, g2, g3, g4, g5\} \tag{4}$$

where N_{rst} denotes the number of tripeptides in groups r, s, and t (Chen Z. et al., 2018). N is the length of a protein.

CTD converts protein sequences into 39-dimensional digital features defined as follows:

$$C(r) = \frac{N_r}{N}, \quad r\{polar, neutral, hydrophoic\} \tag{5}$$

where $N(r)$ represents the number of residue type r in the peptide sequence (Chen Z. et al., 2018). Thus, samples are transformed into 164 -dimensional features.

Support Vector Machine

Support vector machine is a classical machine learning algorithm and has been widely adopted in computational biology (Jiang et al., 2013; Zhao et al., 2015, 2017; Ding H. et al., 2016; Ding et al., 2016a,b; Dao et al., 2018; Feng et al., 2018; Manavalan et al., 2018a,b; Zhang et al., 2018; Chao et al., 2019; Chen et al., 2019a; Wang et al., 2019; Basith et al., 2020). For nonlinear samples, its projects inputted data into high-dimensional spare by a kernel function. There are four kernel functions including Sigmoid function, Gaussian function, line function, and polynomial function, among which Gaussian function is most commonly used. C and g are the most important parameters to adjust performance of Gaussian function. The value of g is related to the partitioning of samples, and the value of C determines the tolerance of the model. In our work, SVC functions in Scikit-learn (Swami and Jain, 2012), based on Python, are used to build models, and Gaussian functions are used as kernel functions, because the Gaussian function can efficiently map small samples with fewer features to high-dimensional space and distinguish positive and negative samples with high Ac. In addition, the GridSearchCV function in Scikit-learn was used to optimize the parameters C and g.

Feature Selection Method

Because one type of feature does not fully represent the characteristics of a protein sequence, the combination of features is a good approach to perform classifications. The combined features could also cause a lot of inconvenience, such as noise, dimension disaster, and so on. Analysis of variance (Feng et al., 2013; Tang et al., 2017; Xianfang et al., 2019), principal component analysis (Dong et al., 2015), minimal redundancy maximal relevance (Ding et al., 2013), maximum relevance maximum distance (Zou et al., 2016b), and increment of diversity (Zuo and Li, 2009; Zhao et al., 2010; Fan and Li, 2012) can solve these problems. In our study, ANOVA is used to screen the best feature set; the idea is to calculate the ratio of the categories to sample variance. Obviously, features with larger ratios are more suitable for classification. The details can be referred from Feng et al. (2013), Tang et al. (2017) and Xianfang et al. (2019).

Performance Evaluation

In statistical prediction, the performance of the model needs to be measured by some methods and parameters (Chen et al., 2017, 2019b; Ding et al., 2017; Tang et al., 2018; Yang et al., 2018). The cross-validation test has been widely used to evaluate methods (Yang et al., 2019; Zhu et al., 2019). To provide a fair comparison, we used the jackknife test in this study. The four parameters, namely, sensitivity (Sn), specificity (Sp), Ac, and Matthew correlation coefficient (MCC), are used to evaluate the performance of the model (Liu et al., 2018; Manavalan et al., 2018c, 2019a,c; Basith et al., 2019), which are defined as follows:

$$\begin{cases} S_n = \frac{TP}{TP+FN} \\ S_p = \frac{TP}{TP+FN} \\ A_c = \frac{TP+FN}{TP+FP+TN+FN} \\ MCC = \frac{(TP\times TN)+(FP\times FN)}{(TP+FN)(TN+FP)(TP+FP)(TN+FN)} \end{cases} \quad (6)$$

where TP and TN are the number of the correctly identified positive samples and the number of the correctly identified negative samples; FP indicates the number of negative samples recognized as positive samples; FN indicates the number of positive samples recognized as negative samples. Also, the area under receiver operating characteristic (ROC) curve (AUC) is often used to evaluate the performance of binary classification models.

RESULTS

Discriminating Phage Enzymes From Nonenzymes

For a new sequenced phage protein, we first need to judge whether the phage protein is an enzyme. Thus, the predictive performances of three combined vectors were investigated by using SVM with jackknife test. First, samples are expressed by three kinds of combinations: GGDC combined with PseAAC, GTPC combined with CTD, and all features. Prediction results are listed in **Table 1**. We observed that all features cannot achieve the best Ac. The reason is maybe noise or redundant information. Thus, we performed feature selection for three feature combinations to discover the best feature subsets. The results are also shown in **Table 1**. After feature selection, the highest Ac was obtained by using 191 features, which was based on all features. **Figure 1A** was drawn to show the *F*-value for all features. The above results implied that the information of phage enzymes requires multiple types of

TABLE 1 | The results by using different features for phage enzymes prediction.

Combined vector features	Original feature		Optimal features	
	Accuracy	Dimensions	Accuracy	Dimensions
GGDC + PseAAC	74.5%	550	83.1%	154
GTPC + CTD	67.8%	164	77.6%	35
GGDC + PseAAC + GTPC + CTD	72.9%	714	85.1%	191

GGDC, g-gap dipeptide composition; CTD, composition transition and distribution; PseAAC, pseudo–amino acid composition; GTPC, grouped tripeptide composition.

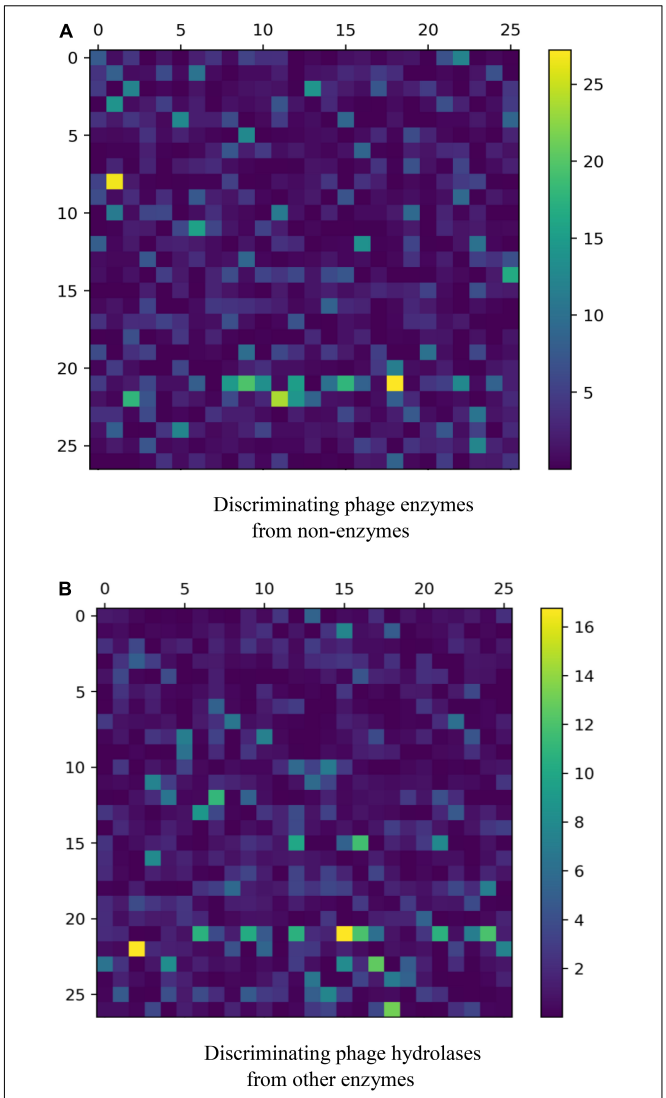


FIGURE 1 | A plot showing the *F*-values for **(A)** discriminating phage enzymes from nonenzymes and **(B)** discriminating phage hydrolases from other enzymes.

TABLE 2 | The comparison of different classifiers for predicting phage enzymes.

Classifier	Sn	Sp	Ac	MCC	AUC
KNN	0.98	0.16	0.702	0.232	0.664
RF	0.73	0.76	0.752	0.490	0.798
SVM	0.83	0.88	0.851	0.703	0.897
MLP	0.77	0.84	0.812	0.610	0.858

SVM, support vector machine; RF, random forest; MLP, multilayer perceptron; KNN, k-nearest neighbors; Sn, sensitivity; Sp, specificity; Ac, accuracy; MCC, Matthew correlation coefficient; AUC, area under receiver operating characteristic (ROC) curve.

TABLE 3 | The results by using different feature for discriminating phage hydrolases from other enzymes.

Combined vector features	Original features		Optimal features	
	Accuracy	Dimensions	Accuracy	Dimensions
GGDC + PseAAC	75.8	550	94.3%	61
GTPC + CTD	76.6%	164	86.4%	37
GGDC + PseAAC + GTPC + CTD	75.8%	714	92.7%	89

GGDC, g-gap dipeptide composition; CTD, composition transition and distribution; PseAAC, pseudo–amino acid composition; GTPC, grouped tripeptide composition.

TABLE 4 | The comparison of different classifiers for discriminating phage hydrolases from other enzymes.

Classifier	Sn	Sp	Ac	MCC	AUC
KNN	0.70	0.89	0.814	0.588	0.863
RF	0.91	0.80	0.86	0.722	0.898
SVM	0.96	0.93	0.943	0.886	0.961
MLP	0.93	0.91	0.927	0.837	0.948

SVM, support vector machine; RF, random forest; MLP, multilayer perceptron; KNN, k-nearest neighbors; Sn, sensitivity; Sp, specificity; Ac, accuracy; MCC, Matthew correlation coefficient; AUC, area under receiver operating characteristic (ROC) curve.

feature expressions. However, noises or redundant information may be results in the poor predictive capabilities of other groups, and the combining vectors of the first and second groups cannot fully express the peculiarity of the samples, which lead to its poor prediction effect. Subsequently, we investigated the performance of four classifiers, including random forest (RF), multilayer perceptron (MLP), k-nearest neighbor (KNN), and SVM, whose input features are the third set of 191-D optimal features. The result parameters of four classifiers have been exhibited in **Table 2**. We found the highest Ac of 85.1% and MCC of 70.3%. The AUC reaches to 89.3% by using SVM. k-Nearest neighbor has achieved the highest Sn of 98% with the lowest Sp of 16%. Moreover, performance of RF has an Sn of 73%, Sp of 76%, Ac of 75.2%, MCC of 0.490, and AUC of 0.798, respectively. Similarity, MLP obtained 77, 84, 81.2, 0.61, and 0.858%, respectively, for Sn, Sp, Ac, MCC, and AUC. These data indicate that SVM is the most suitable for distinguishing phage enzymes.

Discriminating Phage Hydrolases From Other Enzymes

When a phage protein is predicted as a phage enzyme, it is necessary to immediately judge whether the enzyme is a hydrolase. Like phage enzyme prediction, the performances of three combined vectors on phage hydrolase prediction were also examined by using SVM with jackknife cross-validation. As shown in **Table 3**, the three combined vectors were also processed by the feature selection algorithm, which not only improves the Ac but also greatly reduces the dimensions. Obviously, ANOVA can remove redundant information from features. It should be noticed that the optimal features (61-D) obtained from GGDC combined with PseAAC could produce the maximum Ac of 94.3%. This phenomenon indicates that features with a large *F*-value in the second group are not suitable for expressing hydrolases. The heat map for the features is also drawn in **Figure 1B**. Similarly, we compared the performances of different classifiers. In **Table 4**, KNN has yielded Ac of 81.4%, whereas KNN has obtained Ac of 84.64%. The performance of MLP is 93% Sn, 91% Sp, 92.7% Ac, 0.837 MCC, and 0.948 AUC. Support vector machine with Radial Basis Function (RBF) as kernel function gained the best prediction performance (94.3% Ac).

Performance Comparison With Existing Methods

In order to prove that our proposed model performs better than the model by Ding H. et al. (2016), who first used computational methods to predict hydrolases, the performance indexes of the two models were recorded in **Table 5**. In discriminating phage enzymes from nonenzymes, our model is better in Ac and Sp that are 85.1 and 88.0%, respectively. In discriminating phage hydrolases from other enzymes, all the evaluated indexes of our proposed model are better than those of Ding H. et al. (2016). Indeed, hydrolyzing enzymes adopt two types of features to encode samples. Compared with Ding and colleagues' experiment, we have selected more kinds of features in the sample expression, which makes the digital features of the sample more informative.

DISCUSSION

The purpose of this study is to establish a predictive model to predict phage enzymes and hydrolases. In fact, similarity search could be used to perform sequence analysis and function

TABLE 5 | Comparison of predictive performance with exist method.

		Ac	Sp	Sn
Discriminating phage enzymes from nonenzymes	(Ding H. et al., 2016)	84.3%	81.7%	87.1%
	This study	85.1%	88.0%	83.0%
Discriminating phage hydrolases from other enzymes	(Ding H. et al., 2016)	93.5%	92.8%	94.5%
	This study	94.3%	93.0%	96.0%

Sn, sensitivity; Sp, specificity; Ac, accuracy.

prediction. However, the strategy cannot work well on low-similar sequences. Especially, the phage genes display the extreme diversity. Protein functions are inextricably linked to correlation of nucleotides or residues, physicochemical properties, spatial structure, and other information. Therefore, we used multiple characteristics to represent phage and hydrolase, but this method has some problems that multiple features contain too much redundant information; different types of features are suitable for different samples. On the basis of the feature selection technique, promising results for phage enzymes and hydrolases prediction were achieved. In the future, we will pay more attention on deep learning, which has solved several protein prediction problems (Peng et al., 2018; Wei et al., 2018a, 2019; Yu et al., 2018; Lv et al., 2019) and may get well performance on this topic. Moreover, we will establish a free webserver that facilitates users to download data and predict phage hydrolases.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the http://lin-group.cn/server/PHYPred.

AUTHOR CONTRIBUTIONS

H-FL and HT designed the study. H-FL carried out all data collection and drafted the manuscript. X-FW and HT revised the manuscript. All authors approved the final manuscript.

FUNDING

This work has been supported by the National Nature Scientific Foundation of China (61702430).

REFERENCES

Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020). Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med. Res. Rev.* [Epub ahead of print].

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2018). iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.* 16, 412–420. doi: 10.1016/j.csbj.2018.10.007

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2019). SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids* 18, 131–141. doi: 10.1016/j.omtn.2019.08.011

Boopathi, V., Subramaniyam, S., Malik, A., Lee, G., Manavalan, B., and Yang, D. C. (2019). mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.* 20: E1964.

Chao, L., Jin, S., Wang, L., Guo, F., and Zou, Q. (2019). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.* 7:224. doi: 10.3389/fbioe.2019.00224

Chen, W., Feng, P., Liu, T., and Jin, D. (2018). Recent advances in machine learning methods for predicting heat shock proteins. *Curr. Drug Metab.* 20, 224–228. doi: 10.2174/1389200219666181031105916

Chen, W., Feng, P., Song, X., Lv, H., and Lin, H. (2019a). iRNA-m7G: identifying N(7)-methylguanosine sites by fusing multiple features. *Mol. Ther. Nucleic Acids* 18, 269–274. doi: 10.1016/j.omtn.2019.08.022

Chen, W., Lv, H., Nie, F., and Lin, H. (2019b). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 35, 2796–2800. doi: 10.1093/bioinformatics/btz015

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Biomed Res. Int.* 2016:1654623. doi: 10.1155/2016/1654623

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

Cheng, J. H., Yang, H., Liu, M. L., Su, W., Feng, P. M., Ding, H., et al. (2018). Prediction of bacteriophage proteins located in the host cell using hybrid features. *Chemometr. Intell. Lab. Syst.* 180, 64–69. doi: 10.1016/j.chemolab.2018.07.006

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144. doi: 10.1093/nar/gky1051

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2018). Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics* 35, 2075–2083. doi: 10.1093/bioinformatics/bty943

Ding, H., and Li, D. M. (2015). Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids* 47, 329–333. doi: 10.1007/s00726-014-1862-4

Ding, H., Feng, P. M., Chen, W., and Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol. Biosyst.* 10, 2229–2235. doi: 10.1039/c4mb00316k

Ding, H., Guo, S. H., Deng, E. Z., Yuan, L. F., Guo, F. B., Huang, J., et al. (2013). Prediction of Golgi-resident protein types by using feature selection technique. *Chemometr. Intell. Lab.* 124, 9–13. doi: 10.1016/j.chemolab.2013.03.005

Ding, H., Yang, W., Tang, H., Feng, P. M., Huang, J., Chen, W., et al. (2016). PHYPred: a tool for identifying bacteriophage enzymes and hydrolases. *Virol. Sin.* 31, 350–352. doi: 10.1007/s12250-016-3740-6

Ding, Y., Tang, J., and Guo, F. (2016a). Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information. *Int. J. Mol. Sci.* 17:1623. doi: 10.3390/ijms17101623

Ding, Y., Tang, J., and Guo, F. (2016b). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinformatics* 17:398.

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inform. Sci.* 41, 546–560. doi: 10.1016/j.ins.2017.08.045

Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

Dong, W., Han, S., Qu, X., Bao, W., Chen, Y., Fan, Y., et al. (2015). "A novel feature fusion method for predicting protein subcellular localization with multiple sites," in *Proceedings of the International Conference on Informative & Cybernetics for Computational Social Systems 2015*, (Piscataway, NJ: IEEE).

Fan, G.-L., and Li, Q.-Z. (2012). Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* 304, 88–95. doi: 10.1016/j.jtbi.2012.03.017

Feng, C. Q., Zhang, Z. Y., Zhu, X. J., Lin, Y., Chen, W., Tang, H., et al. (2018). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 35, 1469–1477. doi: 10.1093/bioinformatics/bty827

Feng, P. M., Ding, H., Chen, W., and Lin, H. (2013). Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.* 2013:530696. doi: 10.1155/2013/530696

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Hu, B., Zheng, L., Long, C., Song, M., Li, T., Yang, L., et al. (2019). EmExplorer: a database for exploring time activation of gene expression in mammalian embryos. *Open Biol.* 9:190054. doi: 10.1098/rsob.190054

Jiang, Q. H., Wang, G. H., Jin, S. L., Li, Y., and Wang, Y. D. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* 8, 282–293.

Keen, E. C. (2012). Phage therapy: concept to cure. *Front. Microbiol.* 3:238. doi: 10.3389/fmicb.2012.00238

Kimura, K., and Itoh, Y. (2003). Characterization of poly-gamma-glutamate hydrolase encoded by a bacteriophage genome: possible role in phage infection of *Bacillus subtilis* encapsulated with poly-gamma-glutamate. *Appl. Environ. Microbiol.* 69, 2491–2497. doi: 10.1128/aem.69.5.2491-2497.2003

Kovalenko, A. O., Chernyshov, S. V., Kutyshenko, V. P., Molochkov, N. V., and Mikoulinskaia, G. V. (2019). Investigation of the calcium-induced activation of the bacteriophage T5 peptidoglycan hydrolase promoting the host cell lysis. *Metallomics* 11, 799–809. doi: 10.1039/c9mt00020h

Liang, Z. Y., Lai, H. Y., Yang, H., Zhang, C. J., Yang, H., Wei, H. H., et al. (2017). Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics* 33, 467–469. doi: 10.1093/bioinformatics/btw630

Lin, H., and Li, Q. Z. (2011). Eukaryotic and prokaryotic promoter prediction using hybrid approach. *Theory Biosci.* 130, 91–100. doi: 10.1007/s12064-010-0114-8

Liu, B., Han, L., Liu, X., Wu, J., and Ma, Q. (2018). Computational prediction of sigma-54 promoters in bacterial genomes by integrating motif finding and machine learning strategies. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1211–1218. doi: 10.1109/tcbb.2018.2816032

Liu, D., Li, G., and Zuo, Y. (2019). Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Brief. Bioinform.* 20, 1826–1835. doi: 10.1093/bib/bby053

Lv, Z. B., Ao, C. Y., and Zou, Q. (2019). Protein function prediction: from traditional classifier to deep learning. *Proteomics* 19:2.

Ma, Q., Zhang, H., Mao, X., Zhou, C., Liu, B., Chen, X., et al. (2014). DMINDA: an integrated web server for DNA motif identification and analyses. *Nucleic Acids Res.* 42, W12–W19.

Manavalan, B., Basith, S., Shin, T. H., Lee, D. Y., Wei, L., and Lee, G. (2019a). 4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse genome. *Cells* 8:1332. doi: 10.3390/cells8111332

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019b). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35, 2757–2765. doi: 10.1093/bioinformatics/bty1047

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019c). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Manavalan, B., Shin, T. H., and Lee, G. (2018a). DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 9, 1944–1956. doi: 10.18632/oncotarget.23099

Manavalan, B., Shin, T. H., and Lee, G. (2018b). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

Manavalan, B., Subramaniyam, S., Shin, T. H., Kim, M. O., and Lee, G. (2018c). Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.* 17, 2715–2726. doi: 10.1021/acs.jproteome.8b00148

Parfitt, T. (2005). Georgia: an unlikely stronghold for bacteriophage therapy. *Lancet* 365, 2166–2167. doi: 10.1016/s0140-6736(05)66759-1

Parmar, K. M., Dafale, N. A., Tikariha, H., and Purohit, H. J. (2018). Genomic characterization of key bacteriophages to formulate the potential biocontrol agent to combat enteric pathogenic bacteria. *Arch. Microbiol.* 200, 1–12. doi: 10.1007/s00203-017-1471-1

Peng, L., Peng, M. M., Liao, B., Huang, G. H., Li, W. B., and Xie, D. F. (2018). The advances and challenges of deep learning application in biological big data processing. *Curr. Bioinform.* 13, 352–359. doi: 10.2174/1574893612666170707095707

Pirisi, A. (2000). Phage therapy–advantages over antibiotics? *Lancet* 356:1418. doi: 10.1016/s0140-6736(05)74059-9

Qu, K., Wei, L., and Zou, Q. (2019). A review of DNA-binding proteins prediction methods. *Curr. Bioinform.* 14, 246–254. doi: 10.2174/1574893614666181212102030

Riede, I., Schwarz, H., and Jahnig, F. (1987). Predicted structure of tail-fiber proteins of T-even type phages. *FEBS Lett.* 215, 145–150. doi: 10.1016/0014-5793(87)80130-8

Rodriguez-Rubio, L., Quiles-Puchalt, N., Martinez, B., Rodriguez, A., Penades, J. R., and Garcia, P. (2013). The peptidoglycan hydrolase of Staphylococcus aureus bacteriophage 11 plays a structural role in the viral particle. *Appl. Environ. Microbiol.* 79, 6187–6190. doi: 10.1128/AEM.01388-13

Seguritan, V., Alves, N. Jr., Arnoult, M., Raymond, A., Lorimer, D., Burgin, A. B. Jr., et al. (2012). Artificial neural networks trained to detect viral and phage structural proteins. *PLoS Comput. Biol.* 8:e1002657. doi: 10.1371/journal.pcbi.1002657

Shen, C., Jiang, L., Ding, Y., Tang, J., and Guo, F. (2019). LPI-KTASLP: prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* 7, 13486–13496. doi: 10.1109/access.2019.2894225

Song, J. N., Tan, H., Shen, H. B., Mahmood, K., Boyd, S. E., Webb, G. I., et al. (2010). Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* 26, 752–760. doi: 10.1093/bioinformatics/btq043

Song, J., Li, F., Leier, A., Marquez-Lago, T. T., Akutsu, T., Haffari, G., et al. (2018). PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* 34, 684–687. doi: 10.1093/bioinformatics/btx670

Song, L., Li, D., Zeng, X., Wu, Y., Guo, L., and Zou, Q. (2014). nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinformatics* 15:298. doi: 10.1186/1471-2105-15-298

Swami, A., and Jain, R. (2012). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480.

Tang, H., Su, Z. D., Wei, H. H., Chen, W., and Lin, H. (2016). Prediction of cell-penetrating peptides with feature selection techniques. *Biochem. Biophys. Res. Commun.* 477, 150–154. doi: 10.1016/j.bbrc.2016.06.035

Tang, H., Zhang, C. M., Chen, R., Huang, P., Duan, C. G., and Zou, P. (2017). Identification of secretory proteins of malaria parasite by feature selection technique. *Lett. Org. Chem.* 14, 621–624.

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Thiel, K. (2004). Old dogma, new tricks–21st century phage therapy. *Nat. Biotechnol.* 22, 31–36. doi: 10.1038/nbt0104-31

Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151. doi: 10.1093/nar/gkx1096

Wang, G., Wang, Y., Feng, W., Wang, X., Yang, J. Y., Zhao, Y., et al. (2008). Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genomics* 9(Suppl. 2):S22. doi: 10.1186/1471-2164-9-S2-S22

Wang, G., Wang, Y., Teng, M., Zhang, D., Li, L., and Liu, Y. (2010). Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon gamma-stimulated HeLa cells. *PLoS One* 5:e11794. doi: 10.1371/journal.pone.0011794

Wang, Y., Shi, F. Q., Cao, L. Y., Dey, N., Wu, Q., Ashour, A. S., et al. (2019). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinform.* 14, 282–294. doi: 10.2174/1574893614666190304125221

Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018a). Prediction of human protein subcellular localization using deep learning. *J. Parall. Distrib. Comput.* 117, 212–217.

Wei, L., Su, R., Wang, B., Li, X., Zou, Q., and Gao, X. (2019). Integration of deep feature representations and handcrafted features to improve the prediction of

N 6-methyladenosine sites. *Neurocomputing* 324, 3–9. doi: 10.1016/j.neucom.2018.04.082

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018b). ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016.

White, S. W. (1987). Prediction of DNA-binding regulatory proteins in bacteriophage T7. *Protein Eng.* 1, 373–376. doi: 10.1093/protein/1.5.373

Xianfang, W., Hongfei, L., Peng, G., Yifeng, L., and Wenjing, Z. (2019). Combining support vector machine with dual g-gap dipeptides to discriminate between acidic and alkaline enzymes. *Lett. Org. Chem.* 16, 325–331. doi: 10.2174/1570178615666180925125912

Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens. *J. Comput. Biol.* 25, 1266–1277. doi: 10.1089/cmb.2018.0004

Yang, H., Tang, H., Chen, X. X., Zhang, C. J., Zhu, P. P., Ding, H., et al. (2016). Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition. *Biomed Res. Int.* 2016:5413903. doi: 10.1155/2016/5413903

Yang, J., Chen, X., McDermaid, A., and Ma, Q. (2017). DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics* 33, 2586–2588. doi: 10.1093/bioinformatics/btx223

Yang, W., Zhu, X. J., Huang, J., Ding, H., and Lin, H. (2019). A brief survey of machine learning methods in protein sub-Golgi localization. *Curr. Bioinform.* 14, 234–240. doi: 10.2174/1574893613666181113131415

Yu, L., Sun, X., Tian, S. W., Shi, X. Y., and Yan, Y. L. (2018). Drug and nondrug classification based on deep learning with various feature selection strategies. *Curr. Bioinform.* 13, 253–259. doi: 10.2174/1574893612666170125124538

Zhang, J., and Liu, B. (2019). A review on the recent developments of sequence-based protein feature extraction methods. *Curr. Bioinform.* 14, 190–199. doi: 10.2174/1574893614666181212102749

Zhang, N., Sa, Y., Guo, Y., Lin, W., Wang, P., and Feng, Y. M. (2018). Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinform.* 13, 50–56. doi: 10.2174/1574893611666160608102537

Zhao, X., Pei, Z., Liu, J., Qin, S., and Cai, L. (2010). Prediction of nucleosome DNA formation potential and nucleosome positioning using increment of diversity combined with quadratic discriminant analysis. *Chromosome Res.* 18, 777–785. doi: 10.1007/s10577-010-9160-9

Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA promoter identification in arabidopsis using multiple histone markers. *Biomed. Res. Int.* 2015:861402. doi: 10.1155/2015/861402

Zhao, Y., Wang, F., Chen, S., Wan, J., and Wang, G. (2017). Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network. *Biomed. Res. Int.* 2017:7049406. doi: 10.1155/2017/7049406

Zheng, L., Huang, S., Mu, N., Zhang, H., Zhang, J., Chang, Y., et al. (2019). RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. *Database* 2019:190054.

Zhu, X. J., Feng, C. Q., Lai, H. Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016a). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* 10:114. doi: 10.1186/s12918-016-0353-5

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016b). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Zuo, Y. C., and Li, Q. Z. (2009). Using reduced amino acid composition to predict defensin family and subfamily: integrating similarity measure and structural alphabet. *Peptides* 30, 1788–1793. doi: 10.1016/j.peptides.2009.06.032

Zuo, Y. C., Peng, Y., Liu, L., Chen, W., Yang, L., and Fan, G. L. (2014). Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns. *Anal. Biochem.* 458, 14–19. doi: 10.1016/j.ab.2014.04.032

Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* 33, 122–124. doi: 10.1093/bioinformatics/btw564

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 March 2020
doi: 10.3389/fgene.2020.00200

Concordance Study Between IBM Watson for Oncology and Real Clinical Practice for Cervical Cancer Patients in China: A Retrospective Analysis

*Fang-wen Zou[1], Yi-fang Tang[2], Chao-yuan Liu[1], Jin-an Ma[1] and Chun-hong Hu[1]**

[1] Department of Oncology, The Second Xiangya Hospital of Central South University, Changsha, China, [2] Department of Anesthesiology, The Second Xiangya Hospital of Central South University, Changsha, China

Watson for Oncology (WFO) is a artificial intelligence clinical decision-support system with evidence-based treatment options for oncologists. WFO has been gradually used in China, but limited reports on whether WFO is suitable for Chinese patients. This study aims to investigate the concordance of treatment options between WFO and real clinical practice for Cervical cancer patients retrospectively. We retrospectively enrolled 300 cases of cervical cancer patients. WFO provides treatment options for 246 supported cases. Real clinical practice were defined as concordant if treatment options were designated "recommended" or "for consideration" by WFO. Concordance of treatment option between WFO and real clinical practice was analyzed statistically. The treatment concordance between WFO and real clinical practice occurred in 72.8% (179/246) of cervical cancer cases. Logistic regression analysis showed that rural registration residences, advanced age, poor ECOG performance status, stages II-IV disease have a remarkable impact on consistency. The main reasons attributed to the 27.2% (67/246) of the discordant cases were the substitution of nedaplatin for cisplatin, reimbursement plan of bevacizumab, surgical preference, and absence of neoadjuvant/adjuvant chemotherapy and PD-1/PD-L1 antibodies recommendations. WFO recommendations were in 72.8% of concordant with real clinical practice for cervical cancer patients in China. However, several localization and individual factors limit its wider application. So, WFO could be an essential tool but it cannot currently replace oncologists. To be rapidly and fully apply to cervical cancer patients in China, accelerate localization and improvement were needed for WFO.

Keywords: artificial intelligence, Watson for Oncology, cervical cancer, concordance, chian

OPEN ACCESS

Edited by:
Yungang Xu,
The University of Texas Health Science Center at Houston, United States

Reviewed by:
Hauke Busch,
University of Lübeck, Germany
Deli Liu,
Cornell University, United States

****Correspondence:***
Chun-hong Hu
huchunhong@csu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *16 November 2019*
Accepted: *20 February 2020*
Published: *24 March 2020*

Citation:
Zou F, Tang Y, Liu C, Ma J and Hu C (2020) Concordance Study Between IBM Watson for Oncology and Real Clinical Practice for Cervical Cancer Patients in China: A Retrospective Analysis. Front. Genet. 11:200. doi: 10.3389/fgene.2020.00200

INTRODUCTION

Artificial intelligence (AI) is the frontier and dominating terrain of Information Technology which able to simulate human mental status and cognitive function (Jiang et al., 2017). With the development of AI and medical diagnosis technology, clinical decision-support systems (CDSS) with intelligent diagnostic function has become one of the important issues of science for medical information (Meyer et al., 2018). Watson for Oncology (IBM) is a representative AI CDSS that

developed by IBM Co.Ltd in United States. WFO can provide a reasonable individualized treatment plan for cancar patients by obtaining valuable information from medical records. WFO first officially landed in China in 2016, until now, more than 80 hospitals use WFO as an important medical diagnostic tool for individualized treatment of tumor (IBM, 2017). WFO can provide counseling services for almost all cancer patients. However, whether WFO was fit for Chinese cancer patients, especially cervical cancer patients.

Cervical cancer is common in the female genital tract malignant tumors, and the incidence of which is second only to that of breast cancer among women worldwide, making it the second-most serious cancer threatening the health and lives of women (Jassim et al., 2018). Compared to breast cancer, cervical cancer is more common in developing countries due to poor health status, and it is the most common in China (Gu X. Y. et al., 2018). And rural and remote areas are also a prevalent regions for cervical cancer in China. But the current problem of the medical service is that the main hospitals hold too many premium resources, but in the meantime, the primary health agencies are excessively lack of resources (Bao et al., 2018). Cervical cancer patients in rural and remote areas can not reach the effective treatment recommendation, especially at centers where cancer expert resources are limited. So, WFO is of great significance for Chinese patients with cervical cancer, especially patients in rural and remote areas with limited medical resources.

Therefore, we conducted a retrospective and observational study on cervical cancer at The Second Xiangya Hospital Cancer Center to explore consistency between WFO and clinical treatment recommendations supported by an expert panel of cancer specialists for Cervical cancer patients.

MATERIALS AND METHODS

Study Population

This retrospective study was reviewed and approved by the Medical Ethics Committee of The Second Xiangya Hospital of Central south university (approval number was 2017-S104). We retrospectively and randomly selected 300 cases of cervical cancer patients from 05/2016 to 08/2018. All patients with cervical cancer confirmed by pathology at The Second Xiangya Hospital Cancer Center. Untreated Patients and recurrent tumors, rare histology that not yet trained to offer treatment options by WFO system were excluded. A total of 18% (54/300) cases excluded from our study and 82% (246/300) cases were included in our study. The detailed patient selection process is shown in **Figure 1**.

Watson for Oncology

Watson for Oncology (IBM Corporation, United States, version 18.1R) used in our study were provided by Baheal Intelligent Technology Co., Ltd[1]. The clinicopathologic data of supported cases were extracted from medical records and entered into the WFO system. Treatment options recommended by WFO were presented in three categories: Blue represents "Recommended" with a strong evidence supported, Orange represents "For consideration" with a potentially suitable evidence-based alternative considered by oncologists based on their clinical judgment, and Red represents that is "Not recommended" that a treatment with contraindications or strong evidence against its use.

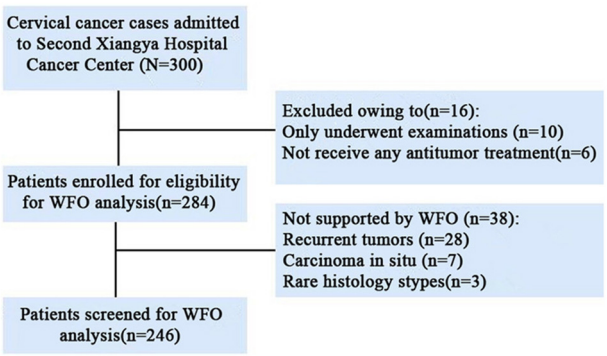


FIGURE 1 | CONSORT diagram. WFO, Watson for Oncology.

Real Clinical Practice for Cervical Cancer

The Second Xiangya Hospital Cancer Center one of the biggest and best oncology departments in the Hunan Province of China. Gynecological Oncology Center is the most important part of The Second Xiangya Hospital Cancer Center and mainly serves cervical cancer, ovarian cancer, endometrial cancer, and other gynecological malignant tumors. Gynecological Oncology Center has a multidisciplinary team (MDT) composed of oncologists, gynecologists, radiologists, pathologists, and nutritionists, et al. MDT forms and implements a comprehensive regimen based on NCCN guidelines and the patient's specific conditions. This comprehensive regimen was considered to be a real clinical practice for cervical cancer.

Data Acquisition and Concordance Judgment

The available clinicopathologic data of 246 patients included a registered residence, age, performance status, pathological type, differentiation degree, FIGO stage, lymphatic and distant metastasis, HPV status, and detailed clinical treatment plan were collected from Second Xiangya Hospital Cancer Center clinical electronic medical records and inputted into WFO system by 2 oncologists manually. Treatment options generated by WFO and recorded through two trained oncologists. It should be noted that in the data analysis process, real clinical practice were categorized as concordant if treatment options were designated "recommended" or "for consideration" by WFO. And if the real clinical practice was not recommended by WFO or if WFO did not provide the same treatment options, the recommendations were considered as discordant. The discordant cases were reevaluated by two senior oncologists provided their

[1]https://www.bsmartd.com

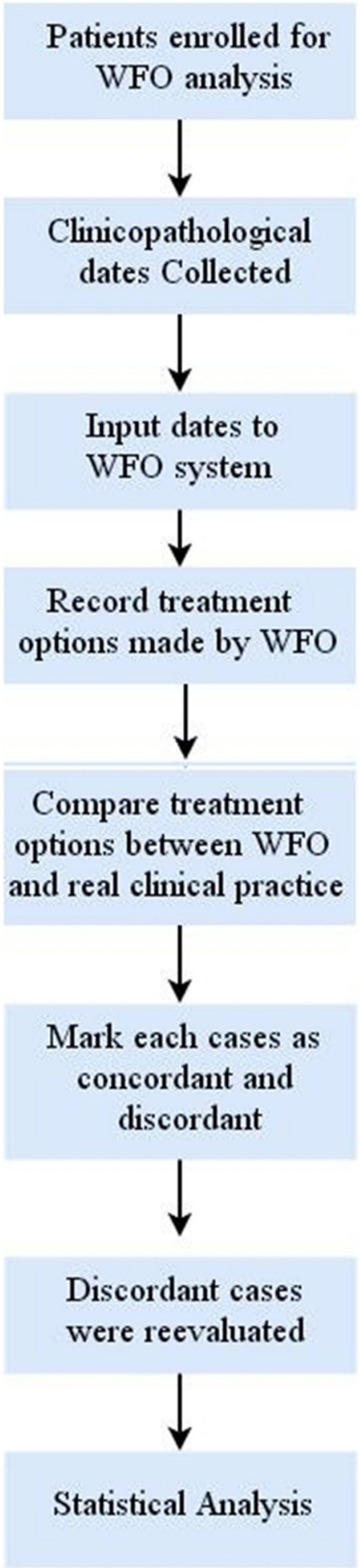


FIGURE 2 | Flow diagram of the study design. WFO, Watson for Oncology.

TABLE 1 | Clinicopathological characteristics of cervical cancer patients (*N* = 246).

Clinicopathological characteristics	Total cases	Concordant cases
Age, years, n (%)		
≤45	29 (11.8)	25 (86.2)
45–65	165 (67.1)	134 (81.2)
≥65	52 (21.1)	20 (38.5)
Median age (range)	53 (35 – 78)	–
Registered residence, n (%)		
Urban registration	81 (33.8)	78 (96.3)
Rural registration	165 (66.2)	101 (61.2)
ECOG[a] performance status, n (%)		
0–1 points	186 (75.6)	145 (77.9)
2 points	47 (19.1)	31 (66.1)
≥3 points	13 (5.3)	3 (23.1)
FIGO stage, n (%)		
I	29 (11.8)	12 (41.4)
II	101 (41.1)	87 (86.1)
III	90 (36.6)	78 (86.7)
IV	26 (10.5)	2 (7.96)
Lymphatic metastasis, n (%)		
Positive	114 (46.3)	82 (71.9)
Negative	132 (53.7)	97 (73.5)
Distant metastasis, n (%)		
Positive	19 (7.7)	10 (52.6)
Negative	227 (92.3)	169 (74.5)
Pathological types, n (%)		
Squamous cell carcinoma	219 (89.0)	159 (72.6)
Adenocarcinoma	16 (6.5)	125 (75)
Adenosscale squamous cell carcinoma	10 (4.1)	7 (70)
Small cell carcinoma	1 (0.4)	1 (100)
Differentiation degrees, n (%)		
High differentiation	48 (19.5)	35 (72.9)
Middle differentiation	90 (36.6)	64 (71.1)
Poorly differentiation	108 (43.9)	80 (74.1)

[a]*Eastern Cooperative Oncology Group, ECOG.*

TABLE 2 | Concordance between WFO and real clinical practice (*N* = 246).

Supported cases	Recommendations	Availability	Total
Concordant cases, n (%)	102 (41.5)[a]	77 (31.3)[b]	179 (72.8)
Discordant cases, n (%)	12 (4.8)[c]	55 (22.4)[d]	67 (27.2)

[a]*Recommended.* [b]*For consideration.* [c]*Not recommended.* [d]*Not available.*

reasons for choosing the real treatment options. The specific study design and procedures and are shown in **Figure 2**.

Statistical Analysis

SPSS20.0 statistics software (SPSS, United States) and Microsoft Excel (2012) were employed to undergo statistical analysis. Descriptive statistics of 246 patients were calculated and presented as means ± standard ($x \pm s$) or median. Differences between the clinicopathological characteristics of the groups were analyzed by Pearson's χ^2 test. Correlation between real clinical practice and WFO recommendations were assessed by the chi-square test. A logistic regression model was estimated with odds

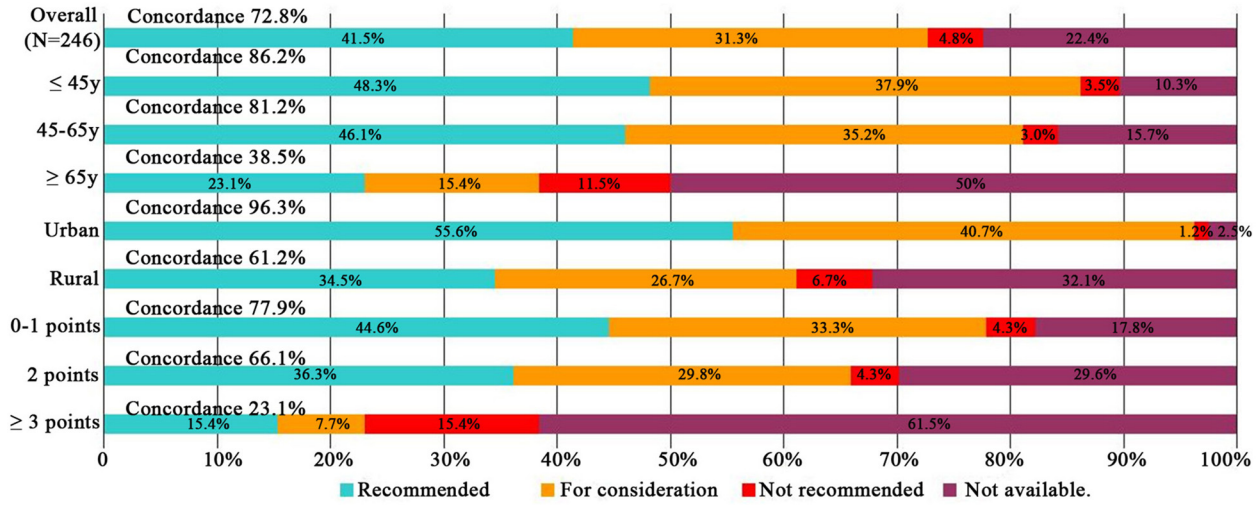


FIGURE 3 | Treatment concordance between WFO and real clinical practice, divided by age, registered residence, and ECOG performance status. WFO, Watson for Oncology.

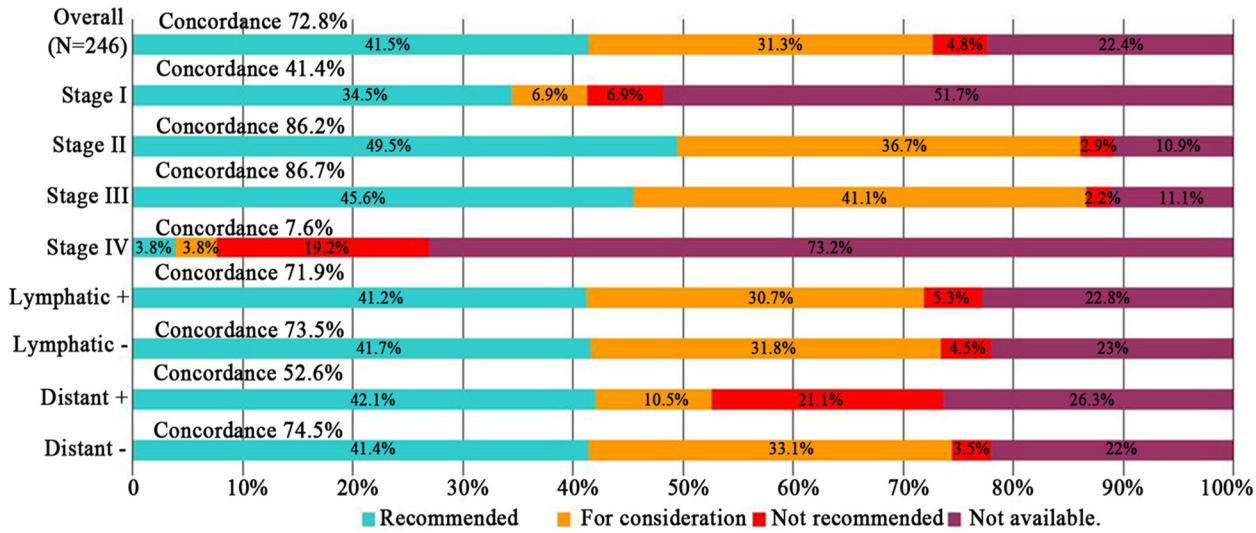


FIGURE 4 | Treatment concordance between WFO and real clinical practice, divided by FIGO stage, lymphatic and distant metastasis. WFO, Watson for Oncology.

ratios (OR) and 95% confidence intervals (CIs). Thep values were designated as $^*P < 0.05$.

RESULTS

Clinicopathological Characteristics of Supported Cases

Of the 300 accrued cervical cancer patients, 246 patients were eligible for WFO analysis. Overall, 82% (246/300) of our enrolled cases were supported by WFO. Clinicopathological characteristics of 246 supported cases are detailed in **Table 1**. Among the 246 supported cases in our study, median age was 53 years (range, 35–78 years), and rural registration patients, stage II/II disease, squamous cell carcinoma, middle/poorly differentiated accounted for 66.2% (165/246), 77.7% (101 + 90/246), 89.0% (219/246), and 80.6% (90 + 108/246), respectively.

Concordance Between WFO and Real Clinical Practice

After reevaluated by two senior oncologists of discordant cases, there was no change to the primary concordance. Overall treatment concordance between WFO and real clinical practice occurred in 72.8% (179/246) of cervical cancer cases, among the concordant cases, treatment options that designated "Recommended" or "For consideration" by WFO accounted for

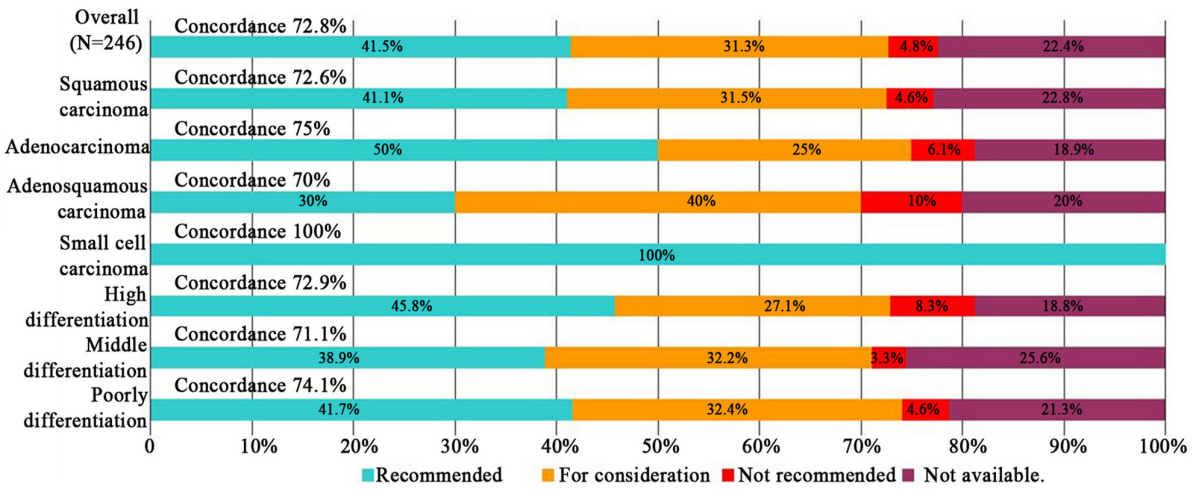


FIGURE 5 | Treatment concordance between WFO and real clinical practice, divided by pathological types and differentiation degrees.

41.5% (102/246) and 31.3% (77/246), respectively. Also, there were 27.2% (67/246) of case cannot consistent with real clinical practice, among the discordant cases, treatment options that not recommended by WFO or did not provided by WFO accounted for 4.8% (12/246) and 22.4% (55/246), respectively. Dates are shown in **Table 2**.

Subgroup Analyses

Subgroup analyses of treatment concordance with clinicopathological characteristics were also carried out. The result showed that urban registration patients [96.3% (78/84)], low age group (≤45 years and 45–65 years groups) [86.2% (25/29), 81.2% (134/165), respectively], good ECOG performance status (0–1 and 2 points groups) [77.9% (145/186), 66.1% (31/47), respectively], and stage II/III disease [80.2% (87/101), 86.7% (78/90), respectively] exhibiting higher concordance than rural registration patients [61.2% (101/165)], advanced age group (≥60 years) [38.5% (20/52)], poor ECOG performance status (≥3 points) [23.1% (3/13)], and stage I/IV disease[41.4% (12/29), 7.96% (2/26), respectively]. While, there were no obvious difference among lymphatic and distant metastasis disease, pathological types, differentiation degrees. Dates shown in **Figures 3–5**.

Logistic Regression Analysis

The logistic regression analysis showed that, compared with patients ≤45 years of age, concordance declined significantly in patients ≥65 years of age and older[0.08 (0.03–0.28), $P = 0.032$]. And Concordance was particularly low for patients with rural registration[0.64 (0.427–0.946), $P = 0.025$],compare with urban patients. Poor ECOG performance status (≥3 points) patients exhibiting lower concordance than good ECOG performance status patients[0.29 (0.083–1.058), $P = 0.048$]. Odds ratios of concordance varied by stage, showed that compared with stage I disease, stages II-III disease were significantly more likely to be concordant ([2.08 (1.002–4.325), $P = 0.046$],[2.09 (1.001–4.381), $P = 0.047$], respectively), whereas, concordance declined remarkably in stages IV disease [0.19 (0.038–0.91), $P = 0.025$]. While, lymphatic and distant metastasis disease, pathological types, differentiation degrees were not found to affect concordance. Dates are shown in **Table 3**.

Analysis of Reasons for Discordant Cases

There were four critical factors attributed to 27.2% (67/246) of the discordant cases. Firstly, Cisplatin is the main chemotherapy drug recommended by WFO, but in our study, of 46.4% (31/67) cases select nedaplatin due to cannot tolerate gastrointestinal reactions of cisplatin. Next, bevacizumab as a routine option recommended by WFO for stage IV stage, but bevacizumab is not in medical reimbursement plan for cervical cancer in China, of 26.9% (18/67) patients reject bevacizumab therapy for the financial burden. Thirdly, for stage Ib2 and IIb disease, only concomitant radiochemotherapy was recommended by WFO, in our study, of 19.4% (13/67) patients prefer surgical therapy instead of concomitant radiochemotherapy. Moreover, neoadjuvant/adjuvant chemotherapy and programmed death-1 and ligand antibodies (PD-1/PD-L1 antibodies) drugs recommendations are not included in the WFO system, in our study, there were 9.1% (6/67),2.8% (2/67) patients chose neoadjuvant/adjuvant chemotherapy and pembrolizumab therapy. Dates are shown in **Table 4**.

DISCUSSION

From 2013, concordance studies between WFO and physicians have been performed in various countries and cancer types. A double-blind study showed that 93% concordance rate for 638 breast cancer patients (Kaur and Singh Mann, 2018; Somashekhar et al., 2018). A retrospective study from India for 1000 consecutive cases showed 80% concordance between multidisciplinary team (MDT) (Baek et al., 2017).

A observational study from Korea showed a 73% concordance rate for colon cancer and a 49% concordance rate for gastric cancer (Somashekhar et al., 2016; Suwanvecho et al., 2017). And, a comparative Study from Korea indicated that WFO without the gene expression assay has limited clinical utility (Kim et al., 2018). It appears that the concordance results varies by countries and cancer types (Zhou et al., 2018). For China, a huge population and regional differences created a different therapeutic experiences and considerations for cancer patients, as well as large differences with Western countries. Also, a retrospective study (Liu et al., 2018) reported by our center revealed that treatment concordance between WFO and MDT occurred in 65.8% (98/149) of lung cancer. Another retrospective study (Zhou et al., 2019) from China showed that Ovarian cancer, lung cancer and breast cancer obtained a high concordance, the concordance of gastric cancer was very low, Incidence and pharmaceuticals may be the major cause of discordance. However, limited reports on whether WFO is suitable for Chinese cervical cancer patients, Zhou et al. reported 14 cervical cancer patients in this study, but the sample size is too small.

TABLE 3 | Logistic regression model of concordance between Watson for Oncology and real clinical practice (*N* = 246).

Clinicopathological characteristics	OR[b] (95%CIs[c])	χ^2	*P* value
Registered residence (Urban and Rural)	0.64 (0.427–0.946)	5.017	*0.025**
Lymphatic metastasis (P[d] and N[e])	1.02 (0.694–1.503)	0.012	*0.913*
Distant metastasis (P[d] and N[e])	1.41 (0.641–3.12)	0.744	*0.388*
Age, years			
≤45 (Reference)	1.00	–	–
45–65	0.94 (0.527–1.685)	0.041	*0.841*
≥65	0.08 (0.03–0.28)	4.609	*0.032**
ECOG[a] performance status			
0–1 points (**Reference**)	1.00	–	–
2 points	0.84 (0.512–1.399)	0.425	*0.514*
≥3 points	0.29 (0.083–1.058)	3.917	*0.048*
FIGO stage			
I (**Reference**)	1.00	–	–
II	2.08 (1.002–4.325)	3.968	*0.046**
III	2.09 (1.001–4.381)	3.958	*0.047**
IV	0.19 (0.038–0.91)	5.036	*0.025**
Pathological types			
Squamous cell carcinoma (**Reference**)	1.00	–	–
Adenocarcinoma	1.03 (0.476–2.244)	0.007	*0.935*
Adenosscale squamous cell carcinoma	0.96 (0.359–2.588)	0.005	*0.942*
Small cell carcinoma	1.38 (0.086–22.187)	0.051	*0.821*
Differentiation degrees			
High differentiation (**Reference**)	1.00	–	–
Middle differentiation	0.97 (0.568–1.675)	0.008	*0.928*
Poorly differentiation	1.01 (0.602–1.714)	0.003	*0.953*

*[a]Eastern Cooperative Oncology Group, ECOG. [b]Odds ratio,OR. [c]Confidence intervals, CIs. [d]Positive and[e]Negative. *P < 0.05.*

TABLE 4 | Analysis of reasons for discordant cases (*N* = 67).

Reasons for discordant cases	Cases, n (%)
Substitution of nedaplatin for cisplatin	28 (41.8)
Reimbursement plan of bevacizumab	18 (26.9)
Surgical preference	13 (19.4)
Neoadjuvant/adjuvant chemotherapy	6 (9.1)
PD-1/PD-L1 antibodies	2 (2.8)

Our retrospective study provides the first evidence that accelerates localization and improvement were needed for WFO before comprehensive application in cervical cancer patients in China. Although treatment options generated by WFO were mostly concordant with real clinical practice, there are still unresolved issues. Firstly, as mentioned in the manual (Gu X. et al., 2018), some clinical settings are not yet supported by WFO system. In our study, of 73.7% (28/38) unsupported cases were recurrent tumors patients. But compare with our center, grass-roots hospitals have a greater proportion of patients with recurrent tumors. So, the cases that cannot be supported by WFO system are very large for cervical cancer patients in China. Secondly, localization factors such as physical of patients, medical reimbursement plan, economic condition, and patient preferences of China were different from western countries, and they ultimately affect the inconsistency. In our study, of 46.4% (31/67) cases select nedaplatin due to cannot tolerate gastrointestinal reactions of cisplatin, of 26.9% (18/67) patients reject bevacizumab therapy for financial burden. of 19.4% (13/67) patients prefer surgical therapy instead of concomitant radiochemotherapy. Moreover, registered residence, age, performance status, FIGO stage have a remarkable impact on consistency. Urban registration patients, low age group, good performance status, and stage II/III disease exhibiting higher concordance than rural registration patients, advanced age group, poor performance status, and stage I/IV disease. These personal factors make WFO unable to achieve individualized treatment and affect the consistency significantly in China. Finally, neoadjuvant/adjuvant chemotherapy (Sun et al., 2018). Chemotherapeutic drugs Goffin et al. (2010) such as gemcitabine, docetaxel, mitomycin, irinotecan, pemetrexed, vinorelbine, and PD-1/PD-L1 antibodies (Kim et al., 2017) drugs recommendations that performed in real clinical practice are not included in the WFO system.

Compared with previous research, our study provides the first evidence that WFO is not suitable for Chinese cervical cancer patients currently, and the sample size of this study was the largest among all cervical cancer studies performed. Also, we not only reported the consistency between WFO and real clinical practice, but also analyzed several influence elements and offered certainly advises for the improvement of WFO to better suit Chinese patients. But, our study contains some limitations. Firstly, this was a retrospective and observational study with control groups lacked, several unmeasured elements may influence the outcome. Secondly, treatment preferences among different experts also affect consistency. Thirdly, the distribution of clinicopathological

characteristics among patients is imbalanced, for example, fewer patients were stage IV diseases may lead to a large disagreement for Stage IV tumors. Finally, molecular parameters, such as mutations, gene expression or protein localization can affect the treatment decision. But, in China, unlike lung cancer and breast cancer, gene detection were lacked for cervical cancer. Although there are some targeted drugs that may be effective for cervical cancer, such as PARP inhibitors (for BRCA1 or BRCA2 mutations patients), EGFR tyrosine kinase inhibitors (for EGFR mutations patients), gene detection is still not widely used in China. So, in our study, Because of the lack of gene detection datas, we cannot observe the effect of molecular parameters on treatment decisions.

For WFO, WFO could be an essential tool for clinicians, provides good references and literature for medical students, or even give some treatment advice to non-specialist (Malin, 2013; Werner et al., 2016). However, we believe that human physicians will not be replaced by AI in the foreseeable future, WFO still has a long way to go to replace oncologists. Medicine is not just a science, but also a social and psychological subject. Any tool and guidelines can only be used as a doctor's reference, localization factors and individual elements should considered for different patients, especially for cancer patients with large heterogeneous (Kemin et al., 2017). Therefore, WFO must be significantly improved to adapt the real clinical practice in different countries. Patient's physical and mental state, economic situation, complications, patient's treatment preference and medical reimbursement plan in different countries should be taken into account and not just provide advice based on existing knowledge. For China, a unique medical database with Chinese characteristics should be created by WFO to adapt and serve Chinese cancer patients.

CONCLUSION

In conclusion, WFO recommendations were in 72.8% of concordant with real clinical practice for cervical cancer patients in China. However, several localization and individual factors limit its wider application. So, WFO cannot replace oncologists for cervical cancer patients in China currently. WFO could be an effective decision-support tool in cancer therapy for Chinese physicians, it also helps to standardize the treatment of cervical cancer. To be rapidly and fully apply to cervical cancer patients in China, accelerate localization and improvement were needed for WFO.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This retrospective study was reviewed and approved by the Medical Ethics Committee of The Second Xiangya Hospital of Central South University (approval number was 2017-S104).

AUTHOR CONTRIBUTIONS

CH was responsible for overall planning for research. FZ was responsible for data collection and statistical analysis. YT and CL were involved in data analysis. JM was participated in the preparation of manuscript.

FUNDING

This work was supported by the Fundamental Research Funds for the Central Universities of Central South University (2019zzts356).

ACKNOWLEDGMENTS

We highly appreciate the kindly help and support from Tianle Li of Qingdao Baheal Corporation and Irene Dankwa-Mullan of IBM Watson Health for their support on the description and use of WFO. Personally, for my parents and wife and daughter, I must say I really appreciate all your love and support up to my work. Thank you for all you did to us, really thanks, I love you-all.

REFERENCES

Baek, J. H., Ahn, S. M., Urman, A., Ahn, H. K., Won, P. K., Lee, W.-S., et al. (2017). Use of a cognitive computing system for treatment of colon and gastric cancer in Korea. *J. Clin. Oncol.* 35(Suppl. 15):e18204A.

Bao, H., Zhang, L., Wang, L., Zhang, M., Zhao, Z., Fang, L., et al. (2018). Significant variations in the cervical cancer screening rate in China by individual-level and geographical measures of socioeconomic status: a multilevel model analysis of a nationally representative survey dataset. *Cancer Med.* 7, 2089–2100. doi: 10.1002/cam4.1321

Goffin, J., Lacchetti, C., Ellis, P. M., Ung Md, Y. C., and Evans Md, W. K. (2010). First-line systemic chemotherapy in the treatment of advanced non-small cell lung cancer: a systematic review. *J. Thorac. Oncol.* 5, 260–274.

Gu, X., Zheng, R., Xia, C., Zeng, H., Zhang, S., Zou, X., et al. (2018). Interactions between life expectancy and the incidence and mortality rates of cancer in China: a population-based cluster analysis. *Cancer Commun.* 38:44. doi: 10.1186/s40880-018-0308-x

Gu, X. Y., Zheng, R. S., and Sun, K. X. (2018). [Incidence and mortality of cervical cancer in China, 2014]. *Zhonghua Zhong Liu Za Zhi* 40, 241–246. doi: 10.3760/cma.j.issn.0253-3766.2018.04.001

IBM (2017). Available at: https://www.ibm.com/us-en/marketplace/ibm-watson-for-oncology.

Jassim, G., Obeid, A., and Nasheet, H. A. A. (2018). Knowledge, attitudes, and practices regarding cervical cancer and screening among women visiting primary health care Centres in Bahrain. *BMC Public Health* 18:128. doi: 10.1186/s12889-018-5023-7

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol.* 2:230. doi: 10.1136/svn-2017-000101

Kaur, J., and Singh Mann, K. D. (2018). *AI Based HealthCare Platform for Real Time, Predictive and Prescriptive Analytics using Reactive Programming* (Singapore: Springer), 138–149.

Kemin, L. I., Yin, R., Wang, D., and Li, Q. (2017). Human papillomavirus subtypes distribution among 2309 cervical cancer patients in West China. *Oncotarget* 8, 28502–28509. doi: 10.18632/oncotarget.16093

Kim, M., Kim, H., Suh, D. H., Kim, K., Kim, H., Kim, Y. B., et al. (2017). Identifying rational candidates for immunotherapy targeting PD-1/PD-L1 in cervical cancer. *Anticancer Res.* 37, 5087–5094.

Kim, Y. Y., Oh, S. J., Chun, Y. S., Lee, W. K., and Park, H. K. (2018). Gene expression assay and Watson for Oncology for optimization of treatment in ER-positive, HER2-negative breast cancer. *PLoS One* 13:e0200100. doi: 10.1371/journal.pone.0200100

Liu, C., Liu, X., Wu, F., Xie, M., Feng, Y., Hu, C., et al. (2018). Using artificial intelligence (Watson for Oncology) for treatment recommendations amongst chinese patients with lung cancer: feasibility study. *J. Med. Internet Res.* 20:e11087. doi: 10.2196/11087

Malin, J. L. (2013). Envisioning Watson as a rapid-learning system for oncology. *J. Oncol. Pract.* 9, 155–157. doi: 10.1200/jop.2013.001021

Meyer, M., Donsa, K., Truskaller, T., Frohner, M., Pohn, B., Felfernig, A., et al. (2018). Development of a protocol for automated glucose measurement transmission used in clinical decision support systems based on the continua design guidelines. *Stud. Health Technol. Inform.* 248, 132–139.

Somashekhar, S., Kumar, S. P. R., Kumar, A., Patil, P., and Rauthan, A. (2016). Validation study to assess performance of IBM cognitive computing system Watson for oncology with Manipal multidisciplinary tumour board for 1000 consecutive cases: an Indian experience. *Ann. Oncol.* 27(Suppl._9):2568.

Somashekhar, S. P., Sepúlveda, M. J., Puglielli, S., Norden, A. D., Shortliffe, E. H., Rohit Kumar, C., et al. (2018). Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann. Oncol.* 29, 418–423. doi: 10.1093/annonc/mdx781

Sun, H., Huang, K., Tang, F., Li, X., Wang, X., Long, S., et al. (2018). Adjuvant chemotherapy after surgery can improve clinical outcomes for patients with IB2-IIB cervical cancer with neoadjuvant chemotherapy followed by radical surgery. *Sci. Rep.* 8 :6443.

Suwanvecho, S., Suwanrusme, H., Sangtian, M., Norden, A. D., Urman, A., Hicks, A., et al. (2017). Concordance assessment of a cognitive computing system in Thailand. *J. Clin. Oncol.* 35:6589. doi: 10.1200/jco.2017.35.15_suppl.6589

Werner, N. E., Gurses, A. P., Leff, B., and Arbaje, A. I. (2016). Improving care transitions across healthcare settings through a human factors approach. *J. Healthc. Qual.* 38, 328–343. doi: 10.1097/jhq.0000000000000025

Zhou, N., Zhang, C. T., Lv, H. Y., Hao, C. X., Li, T. J., Zhu, J. J., et al. (2018). Concordance study between IBM Watson for oncology and clinical practice for patients with cancer in China. *Oncologist* 23, 1–8. doi: 10.1634/theoncologist.2018-0255

Zhou, N., Zhang, C. T., Lv, H. Y., Hao, C. X., Li, T. J., Zhu, J. J., et al. (2019). Concordance study between IBM Watson for oncology and clinical practice for patients with cancer in China. *Oncologist* 24, 812–819.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 26 March 2020
doi: 10.3389/fbioe.2020.00244

Identifying Antifreeze Proteins Based on Key Evolutionary Information

Shanwen Sun[1], Hui Ding[2], Donghua Wang[3] and Shuguang Han[2]**

[1] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [2] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China, [3] Department of General Surgery, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China

Antifreeze proteins are important antifreeze materials that have been widely used in industry, including in cryopreservation, de-icing, and food storage applications. However, the quantity of some commercially produced antifreeze proteins is insufficient for large-scale industrial applications. Further, many antifreeze proteins have properties such as cytotoxicity, severely hindering their applications. Understanding the mechanisms underlying the protein–ice interactions and identifying novel antifreeze proteins are, therefore, urgently needed. In this study, to uncover the mechanisms underlying protein–ice interactions and provide an efficient and accurate tool for identifying antifreeze proteins, we assessed various evolutionary features based on position-specific scoring matrices (PSSMs) and evaluated their importance for discriminating of antifreeze and non-antifreeze proteins. We then parsimoniously selected seven key features with the highest importance. We found that the selected features showed opposite tendencies (regarding the conservation of certain amino acids) between antifreeze and non-antifreeze proteins. Five out of the seven features had relatively high contributions to the discrimination of antifreeze and non-antifreeze proteins, as revealed by a principal component analysis, i.e., the conservation of the replacement of Cys, Trp, and Gly in antifreeze proteins by Ala, Met, and Ala, respectively, in the related proteins, and the conservation of the replacement of Arg in non-antifreeze proteins by Ser and Arg in the related proteins. Based on the seven parsimoniously selected key features, we established a classifier using support vector machine, which outperformed the state-of-the-art tools. These results suggest that understanding evolutionary information is crucial to designing accurate automated methods for discriminating antifreeze and non-antifreeze proteins. Our classifier, therefore, is an efficient tool for annotating new proteins with antifreeze functions based on sequence information and can facilitate their application in industry.

Keywords: antifreeze proteins, support vector machine, evolution, machine learning, position-specific scoring matrix

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Ying Wang,
Xiamen University, China
Liang Yu,
Xidian University, China

****Correspondence:***
Donghua Wang
wangdonghua7885@163.com
Shuguang Han
shughan@uestc.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *14 January 2020*
Accepted: *09 March 2020*
Published: *26 March 2020*

Citation:
Sun S, Ding H, Wang D and Han S (2020) Identifying Antifreeze Proteins Based on Key Evolutionary Information. Front. Bioeng. Biotechnol. 8:244. doi: 10.3389/fbioe.2020.00244

INTRODUCTION

Antifreeze proteins can protect cells and body fluids from freezing by hindering the nucleation, inhibiting the growth of ice crystals, and impeding the recrystallization of ice (Kandaswamy et al., 2011) and are thus important natural antifreeze materials that are widely used in food preservation (Zhan et al., 2018; Provesi et al., 2019; Song et al., 2019), medicine (Lee et al., 2012; Khan et al., 2019), and biotechnological applications (Naing and Kim, 2019). They were first found in the

blood of Antarctic fishes about 50 years ago (DeVries and Wohlschlag, 1969; DeVries et al., 1970). Later studies revealed their existence in other living organisms that have to withstand sub-zero temperatures in their lifetimes, including plants (Griffith et al., 1992; Duman and Olsen, 1993), insects (Husby and Zachariassen, 1980), fungi (Duman and Olsen, 1993), and bacteria (Duman and Olsen, 1993). However, despite their superior performance at the molecular level, the quantity of many proteins that can be commercially produced is insufficient for large-scale industrial applications (Nishimiya et al., 2008). Further, some important antifreeze proteins are cytotoxic, which severely limits their potential applications (Naing and Kim, 2019). Therefore, developing tools to identify novel proteins with antifreeze functions is urgently needed.

However, in spite of similar functions among antifreeze proteins, traditional tools that search for homologous proteins based on sequence similarity, such as Basic Local Alignment Search Tool (BLAST) and Position-Specific Iterative (PSI)-BLAST, perform poorly when attempting to identify antifreeze proteins (Kandaswamy et al., 2011; Eslami et al., 2018; Nath and Subbiah, 2018), because antifreeze proteins exhibit a great diversity among species in their structures and sequence properties. For example, the ice-binding sites in fishes are moderately hydrophobic (Jia and Davies, 2002), while in plants they are mostly hydrophilic (Ramya, 2017). Distinct physicochemical and structural properties are also evident even among phylogenetically related species. Previous research on teleost fishes identified four unrelated types of antifreeze proteins, categorized by their differences in sequence and structural characteristics (Ewart et al., 1999). Type I antifreeze proteins are alanine-rich α-helical proteins; type II have C-type lectin folds of mixed α-helices and β-strands and are composed mainly of Cys, Ala, Asn, Gln, and Thr; type III are globular proteins with no particular repeated structure; type IV mainly consist of Glu and Gln and have folded α-helical bundles (Cheung et al., 2017). In insects, there are two types of antifreeze proteins that are fundamentally different in their primary, secondary, and tertiary structures despite both containing two rows of Thr residues that form β-helices (Jia and Davies, 2002). Similarly, in plants, 15 antifreeze proteins have been purified and characterized (Gupta and Deswal, 2014), and they have low homology and highly diverse properties regarding amino acid sequences (Atici and Nalbantoglu, 2003). Overall, these results suggest that antifreeze proteins may have independently evolved their ice-binding capacities (Cheung et al., 2017) and this has impeded our understanding of the relationship between sequence and function.

Despite these challenges, some researchers have attempted to build classifiers to identify antifreeze proteins based mostly on sequence-derived properties (Doxey et al., 2006; Kandaswamy et al., 2011; Zhao et al., 2012; Appels et al., 2018). For example, Doxey et al. (2006) established an algorithm to predict antifreeze proteins based on physicochemical surface features. Their method, unfortunately, is not suitable for the majority of proteins, as 3D crystallographic structures are unavailable for most proteins. Later studies on predicting antifreeze proteins used modern machine learning algorithms, which have demonstrated their ability in other protein-related research, such as identifying membrane proteins and their subcategories (Chou and Shen, 2007), predicting subcellular localization of multi-label proteins (Javed and Hayat, 2019), and classifying protein secondary structures (Ge et al., 2019). Most of these studies focused on amino acid composition-related features, and various physicochemical properties of amino acid sequences have been extensively used to identify antifreeze proteins (Kandaswamy et al., 2011; Yu and Lu, 2011; Mondal and Pai, 2014; Pratiwi et al., 2017). In contrast, despite the presumed convergent evolution of antifreeze proteins, Zhao et al. (2012) built a classifier with high performance solely based on evolutionary features derived from position-specific scoring matrices (PSSMs), suggesting that evolutionary information is also important for identifying antifreeze proteins. He et al. (2015) further compared the performances of evolutionary features with two amino acid composition metrics (i.e., amino acid composition and pseudo amino acid composition), and showed that features derived from PSSMs achieved higher performance. Similarly, Yang et al. (2015) reported that among various features pertinent to identifying antifreeze proteins, features derived from PSSMs accounted for the largest proportion, though another study showed that physicochemical properties were more important (Eslami et al., 2018). Nevertheless, these results suggest that identifying the evolutionary information underlying the differentiation between antifreeze and non-antifreeze proteins is important for increasing our understanding of protein–ice interactions.

In this study, to uncover the mechanisms of protein–ice interactions and provide an efficient and accurate automated tool for identifying antifreeze proteins, we identified key evolutionary information underlying the differentiation between antifreeze and non-antifreeze proteins. We first derived evolutionary features from PSSMs. A problem that was not resolved in most previous studies on building classifiers based on machine learning algorithms is that antifreeze proteins are rare compared to non-antifreeze proteins. This can lead the models to focusing on non-antifreeze proteins, thus impairing the training process and the assessment of model accuracy (ACC) (Yang et al., 2015). Therefore, we created a pre-processed training data set by using the Majority Weighted Minority Oversampling TEchnique (MWMOTE) to generate synthetic antifreeze proteins based on the weighted informative antifreeze proteins in the raw training data set to remedy the imbalanced training problem (Barua et al., 2014). This method uses a clustering approach to ensure that all generated antifreeze proteins are within some raw antifreeze protein clusters and has been shown to outperform several other methods (Barua et al., 2014). Thereafter, we parsimoniously selected key features to reduce redundant and noisy information based on a feature selection procedure. A classifier based on the selected key features was then trained using the support vector machine (SVM) method to discriminate antifreeze and non-antifreeze proteins.

MATERIALS AND METHODS

Data Sets

The benchmark data sets of antifreeze and non-antifreeze proteins were obtained from Kandaswamy et al. (2011). Previously, 481 antifreeze and 9439 non-antifreeze proteins with low similarity (≤40%) were selected in the study by Kandaswamy et al. (2011), and 221 antifreeze and all the non-antifreeze protein sequences were retrieved from seed proteins in the Pfam database (Sonnhammer et al., 1997). In this study, we further removed sequences containing ambiguous residues, i.e., "X", "B", "U", and "O". In total, 479 antifreeze and 9139 non-antifreeze protein sequences were retained to derive features from PSSMs.

PSI-BLAST was used to assess the PSSM for each sequence based on sequences in the non-redundant Swiss-PROT database that share significant similarity, with three iterations and an e-value threshold of 0.0001 (Bhagwat and Aravind, 2007; Zhu et al., 2019). The raw PSSMs are $n \times 20$ matrices; n rows indicate the query protein residues with n being the length of the protein sequence and 20 columns represent the 20 standard amino acids that may exist in the related protein sequences. The element in *i*th row and *j*th column assesses the frequencies of a specific amino acid (X) at position *i* in the query sequence mutating to the *j*th alternative amino acid (Z) in the related protein sequences during the evolution process. Some amino acids in the rows of each raw PSSM may appear multiple times. The rows of the same amino acids were then summed to form a 20×20 matrix. Thereafter, the matrix was transformed into a vector with 400 dimensions [features; for details see Zhao et al. (2012)]. Thus, each element in the vector is the occurrence of the replacement of a specific amino acid (X) in the query protein by an alternative amino acid (Z) in the related proteins, which indicates the conservation of amino acid X in each query protein. A negative (low) value of X–Z, or a positive (high) value of X–X, suggests that the mutation rate of amino acid X to Z or other amino acids is lower than expected by chance and thus X is conserved. Some sequences could not be assessed in the PSSM analysis and were, therefore, excluded. Finally, vectors based on 398 antifreeze and 7423 non-antifreeze proteins were combined into a single data set, and 80% of the antifreeze and non-antifreeze proteins were used as the training data set while the remaining 20% were used as the test data set.

The training data set was then pre-processed based on MWMOTE using the "imbalance" R package (Cordn et al., 2018) with a ratio of 0.78 being achieved between antifreeze and non-antifreeze proteins.

Feature Selection

Features were first ranked based on the mutual information using an ensemble minimum redundancy–maximum relevance (mRMR) approach (De Jay et al., 2013; Wang et al., 2018; Yuan et al., 2018). The top ranked features were thus both the most relevant for the discrimination of antifreeze and non-antifreeze proteins and complementary to each other (Ding and Peng, 2003). Features were then added to the models sequentially starting with the one with the highest rank and the classifier was trained and evaluated based on five-fold cross-validation and the independent test data set using the SVM method (see below). To parsimoniously select key features to build the classifier to discriminate antifreeze and non-antifreeze proteins, the model preceding the one with decreased performance in the independent test data set was retained.

Model Training and Evaluation

Support vector machine is a popular classifier which has solved several bioinformatics problems (Li et al., 2016; Chen et al., 2017; Bu et al., 2018; Zhang et al., 2018; Chao et al., 2019a,b; Sun et al., 2019; Wang et al., 2019). The "caret" R package was used to train models and tune the model hyperparameters based on SVM (Kuhn, 2008). Model performances were assessed based on ACC, sensitivity (SN), specificity (SP), and the area under the receiver operating characteristics curve (AUC) using five-fold cross-validation and the independent test data set (Tan et al., 2019). ACC is the ratio of the number of correctly discriminated proteins relative to the total number of proteins, assessing the model's overall performance. SN is the ratio of the number of correctly discriminated antifreeze proteins relative to the number of all true antifreeze proteins. SP is the ratio of the number of correctly discriminated non-antifreeze proteins relative to the number of all true non-antifreeze proteins. In contrast, AUC considers both SN and SP, evaluating the model's capacity to recognize antifreeze proteins among unlabeled antifreeze proteins, and non-antifreeze proteins among unlabeled non-antifreeze proteins. It is thus robust to imbalanced data. Higher AUC values indicate that a model is better at discriminating antifreeze and non-antifreeze proteins.

Additionally, to compare the performances of classifiers based on the raw data set with classifiers based on the pre-processed data set (created using MWMOTE) and the performances of classifiers based on our parsimoniously selected key features with classifiers based on all features, classifiers were also trained and evaluated using the raw data set and the pre-processed data set with all features. Additionally, principal component (PC) analysis was used to further reduce the dimensionality in all data sets and classifiers based on the first two PCs were then trained and their performances were plotted to visually illustrate the model performances. To assess the importance of each selected key feature for the first two PCs, their contributions were assessed based on the following equation:

$$\text{Contribution} = r_{ij}^2 / \sum r_{ij}^2$$

where r_{ij}^2 is the correlation coefficient between the *i*th key feature and the *j*th PC.

RESULTS

Selection of Key Features for Discriminating Antifreeze and Non-antifreeze Proteins

Seven features derived from PSSMs were parsimoniously selected as key features for discriminating antifreeze and non-antifreeze

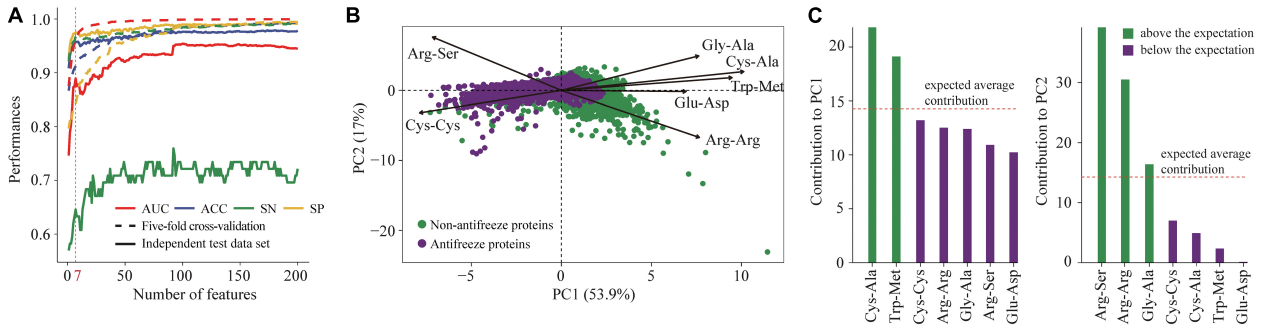


FIGURE 1 | (A) Selection of key features derived from position-specific scoring matrices (PSSMs) for discriminating antifreeze and non-antifreeze proteins. Features were first ranked based on the mutual information using an ensemble minimum redundancy–maximum relevance (mRMR) approach. Starting with the highest ranked feature, the top 200 features were then sequentially added to the models. Model performances were assessed using five-fold cross-validation and an independent test data set based on the AUC, ACC, SN, and SP. The top seven features were parsimoniously selected to build the classifier to discriminate antifreeze and non-antifreeze proteins, and AUC, ACC, and SN then decreased in the independent test data set. **(B)** Distribution of antifreeze and non-antifreeze proteins along the first two principal components (PCs). Arrows indicate the correlations between each of the seven features and PC1 and PC2. **(C)** Contribution of each of the seven features to PC1 and PC2. Features are sorted in a descending order based on their contributions. The expected average contribution was 1/7, as there were seven features and the contribution of each feature was assumed to be uniform (Kassambara and Mundt, 2017).

proteins (**Figure 1A**). Adding more features resulted in initial reductions in performances in the independent test data set regarding AUC, ACC, and SN, although with even more features being included, the performances increased (**Figure 1A**). Based on the seven features, most of the proteins were correctly discriminated in the training data set, that is 96% and 97% antifreeze proteins and non-antifreeze proteins were correctly identified, respectively (**Table 1**). The overall ACC and AUC were 0.91 and 0.96, respectively (**Table 1**). In the independent test data set, a slightly lower proportion (63%) of antifreeze proteins were successfully identified, and 97% of non-antifreeze proteins were correctly predicted, which led to an increase in ACC but a decrease in AUC compared to the training data set (**Table 1**).

The first two PCs derived from the seven selected key features accounted for 70% of the variation among features (**Figure 1B**). Along PC1, the replacements of Cys and Trp in non-antifreeze proteins by Ala and Met, respectively, in the related proteins increased in line with increasing occurrences of non-antifreeze proteins (**Figures 1B,C**). Similarly, along PC2, Gly and Arg in non-antifreeze proteins were more frequently replaced by Ala and Arg, respectively, in the related proteins. In contrast, there were fewer replacements of Cys, Trp, and Gly in antifreeze proteins, but more Arg was replaced by Ser (**Figures 1B,C**). With only the first two PCs, relatively high performances regarding discriminating antifreeze and non-antifreeze proteins were achieved (**Table 1** and **Figure 2C**). The classifier correctly identified 94% of antifreeze proteins and 78% of non-antifreeze proteins in the training data set and 61% of antifreeze proteins and 95% of non-antifreeze proteins in the independent test data set (**Table 1**). The ACC and AUC were 0.87 and 0.90 in the training data set, respectively, and 0.93 and 0.82 in the independent test data set, respectively (**Table 1**).

Performance of MWMOTE Method

Using the MWMOTE method to create the pre-processed data set greatly enhanced model performances. When using all features, almost every protein was correctly identified in the training data set, with SN and SP values of 1.00 and, in the independent test data set, 70% of the antifreeze proteins and 100% of the non-antifreeze proteins were correctly discriminated (**Table 1** and **Figure 2B**). In contrast, although the classifier trained with all features and the raw data set showed overall high performances in terms of AUC, ACC, and SP, this was at the expense of correctly identifying the antifreeze proteins, i.e., a low SN (**Table 1**). Most of the proteins were predicted to be non-antifreeze proteins and only 65% and 67% of the antifreeze proteins were correctly recognized in the training and independent test data sets, respectively (**Table 1** and **Figure 2A**).

DISCUSSION

We found that pre-processing based on the MWMOTE method improved our capacity to discriminate antifreeze and non-antifreeze proteins. Seven out of 400 features derived from PSSMs were parsimoniously selected as the key features that led to relatively high performances. There was still redundant and noisy information among these features that were minimized using a PC analysis, with a minor loss of discrimination ability. These results suggest that antifreeze and non-antifreeze proteins could be differentiated based on a few features derived from PSSMs and thus a little evolutionary information.

Differentiation of Antifreeze and Non-antifreeze Proteins

Antifreeze proteins have been shown to have convergently evolved from different protein families (Ewart et al., 1999; Nath et al., 2013; Nath and Subbiah, 2018). Here, we found that common evolutionary relationships among antifreeze proteins may exist, i.e., Cys, Trp, and Gly are conservative and their replacements by Ala, Met, and Ala, respectively, are rare in antifreeze proteins. This result is surprising because Cys, Trp,

TABLE 1 | Performances regarding discriminating antifreeze and non-antifreeze proteins based on the support vector machine (SVM) method in different data sets.

	Features	Five-fold cross-validation				Independent test data set			
		AUC	ACC	SN	SP	AUC	ACC	SN	SP
Raw data set	400 features	0.97	0.98	0.65	1.00	0.98	0.98	0.67	1.00
	First two PCs	0.97	0.83	0.54	1.00	0.78	0.97	0.47	1.00
Pre-processed data set[a]	400 features	1.00	0.99	1.00	1.00	0.96	0.98	0.70	1.00
	First two PCs	0.88	0.86	0.95	0.75	0.81	0.94	0.61	0.96
Pre-processed data set[a]	Seven key features	0.96	0.91	0.97	0.84	0.89	0.96	0.63	0.97
	First two PCs	0.90	0.87	0.94	0.78	0.82	0.93	0.61	0.95

"400 features" refers to all features derived from position-specific scoring matrices (PSSMs), "first two PCs" refers to the corresponding first two principal components (PCs), and "seven key features" refers to the seven parsimoniously selected key features. [a]Data set based on the Majority Weighted Minority Oversampling TEchnique (MWMOTE). AUC, area under the receiver operating characteristic curve; ACC, accuracy; SN, sensitivity, SP, specificity.

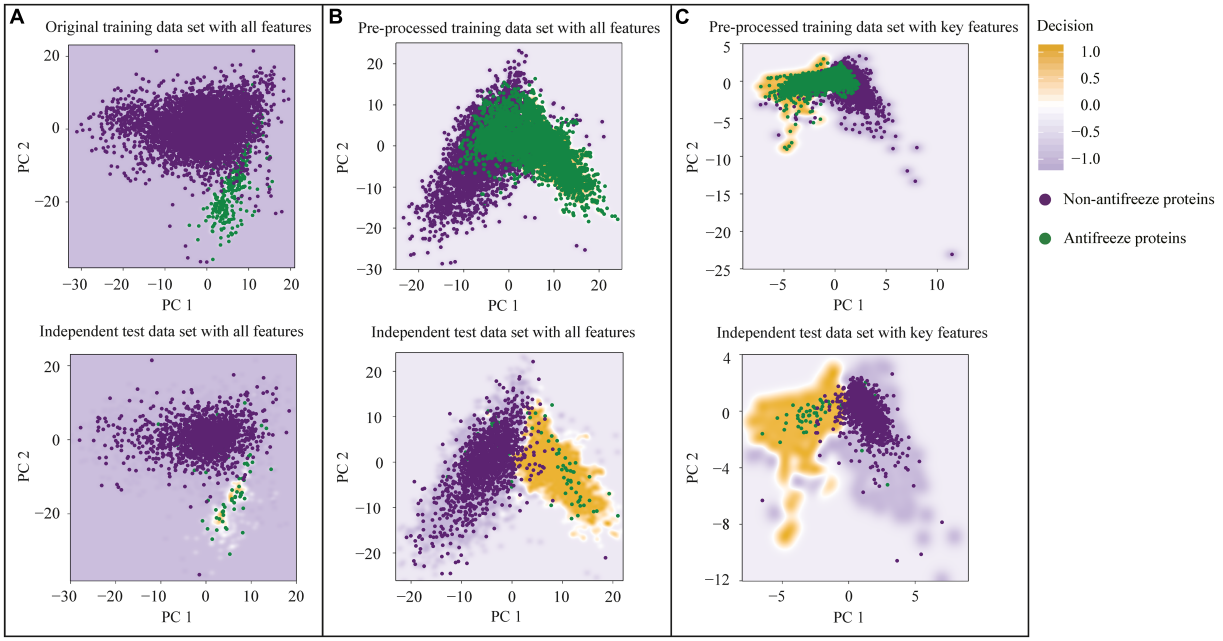


FIGURE 2 | Performances of models for discriminating antifreeze and non-antifreeze proteins based on the first two principal components (PCs) derived from **(A)** all features derived from position-specific scoring matrices (PSSMs) using the raw data set, **(B)** all features derived from PSSMs using the pre-processed data set based on the Majority Weighted Minority Oversampling TEchnique (MWMOTE), and **(C)** the seven selected key features using the pre-processed data set. The upper figures are based on five-fold cross-validation and the lower figures are based on the independent test data set. See **Table 1** for exact performance values. Additionally, the decision values that were used to predict the antifreeze and non-antifreeze proteins are shown.

Gly, Met, and Ala are the most hydrophobic amino acid residues (Rose et al., 1985), have been shown to have high similarities among each other in terms of hydrophobicity (Riek et al., 1995), and thus the mutation rates or replacements of Cys, Trp, and Gly by Ala, Met, and Ala, respectively, should be high (Riek et al., 1995). The conservation of Cys, Trp, and Gly in antifreeze proteins, therefore, suggests that evolutionary pressure may have existed to keep these amino acids in antifreeze proteins, and the conservation of Cys, Trp, and Gly may confer the antifreeze function on proteins, although the underlying mechanisms are still unclear. Similarly, Graham and Davies (2005) showed that, despite the surprising divergency in primary sequences, both isoforms of a highly effective antifreeze protein found in snow fleas start with Gly. Gly is thought to be very unique and highly conformationally flexible and it can occupy positions, such as tight turns, that are impossible for all other amino acids (Betts and Russell, 2003). The existence of Gly may be essential for forming various ice-binding surfaces in antifreeze proteins (Jia and Davies, 2002; Doxey et al., 2006). Moreover, the disulfide bonds formed by paired Cys residues are ubiquitous among antifreeze proteins in various taxa, including insects (Li et al., 1998; Graether et al., 2000), bacteria (Bar et al., 2006), plants (Hon et al., 1994; Bar et al., 2006), and fishes (Davies and Hew, 1990), which may enable proteins to resist destruction due to ice adsorption or denaturation stress during freezing (Li et al., 1998). Trp is an aromatic amino acid with a hydrophobic side chain, and it tends to be buried in protein hydrophobic cores, potentially forming ice-binding sites (Betts and Russell, 2003). Another

possible explanation for the conservation of Cys, Trp, and Gly in antifreeze proteins is that these amino acids have higher propensities to form α-helixes (Koehl and Levitt, 1999), which is important for inhibiting the growth of ice crystals (Knight et al., 1991). In contrast to the conservation of Cys, Trp, and Gly in antifreeze proteins, Arg in antifreeze proteins was more frequently replaced by Ser and less frequently replaced by itself in the related proteins, which suggests a lack of conservation of Arg in antifreeze proteins. Similarly, Nath et al. (2013) compared the evolutionary differences between three types of antifreeze proteins in fishes and their corresponding homologous non-antifreeze proteins, and they found that Arg is commonly avoided in all types of antifreeze proteins. However, it is important to note that the PSSMs of our antifreeze proteins were based on comparing sequence similarities with related proteins but not necessarily proteins with antifreeze function. Antifreeze proteins are rare and dissimilar in their sequences, and PSI-BLAST and BLAST have difficulty using an antifreeze protein as the query sequence to search for new antifreeze proteins based on similarity (Kandaswamy et al., 2011; Eslami et al., 2018; Nath and Subbiah, 2018). Thus, some of the sequences that were used to calculate the PSSMs of our antifreeze proteins may have been non-antifreeze protein sequences. If this is the case, the high frequency of the replacement of Arg in antifreeze proteins with Ser in non-antifreeze proteins (or, in other words, the high frequency of the replacement of Ser in non-antifreeze proteins with Arg in antifreeze proteins) may indicate an important mutation contributing to antifreeze function. More stringent selection of proteins during the assessment of PSSMs could help to clarify this. Nevertheless, our results as well as the results from previous studies indicate that identifying key evolutionary information is important for understanding protein–ice interactions and for understanding the development of antifreeze proteins from pre-existing non-antifreeze proteins.

Comparison of Our Seven Key Features With State-of-the-Art Tools for Discriminating Antifreeze and Non-antifreeze Proteins

With the advancements of genome sequencing, a large number of sequenced proteins have been accumulated and need to be functionally annotated. Many auto-annotation tools exist to identify antifreeze proteins, such as TargetFreeze (He et al., 2015), AFP_PSSM (Zhao et al., 2012), CryoProtect (Pratiwi et al., 2017), and afpCOOL (Eslami et al., 2018). However, these tools use too many features (**Table 2**), which may often be redundant and lead to overfitting. We found that high performances were achieved using only seven key features derived from PSSMs. Compared with other methods, our method used the smallest number of features while achieving the highest Matthews correlation coefficient (MCC), which is the correlation between predicted and true classifications and is robust to imbalanced data (Boughorbel et al., 2017), and ACC values, as well as high SN and SP (**Table 2**). These results indicate that our model outperforms the state-of-the-art tools and so could be more appropriate for discriminating antifreeze and non-antifreeze proteins.

TABLE 2 | Comparison of our seven key features derived from position-specific scoring matrices (PSSMs) with existing machine learning methods for discriminating antifreeze and non-antifreeze proteins using independent test data set(s).

Method	Number of features	ACC	SN	SP	MCC
Seven key features	7	0.96	0.63	0.97	0.57
iAFP[a]	13	0.95	0.13	0.97	0.09
AFP-Pred[a]	25	0.77	0.91	0.77	0.23
AFP-PseAAC[a]	30	0.85	0.85	0.85	0.27
TargetFreeze[a]	300	0.91	0.92	0.91	0.04
CryoProtect[a]	420	0.88	0.87	0.88	0.31
AFP_PSSM[b]	400	0.93	0.76	0.93	N/A
afpCOOL[c]	641	0.96	0.72	0.98	N/A

[a]Results were obtained from a study by Pratiwi et al. (2017). [b]Results were obtained from a study by Zhao et al. (2012). [c]Results were obtained from a study by Eslami et al. (2018). AUC, area under the receiver operating characteristic curve; ACC, accuracy; SN, sensitivity; SP, specificity; MCC, Matthews correlation coefficient. N/A: not available.

CONCLUSION

Understanding the evolution of antifreeze proteins is important for uncovering the interactions between proteins and ice, and, more broadly, the adaptation of organisms to their environments. We found that the conservation of several key amino acids showed opposite tendencies in antifreeze and non-antifreeze proteins, suggesting that there has been strong selection pressure related to these amino acids leading to the differentiation between antifreeze and non-antifreeze proteins regarding their ice-binding capacities. Moreover, we showed that evolutionary information is crucial for designing accurate automated tools for discriminating antifreeze and non-antifreeze proteins. Therefore, our model, which is based on seven key features derived from PSSMs and outperforms the state-of-the-art tools, is an efficient and crucial tool to help to identify new antifreeze proteins and facilitate their use.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found in Kandaswamy et al. (2011).

AUTHOR CONTRIBUTIONS

SS, HD, DW, and SH: conceptualization. SS: formal analysis and writing and preparation of the original draft. SS, HD, DW, and SH: writing-review and editing. All authors have read and agreed to the published version of the manuscript.

FUNDING

The work was supported by the Natural Science Foundation of China (No. 61772119).

REFERENCES

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:eaar7191. doi: 10.1126/science.aar7191

Atici, O., and Nalbantoglu, B. (2003). Antifreeze proteins in higher plants. *Phytochemistry* 64, 1187–1196. doi: 10.1016/s0031-9422(03)00420-5

Bar, M., Bar-Ziv, R., Scherf, T., and Fass, D. (2006). Efficient production of a folded and functional, highly disulfide-bonded β-helix antifreeze protein in bacteria. *Protein Express. Purif.* 48, 243–252. doi: 10.1016/j.pep.2006.01.025

Barua, S., Islam, M. M., Yao, X., and Murase, K. (2014). MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans. Knowl. Data Eng.* 26, 405–425. doi: 10.1109/TKDE.2012.232

Betts, M. J., and Russell, R. B. (2003). "Amino acid properties and consequences of subsitutions," in *Bioinformatics for Geneticists*, eds M. R. Barnes and I. C. Gray (London: Wiley).

Bhagwat, M., and Aravind, L. (2007). "PSI-BLAST Tutorial," in *Comparative Genomics*, ed. N. H. Bergman (Totowa, NJ: Humana Press).

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* 12:e0177678. doi: 10.1371/journal.pone.0177678

Bu, H. D., Hao, J. Q., Guan, J. H., and Zhou, S. G. (2018). Predicting enhancers from multiple cell lines and tissues across different developmental stages based On SVM method. *Curr. Bioinformat.* 13, 655–660. doi: 10.2174/1574893613666180726163429

Chao, L., Jin, S., Wang, L., Guo, F., and Zou, Q. (2019a). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.* 7:224. doi: 10.3389/fbioe.2019.00224

Chao, L., Wei, L., and Zou, Q. (2019b). SecProMTB: a SVM-based classifier for secretory proteins of mycobacterium tuberculosis with imbalanced data set. *Proteomics* 19:e1900007. doi: 10.1002/pmic.201900007

Chen, W., Xing, P., and Zou, Q. (2017). Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. *Sci. Rep.* 7:40242. doi: 10.1038/srep40242

Cheung, R. C. F., Ng, T. B., and Wong, J. H. (2017). Antifreeze proteins from diverse organisms and their applications: an overview. *Curr. Prot. Peptide Sci.* 18, 262–283. doi: 10.2174/1389203717666161013095027

Chou, K. C., and Shen, H. B. (2007). MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* 360, 339–345. doi: 10.1016/j.bbrc.2007.06.027

Cordn, I., Garca, S., Fernndez, A., and Herrera, F. (2018). *imbalance: Preprocessing Algorithms for Imbalanced Datasets. R package verion* 1.0.2. Available online at: https://rdrr.io/cran/imbalance/ (accessed July 21, 2019).

Davies, P. L., and Hew, C. L. (1990). Biochemistry of fish antifreeze proteins. *FASEB J.* 4, 2460–2468. doi: 10.1096/fasebj.4.8.2185972

De Jay, N., Papillon-Cavanagh, S., Olsen, C., El-Hachem, N., Bontempi, G., and Haibe-Kains, B. (2013). mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics* 29, 2365–2368. doi: 10.1093/bioinformatics/btt383

DeVries, A. L., Komatsu, S. K., and Feeney, R. E. (1970). Chemical and physical properties of freezing point-depressing glycoproteins from Antarctic fishes. *J. Biol. Chem.* 245, 2901–2908.

DeVries, A. L., and Wohlschlag, D. E. (1969). Freezing resistance in some Antarctic fishes. *Science (New York, N.Y.)* 163, 1073–1075. doi: 10.1126/science.163.3871.1073

Ding, C., and Peng, H. C. (2003). "Minimum redundancy feature selection from microarray gene expression data," in *Proceedings of the 2003 Ieee Bioinformatics Conference*, Los Alamitos, 523–528. doi: 10.1109/csb.2003.1227396

Doxey, A. C., Yaish, M. W., Griffith, M., and McConkey, B. J. (2006). Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. *Nat. Biotechnol.* 24, 852–855. doi: 10.1038/nbt1224

Duman, J. G., and Olsen, T. M. (1993). Thermal hysteresis protein-activity in bacteria, fungi, and phylogenetically diverse plants. *Cryobiology* 30, 322–328. doi: 10.1006/cryo.1993.1031

Eslami, M., Shirali Hossein, Zade, R., Takalloo, Z., Mahdevar, G., Emamjomeh, A., et al. (2018). afpCOOL: a tool for antifreeze protein prediction. *Heliyon* 4:e00705. doi: 10.1016/j.heliyon.2018.e00705

Ewart, K. V., Lin, Q., and Hew, C. L. (1999). Structure, function and evolution of antifreeze proteins. *Cell. Mol. Life Sci.* 55, 271–283. doi: 10.1007/s000180050289

Ge, Y., Zhao, S., and Zhao, X. (2019). A step-by-step classification algorithm of protein secondary structures based on double-layer SVM model. *Genomics* 112, 1941–1946. doi: 10.1016/j.ygeno.2019.11.006

Graether, S. P., Kuiper, M. J., Gagné, S. M., Walker, V. K., Jia, Z., Sykes, B. D., et al. (2000). β-Helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. *Nature* 406, 325–328. doi: 10.1038/35018610

Graham, L. A., and Davies, P. L. (2005). Glycine-rich antifreeze proteins from snow fleas. *Science* 310, 461–461. doi: 10.1126/science.1115145

Griffith, M., Ala, P., Yang, D. S. C., Hon, W. C., and Moffatt, B. A. (1992). Antifreeze protein produced endogenously in winter rye leaves. *Plant Physiol.* 100, 593–596. doi: 10.1104/pp.100.2.593

Gupta, R., and Deswal, R. (2014). Antifreeze proteins enable plants to survive in freezing conditions. *J. Biosci.* 39, 931–944. doi: 10.1007/s12038-014-9468-2

He, X., Han, K., Hu, J., Yan, H., Yang, J.-Y., Shen, H.-B., et al. (2015). TargetFreeze: identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition. *J. Membr. Biol.* 248, 1005–1014. doi: 10.1007/s00232-015-9811-z

Hon, W. C., Griffith, M., Chong, P., and Yang, D. S. C. (1994). Extraction and isolation of antifreeze proteins from winter rye (*Secale cereale* L.) Leaves. *Plant Physiol.* 104, 971–980. doi: 10.1104/pp.104.3.971

Husby, J. A., and Zachariassen, K. E. (1980). Antifreeze agents in the body fluid of winter active insects and spiders. *Experientia* 36, 963–964. doi: 10.1007/BF01953821

Javed, F., and Hayat, M. (2019). Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC. *Genomics* 111, 1325–1332. doi: 10.1016/j.ygeno.2018.09.004

Jia, Z. C., and Davies, P. L. (2002). Antifreeze proteins: an unusual receptor-ligand interaction. *Trends Biochem. Sci.* 27, 101–106. doi: 10.1016/s0968-0004(01)02028-x

Kandaswamy, K. K., Chou, K. C., Martinetz, T., Moller, S., Suganthan, P. N., Sridharan, S., et al. (2011). AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol.* 270, 56–62. doi: 10.1016/j.jtbi.2010.10.037

Kassambara, A., and Mundt, F. (2017). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5*. Available online at: https://cloud.r-project.org/web/packages/factoextra/index.html (accessed June 09, 2019).

Khan, M. S., Ibrahim, S. M., Adamu, A. A., Rahman, M. B. A., Bakar, M. Z. A., Noordin, M. M., et al. (2019). Pre-grafting histological studies of skin grafts cryopreserved in α helix antarctic yeast oriented antifreeze peptide (Afp1m). *Cryobiology* [in press]. doi: 10.1016/j.cryobiol.2019.09.012

Knight, C. A., Cheng, C. C., and DeVries, A. L. (1991). Adsorption of alpha-helical antifreeze peptides on specific ice crystal surface planes. *Biophys. J.* 59, 409–418. doi: 10.1016/s0006-3495(91)82234-2

Koehl, P., and Levitt, M. (1999). Structure-based conformational preferences of amino acids. *Proc. Natl. Acad. Sci. U.S.A.* 96, 12524–12529. doi: 10.1073/pnas.96.22.12524

Kuhn, M. (2008). Building predictive models in R Using the caret package. *J. Stat. Softw.* 28, 1–26. doi: 10.18637/jss.v028.i05

Lee, S. G., Koh, H. Y., Lee, J. H., Kang, S. H., and Kim, H. J. (2012). Cryopreservative effects of the recombinant ice-binding protein from the arctic yeast leucosporidium sp on red blood cells. *Appl. Biochem. Biotechnol.* 167, 824–834. doi: 10.1007/s12010-012-9739-z

Li, D., Ju, Y., and Zou, Q. (2016). Protein folds prediction with hierarchical structured SVM. *Curr. Proteom.* 13, 79–85. doi: 10.2174/1570164613021605140009 40

Li, N., Kendrick, B. S., Manning, M. C., Carpenter, J. F., and Duman, J. G. (1998). Secondary structure of antifreeze proteins from overwintering larvae of the beetle *Dendroides canadensis*. *Arch. Biochem. Biophys.* 360, 25–32. doi: 10.1006/abbi.1998.0930

Mondal, S., and Pai, P. P. (2014). Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J. Theor. Biol.* 356, 30–35. doi: 10.1016/j.jtbi.2014.04.006

Naing, A. H., and Kim, C. K. (2019). A brief review of applications of antifreeze proteins in cryopreservation and metabolic genetic engineering. *3 Biotech* 9:9. doi: 10.1007/s13205-019-1861-y

Nath, A., Chaube, R., and Subbiah, K. (2013). An insight into the molecular basis for convergent evolution in fish antifreeze Proteins. *Comput. Biol. Med.* 43, 817–821. doi: 10.1016/j.compbiomed.2013.04.013

Nath, A., and Subbiah, K. (2018). The role of pertinently diversified and balanced training as well as testing data sets in achieving the true performance of classifiers in predicting the antifreeze proteins. *Neurocomputing* 272, 294–305. doi: 10.1016/j.neucom.2017.07.004

Nishimiya, Y., Mie, Y., Hirano, Y., Kondo, H., Miura, A., and Tsuda, S. (2008). Mass preparation and technological development of an antifreeze protein. *Synthesiol. Engl. Ed.* 1, 7–14. doi: 10.5571/syntheng.1.7

Pratiwi, R., Malik, A. A., Schaduangrat, N., Prachayasittikul, V., Wikberg, J. E. S., Nantasenamat, C., et al. (2017). CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins. *J. Chem.* 2017:15. doi: 10.1155/2017/9861752

Provesi, J. G., Neto, P. A. V., Arisi, A. C. M., and Amante, E. R. (2019). Extraction of antifreeze proteins from cold acclimated leaves of Drimys angustifolia and their application to star fruit (*Averrhoa carambola*) freezing. *Food Chem.* 289, 65–73. doi: 10.1016/j.foodchem.2019.03.055

Ramya, L. (2017). Physicochemical properties of insect and plant antifreeze proteins: a computational study. *Curr. Sci.* 112, 1512–1520.

Riek, R. P., Handschumacher, M. D., Sung, S. S., Tan, M., Glynias, M. J., Schluchter, M. D., et al. (1995). Evolutionary conservation of both the hydrophilic and hydrophobic nature of transmembrane residues. *J. Theor. Biol.* 172, 245–258. doi: 10.1006/jtbi.1995.0021

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science* 229, 834–838. doi: 10.1126/science.4023714

Song, D. H., Kim, M., Jin, E. S., Sim, D. W., Won, H. S., Kim, E. K., et al. (2019). Cryoprotective effect of an antifreeze protein purified from Tenebrio molitor larvae on vegetables. *Food Hydrocolloids* 94, 585–591. doi: 10.1016/j.foodhyd.2019.04.007

Sonnhammer, E. L. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Prot. Struct. Funct. Bioinform.* 28, 405–420.

Sun, S., Wang, C., Ding, H., and Zou, Q. (2019). Machine learning and its applications in plant molecular studies. *Brief. Funct. Genom.* 19, 40–48. doi: 10.1093/bfgp/elz036

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Mathemat. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Wang, S. P., Zhang, Q., Lu, J., and Cai, Y. D. (2018). Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm. *Curr. Bioinform.* 13, 3–13. doi: 10.2174/1574893611666160608075753

Wang, Y., Shi, F. Q., Cao, L. Y., Dey, N., Wu, Q., Ashour, A. S., et al. (2019). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinform.* 14, 282–294. doi: 10.2174/1574893614666190304125221

Yang, R., Zhang, C., Gao, R., and Zhang, L. (2015). An effective antifreeze protein predictor with ensemble classifiers and comprehensive sequence descriptors. *Int. J. Mol. Sci.* 16, 21191–21214. doi: 10.3390/ijms160921191

Yu, C. S., and Lu, C. H. (2011). Identification of antifreeze proteins and their functional residues by support vector machine and genetic algorithms based on n-peptide compositions. *PLoS ONE* 6:8. doi: 10.1371/journal.pone.0020445

Yuan, F., Lu, L., Zhang, Y. H., Wang, S. P., and Cai, Y. D. (2018). Data mining of the cancer-related lncRNAs GO terms and KEGG pathways by using mRMR method. *Mathmat. Biosci.* 304, 1–8. doi: 10.1016/j.mbs.2018.08.001

Zhan, X. M., Sun, D. W., Zhu, Z. W., and Wang, Q. J. (2018). Improving the quality and safety of frozen muscle foods by emerging freezing technologies: a review. *Crit. Rev. Food Sci. Nutr.* 58, 2925–2938. doi: 10.1080/10408398.2017.1345854

Zhang, N., Yu, S., Guo, Y., Wang, L., Wang, P., and Feng, Y. (2018). Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinform.* 13, 50–56. doi: 10.2174/1574893611666160608102537

Zhao, X. W., Ma, Z. Q., and Yin, M. H. (2012). Using support vector machine and evolutionary profiles to predict antifreeze protein sequences. *Int. J. Mol. Sci.* 13, 2196–2207. doi: 10.3390/ijms13022196

Zhu, X. J., Feng, C. Q., Lai, H. Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 24 April 2020
doi: 10.3389/fgene.2020.00400

A Literature Review of Gene Function Prediction by Modeling Gene Ontology

Yingwen Zhao[1], *Jun Wang*[1], *Jian Chen*[2], *Xiangliang Zhang*[3], *Maozu Guo*[4]* and *Guoxian Yu*[1,3]*

[1] College of Computer and Information Science, Southwest University, Chongqing, China, [2] State Key Laboratory of Agrobiotechnology and National Maize Improvement Center, China Agricultural University, Beijing, China, [3] CBRC, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, [4] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China

Annotating the functional properties of gene products, i.e., RNAs and proteins, is a fundamental task in biology. The Gene Ontology database (GO) was developed to systematically describe the functional properties of gene products across species, and to facilitate the computational prediction of gene function. As GO is routinely updated, it serves as the gold standard and main knowledge source in functional genomics. Many gene function prediction methods making use of GO have been proposed. But no literature review has summarized these methods and the possibilities for future efforts from the perspective of GO. To bridge this gap, we review the existing methods with an emphasis on recent solutions. First, we introduce the conventions of GO and the widely adopted evaluation metrics for gene function prediction. Next, we summarize current methods of gene function prediction that apply GO in different ways, such as using hierarchical or flat inter-relationships between GO terms, compressing massive GO terms and quantifying semantic similarities. Although many efforts have improved performance by harnessing GO, we conclude that there remain many largely overlooked but important topics for future research.

Keywords: gene ontology, gene function prediction, functional genomics, directed acyclic graph, inter-relationships, semantic similarity

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Lei Deng,
Central South University, China
Jiajie Peng,
Northwestern Polytechnical University, China

***Correspondence:**
Maozu Guo
guomaozu@bucea.edu.cn
Guoxian Yu
gxyu@swu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *02 January 2020*
Accepted: *30 March 2020*
Published: *24 April 2020*

Citation:
Zhao Y, Wang J, Chen J, Zhang X, Guo M and Yu G (2020) A Literature Review of Gene Function Prediction by Modeling Gene Ontology. Front. Genet. 11:400. doi: 10.3389/fgene.2020.00400

1. INTRODUCTION

Functional annotations of gene products, i.e., proteins and RNAs, can promote the progress of drug development (Barabási et al., 2011; Xuan et al., 2019), disease analysis (Kissa et al., 2015; Zeng et al., 2015; Zhang et al., 2019), gene set enrichment analysis (Zheng and Wang, 2008; Mi et al., 2013), and many other domains (Radivojac et al., 2013; Jiang et al., 2016; Shehu et al., 2016; Zhou et al., 2019). Advances in bio-technology make it possible to perform high-throughput experiments, which yield diverse functional information about gene products, at decreasing costs. The key task has shifted from collecting such data to analyzing the data with a unified functional description scheme. To address this problem, some paradigms (Ashburner et al., 2000; Ruepp et al., 2004; Dessimoz and Škunca, 2017) aim to describe the functional properties of gene products in a formal and species neutral way, as well as to assist computational gene function prediction. Among these paradigms, Gene Ontology (GO) (Ashburner et al., 2000) and MIPS Functional Catalog (FunCat) (Ruepp et al., 2004) are the most often used. Compared with FunCat, GO is more comprehensive, is continuously

updated, has more affiliated functional annotations, and is more widely used. Therefore, we focus on function prediction methods using GO.

GO is composed of three ontologies: molecular functional ontology (MFO), biological process ontology (BPO), and cellular component ontology (CCO) (Ashburner et al., 2000). MFO describes the elemental activities of a gene product at the molecular level (i.e., binding and catalysis); BPO captures the beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms; CCO describes the parts of cells and their extracellular environments. Each ontology consists of a set of ontological terms (GO terms), which are organized in a hierarchy, or directed acyclic graph (DAG), as shown in **Figure 1**. This DAG can be generated from the ontology file with moderate scripts (i.e., Matlab, R, and Python). In the **Supplementary Material**, we provide some exemplar codes for generating an association matrix from GO and to visualize the Ontology. Each GO term is defined by a unique alphanumeric identifier and can be viewed as a vertex of the graph, and the function is described using controlled words. The edge encodes the relationships (*is a*, *part of*, and *regulate*) between GO terms. For example, "GO:0043473" represents the pigmentation, and "GO:0048066" describes the developmental pigmentation; the two terms are connected by a line with "I," which means that the developmental pigmentation *is a* subtype of pigmentation.

GO annotation is another component of GO, and it stores the *currently* known functional knowledge of gene products. Each positive annotation relates a gene with a GO term, and indicates the gene product carries out the function described by this term. Similarly, each negative annotation indicates the gene product does not perform the function described by this term. The GO consortium (Ashburner et al., 2000) independently or collaboratively annotate genes with GO terms from model organisms (or species) of wide interest among biologists, such as *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, and so on. However, our current knowledge about the functional taxonomy of gene products is still immature. Therefore, both the GO hierarchy and annotations are regularly updated with new knowledge and archived for reference. The collected GO annotations are still quite incomplete, imbalanced, and rather shallow (Rhee et al., 2008; Thomas et al., 2012; Dessimoz and Škunca, 2017). For example, different species have different distributions of GO annotations; *zebrafish* is heavily studied in terms of developmental biology and embryogenesis, while *rat* is the standard model for toxicology (Dessimoz and Škunca, 2017). The portion of negative annotations is much smaller than positive ones, because a negative result may be due to inadequate experimental conditions and is often deemed as less useful and publishable than a positive annotation. By December 2019, GO included more than 45,000 terms, and each gene was only annotated with several or dozens of these terms. Therefore, it is rather difficult to accurately infer the associations between the genes and the many GO terms.

Each GO term can be modeled as a semantic label and, thus, the gene function prediction task can be treated as a classification problem to determine whether the label is positive for the gene or not. Early gene function prediction solutions simply utilized this annotation information (Schwikowski et al., 2000; Hvidsten et al., 2001; Raychaudhuri et al., 2002; Schug et al., 2002; Troyanskaya et al., 2003; Karaoz et al., 2004), and converted the problem into a plain binary (or multi-class) classification task (Hua and Sun, 2001; Lanckriet et al., 2003; Leslie et al., 2004). Such methods ignored the correlations between the GO terms and the imbalanced characteristics of terms; therefore, their accuracy was low. Since a gene is often simultaneously annotated with a set of structurally organized GO terms, some researchers model

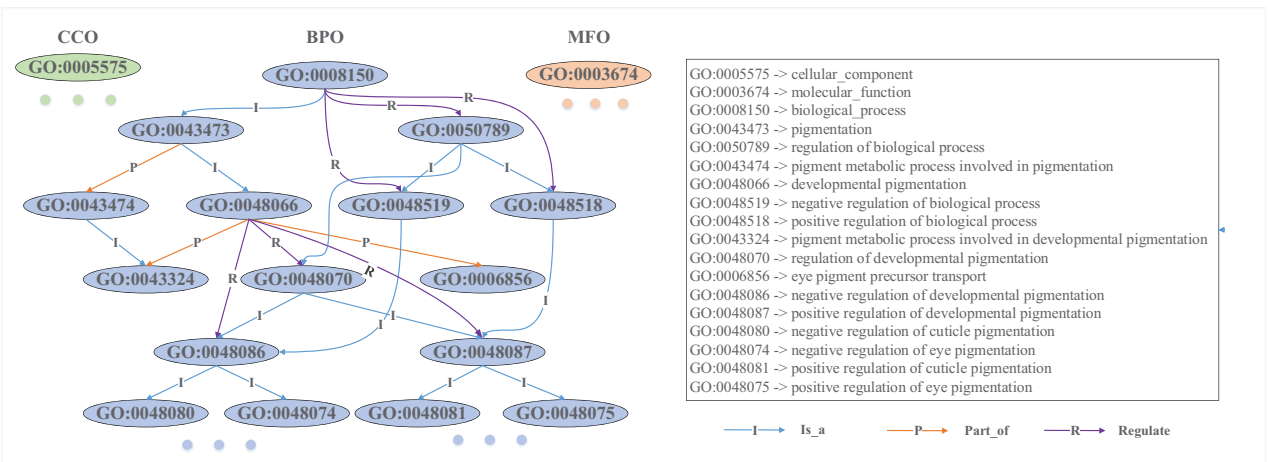


FIGURE 1 | Snapshot of a directed acyclic graph from Gene Ontology. Each ontological term is represented by an alphanumeric identifier, and its biological function is described by controlled words. These GO terms are hierarchically connected with different types of directed edges. The level of a GO term in the DAG is determined by its furthest distance to the root GO term ("GO:0008150" in BPO, "GO:0005575" in CCO, and "GO:0003674" in MFO). For example, "GO:0048087" is a direct child and also a grandson of "GO:0048066," and its furthest distance to the root term is 5, while "GO:0006856" is another direct child of "GO:0048066" and its furthest distance to the root is 4, so "GO:0006856" is plotted at a higher level than "GO:0048087".

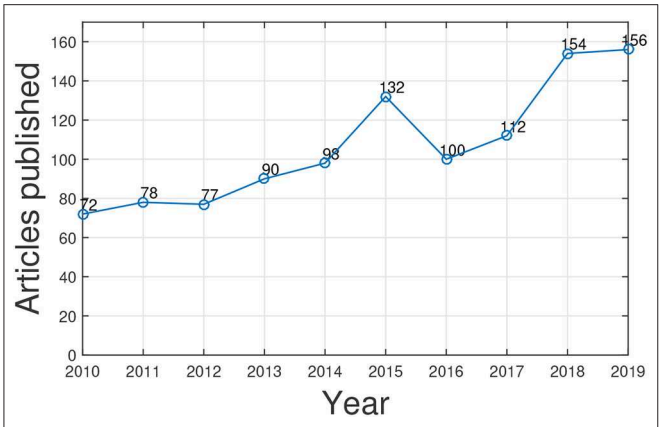


FIGURE 2 | The number of published papers related to GO-based gene function prediction over 10 years.

gene function prediction as a multi-label or structural output prediction task (Barutcuoglu et al., 2006; Obozinski et al., 2008; Zhang and Zhou, 2014; Kahanda and Ben-Hur, 2017). Others attempted to use the inter-relationships among GO terms, and introduced a variety of solutions based on multi-label learning. These generally obtained an improved accuracy (Mostafavi et al., 2008; Mostafavi and Morris, 2010; Yu et al., 2012a, 2015a).

We utilized Web of Science[1] to search articles related to gene function prediction using GO published in the past 10 years through a keyword search: "gene ontology and gene function prediction." The statistic counts are shown in **Figure 2**. We can see that research interest in this topic is increasing. As the need of human knowledge (i.e., GO and its annotations) for artificial intelligence in biology increases, we believe the study of GO for gene function prediction and for other biomedical data mining tasks will be fast growing. Several excellent surveys provide a comprehensive literature summation of the progress in gene function prediction (a.k.a. protein function prediction) and the studies of GO from different perspectives (Pandey et al., 2006; Tiwari and Srivastava, 2014; Valentini, 2014; Mazandu et al., 2016; Shehu et al., 2016; Dessimoz and Škunca, 2017). However, to the best of our knowledge, none of them focus on harnessing GO for gene function prediction.

Therefore, we give a comprehensive review of GO-based gene function prediction methods (categorized in **Figure 3**). The three main issues in gene function prediction are summarized on the left side of **Figure 3**. Categories of computational methods that combat one or two of these issues are on the right side of **Figure 3**. Each of these methods is detailed in the following sections.

The rest of this review is organized as follows. We introduce the workflow of gene function prediction, conventions in GO and typical evaluation metrics in section 2. In section 3, we categorize the existing GO-based gene function prediction methods. In section 4, we summarize remaining issues, as well as some interesting but less explored topics in gene function prediction. Section 5 concludes the survey.

[1]webofknowledge.com

2. RELATED KNOWLEDGE

Gene function prediction methods mainly utilize the structure of GO and biological features (including nucleotide/amino acids sequences, gene expression, and interaction data, etc.) of genes. Therefore, we first review the basic workflow of gene function prediction, introduce the *True Path Rule*, and evidence codes from GO, and then present the widely-used evaluation metrics for gene function prediction.

2.1. The Workflow of Gene Function Prediction

The GO file and annotation files are publicly accessible at http://geneontology.org/. They are regularly updated and archived. GO can be represented by a DAG ($\mathbf{G} \in \mathbb{R}^{m\times m}$ for m terms). The GO annotations are usually encoded by a gene-term association matrix ($\mathbf{Y} \in \mathbb{R}^{n\times m}$ for n genes with respect to m GO terms). If gene i is annotated with t or t's descendants, then $\mathbf{Y}(i,t) = 1$; if this gene is not annotated with t or its ancestor, then $\mathbf{Y}(i,t) = -1$; otherwise, $\mathbf{Y}(i,t) = 0$. We want to remark that $\mathbf{Y}(i,t) = 0$ simply indicates that till now there is no evidence that this gene does or does not carry out the function related to term t. This specification is based on the incompleteness and open-world assumption of GO annotations (Schnoes et al., 2013; Dessimoz and Škunca, 2017). If $\mathbf{X} \in \mathbb{R}^{n\times d}$ stores the numeric features of these genes, then the function prediction task can be seen as a classification task that makes use of $\mathbf{Y}$ and input pattern $\mathbf{X}$ to train a model, which can predict the association probabilities between these (or new) genes and GO terms.

Existing methods of computational gene function prediction generally focus on the three tasks (illustrated in **Figure 4**): (i) predicting *missing* (new) annotations, which updates some entries in $\mathbf{Y}$ with value 0 into 1 to identify new functional annotations of genes; (ii) identifying *noisy* annotations, which updates some entries in $\mathbf{Y}$ with value 1 into -1 to remove these false positive annotations; (iii) predicting *negative* examples, which updates some entries in $\mathbf{Y}$ with value 0 into -1 to state that the gene clearly does not carry out this function. The first task has been extensively studied, while the latter two tasks are attracting research interest.

The evaluation protocol for gene function prediction is generally performed one of two ways. One way is called *history to recent*, which takes advantage of previously archived GO annotations to train a model and evaluate the model's predictions by referring to more recent GO annotations. The second way is called *dataset partition* (or cross-validation), which divides the archived GO annotations into two (or three) sets, the first one (or two) sets for training (or tuning) the predictor, and the remaining set for testing the predictor. There are three main differences between the two ways. First, from the view of selecting training and testing sets, the *history to recent* evaluation is affected by the time span, since GO annotations are regularly updated. A time span of one or 2 years is often adopted. The *dataset partition* evaluation is influenced by the proportion of training and testing sets; a higher proportion of training sets generally gives better results. Second, from the prediction results, the *history to recent* way evaluates the fixed, recent annotations and, thus, it does

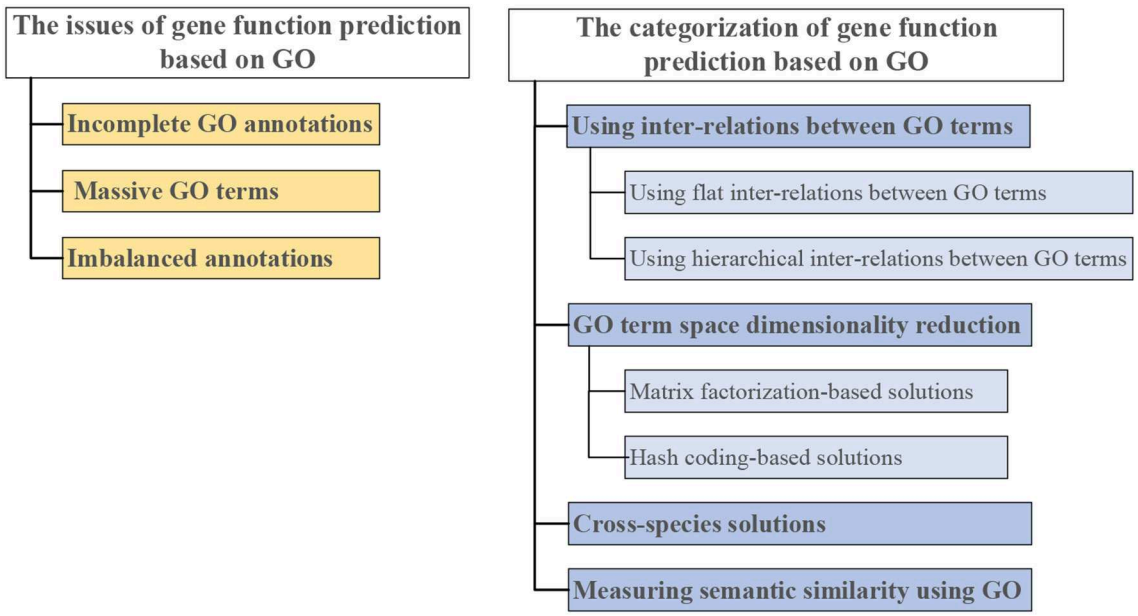


FIGURE 3 | Three issues in gene function prediction **(left)**, and categorization of existing computational solutions based on GO **(right)**.

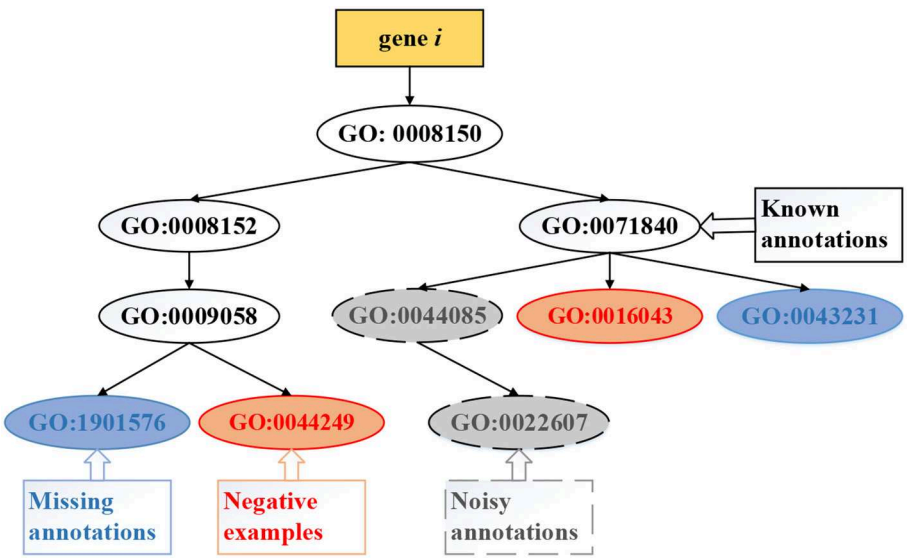


FIGURE 4 | Exemplar tasks of gene function prediction, which include predicting *missing*, *negative*, and *noisy* annotations.

not have a variance. In contrast, the *dataset partition* evaluation has to repeat multiple, independent runs to avoid the impact of random partition, and the average results and variances are both influenced. The results obtained in the *history to recent* evaluation are generally better than those obtained by the *dataset partition* evaluation. That is because *history to recent* evaluation uses all the genes and annotations for training, while *dataset partition* only uses genes in the training set and excludes genes in the testing set. Third, from the application view, the *history to recent* evaluation is deemed as more realistic and is more popular. Since GO annotations are regularly updated, the *history to recent* can reflect the potential of the model with up-to-data annotations. In contrast, the *dataset partition* may suffer from a circular prediction caused by the complex inter-connections between the partitioned training and testing sets.

2.2. Conventions in GO

2.2.1. True Path Rule

The *True Path Rule* is one of the most important rules in GO (Blake, 2013), and should be respected in gene function

prediction. If a gene is annotated with GO term t, then this gene is also annotated with t's ancestor terms. Conversely, if this gene does not have the function described by t, then it should not be annotated with t's descendant terms other. From this rule, we have

$$p(t|par(t)) \geq p(t|gpar(t)) \tag{1}$$

$$p(t|gpar(t)) \geq p(t|uncle(t)) \tag{2}$$

where $par(t)$ denotes the parent term of term t, $gpar(t)$ is the grandparent term of t, and $uncle(t)$ is the uncle (parent's sibling) term of t. $p(t|par(t))$ is the conditional probability that a gene is annotated with t given this gene is already annotated with $par(t)$. These equations imply that if a gene is annotated with GO terms $par(t)$ [or $uncle(t)$], then this gene is also annotated with $gpar(t)$ (if any), but not vice versa.

Given the structural relationships between terms, gene function prediction can be viewed as a structure output or multi-label learning problem (Barutcuoglu et al., 2006; Obozinski et al., 2008; Yu et al., 2012a; Zhang and Zhou, 2014; Kahanda and Ben-Hur, 2017; Kulmanov et al., 2017). The structure or multi-label predictions are consistent if they obey the *True Path Rule* or satisfy Equations (1, 2). According to this rule, a positive prediction for a term but a negative prediction for its ancestor terms with respect to the same gene are inconsistent predictions. In other words, a positive prediction for a term implies positive predictions for all the ancestors, and a negative prediction implies negative associations for all the descendant terms.

2.2.2. Evidence Code

Each GO annotation is tagged with one or more evidence codes, which state the type of evidence (or source) from which the annotation is collected. GO adopts 21 evidence codes and groups them into four categories: (i) *Experimental*: EXP (Inferred from Experiment), IDA (Inferred from Direct Assay), IPI (Inferred from Physical Interaction), IMP (Inferred from Mutant Phenotype), IGI (Inferred from Genetic Interaction), and IEP (Inferred from Expression Pattern); (ii) *Computational*: ISS (Inferred from Sequence or structural Similarity), ISO (Inferred from Sequence Orthology), ISA (Inferred from Sequence Alignment), ISM (Inferred from Sequence Model), IGC (Inferred from Genomic Context), IBA (Inferred from Biological aspect of Ancestor), IBD (Inferred from Biological aspect of Descendant), IKR (Inferred from Key Residues), IRD (Inferred from Rapid Divergence), RCA (Inferred from Reviewed Computational Analysis), and IEA (Inferred from Electronic Annotation); (iii) *Author*: TAS (Traceable Author Statement) and NAS (Non-traceable Author Statement); (iv) *Curatorial*: IC (Inferred by Curator) and ND (No biological Data Available) (Consortium et al., 2017). The specific meanings of these evidence codes can be found at http://www.geneontology.org/page/guide-go-evidence-codes.

Except IEA, all other evidence codes are curated by curators. Several studies investigate the quality of GO annotations from the perspective of evidence codes. Thomas et al. (2007) proposed to apply evidence codes as indicator for the reliability of annotations, and found that the annotations achieved by experimental and author statement are more reliable than others. Clark and Radivojac (2011) investigated the quality of NAS and IEA annotations, and found IEA annotations were much more reliable than NAS ones in MFO branch. Gross et al. (2009) considered evolutionary changes to evaluate stability and quality of different evidence codes. Buza (2008) estimated the annotation quality with respect to terms in BPO via a rank of evidence codes. Jones et al. (2007) found that a high false positive rate is obtained when leveraging ISS annotations and sequence data as the basis for prediction. Yu et al. (2017c) adopted evidence codes to weight the annotations and to identify the noisy annotations.

2.3. Evaluation Metrics

Multiple evaluation metrics can be adopted to quantify the results of gene function prediction. Given the complexity of gene function prediction, these metrics aim to evaluate the performance from different aspects (Radivojac et al., 2013; Jiang et al., 2016). For recent gene function prediction, *AUC*, *Fmax*, and *Smin* are recommended by CAFA (Critical Assessment of protein Function Annotation algorithms) (Radivojac et al., 2013; Jiang et al., 2016; Zhou et al., 2019). *AUC* defines different thresholds to plot the receiver-operating characteristics curve of each GO term, and then calculates the average-area value of these terms.

Fmax is the overall maximum harmonic mean of precision and recall across all possible thresholds on the predicted gene-term association matrix (Jiang et al., 2016). The formal definition of *Fmax* is

$$Fmax = \max_{\theta} \frac{2pre(\theta)rec(\theta)}{pre(\theta) + rec(\theta)} \tag{3}$$

$$pre(\theta) = \frac{1}{m(\theta)} \sum_{i=1}^{m(\theta)} \frac{TP_i}{TP_i + FP_i} \tag{4}$$

$$rec(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \tag{5}$$

where $m(\theta)$ is the number of genes, which have at least one predicted score $\geq \theta$. TP_i counts the number of true positive predictions, FP_i is the number of false positive predictions and FN_i counts the number of false negative predictions for gene i.

Smin utilizes information theoretic analogs based on the GO hierarchy to evaluate the minimum semantic distance between the predictions and ground-truths across all possible thresholds (Jiang et al., 2014). The formal definition of *Smin* is

$$Smin = \min_{\theta} \sqrt{ru(\theta)^2 + mi(\theta)^2} \tag{6}$$

$$ru(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t} IC(t) l(t \notin p_i(\theta) \wedge t \in T_i) \tag{7}$$

$$mi(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t} IC(t) l(t \in p_i(\theta) \wedge t \notin T_i) \tag{8}$$

where $IC(t)$ is the information content of the term t, which estimates a term's specificity by its frequency of annotation to genes (Lin, 1998). $p_i(\theta)$ denotes the set of terms with predicted scores $\geq \theta$ for gene i, and T_i denotes the set of terms annotated to that gene. In addition, the area under the precision-recall curve (AUPRC) is also widely used as an evaluation metric. Unlike AUC, it accounts for the imbalance in the GO terms and is also more discriminant than AUC (Guan et al., 2008; Peña-Castillo et al., 2008).

Gene function prediction can be viewed as a multi-label classification problem (Yu et al., 2012a; Zhang et al., 2012). Evaluation metrics for multi-label learning are also used to quantify the performance of gene function prediction, such as *MicroAvgF1*, *MacroAvgF1*, *RankingLoss*, *Coverage*, and *AvgPrecision*. *MicroAvg-F1* calculates the F1 measure from the predictions of different GO terms as a whole; it is more affected by the performance of terms that have more relevant genes. *MacroAvgF1* averages the F1 scores of different GO terms, and is more affected by the performance of sparse GO terms with fewer relevant genes. *RankingLoss* evaluates the average fraction of GO-term pairs that are incorrectly ranked. *Coverage* examines the search steps to cover all relevant annotations from a predicted gene-term association matrix. *AvgPrecision* evaluates the average fraction of GO terms ranked above a particular GO term. The formal definitions of these multi-label evaluation metrics can be found elsewhere (Zhang and Zhou, 2014; Gibaja and Ventura, 2015). Here, we want to highlight that these metrics quantify the results of gene function prediction from different perspectives. Any single prediction model generally cannot consistently outperform all others across each of these metrics.

3. CATEGORIZATION OF EXISTING SOLUTIONS

It is difficult to give a pure categorization of GO-based gene function prediction solutions since there are always overlaps. In this paper, we classify the existing solutions according to whether hierarchical inter-relations are used between the GO terms, and whether the massive GO terms are compressed.

3.1. Gene Function Prediction Using Inter-Relations Between GO Terms

GO uses a DAG to hierarchically organize the GO terms. This DAG encodes domain knowledge of biology. Evidence suggests that using the inter-relations between GO terms can boost the performance of gene function prediction (Tao et al., 2007; Pandey et al., 2009; Done et al., 2010). The inter-relations between GO terms can be measured from different viewpoints (Teng et al., 2013; Peng et al., 2018), and can be roughly grouped into two categories, *flat* and *hierarchical*. The flat inter-relations simply consider the occurrence of two GO terms annotated to the same genes, without explicitly using the hierarchical structure between the terms. The hierarchical inter-relations additionally account for the ontology structure. Based on the target tasks, we further divide those two methods into three subtypes based on whether they predict missing, noisy or negative annotations of genes, as listed in **Table 1**.

3.1.1. Flat Inter-Relations-Based Solutions

Early solutions simply treated gene function prediction as a binary (or multi-class) classification problem (Hua and Sun, 2001; Lanckriet et al., 2003; Leslie et al., 2004). These solutions

TABLE 1 | Categories of solutions that use different inter-relations between GO terms.

	Solutions	Inter-relations	Basic techniques
Predicting *missing* annotations	ProWL (Yu et al., 2012b)	Flat	Weak label learning
	ProDM (Yu et al., 2013a)	Flat	Weak label learning
	ProHG (Liu et al., 2016)	Flat	Random walks
	ITSS (Tao et al., 2007)	Hierarchical	Semantic similarity
	NtN (Done et al., 2010)	Hierarchical	Singular value decomposition
	dRW (Yu et al., 2015d)	Hierarchical	Random walks
	PILL (Yu et al., 2015b)	Hierarchical	Random walks
	DeepGO (Kulmanov et al., 2017)	Hierarchical	Deep learning
	NewGOA (Yu et al., 2018a)	Hierarchical	Bi-random walks
	AsyRW (Zhao et al., 2019b)	Hierarchical	Bi-random walks
Identifying *noisy* annotations	NoisyGOA (Lu et al., 2016)	Hierarchical	Semantic-based kNN
	NoGOA (Yu et al., 2017c)	Hierarchical	Sparse representation
	NFA (Lu et al., 2018)	Hierarchical	Sparse representation
Selecting *negative* annotations	ALBias (Youngs et al., 2013)	Flat	Bayesian model
	ProPN (Fu et al., 2016b)	Flat	Random walks
	SNOB (Youngs et al., 2014)	Hierarchical	Bayesian model
	NETL (Youngs et al., 2014)	Hierarchical	Topic model
	IFDR (Yu et al., 2017b)	Hierarchical	Semi-supervised linear regression
	NegGOA (Fu et al., 2016a)	Hierarchical	Random walks

accounted for neither the flat nor the hierarchical inter-relations between GO terms. As a result, they are generally less accurate than more advanced solutions (Tao et al., 2007; Pandey et al., 2009; Done et al., 2010; Liu et al., 2016), which take into account the various inter-relation among GO terms.

To predict new GO annotations of genes, Elisseeff and Weston (2002) pioneered a rank-based support vector machine that ranked relevant annotations of genes ahead of irrelevant ones. Yu et al. (2012a) and Zhang et al. (2012) used the empirical co-occurrence of two GO terms annotated to the same genes to predict new annotations of genes, and Yu et al. (2013b, 2015a) further selectively fused multiple functional networks for gene function prediction. To replenish the missing annotations of partially annotated genes, Yu et al. (2012b) proposed a gene function prediction model based on weak label learning (ProWL), in which the labels of the annotated training data were incomplete. ProWL performs the prediction for one GO term at a time. To solve this problem, Yu et al. (2013a) presented an algorithm called ProDM, which uses the maximized dependency between the features and GO annotations of genes to predict missing (or new) GO annotations of genes. Chicco et al. (2014) took advantage of the equivalence between a truncated singular value decomposition and an autoencoder neural network, and employed an autoencoder on the gene-term association matrix to predict missing annotations of genes.

To identify negative examples (or negative annotations with respect to a GO term/gene), some models (Mostafavi and Morris, 2009; Cesa-Bianchi et al., 2012) utilized heuristics to determine negative examples first and, thus, reduce the impact of an absence of negative examples in discriminative learning. Next, these models merged the selected negative examples to make a prediction. For example, Guan et al. (2008) assumed that the negative examples of a given term were all genes not annotated with that term. Mostafavi and Morris (2009) and Cesa-Bianchi et al. (2012) presumed that negative examples of a target term came from the genes which were not annotated with sibling terms of that term. This hypothesis may be often violated, since a gene may be annotated with one or more of those sibling terms as more experimental evidence becomes available. Youngs et al. (2013) introduced a model called ALBias, which assumed that the negative examples of a gene should root in the terms with the smallest probability of being annotated to that gene. The negative examples selected by ALBias can boost the performance of gene function predictions. To take advantage of information about features of genes and the available-but-scanty negative examples, Fu et al. (2016b) proposed a gene function prediction approach using positive and negative examples (ProPN). In ProPN, a signed hybrid directed graph encodes the positive and negative examples, the interactions between genes and the flat inter-relations between terms. Then, label propagation on the graph identifies the negative examples.

Irrespective of the target task, these solutions generally focus on using the co-occurrence of GO terms annotated to the same genes. Although some of them also use the annotations augmented by *True Path Rule*, they still do not explicitly include the important hierarchical inter-relations among the GO terms.

3.1.2. Hierarchical Inter-Relations-Based Solutions

Many models use the hierarchical inter-relations between GO terms and prove that the appropriate use of inter-relations can improve the gene function prediction (Tao et al., 2007; Done et al., 2010; Yu et al., 2015b). For example, Barutcuoglu et al. (2006) organized the predictions obtained from multiple binary classifiers for different terms in a Bayesian network derived from the GO hierarchy. Valentini (2011) and Cesa-Bianchi et al. (2012) further introduced a bi-directional asymmetric flow of information based on the GO hierarchy using an ensemble method , in which the positive predictions for a node propagated to its ancestors in a recursive way, while the negative predictions propagated to its offsprings. Obozinski et al. (2008) focused on calibrating and combining independent predictions to obtain a set of probabilistic predictions that are consistent with the topology of the ontology. Kahanda and Ben-Hur (2017) proposed a structured output solution that adopted a structural kernel function.

King et al. (2003) directly applied the annotation patterns of genes to induce a decision tree or Bayesian classifier to predict gene functions. However, neither classifiers was reliable for sparse GO terms, which are annotated with too few ($\leq$10) genes. Tao et al. (2007) quantified the semantic similarity between genes by combing the hierarchical relationships between terms and known GO annotations of genes, then using a k nearest neighbor (kNN) classifier with the semantic similarity to predict unknown annotations of genes. Pandey et al. (2009) employed Lin's similarity (Lin, 1998) to capture the inter-relations between hierarchically organized terms and to infer annotations of genes. Done et al. (2010) introduced a method called NtN, which applies singular value decomposition (SVD) (Golub and Reinsch, 1971) on the gene-term association matrix, whose entries are weighted by the term frequency-inverse document frequency and GO hierarchy; thus, the semantic relationships between genes and between terms were explored and the missing associations between genes and terms were completed. Yu et al. (2015b) utilized the hierarchical and flat inter-relations among terms to predict additional annotations of partially annotated genes. However, this solution ignored GO terms in the GO hierarchy that were not yet annotated to studied genes. To solve this problem, Yu et al. (2015d) introduced a downward Random Walks model (dRW), which performed random walks on the GO hierarchy while taking the terms annotated to a gene as the initial nodes. Given the structural difference between the GO terms subgraph and the genes subgraph, Yu et al. (2018a) proposed a method called NewGOA, which used a bi-random walk strategy on a hybrid graph to predict new annotations of genes. Zhao et al. (2019b) quantified the individual walk-lengths for each node of a hybrid network composed of genes, GO terms and their hierarchical relations; then, a random walk with individual walk-lengths on the network was performed to achieve cross-species gene function prediction. Kulmanov et al. (2017) developed a deep learning-based approach that utilized the GO structure as background information to optimize the predictions.

To select negative examples, Youngs et al. (2014) proposed two algorithms: selection of negatives through observed bias

(SNOB) and negative examples from topic likelihood (NETL). SNOB approximated the empirical conditional probability between terms using both direct and GO-hierarchy augmented annotations. NTEL assumed a gene is a document and all terms affiliated with that gene are words of that document; then it used a Latent Dirichlet Allocation topic model (Blei et al., 2003) to select negative examples. Fu et al. (2016a) proposed a negative GO annotations selection approach (NegGOA) that leveraged GO hierarchy, random walks, and co-occurrence patterns of annotations to select negative examples of a gene. Experimental study has demonstrated that NegGOA suffered less from incomplete annotations than NETL or SNOB, and that the selected negative examples improved the performance of gene function prediction. Yu et al. (2017b) applied a random walk on the GO hierarchy and biological network to enrich the links between nodes, and then factorized the updated relational matrices of hierarchy and the network into two low-rank numeric matrices (one for the feature data matrix and the other for the GO label matrix), and finally imposed a semi-supervised classification on the two low-rank matrices to infer positive or negative annotations of genes.

The GO hierarchical structure has also been used to identify noisy annotations, which is a less-studied but practical topic of gene function prediction. Since GO annotations of genes are collected from different sources (like crowdsourcing), these annotations are inevitably inaccurate (Huntley et al., 2014). Lu et al. (2016) proposed a novel model (NoisyGOA) that measured the taxonomic similarity between ontological terms using the GO hierarchy and the semantic similarity between genes using annotations. Next, NoisyGOA utilized the GO annotations of a gene's neighbors to aggregate annotations of the gene. Then, it takes the positive annotations with the lowest aggregated scores as noisy annotations. However, NoisyGOA does not evaluate the reliability of different annotations, and includes noisy annotations when quantifying the semantic similarity between genes. To address that, Lu et al. (2018) preset weights for different evidence codes and upward-propagated weights to ancestor annotations via the GO hierarchy. Next, they measured the semantic similarity between genes by l_1-norm regularized sparse representation on the weighted gene-term association matrix, and took advantage of annotations of semantic neighbors to identify noisy annotations of a gene. Further, Yu et al. (2017c) introduced a more advanced and adaptive approach (NoGOA), which used evidence codes of annotations to deferentially weight annotations and sparse representation to quantify the similarity between genes to identify noisy annotations.

Overall, these solutions each model GO by using the pattern of GO annotations and/or GO hierarchy. Therefore, they generally obtain a better performance than counterparts without such modeling.

3.2. Gene Function Prediction by Compressing Massive GO Terms

GO now includes more than 45,000 GO terms, and most GO annotations of genes are sparse and incomplete. As such, predicting the associations between genes and massive terms is rather difficult. Some solutions (Emmert-Streib and Dehmer, 2009; Li et al., 2009; Yu et al., 2018a) use different techniques to utilize the GO hierarchy graph and to boost performance with respect to sparse GO terms, which are annotated to too few genes. However, they still have to handle massive GO terms. In actual fact, the huge number of GO terms also causes a heavy computation burden for GO-based semantic similarity studies (Mistry and Pavlidis, 2008; Yu et al., 2015d). To alleviate this difficulty, researchers have tried to compress massive terms, and predict gene functions in a compressed label space. Based on the adopted techniques, existing solutions can be divided into two types: (i) *matrix factorization-based* and (ii) *hashing coding-based* techniques. These methods are summarized in **Table 2**. Obviously, these solutions have some overlaps with the ones introduced in the previous subsections. These solutions demonstrate that compressing GO terms improves accuracy and may even boost efficiency (Wang et al., 2015; Yu et al., 2017e; Zhao et al., 2019a).

3.2.1. Matrix Factorization-Based Solutions

Some efforts have been made toward applying matrix factorization-based solutions to compress sparse GO terms and to infer annotations of genes (Done et al., 2010; Wang et al., 2015; Yu et al., 2017b). NtN (Done et al., 2010) and IFDR (Yu et al., 2017b) are methods already mentioned in section 3.1.2. In addition, Yu et al. (2017d) proposed ProCMF to explore the latent relationships between genes and GO terms by matrix factorization. ProCMF factorized the gene-term association matrix into two low-rank matrices, and then defined two smoothness terms on these two matrices to use multiple functional association networks of genes and flat inter-relations between GO terms. These two terms also guide the matrix factorization and the approximation of the to-be-predicted gene-term association matrix. Wang et al. (2015) introduced a method called clusDCA based on Diffusion Component Analysis (DCA) (Cho et al., 2015). clusDCA individually performed a random walk on the GO DAG and on the biological networks to capture information about the underlying structure, then

TABLE 2 | Exemplar solutions based on compressing GO terms.

	Solutions	Inter-relations
Matrix factorization	ProCMF (Yu et al., 2017d)	Flat
	clusDCA (Wang et al., 2015)	Hierarchical
	NtN (Done et al., 2010)	Hierarchical
	clusDCA (Wang et al., 2015)	Hierarchical
	ProsNet (Wang et al., 2017)	Hierarchical
	IFDR (Yu et al., 2017b)	Hierarchical
	NMFGO (Yu et al., 2020b)	Hierarchical
	ZOMF (Zhao et al., 2019c)	Hierarchical
	LSDRs (Makrodimitris et al., 2019)	Hierarchical
Hash learning	HashGO (Yu et al., 2017e)	Hierarchical
	HPHash (Zhao et al., 2019a)	Hierarchical

obtained two updated adjacency matrices. To reduce noise, it applied SVD on the two matrices to compress them into two low-dimensional matrices. After that, clusDCA optimized a relational matrix between low-dimensional matrices to explore the latent relations, and to predict the associations between genes and GO terms. clusDCA manifested a significantly improved performance on sparse terms. Yu et al. (2020b) introduced a method called NMFGO, which combined non-negative matrix factorization (NMF) (Lee and Seung, 1999) with a GO DAG regularization term to factorize the gene-term association matrix into two low-rank matrices. Next, NMFGO used the low-rank matrices to explicitly calculate the semantic similarity between genes. After that, NMFGO predicted the low-rank labels of a gene based on the low-rank labels of its semantic neighbors. Then, it restored the predictions to the original GO terms. Makrodimitris et al. (2019) recently experimentally evaluated a series of label-compression solutions based on matrix factorization and proved that compressed labels can boost the prediction performance.

However, the matrix factorization-based methods above lack interpretability of the compressed labels, and suffer from an inherent problem of thresholding both the relevant and irrelevant GO annotations from the predicted numeric gene-term association matrix. This problem is also found in multi-label learning (Pillai et al., 2013). To solve these problems, Zhao et al. (2019c) introduced a method based on zero-one matrix factorization (ZOMF). ZOMF decomposed the gene-term association matrix into two low-rank matrices with entry values restricted to one or zero, then explored the inner latent relationships between the genes and terms. Next, it defined two smoothness terms on these two low-rank matrices with respect to the gene-gene interactions and the structural relationships between terms, thus guiding the optimization of low-rank matrices. Finally, it reconstructed the association matrix using the optimized two low-rank matrices to predict gene functions. ZOMF did not need to threshold the reconstructed association probability matrix, and the compressed zero-one labels had a more intuitive explanation than compressed labels.

3.2.2. Hashing-Based Solutions

To achieve low storage and fast retrieval, hashing has been widely used in big data applications (Wang et al., 2016; Liu et al., 2019). For example, Tian et al. (2016) used hash tables to store essential information learned from GO DAG and to efficiently compute the semantic similarity of genes. Empirical studies show that hash tables-based solutions can speed up diverse semantic similarity metrics, e.g., the group-based one (Teng et al., 2013) and Best Match Average (Pesquita et al., 2008). Researchers also recently employed hashing learning techniques to convert the typical one-hot coding of massive GO terms into short binary hashing codes. For example, Yu et al. (2017e) adopted a hashing technique that preserved the graph structure from Liu et al. (2011) to represent a large set of GO terms with compact binary codes, and then computed semantic similarity between the genes using the Hamming distance to predict gene functions. However, this method did not obey the GO hierarchy very well. To solve this problem, Zhao et al. (2019a) introduced a hashing method that preserved the ontology hierarchy (HPHash), which sought a set of hash functions to maintain the GO hierarchy order and the taxonomic similarity between the terms. Then, HPHash used the hash functions to compress a high-dimensional gene-term association matrix into a low-dimensional binary matrix, and predicted the gene functions therein. HPHash improved the prediction accuracy, and can be used as a plugin to boost the BLAST-based gene function prediction (Zhang et al., 1997; You et al., 2018).

3.3. Cross-Species Solutions

GO is a community-collaborative effort in functional genomics, and GO terms are generally organized in a species-neutral way to reflect the broad domain knowledge of biology. Due to differences in the preferences of biologist and in research ethics for experiments involving humans, animals, and plants, the curated annotations of genes for different species are biased, incomplete, and imbalanced (Schnoes et al., 2013; Dessimoz and Škunca, 2017; Zhao et al., 2019b). Two species with high homology have a large number of homologous genes, which should share similar (or even identical) GO annotations (Schnoes et al., 2013). Unfortunately, contemporary homologous genes are associated with different GO terms, due to the bias of biologists and diverse focuses on different species. Therefore, it is interesting to leverage the shared GO structure and complementary annotations of genes for cross-species gene function prediction.

In the early stages, typical cross-species solutions only involved the sequence data along with BLAST and PSI-BLAST (Zhang et al., 1997), but these solutions were unreliable, and the sequence identification was <25% (Shehu et al., 2016). Eisen (1998) found that utilizing evolutionary information improved gene function prediction. Guided by this observation, some databases based on the phylogenetic trees of animal-gene families appeared, such as TreeFam (Li et al., 2006). Chikina and Troyanskaya (2011) leveraged gene sequence and expression data to identify function analogous genes, and obtained an improved accuracy. However, these solutions ignored GO. To consider GO, Mitrofanova et al. (2011) presented a GO chain-graph-based approach to improve gene function prediction, which utilized high inter-species sequence homology, the PPIs of two or more species together and the GO hierarchy to construct a heterogeneous network. But this inter-species method only considered a small number of GO terms. Park et al. (2013) demonstrated that comparing the sequences of just two genes participating in the same biological processes is somewhat inaccurate. Using other genomic data, such as gene expression, can supplement traditional sequence-similarity measures to boost the performance when evalusting biological-process functions. Some other solutions attempted more advanced sequence or physical-chemical similarity metrics to improve the function prediction (Vidulin et al., 2016; Kulmanov et al., 2017; You et al., 2018; Kulmanov and Hoehndorf, 2020). For example, You et al. (2018) recently presented the GOLabeler, which separately trained five different classifiers from five different feature descriptors on sequence data, and then combined these classifiers to make a prediction. These attempts typically assumed that the annotations of the "well-annotated" species were

complete, which is not true (Jiang et al., 2014). Moreover, they neglected the dynamic, mutually supplementary GO annotations of the close-homology species. Yu et al. (2016b) studied cross-species gene function prediction based on semantic similarity. They separately explored the prediction performance for two species with high or low homology, finding that annotations of highly-homologous species were complementary, while those of less homologous species did not complement each other. Kulmanov et al. (2017) developed a deep learning-based method (DeepGO) to predict gene function from sequences. In DeepGO, the deep learning model predicted the GO annotations of genes based on gene sequences and dependencies between GO terms. To leverage the GO annotations of different species, Zhao et al. (2019b) constructed a heterogeneous network including the GO hierarchy, intra- and inter-species subnetworks. Then, they introduced an asynchronous random work on the heterogeneous network to predict gene functions.

3.4. GO-Based Semantic-Similarity Measures and Applications

The semantic similarity between genes is quantified using GO annotations and/or GO hierarchy. It is positively correlated with the feature similarity between them, which is computed from other biological data (Pesquita et al., 2009; Yu et al., 2015d). Therefore, semantic-based (and also sequence similarity- or interaction network-based) gene function prediction has been popular in recent years (Tao et al., 2007; Yu et al., 2015d, 2016a, 2017c,e).

Semantic similarity-based methods typically use the semantic similarity to select the neighborhood genes and predict the annotations of a gene based on annotations of those neighborhood genes. ITSS (Tao et al., 2007), dRW (Yu et al., 2015d), HashGO (Yu et al., 2017e), HPHash (Zhao et al., 2019a), and NMFGO (Yu et al., 2020b) are some representative methods introduced in sections 3.1.2, 3.2.2. In addition, the semantic similarity is integrated with other feature similarities for gene function prediction (Yu et al., 2015c, 2016a). For example, Yu et al. (2016a) proposed a semantic data fusion method (SimNet), which optimized the weights of multiple functional association networks to align with a semantic-similarity kernel matrix induced from the GO annotations of genes. After that, SimNet applied these weights to fuse the networks into a composite network, and then performed random walks on the composite network to make a prediction.

Measures of the similarity between genes can be extended from taxonomic similarity measures between GO terms. Existing similarity measures between genes can be further divided into two categories (Pesquita et al., 2009), pairwise and groupwise. Pairwise measures generally employ an average combination (Lord et al., 2003), maximum combination (Sevilla et al., 2005), or best match average combination (BMA) to integrate the proximity between pairwise terms. Among them, BMA provides a good balance between the maximum and average measure, since the latter two measures are inherently influenced by the number of terms being combined (Pesquita et al., 2009). Groupwise measures directly apply set (Mistry and Pavlidis, 2008), graph (Pesquita et al., 2008; Teng et al., 2013), or vector operator to compute the similarity between two sets of terms. For example, Mistry and Pavlidis (2008) introduced a set based metric called term overlap (TO), which takes into account the ratio between the number of shared annotations and minimum number of annotations of two genes. Graph-based measures organize terms annotated to a gene by a subgraph of DAG and then use graph comparing techniques to quantify the similarity between genes, i.e., simGIC (Pesquita et al., 2008) and SORA (Teng et al., 2013). The associations between a gene and all its terms can be encoded as a binary vector; vector-based measures then directly calculate the similarity between genes on the binary vectors using traditional similarity metrics (i.e., cosine and Hamming distances). The methods mentioned above use only the GO annotations and structure, whereas Peng et al. (2018) presented a similarity measure that integrated information from gene co-function networks, the GO structure and annotations.

To facilitate effective exploration of these semantic measures, some online tools or packages have been developed for the community. Yu et al. (2010) introduced an R package called GOSemSim to efficiently compute the semantic similarity between individual GO terms, sets of GO terms, genes or gene clusters. Peng et al. (2016) developed a web tool called InteGO2 to select the most appropriate measure from a set of measures using a voting method, or to integrate measures via a meta-heuristic search method. Mazandu et al. (2015) introduced a Python portable application called A-DaGO-Fun, which assembled diverse semantic measures and biological applications using these measures.

However, most solutions based on semantic similarity are still impacted by incomplete GO annotations. For a gene without any GO annotations, its semantic similarity with other genes is zero. Another limitation of semantic similarity-based solutions is that they cannot predict new annotations for a gene without any annotations. Furthermore, semantic measures are computed with respect to massive GO terms and, thus, are less reliable with sparse annotations. To address the last issue, some efforts have been made toward compressing these terms before measuring the semantic similarity (Done et al., 2010; Yu et al., 2017e, 2020b; Zhao et al., 2019a); these were reviewed in previous subsections.

4. REMAINING CHALLENGES AND POTENTIAL TOPICS

Despite much progress, the intrinsic complexity of GO-based gene function prediction, the evolution of GO and the importance of reliable GO annotations for various domains mean that there are still interesting and challenging research directions, which deserve further efforts.

First, the GO annotations of genes are still incomplete, shallow, imbalanced across species and even noisy (Thomas et al., 2012; Dessimoz and Škunca, 2017). Since the semantic similarity between genes may not faithfully reflect the actual similarity between genes or terms with incomplete annotations, semantic similarity-based solutions can only be applied for species with sufficient annotations. Although several semantic

similarity-based solutions make specific use of the GO hierarchy, GO annotations (Tao et al., 2007; Done et al., 2010; Xu et al., 2013; Yu et al., 2015b,d) and additional data sources (Peng et al., 2018; Yu et al., 2020b) to obtain an improved performance, they are mostly based on the assumption of complete annotations. In addition, many solutions suffer from an overwhelming computational load when handling massive GO terms. Hence, more efficient and effective models are still welcomed.

Second, for massive GO terms, the models based on compressed GO terms (Done et al., 2010; Wang et al., 2015; Yu et al., 2017e, 2020b; Zhao et al., 2019a) have attracted increasing interest. Although the compressed labels allow researchers to explore and employ potential relationships between terms, more theoretically sound label-compression solutions, which enable efficient gene function prediction with improved efficiency and reliability, are still anticipated.

Third, multi-omics data can reflect gene function from different aspects and they complement each other. Some efforts have been made to combine GO and heterogeneous proteomics/genomics data (Cho et al., 2016; Yu et al., 2016a, 2017d), but they often suffer from a large number of GO terms. Therefore, they have to project heterogeneous data onto the common latent feature space, which obscures the intrinsic structures of the respective data sources. More advanced integrative solutions must integrate these heterogeneous biological data and the GO knowledge more effectively.

Fourth, due to the research priorities of biologists and animal/plant ethics, the collected GO annotations of genes are imbalanced across different species (Schnoes et al., 2013). Many species have scarce annotations, and their annotations must be electronically inferred from those of relatively well-annotated species. Some studies show that the GO annotations of homologous genes across species are complementary. One fruitful direction would be to credibly transfer annotations from several well-annotated and curated species to less-studied species.

Fifth, most existing solutions focus on predicting the new annotations of a newly-sequenced gene or the missing annotations of a gene with sparse annotations. In fact, gene function prediction relies on the known positive and negative annotations of a gene, but conventionally only the positive annotations of genes are reported and, thus, recorded in GO. Therefore, it lacks negative annotations, which limits the discriminative ability of function prediction models (Youngs et al., 2014; Fu et al., 2016a). Noisy annotations are also still largely overlooked by the community, which may mislead wet-lab experimental verification, GO enrichment analysis, and more. More efforts can be devoted into identifying noisy annotations and irrelevant (or negative) annotations of genes.

Last but not least, beside proteins, other gene products like miRNAs and lncRNAs also play important roles in many life processes and have associations with different complex diseases (Lu et al., 2008; Chen et al., 2012; Deng et al., 2019; Zou et al., 2019). Our preliminary studies (Yu et al., 2017a, 2018b; Fu et al., 2018; Wang et al., 2019) show that using GO appropriately can boost the prediction of lncRNA-disease associations, and GO has some overlaps with Disease Ontology (Schriml et al., 2011), which also adopts a DAG to hierarchically organize disease terms. For example, GO has been used to find functional similarities in genes that are overexpressed or underexpressed in diseases (Chen et al., 2013), and our empirical results showed that the exclusion of GO annotations of genes significantly compromised the precision of an lncRNA-disease association prediction (Yu et al., 2017a; Fu et al., 2018). Another issue is that alternative splicing causes a gene to be translated into different isoforms or protein variants, but GO collectively stores the associations between GO terms and genes irrespective these variants. Differentiating the GO annotations of individual isoforms can provide a deeper analysis of living processes (Li et al., 2014). Our recent study confirmed that considering the GO hierarchy also helps to identify the functions of individual isoforms (Wang et al., 2020; Yu et al., 2020a). The accumulated experiences of using GO for gene function prediction are expected to shed light on the predicted functions of other molecules (i.e., ncRNAs).

5. CONCLUSIONS

Identifying the functional roles of gene products such as proteins and RNAs is one of the fundamental tasks in the post-genomic era. Given the incomplete functional knowledge of genes, we have to admit that existing gene function prediction solutions are still no substitute for wet-lab experiments. Rather, they are an important supplementary technique. As more evidence of gene functions is accumulated from experiments, the gene function prediction solutions will become more competent.

Our survey reviews the literature of ongoing studies of gene function prediction using GO, with the aim of expediting research into reliable gene function prediction. We may neglect some important work related to GO-based computational gene function prediction, given multiplicity and diverse progress in various areas. The main challenges of gene function prediction are: (i) GO annotations that are incomplete, sparse, shallow, and imbalanced within and between species; (ii) massive structurally organized GO terms; and (iii) increasing relevant and irrelevant multi-type biological data. In summary, although various computational methods based on GO have been proposed, there are still promising topics and challenges that deserve further efforts.

AUTHOR CONTRIBUTIONS

YZ and GY drafted the manuscript. MG and GY conceived the whole program, extensively revised the manuscript, and finally approved the final manuscript. JW, JC, and XZ participated in the discussion and revision of this manuscript.

FUNDING

This work was financially supported by Natural Science Foundation of China (61872300), Fundamental Research

Funds for the Central Universities (XDJK2019B024 and XDJK2020B028), Natural Science Foundation of CQ CSTC (cstc2018-jcyjAX0228), and King Abdullah University of Science and Technology, under award number FCC/1/1976-19-01.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00400/full#supplementary-material

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet*. 25, 25–29. doi: 10.1038/75556

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet*. 12, 56–68. doi: 10.1038/nrg2918

Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. *Bioinformatics* 22, 830–836. doi: 10.1093/bioinformatics/btk048

Blake, J. A. (2013). Ten quick tips for using the gene ontology. *PLoS Comput. Biol*. 9:e1003343. doi: 10.1371/journal.pcbi.1003343

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res*. 3, 993–1022. doi: 10.1162/jmlr.2003.3.4-5.993

Buza, T. J. (2008). Gene ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Res*. 36:e12. doi: 10.1093/nar/gkm1167

Cesa-Bianchi, N., Re, M., and Valentini, G. (2012). Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. *Mach. Learn*. 88, 209–241. doi: 10.1007/s10994-011-5271-6

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012). LncRNAdisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res*. 41, D983?D986. doi: 10.1093/nar/gks1099

Chen, W.-H., Zhao, X.-M., van Noort, V., and Bork, P. (2013). Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput. Biol*. 9:e1003073. doi: 10.1371/journal.pcbi.1003073

Chicco, D., Sadowski, P., and Baldi, P. (2014). "Deep autoencoder neural networks for gene ontology annotation predictions?" in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (Newport Beach, CA), 533–540. doi: 10.1145/2649387.2649442

Chikina, M. D., and Troyanskaya, O. G. (2011). Accurate quantification of functional analogy among close homologs. *PLoS Comput. Biol*. 7:e1001074. doi: 10.1371/journal.pcbi.1001074

Cho, H., Berger, B., and Peng, J. (2015). "Diffusion component analysis: unraveling functional topology in biological networks?" in *International Conference on Research in Computational Molecular Biology* (Warsaw), 62–64. doi: 10.1007/978-3-319-16706-0_9

Cho, H., Berger, B., and Peng, J. (2016). Compact integration of multi-network topology for functional analysis of genes. *Cell Syst*. 3:540. doi: 10.1016/j.cels.2016.10.017

Clark, W. T., and Radivojac, P. (2011). Analysis of protein function and its prediction from amino acid sequence. *Proteins* 79, 2086–2096. doi: 10.1002/prot.23029

Deng, L., Wang, J., and Zhang, J. (2019). Predicting gene ontology function of human micrornas by integrating multiple networks. *Front. Genet*. 10:3. doi: 10.3389/fgene.2019.00003

Dessimoz, C., and Skunca, N. (2017). The gene ontology handbook. *Methods Mol. Biol*. 1446,3–68. doi: 10.1007/978-1-4939-3743-1

Done, B., Khatri, P., Done, A., and Draghici, S. (2010). Predicting novel human gene ontology annotations using semantic analysis. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 7, 91–99. doi: 10.1109/TCBB.2008.29

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*. 8, 163–167. doi: 10.1101/gr.8.3.163

Elisseeff, A., and Weston, J. (2002). "A kernel method for multi-labelled classification?" in *Advances in Neural Information Processing Systems* (Vancouver, BC), 681–687.

Emmert-Streib, F., and Dehmer, M. (2009). Predicting cell cycle regulated genes by causal interactions. *PLoS ONE* 4:e6633. doi: 10.1371/journal.pone.0006633

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* 34, 1529–1537. doi: 10.1093/bioinformatics/btx794

Fu, G., Wang, J., Yang, B., and Yu, G. (2016a). NegGOA: Negative go annotations selection using ontology structure. *Bioinformatics* 32, 2996–3004. doi: 10.1093/bioinformatics/btw366

Fu, G., Yu, G., Wang, J., and Maozu, G. (2016b). Protein function prediction using positive and negative example. *J. Comput. Res. Dev*. 53, 1753–1765. doi: 10.7544/issn1000-1239.2016.20160196

Gibaja, E., and Ventura, S. (2015). A tutorial on multilabel learning. ACM Comput. Surveys 47:52. doi: 10.1145/2716262

Golub, G. H., and Reinsch, C. (1971). "Singular value decomposition and least squares solutions?", in *Handbook for Automatic Computation. Die Grundlehren der mathematischen Wissenschaften (in Einzeldarstellungen mit besonderer Berï£¡cksichtigung der Anwendungsgebiete)*, Vol. 186, eds F. L. Bauer, A. S. Householder, F. W. J. Olver, H. Rutishauser, K. Samelson, and E. Stiefel (Berlin; Heidelberg: Springer), 134–151. doi: 10.1007/978-3-662-39778-7_10

Gross, A., Hartung, M., Kirsten, T., and Rahm, E. (2009). 11Estimating the quality of ontology-based annotations by considering evolutionary changes?" in *International Workshop on Data Integration in the Life Sciences* (Manchester), 71–87. doi: 10.1007/978-3-642-02879-3_7

Guan, Y., Myers, C. L., Hess, D. C., Barutcuoglu, Z., Caudy, A. A., and Troyanskaya, O. G. (2008). Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biol*. 9:S3. doi: 10.1186/gb-2008-9-s1-s3

Hua, S., and Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17, 721–728. doi: 10.1093/bioinformatics/17.8.721

Huntley, R. P., Sawford, T., Martin, M. J., and Donovan, C. (2014). Understanding how and why the gene ontology and its annotations evolve: the go within uniprot. *GigaScience* 3, 2047–217X. doi: 10.1186/2047-217X-3-4

Hvidsten, T. R., Komorowski, J., Sandvik, A. K., and Laegreid, A. (2001). Predicting gene function from gene expressions and ontologies,? in *Pacific Symposium on Biocomputing* (Hawaii: World Scientific), 299–310.

Jiang, Y., Clark, W. T., Friedberg, I., and Radivojac, P. (2014). The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. *Bioinformatics* 30, i609?i616. doi: 10.1093/bioinformatics/btu472

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., et al. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol*. 17:184. doi: 10.1186/s13059-016-1037-6

Jones, C. E., Brown, A. L., and Baumann, A. U. (2007). Estimating the annotation error rate of curated go database sequence annotations. *BMC Bioinformatics* 8:170. doi: 10.1186/1471-2105-8-170

Kahanda, I., and Ben-Hur, A. (2017). "Gostruct 2.0: Automated protein function prediction for annotated proteins?" in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Boston, MA), 60–66. doi: 10.1145/3107411.3107417

Karaoz, U., Murali, T., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R., et al. (2004). Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. U.S.A*. 101, 2888–2893. doi: 10.1073/pnas.0307326101

King, O. D., Foulger, R. E., Dwight, S. S., White, J. V., and Roth, F. P. (2003). Predicting gene function from patterns of annotation. *Genome Res*. 13, 896–904. doi: 10.1101/gr.440803

Kissa, M., Tsatsaronis, G., and Schroeder, M. (2015). Prediction of drug gene associations via ontological profile similarity with application to drug repositioning. *Methods* 74, 71–82. doi: 10.1016/j.ymeth.2014.11.017

Kulmanov, M., and Hoehndorf, R. (2020). Deepgoplus: improved protein function prediction from sequence. *Bioinformatics* 36, 422–429. doi: 10.1101/615260

Kulmanov, M., Khan, M. A., and Hoehndorf, R. (2017). DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* 34, 660–668. doi: 10.1093/bioinformatics/btx624

Lanckriet, G. R., Deng, M., Cristianini, N., Jordan, M. I., and Noble, W. S. (2003). "Kernel-based data fusion and its application to protein function prediction in yeast?" in *Pacific Symposium on Biocomputing* (Hawaii: World Scientific), 300–311. doi: 10.1142/9789812704856_0029

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Leslie, C. S., Eskin, E., Cohen, A., Weston, J., and Noble, W. S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics* 20, 467–476. doi: 10.1093/bioinformatics/btg431

Li, H., Coghlan, A., Ruan, J., Coin, L. J., Heriche, J.-K., Osmotherly, L., et al. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*. 34(Suppl. 1), D572?D580. doi: 10.1093/nar/gkj118

Li, H.-D., Menon, R., Omenn, G. S., and Guan, Y. (2014). The emerging Era of genomic data integration for analyzing splice isoform function. *Trends Genet*. 30, 340–347. doi: 10.1016/j.tig.2014.05.005

Li, X., Chen, H., Li, J., and Zhang, Z. (2009). Gene function prediction with gene interaction networks: a context graph kernel approach. *IEEE Trans. Inform. Technol. Biomed*. 14, 119–128. doi: 10.1109/TITB.2009.2033116

Lin, D. (1998). "An information-theoretic definition of similarity?" in *Proceedings of 15th International Conference on Machine Learning* (Madison, WI), 296–304.

Liu, J., Wang, J., and Yu, G. (2016). Protein function prediction by random walks on a hybrid graph. *Curr. Proteomics* 13, 130–142. doi: 10.2174/1570164613021605140043 07

Liu, W., Wang, J., Kumar, S., and Chang, S.-F. (2011). "Hashing with graphs?" in *Proceedings of the 28th International Conference on Machine Learning* (Bellevue, WA), 1–8.

Liu, X., Yu, G., Domeniconi, C., Wang, J., Ren, Y., and Guo, M. (2019). "Ranking-based deep cross-modal hashing?" in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33* (Hawaii), 4400–4407. doi: 10.1609/aaai.v33i01.33014400

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics* 19, 1275–1283. doi: 10.1093/bioinformatics/btg153

Lu, C., Chen, X., Wang, J., Yu, G., and Yu, Z. (2018). Identifying noisy functional annotations of proteins using sparse semantic similarity. *Sci. Sin. Inform*. 48, 1035–1050. doi: 10.1360/N112017-00105

Lu, C., Wang, J., Zhang, Z., Yang, P., and Yu, G. (2016). NoisyGOA: Noisy GO annotations prediction using taxonomic and semantic similarity. *Comput. Biol. Chem*. 65, 203–211. doi: 10.1016/j.compbiolchem.2016.09.005

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., et al. (2008). An analysis of human microRNA and disease associations. *PLoS ONE* 3:e3420. doi: 10.1371/journal.pone.0003420

Makrodimitris, S., van Ham, R. C., and Reinders, M. J. (2019). Improving protein function prediction using protein sequence and GO-term similarities. *Bioinformatics* 35, 1116–1124. doi: 10.1093/bioinformatics/bty751

Mazandu, G. K., Chimusa, E. R., Mbiyavanga, M., and Mulder, N. J. (2015). A-DaGO-Fun: an adaptable gene ontology semantic similarity-based functional analysis tool. *Bioinformatics* 32, 477–479. doi: 10.1093/bioinformatics/btv590

Mazandu, G. K., Chimusa, E. R., and Mulder, N. J. (2016). Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. *Brief. Bioinformatics* 18, 886–901. doi: 10.1093/bib/bbw067

Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2013). Large-scale gene function analysis with the panther classification system. *Nat. Protoc*. 8, 1551–1566. doi: 10.1038/nprot.2013.092

Mistry, M., and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics* 9:327. doi: 10.1186/1471-2105-9-327

Mitrofanova, A., Pavlovic, V., and Mishra, B. (2011). Prediction of protein functions with gene ontology and interspecies protein homology data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 8, 775–784. doi: 10.1109/TCBB.2010.15

Mostafavi, S., and Morris, Q. (2009). "Using the gene ontology hierarchy when predicting gene function?" in *Conference on Uncertainty in Artificial Intelligence* (Montreal, QC), 419–427.

Mostafavi, S., and Morris, Q. (2010). Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics* 26, 1759–1765. doi: 10.1093/bioinformatics/btq262

Mostafavi, S., Ray, D., Wardefarley, D., Grouios, C., and Morris, Q. (2008). Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*. 9(Suppl. 1), 1–15. doi: 10.1186/gb-2008-9-s1-s4

Obozinski, G., Lanckriet, G., Grant, C., Jordan, M. I., and Noble, W. S. (2008). Consistent probabilistic outputs for protein function prediction. *Genome Biol*. 9:S6. doi: 10.1186/gb-2008-9-s1-s6

Pandey, G., Kumar, V., and Steinbach, M. (2006). *Computational Approaches for Protein Function Prediction: A Survey*. Twin Cities: Department of Computer Science and Engineering; University of Minnesota.

Pandey, G., Myers, C. L., and Kumar, V. (2009). Incorporating functional inter-relationships into protein function prediction algorithms. *BMC Bioinformatics* 10:142. doi: 10.1186/1471-2105-10-142

Park, C. Y., Wong, A. K., Greene, C. S., Rowland, J., Guan, Y., Bongo, L. A., et al. (2013). Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. *PLoS Comput. Biol*. 9:e1002957. doi: 10.1371/journal.pcbi.1002957

Pe na-Castillo, L., Tasan, M., Myers, C. L., Lee, H., Joshi, T., Zhang, C., et al. (2008). A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol*. 9:S2. doi: 10.1186/gb-2008-9-s1-s2

Peng, J., Li, H., Liu, Y., Juan, L., Jiang, Q., Wang, Y., et al. (2016). InteGO2: a web tool for measuring and visualizing gene semantic similarities using gene ontology. *BMC Genomics* 17:553. doi: 10.1186/s12864-016-2828-6

Peng, J., Zhang, X., Hui, W., Lu, J., Li, Q., Liu, S., et al. (2018). Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. *BMC Syst. Biol*. 12:18. doi: 10.1186/s12918-018-0539-0

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falc ao, A. O., and Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* 9:S4. doi: 10.1186/1471-2105-9-S5-S4

Pesquita, C., Faria, D., Falcao, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput. Biol*. 5:e1000443. doi: 10.1371/journal.pcbi.1000443

Pillai, I., Fumera, G., and Roli, F. (2013). Threshold optimisation for multi-label classifiers. *Pattern Recogn*. 46, 2055–2065. doi: 10.1016/j.patcog.2013.01.012

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., et al. (2013). A large-scale evaluation of computational protein function prediction. *Nat. Methods* 10, 221–227. doi: 10.1038/nmeth.2340

Raychaudhuri, S., Chang, J. T., Sutphin, P. D., and Altman, R. B. (2002). Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res*. 12, 203–214. doi: 10.1101/gr.199701

Rhee, S. Y., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nat. Rev. Genet*. 9, 509–515. doi: 10.1038/nrg2363

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., et al. (2004). The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*. 32, 5539–5545. doi: 10.1093/nar/gkh894

Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., and Friedberg, I. (2013). Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput. Biol*. 9:e1003063. doi: 10.1371/journal.pcbi.1003063

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., et al. (2011). Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res*. 40, D940?D946. doi: 10.1093/nar/gkr972

Schug, J., Diskin, S., Mazzarelli, J., Brunk, B. P., and Stoeckert, C. J. (2002). Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res*. 12, 648–655. doi: 10.1101/gr.222902

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261. doi: 10.1038/82360

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A., et al. (2005). Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 2, 330–338. doi: 10.1109/TCBB.2005.50

Shehu, A., Barbarï£¡, D., and Molloy, K. (2016). "A survey of computational methods for protein function prediction?", in *Big Data Analytics in Genomics*, ed K. C. Wong (Cham: Springer), 225–298. doi: 10.1007/978-3-319-41279-5_7

Tao, Y., Li, J., Friedman, C., and Lussier, Y. A. (2007). Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics* 23, i529?i538. doi: 10.1093/bioinformatics/btm195

Teng, Z., Guo, M., Liu, X., Dai, Q., Wang, C., and Xuan, P. (2013). Measuring gene functional similarity based on group-wise comparison of go terms. *Bioinformatics* 29, 1424–1432. doi: 10.1093/bioinformatics/btt160

The Gene Ontology Consortium (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331?D338. doi: 10.1093/nar/gkw1108

Thomas, P. D., Mi, H., and Lewis, S. (2007). Ontology annotation: mapping genomic regions to biological function. *Curr. Opin. Chem. Biol.* 11, 4–11. doi: 10.1016/j.cbpa.2006.11.039

Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., and Blake, J. A. (2012). On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report. *PLoS Comput. Biol.* 8:e1002386. doi: 10.1371/journal.pcbi.1002386

Tian, Z., Wang, C., Guo, M., Liu, X., and Teng, Z. (2016). SGFSC: speeding the gene functional similarity calculation based on hash tables. *BMC Bioinformatics* 17:445. doi: 10.1186/s12859-016-1294-0

Tiwari, A. K., and Srivastava, R. (2014). A survey of computational intelligence techniques in protein function prediction. *Int. J. Proteomics* 2014:845479. doi: 10.1155/2014/845479

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003). A bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. U.S.A.* 100, 8348–8353. doi: 10.1073/pnas.0832373100

Valentini, G. (2011). True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 8, 832–847. doi: 10.1109/TCBB.2010.38

Valentini, G. (2014). Hierarchical ensemble methods for protein function prediction. *ISRN Bioinformatics* 2014:901419. doi: 10.1155/2014/901419

Vidulin, V., Šmuc, T., and Supek, F. (2016). Extensive complementarity between gene function prediction methods. *Bioinformatics* 32, 3645–3653. doi: 10.1093/bioinformatics/btw532

Wang, J., Liu, W., Kumar, S., and Chang, S. F. (2016). Learning to hash for indexing big data - a survey. *Proc. IEEE* 104, 34–57. doi: 10.1109/JPROC.2015.2487976

Wang, K., Wang, J., Domeniconi, C., Zhang, X., and Yu, G. (2020). Isoform function prediction based on bi-random walks on a heterogeneous network. *Bioinformatics* 36, 1864–1871.

Wang, S., Cho, H., Zhai, C., Berger, B., and Peng, J. (2015). Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* 31, i357?i364. doi: 10.1093/bioinformatics/btv260

Wang, S., Qu, M., and Peng, J. (2017). "ProSNet: Integrating homology with molecular networks for protein function prediction?" in *Pacific Symposium on Biocomputing* (Hawaii), 27–38. doi: 10.1142/9789813207813_0004

Wang, Y., Yu, G., Domeniconi, C., Wang, J., Zhang, X., and Guo, M. (2019). Selective matrix factorization for multi-relational data fusion,? in *International Conference on Database Systems for Advanced Applications*, 313–329. doi: 10.1007/978-3-030-18576-3_19

Xu, Y., Guo, M., Shi, W., Liu, X., and Wang, C. (2013). A novel insight into gene ontology semantic similarity. *Genomics* 101, 368–375. doi: 10.1016/j.ygeno.2013.04.010

Xuan, P., Sun, C., Zhang, T., Ye, Y., Shen, T., and Dong, Y. (2019). A gradient boosting decision tree-based method for predicting interactions between target genes and drugs. *Front. Genet*, 10:459. doi: 10.3389/fgene.2019.00459

You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., and Zhu, S. (2018). GOLabeler: Improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* 34, 2465–2473. doi: 10.1093/bioinformatics/bty130

Youngs, N., Penfold-Brown, D., Bonneau, R., and Shasha, D. (2014). Negative example selection for protein function prediction: the NoGo database. *PLoS Comput. Biol.* 10:e1003644. doi: 10.1371/journal.pcbi.1003644

Youngs, N., Penfold-Brown, D., Drew, K., Shasha, D., and Bonneau, R. (2013). Parametric Bayesian priors and better choice of negative examples improve protein function prediction. *Bioinformatics* 29, 1190–1198. doi: 10.1093/bioinformatics/btt110

Yu, G., Domeniconi, C., Rangwala, H., and Zhang, G. (2013a). "Protein function prediction using dependence maximization?" in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Prague: Springer), 574–589. doi: 10.1007/978-3-642-40988-2_37

Yu, G., Domeniconi, C., Rangwala, H., Zhang, G., and Yu, Z. (2012a). "Transductive multi-label ensemble classification for protein function prediction?" in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing), 1077–1085. doi: 10.1145/2339530.2339700

Yu, G., Fu, G., Lu, C., Ren, Y., and Wang, J. (2017a). BRWLDA: bi-random walks for predicting lncRNA-disease associations. *Oncotarget* 8:60429. doi: 10.18632/oncotarget.19588

Yu, G., Fu, G., Wang, J., and Guo, M. (2017b). Predicting irrelevant functions of proteins based on dimensionality reduction. *Sci. Sin. Inform.* 47, 1349–1368. doi: 10.1360/N112017-00009

Yu, G., Fu, G., Wang, J., and Zhao, Y. (2018a). NewGOA: Predicting new go annotations of proteins by bi-random walks on a hybrid graph. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 15, 1390–1402. doi: 10.1109/TCBB.2017.2715842

Yu, G., Fu, G., Wang, J., and Zhu, H. (2016a). Predicting protein function via semantic integration of multiple networks. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 13, 220–232. doi: 10.1109/TCBB.2015.2459713

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among go terms and gene products. *Bioinformatics* 26, 976–978. doi: 10.1093/bioinformatics/btq064

Yu, G., Lu, C., and Wang, J. (2017c). NoGOA: predicting noisy GO annotations using evidences and sparse representation. *BMC Bioinformatics* 18:350. doi: 10.1186/s12859-017-1764-z

Yu, G., Luo, W., Fu, G., and Wang, J. (2016b). Interspecies gene function prediction using semantic similarity. *BMC Syst. Biol.* 10:361. doi: 10.1186/s12918-016-0361-5

Yu, G., Rangwala, H., Domeniconi, C., Zhang, G., and Zhang, Z. (2013b). "Protein function prediction by integrating multiple kernels?" in *Twenty-Third International Joint Conference on Artificial Intelligence* (Beijing), 1869–1875.

Yu, G., Rangwala, H., Domeniconi, C., Zhang, G., and Zhang, Z. (2015a). Predicting protein function using multiple kernels. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 12, 219–233. doi: 10.1109/TCBB.2014.2351821

Yu, G., Wang, K., Domeniconi, C., Guo, M., and Wang, J. (2020a). Isoform function prediction based on bi-random walks on a heterogeneous network. *Bioinformatics* 36, 303–310. doi: 10.1093/bioinformatics/btz535

Yu, G., Wang, K., Fu, G., Guo, M., and Wang, J. (2020b). NMFGO: Gene function prediction via nonnegative matrix factorization with gene ontology. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 17, 238–249. doi: 10.1109/TCBB.2018.2861379

Yu, G., Wang, K., Fu, G., Wang, J., and Zeng, A. (2017d). Protein function prediction based on multiple networks collaborative matrix factorization. *J. Comput. Res. Dev.* 54, 2660–2673. doi: 10.7544/issn1000-1239.2017.20170644

Yu, G., Wang, Y., Wang, J., Fu, G., Guo, M., and Domeniconi, C. (2018b). "Weighted matrix factorization based data fusion for predicting lncRNA-disease associations?" in *IEEE International Conference on Bioinformatics and Biomedicine* (Madrid), 572–577. doi: 10.1109/BIBM.2018.8621081

Yu, G., Zhang, G., Rangwala, H., Domeniconi, C., and Yu, Z. (2012b). "Protein function prediction using weak-label learning?" in *Conference on Bioinformatics, Computational Biology and Biomedicine* (Orlando, FL), 202–209. doi: 10.1145/2382936.2382962

Yu, G., Zhao, Y., Lu, C., and Wang, J. (2017e). HashGO: hashing gene ontology for protein function prediction. *Comput. Biol. Chem.* 71, 264–273. doi: 10.1016/j.compbiolchem.2017.09.010

Yu, G., Zhu, H., and Domeniconi, C. (2015b). Predicting protein functions using incomplete hierarchical labels. *BMC Bioinformatics* 16:1. doi: 10.1186/s12859-014-0430-y

Yu, G., Zhu, H., Domeniconi, C., and Guo, M. (2015c). Integrating multiple networks for protein function prediction. *BMC Syst. Biol.* 9:S3. doi: 10.1186/1752-0509-9-S1-S3

Yu, G., Zhu, H., Domeniconi, C., and Liu, J. (2015d). Predicting protein function via downward random walks on a gene ontology. *BMC Bioinformatics* 16:271. doi: 10.1186/s12859-015-0713-y

Zeng, X., Zhang, X., and Zou, Q. (2015). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinformatics* 17, 193–203. doi: 10.1093/bib/bbv033

Zhang, J., Zhang, Z., Chen, Z., and Deng, L. (2019). Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 16, 396–406. doi: 10.1109/TCBB.2017.2701379

Zhang, M.-L., and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* 26, 1819–1837. doi: 10.1109/TKDE.2013.39

Zhang, X. F., Dai, D. Q., and Li, X. X. (2012). Protein complexes discovery based on protein-protein interaction data via a regularized sparse generative network model. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 9, 857–870. doi: 10.1109/TCBB.2012.20

Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Zhao, Y., Fu, G., Wang, J., Guo, M., and Yu, G. (2019a). Gene function prediction based on gene ontology hierarchy preserving hashing. *Genomics* 111, 334–342. doi: 10.1016/j.ygeno.2018.02.008

Zhao, Y., Wang, J., Guo, M., Zhang, X., and Yu, G. (2019b). Cross-species protein function prediction with asynchronous-random walk. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 99, 1–12. doi: 10.1109/TCBB.2019.2943342

Zhao, Y., Wang, J., Guo, M., Zhang, Z., and Yu, G. (2019c). Protein function prediction based on zero-one matrix factorixation. *Sci. Sin. Inform.* 49, 1159–1174. doi: 10.1360/N112018-00331

Zheng, Q., and Wang, X.-J. (2008). GOEAST: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic Acids Res.* 36, W358?W363. doi: 10.1093/nar/gkn276

Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., et al. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* 20, 1–23. doi: 10.1186/s13059-019-1835-8

Zou, Q., Sangaiah, A. K., and Mrozek, D. (2019). Machine learning techniques on gene function prediction. *Front. Genet.* 10:938. doi: 10.3389/978-2-88963-214-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 29 April 2020
doi: 10.3389/fbioe.2020.00350

Identifying Cell-Type Specific Genes and Expression Rules Based on Single-Cell Transcriptomic Atlas Data

Fei Yuan[1,2†], XiaoYong Pan[3†], Tao Zeng[4], Yu-Hang Zhang[5], Lei Chen[6,7], Zijun Gan[5], Tao Huang[5] and Yu-Dong Cai[1]**

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Department of Science and Technology, Binzhou Medical University Hospital, Binzhou, China, [3] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China, [4] Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China, [5] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [6] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [7] Shanghai Key Laboratory of Pure Mathematics and Mathematical Practice, East China Normal University, Shanghai, China

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Xiao Chang,
Children's Hospital of Philadelphia, United States
Lin Lu,
Columbia University, United States

***Correspondence:**
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 31 January 2020
Accepted: 30 March 2020
Published: 29 April 2020

Citation:
Yuan F, Pan X, Zeng T, Zhang Y-H, Chen L, Gan Z, Huang T and Cai Y-D (2020) Identifying Cell-Type Specific Genes and Expression Rules Based on Single-Cell Transcriptomic Atlas Data. Front. Bioeng. Biotechnol. 8:350. doi: 10.3389/fbioe.2020.00350

Single-cell sequencing technologies have emerged to address new and longstanding biological and biomedical questions. Previous studies focused on the analysis of bulk tissue samples composed of millions of cells. However, the genomes within the cells of an individual multicellular organism are not always the same. In this study, we aimed to identify the crucial and characteristically expressed genes that may play functional roles in tissue development and organogenesis, by analyzing a single-cell transcriptomic atlas of mice. We identified the most relevant gene features and decision rules classifying 18 cell categories, providing a list of genes that may perform important functions in the process of tissue development because of their tissue-specific expression patterns. These genes may serve as biomarkers to identify the origin of unknown cell subgroups so as to recognize specific cell stages/states during the dynamic process, and also be applied as potential therapy targets for developmental disorders.

Keywords: cell type, expression rule, single-cell transcriptomics, tissue development, multi-class classification

INTRODUCTION

The increasing development of next-generation sequencing technologies has prompted great research progress in the areas of genomics, epigenomics, and transcriptomics (Schuster, 2007). Numerous notable achievements have been made through macro-scale studies. Nevertheless, scientists have begun to focus on the subtle differences among individual cells originating from the same organ or tissue to identify cellular heterogeneity, which plays crucial functional roles in cancers or other complex diseases (Meacham and Morrison, 2013). Cutting-edge single-cell sequencing technologies have emerged to address longstanding biological and biomedical questions.

The human body is composed of approximately 10^{13} single cells that live harmoniously in various sites and tissues (Bianconi et al., 2013). Each single cell is the fundamental unit of living organisms, and it plays a unique role in maintaining normal biological processes. In diseases such as cancer, the abnormal alteration of one single cell can initiate the progression of tumorigenesis and the subsequent downfall of the entire organism (Nowell, 1976). Previous studies usually

focused on the analysis of bulk tissue samples, which are composed of millions of cells, to elucidate the mechanism and establish therapeutic strategies for treating diseases. However, the genomes within the cells of an individual multicellular organism are not always the same. Hence, identifying the key factors from averaged data sets is difficult. The recent developments in single cell sequencing techniques have provided insights into the detailed and comprehensive research of individual cells (Grün and van Oudenaarden, 2015).

Identifying cell components and cell types to understand cell functions is important because many organs comprise cells of various types and with interdependent functions. In addition, cell functions vary depending on the cells' active or inhibited state, and they cause changes during organ development (Serewko et al., 2002). These factors cause huge challenges in classifying and cataloging the various cells in the human body. All adult diverse cells originate from a single zygote through a series of cell divisions and fate decisions in which one cell transitions from one type to another. The changes during embryonic development are driven by intricate gene expression programming (Maston et al., 2006), which reveals specific expression patterns in different types of cells at different development stages. At present, we can assay the expression profiles of every gene within genomes across thousands of individual cells in one experiment. Hence, we are capable of rigorously classifying cell types, defining the potential function of each cell type, and predicting the behavior of cells during biological development.

Many important genes play crucial roles in tissue development or cell differentiation with specific expression patterns. For instance, laminin can mediate tissue-specific gene expression in mammary epithelia in the presence of lactogenic hormones (Streuli et al., 1995). The expression level of transcription factor from zinc finger family turns out to be stable in hematopoietic stem cells but they turns out to have quite different expression patterns in the differentiated cells like erythroid cells, and megakaryocytes (Orkin, 2004). In various mesoderm- and endoderm-derived tissues, genes in the GATA family play a critical role in adjusting tissue-specific gene expression (Kelley et al., 1993; Laverriere et al., 1994). The expression levels of toll-like receptors and some related genes, such as CD14, MyD88, and LY96, vary across different adult human tissues, including the brain, heart, placenta, prostate, and trachea (Nishimura and Naito, 2005). These genes and their specific expression patterns during development and differentiation may be applied as biomarkers to recognize specific cell stages/states during the dynamic process.

On the basis of existing single-cell profiling datasets from a transcriptomic atlas of mice (Tabula Muris Consortium, 2018), we applied our newly presented computational approach to select crucial and characteristically expressed genes, which may perform essential functions in tissue development and organogenesis. We constructed some accurate classifiers that can group millions of cells into 18 tissue types depending on their gene expression profiles. We applied the minimum redundancy maximum relevance (mRMR) (Peng et al., 2005) and Monto Carlo feature selection (MCFS) (Draminski et al., 2008) methods to identify the most relevant gene features and decision rules classifying 18 cell categories and then ranked the features characterizing gene expression levels (Peng et al., 2005; Draminski et al., 2008). The selected features provided a meaningful list of genes that may have important functions during tissue development because of their specific expression patterns in distinct tissues. Further research of these genes may clarify the detailed mechanism of tissue development. In addition, these genes can be used as biomarkers to identify the origin of some unknown subgroups of cells. They can also be applied as potential targets for developmental disorders.

MATERIALS AND METHODS

Datasets

We downloaded the single-cell expression profiles of 53,760 mouse cells in 18 tissues from Gene Expression Omnibus under accession number GSE109774 (Tabula Muris Consortium, 2018). The sample sizes of the tissues are listed in **Table 1**. The expression levels of 23,433 genes were measured using NovaSeq. We aimed to investigate the tissue differences at the single-cell level.

Feature Selection

We designed a rigorous feature selection procedure for evaluating features. The purpose was to remove unimportant features for classifying cells from different tissues and rank remaining features according to their importance. First, each cell was represented in a vector of expression values of 23,433 genes, which were reduced to 5,451 by discarding features with low mutual information (MI) to targets. Second, remaining features were further reduced to 3,384 by using Boruta feature selection (BFS) (Kursa and Rudnicki, 2010). Third, these features were ranked by using mRMR (Peng et al., 2005) and MCFS (Draminski

TABLE 1 | Sample size of each tissue.

Index	Tissue	Sample size
1	Bladder	1638
2	Brain microglia	4762
3	Brain neurons	5799
4	Colon	4149
5	Fat	5862
6	Heart	7115
7	Kidney	865
8	Liver	981
9	Lung	1923
10	Mammary	2663
11	Marrow	5355
12	Muscle	2102
13	Pancreas	1961
14	Skin	2464
15	Spleen	1718
16	Thymus	1580
17	Tongue	1432
18	Trachea	1391

et al., 2008), resulting in two feature lists, respectively. Finally, on the basis of the ranked feature lists, incremental feature selection (IFS) (Liu and Setiono, 1998) with a supervised classifier was used to select the optimum features for classifying different cell types.

Evaluating Features by MI

Important criteria should be designed to determine important features according to meaningful correlations between variables and outputs. The direct way to measure the importance of features was to evaluate their correlations to targets. MI is a widely used and accepted measurement to assess features in this regard. The MI value for two variables x and y can be calculated by

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \quad (1)$$

where $p(x)$ and $p(y)$ stand for marginal probabilistic density, and $p(x, y)$ stands for joint probabilistic density. Here, for each feature, we calculated its MI value to targets (class labels) and selected those with MI values larger than 0.02. Remaining features would be poured into the following feature selection steps.

Boruta Feature Selection

In this step, features with MI values > 0.02 were analyzed by BFS (Kursa and Rudnicki, 2010). It is a wrapper feature selection method based on random forest (RF) (Breiman, 2001) that evaluates feature importance by comparing the features with randomized ones. BFS is different from most of the other wrapper feature selection algorithms that achieve minimal errors for a supervised classifier on a small subset of features, that is, BFS selects all features that may be either strongly or weakly relevant to outcome variables.

BFS mainly creates a shuffled version of original features and then uses an RF classifier to measure the importance score of the combined shuffled and original features. Only those features with importance scores higher than those of the randomized features are selected, and these significantly correlated features are considered relevant to the outcome variables. The difference between the RF and BFS importance scores lies in the introduction of the statistical significance of variable importance. A random permutation procedure is repeated to obtain statistically robust important features. BFS proceeds as follows by repeating multiple iterations:

1. Randomness is added to the given data set by shuffling original features.
2. The shuffled data set and original data set are combined.
3. An RF classifier is trained on the combined data set, and the importance of each feature is evaluated.
4. The Z-scores of the original and shuffled features are calculated. The Z-scores of individual features are calculated as the mean of the importance scores divided by the standard error. Each real feature is evaluated in terms of whether it has a higher Z-score than the maximum shuffled feature. If so, this feature is tagged as important; otherwise, it is unimportant.
5. Finally, the algorithm stops when one of the two following conditions is met: (1) all features are either tagged as "unimportant" or "important"; (2) a predefined number of iterations is reached.

In this study, we used the Python implementation of BFS from https://github.com/scikit-learn-contrib/boruta_py, along with the default parameters. Selected features were evaluated by mRMR and MCFS methods, respectively.

Minimum Redundancy Maximum Relevance

mRMR (Peng et al., 2005; Chen et al., 2017, 2018; Li et al., 2019) is a feature selection method based on MI. The merit of this method is that it considers both the relevance between input features and targets and the redundancy between features themselves. To indicate the importance of features, they are ranked in a feature list, named mRMR feature list. The list is generated by repeatedly selecting features from the feature pool until all features have been selected. In detail, for any feature in the feature pool, calculate its MI value to targets and its average MI value to already-selected features. Then, the difference of above-mentioned two values is computed. The feature with maximum difference is selected and appended to the list. In this study, the mRMR feature list was denoted by F_m.

Monto Carlo Feature Selection

Different from mRMR method, MCFS (Draminski et al., 2008; Cai et al., 2018; Li et al., 2018; Chen et al., 2019) method evaluates the importance of features in a completely different way. This method is based on decision trees. First, it generates m bootstrap sets and t feature subsets from the original dataset. Then, one tree is grown for each combination of m bootstrap sets and t feature subsets. In total, $m \times t$ decision trees are grown. On the basis of these decision trees, we calculated the relative importance (RI) score for each input feature. The RI score is calculated in terms of how frequent a feature is involved in growing the decision trees, which can be computed by:

$$RI_f = \sum_{\tau=1}^{mt} (wAcc)^u IG(n_f(\tau)) \left(\frac{no.in\ n_f(\tau)}{no.in\ \tau} \right)^v \quad (2)$$

where f stands for a feature, $wAcc$ indicates the weighted accuracy of the decision tree τ, $IG(n_f(\tau))$ is the information gain of node $n_f(\tau)$, $no.in\ n_f(\tau)$ is the number of samples in $n_f(\tau)$, $(no.\ in\ \tau)$ represents the number of samples in tree τ. u and v are weighted factors, which is set to 1. Clearly, features with high RI values are more important than others. Accordingly, features were ranked in another feature list with the decreasing order of their RI values. For convenience, this list was denoted as F_M.

Incremental Feature Selection

Although, according to the results of mRMR and MCFS methods, we can obtain two feature lists, it is still difficult to access the optimum feature subspace for a given classifier. In view of this, IFS (Liu and Setiono, 1998) integrated with a supervised classifier was employed to select the optimum number of features for the classifier, thereby constructing the optimum classifier. On the basis of the feature list (F_m or F_M), a series of feature subsets with step 5 is generated, that is, the first feature subset has the top 5 features, the second feature subset has the top 10

features, and so on. Then, for each feature subset, a supervised classifier (e.g., RF) is trained on the samples consisting of the features from this feature subset, and the classifier is evaluated using 10-fold cross-validation (Kohavi, 1995). The classifier with the best performance is selected and termed the optimum classifier, and the features used for this classifier are called the optimum features.

Random Forest

RF (Breiman, 2001) is a supervised classifier comprising multiple decision trees, each of which is grown from a bootstrap set and a feature subset randomly selected from original features. RF has been widely used for many biological applications (Pan et al., 2010; Zhao et al., 2018; Zhao R. et al., 2019; Zhao X. et al., 2019; Zhang et al., 2019). One advantage of RF is that it does not require much effort in hyperparameter optimization; in general, only default parameters are necessary.

PART Rule Learning

Contrary to black-box machine learning models, rule learning methods can learn rules about making a prediction from the data, and these rules are easy to understand. The most widely used rules is the if–then rule; IF one condition is met, THEN a prediction is generated. These simple rules can assist experts in analyzing learned knowledge so that it is aligned with established facts.

In comparison with another widely used rule learning method RIPPER, PART (Frank and Witten, 1998) learns a rule at a time without global optimization, and it is considerably simple. PART generates multiple partial decision trees and combines the rules from the decision trees using the separate-and-conquer technique. A pruned decision tree is built, and then a rule set is generated. Under this rule set, each rule walks along each path from the root to a leaf. The separate-and-conquer technique generates a rule at a time. Then, the instances aligned with this rule are removed from the training set until all instances are covered by the learned rules. PART repeatedly grows partial decision trees instead of a fully explored tree, and each partial tree is grown as follows: (1) dividing the samples into subsets; (2) expanding all subsets until each subset is expanded to a leaf in the same way as C4.5, with the only difference being the selection of the node with the lowest entropy for expansion; and (3) backtracking is intrigued when all child nodes of internal nodes are expanded into a leaf. PART prunes the trees by checking if an internal node can be replaced with a leaf. Once a tree is built, a rule can be extracted from its leaf to the root.

RESULTS

In this study, we used several machine learning algorithms to analyze the single-cell expression profiles of mouse cells in 18 tissues. The whole procedures are illustrated in **Figure 1**.

Results of Feature Selection Procedure

There were more than 50,000 features to encode each mouse cell in 18 tissues. A rigorous feature selection procedure was necessary to analyze them. First, we evaluated the importance of each feature by its MI value to targets. Those with MI values larger than 0.02 were picked up, resulting in 5,451 features. Then, the BFS method was applied on the remaining features to further select relevant features, producing 3,384 features.

Above-obtained features were fed into mRMR and MCFS methods, respectively. Accordingly, we obtained two feature lists, which are summarized in **Supplementary Tables S1, S2**, respectively.

TABLE 2 | Performance and optimum number of features of IFS with RF when using different feature ranking methods.

Feature ranking	Number of optimum features	MCC	Overall accuracy
mRMR	2265	0.882	0.890
MCFS	1170	0.892	0.899

Results of IFS With RF

The mRMR and MCFS methods provided different rankings of the remaining 3,384 features. We used IFS with RF to analyze the ranked features and thereby obtain the optimum features for classifying different cells with RF.

First, we applied IFS with RF to select the optimum features on the basis of the mRMR feature list yielded by mRMR method. Step five was adopted to construct a series of feature subsets. On each feature subset, one RF classifier was trained and evaluated on the samples consisting of the features from this feature subset by using 10-fold cross-validation (Kohavi, 1995; Che et al., 2019; Cui and Chen, 2019; Zhou et al., 2019). The performance corresponding to the different numbers of features is given in **Supplementary Table S3**. For an easy observation, an IFS curve was plotted in **Figure 2** with Matthew's correlation coefficient (MCC) (Matthews, 1975) as Y-axis and number of features as X-axis. We can see that when the top 2,265 features were used, the RF classifier yielded a maximum MCC value of 0.882 and an overall accuracy of 0.890 (**Table 2**). The performance of such optimum classifier on 18 tissues is shown in **Figure 3**. 12 tissues received accuracies over 0.900, suggesting the good performance of such classifier.

We also applied IFS with RF to select the optimum features from the feature list produced by MCFS. The performance corresponding to the different numbers of features is provided in **Supplementary Table S4**. An IFS curve was also plotted in **Figure 2** for clearly displaying the performance of RF classifier on different numbers of top features. When top 1,170 features were adopted, the RF classifier generated the highest MCC of 0.892 and overall accuracy of 0.899 (**Table 2**), which were a litter better than those of the optimum RF classifier on the feature list yielded by mRMR method. The detailed performance of such classifier on 18 tissues is illustrated in **Figure 3**. 13 tissues were assigned accuracies exceeding 0.900. These results indicate that this optimum RF classifier yielded better performance when using much fewer features from MCFS than from mRMR.

As analyzed above, the optimum features for RF on the list yielded by mRMR method were top 2,265 features, and they

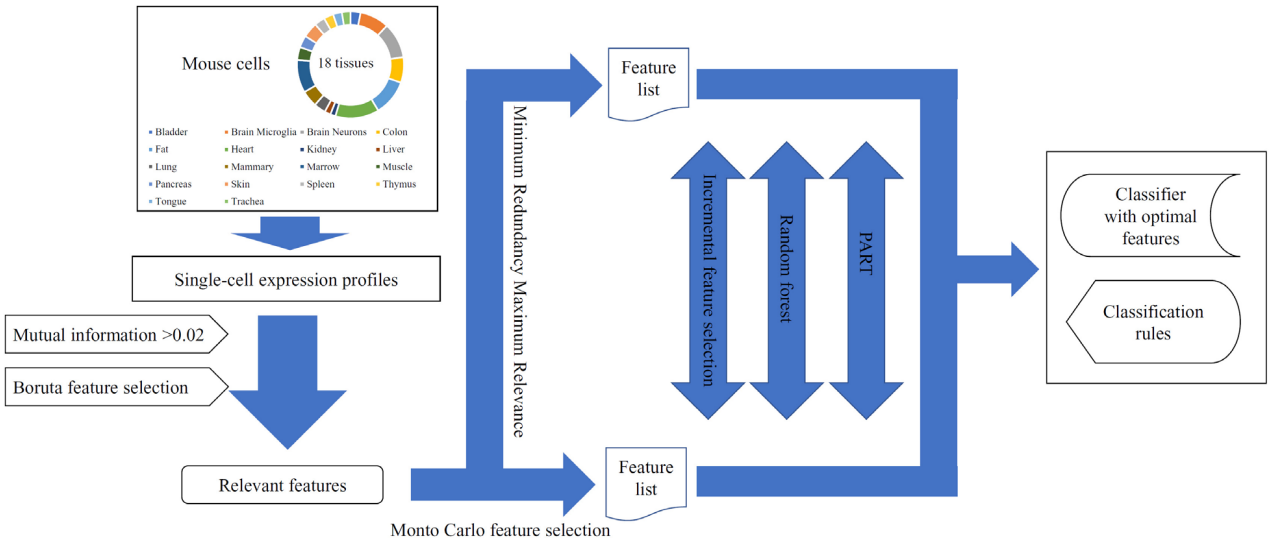


FIGURE 1 | The entire procedures for analyzing the single-cell expression profiles of mouse cells in 18 tissues.

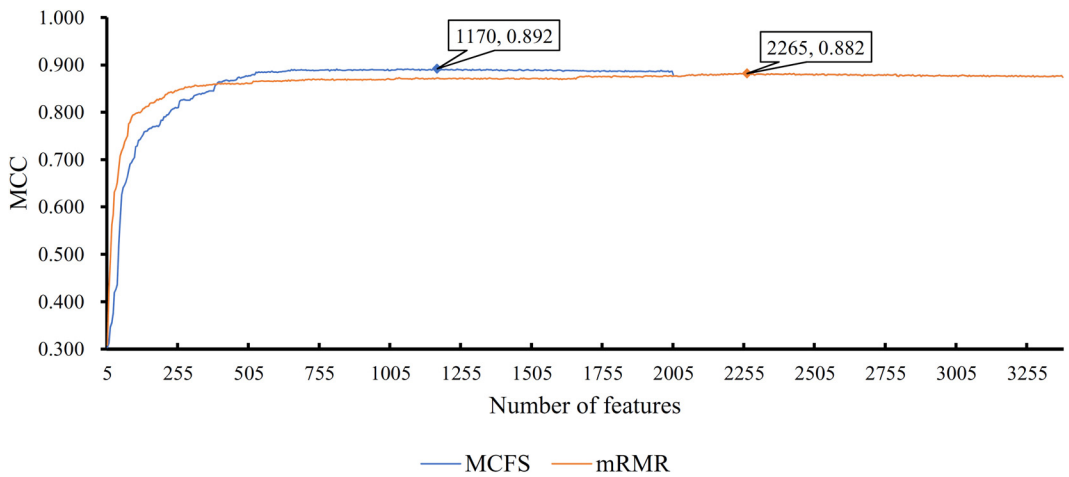


FIGURE 2 | IFS curves for IFS with RF on the feature list yielded by mRMR and MCFS methods, respectively. The best MCC for RF on the list yielded by mRMR method is 0.882 when top 2265 features are used. The highest MCC for RF on the list yielded by MCFS method is 0.892 when top 1170 features are adopted.

were top 1,170 features for RF on the list yielded by MCFS method. A Venn diagram was plotted in **Figure 4A** to show the intersection of two optimum feature sets. There were 957 common feature (genes). We used hypergeometric test to assess their overlapping significance, obtaining *P*-value less than 0.05. Thus, these two feature select methods tend to output the same important features.

Results of IFS With PART

In addition to the use of the black-box classifier RF as the supervised classifier, the rule learning classifier PART is also utilized to select the optimum features for classifying different cells. Because PART is a rule learning algorithm with low efficiency, we only tried the top 200 features on the list of mRMR method. The 10-fold cross-validation results of PART classifier on different numbers of top features is listed in **Supplementary Table S5**. An IFS curve was plotted in **Figure 5**, from which we can see that the highest MCC was 0.709 when top 200 features were used. The overall accuracy was 0.730 (**Table 3**) and the detailed performance on 18 tissues is displayed in **Figure 6**. There were four tissues receiving accuracies higher than 0.900. All these suggest that such classifier provided an acceptable performance. Thus, the PART used these 200 features to construct rules based on all mouse cells, resulting in 7085 classification rules. These rules are listed in **Supplementary Table S6**.

Similarly, we performed IFS with PART on the feature list from MCFS. We tried top 400 features this time. The performance of PART classifier corresponding to different numbers of top features is summarized in **Supplementary**

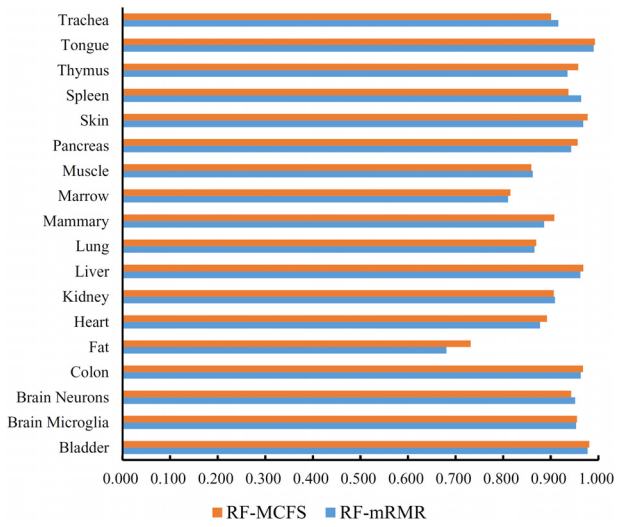


FIGURE 3 | Bar chart to show accuracies on 18 tissues yielded by the optimum RF classifiers on the feature lists of mRMR and MCFS methods.

Table S7. An IFS curve was plotted in **Figure 5**. It can be observed that when top 400 features were used, the PART classifier yielded the best MCC value of 0.781 and an overall accuracy of 0.798 (**Table 3**), which were higher than those of the PART classifier on the feature list of mRMR method. The detailed performance of such classifier on 18 tissues is shown in **Figure 6**. The accuracies on six tissues were higher than 0.900, also better than those of PART classifier generated by mRMR results. Furthermore, PART used obtained 400 features to build classification rules with all cells, generating 7,413 classification rules, which are listed in **Supplementary Table S8**.

Of the top 200 features in the mRMR feature list and top 400 features in the list of MCFS method, exactly 122 genes were common (**Figure 4B**). The overlapping significance on these two feature sets was at $P < 0.05$. Therefore, these two methods also tended to robustly select the same important features for PART.

DISCUSSION

In this study, the single-cell expression profiles of mouse cells in 18 tissues were analyzed by several machine learning algorithms. With two feature selection methods, mRMR and MCFS, two optimum RF classifiers were built and important genes were listed in two feature lists. However, the optimum RF classifiers were black-box classifiers, which can not reveal the different expression patterns of cells in different tissues. Thus, we further employed the rule learning algorithm, PART. With different feature selection methods, we obtained two groups of classification rules, which are provided in **Supplementary Tables S6, S8**. The first rule group (**Supplementary Table S6**) contained 7085 rules, involving 95 crucial features (genes) and the second group consisted of 7413 rules, using 130 crucial features (genes). In this section, we focused on some crucial features and decision rules with classification significance. These characteristics of gene expressions play key roles in tissue-specific differentiation or organ specificity.

Analysis of Top Gene Features and Decision Rules Identified Using mRMR

We identified 7085 decision rules involving 95 features via the mRMR method to distinguish 18 different types of tissues. Here, we briefly summarized some experimental evidence for the most significant features and rules in the classifier to validate the efficacy and accuracy of our prediction.

The protein coding gene **Hexb**, which was identified as the most relevant feature through the mRMR method, produced the beta subunit of the lysosomal enzyme beta-hexosaminidase that can degrade various substrates containing N-acetylgalactosamine residues. Hexb transcripts distribute widespread tissues, thus playing a housekeeping role in the enzyme. However, the expression patterns of Hexb exhibit tissue-specific differences with relatively low levels in the lung, liver, and testis, which imply its unique biological function in tissue differentiation (Yamanaka et al., 1994). Similarly, another study analyzed the tissue distribution of the Hexb mRNA in mice and revealed remarkable tissue-specific variations, with the kidney showing the highest gene expression, which are consistent with past research (Triggs-Raine et al., 1994). These findings are consistent with our expectation that Hexb displays a restricted pattern in distinct tissues and is thus an effective feature in classification.

Lgals7, also known as Galectin7, is a member of beta-galactoside-binding proteins that are implicated in modulating cell–cell and cell–matrix interactions. Differential studies indicate that lectin is specifically expressed in keratinocytes and is mainly found in stratified squamous epithelium (Magnaldo et al., 1998; Saussez and Kiss, 2006). This finding confirms our decision rules that the high expression of Lgals7 leads to the identification of skin tissues. Meanwhile, the increased expression of Lgals7 plays a positive role in cell growth and dispersal by inducing MMP9 (Demers et al., 2005). However, the functional effects of Lgals7 vary across different tissue types, and thus, the multiple roles of Lgals7 may be tissue-type dependent (Shadeo et al., 2007).

Protein coding gene **Lgals4** or galection4, as another member of the beta-galactoside-binding protein family, has a similar function to galectin7 in protein interactions, but it shows a differential expression pattern that is restricted to the intestine, colon, and rectum (Huflejt et al., 1997). It is consistent with our decision rules, which require a high level of Lgals4 expression to classify cells into the category of the colon. Galectin4 is overexpressed mainly in cells with highly differentiated polarized monolayers but is absent in less differentiated ones, suggesting its crucial roles in organogenesis and its potential as a tissue-specific marker (Huflejt and Leffler, 2003).

The protein encoded by **Krt5** (keratin 5) is a member of the keratin gene family, which comprises cytoplasmic intermediate filament proteins that are usually expressed in epithelial tissues in a differentiation-dependent manner. Keratins display a complex expression pattern that is tightly regulated by the differentiation progress of the tissue in stratified epithelia (Alam et al., 2011).

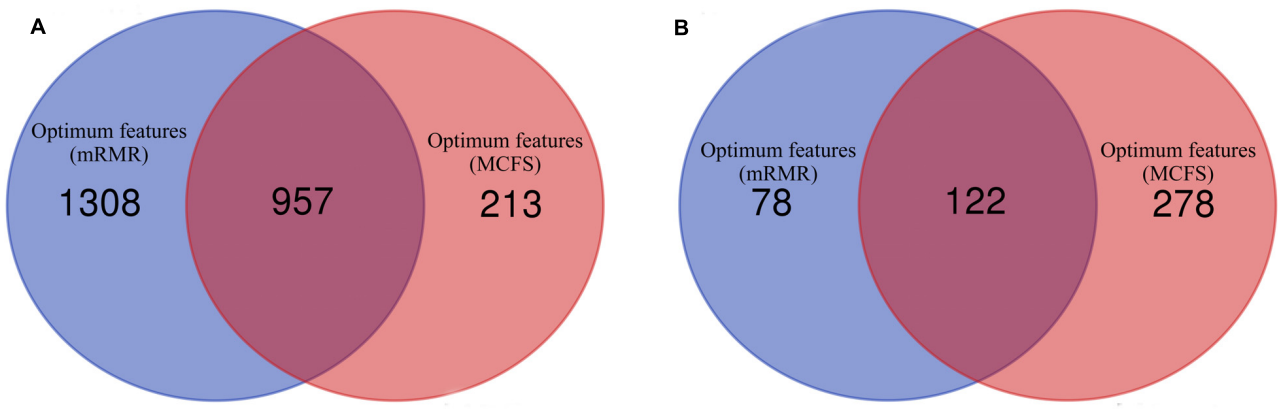


FIGURE 4 | Venn diagrams to show the intersection of optimum features for RF and PART based on the feature lists of mRMR and MCFS methods. **(A)** Venn diagram to show the intersection of optimum features for RF; **(B)** Venn diagram to show the intersection of optimum features for PART.

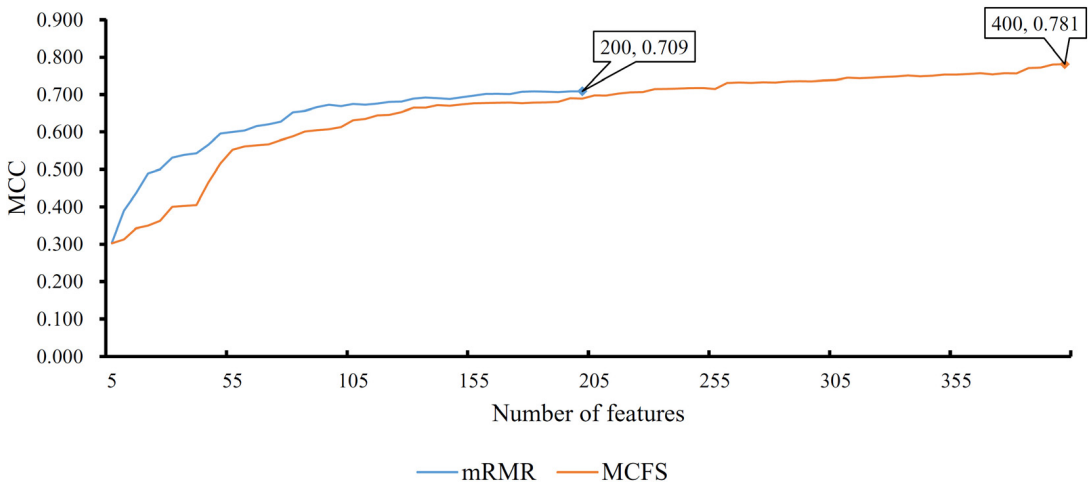


FIGURE 5 | IFS curves for IFS with PART on the feature list yielded by mRMR and MCFS methods, respectively. The best MCC for PART on the list yielded by mRMR method is 0.709 when top 200 features are used. The highest MCC for PART on the list yielded by MCFS method is 0.781 when top 400 features are adopted.

TABLE 3 | Performance and optimum number of features of IFS with PART when using different feature ranking methods.

Feature ranking	Number of optimum features	Number of classification rules	MCC	Overall accuracy
mRMR	200	7085	0.709	0.730
MCFS	400	7413	0.781	0.798

Gene ontology annotations related to Krt5 contain structural molecule activity, and mutations in this gene are associated with epidermolysis bullosa simplex (Schuilenga-Hut et al., 2003). KRT5 is one of the basal epithelial cell markers similar to KRT7 and EGFR, which follow several rules in our prediction in which Krt5 should have a low expression or even absent expression in fat tissue.

The purinergic receptor P2Y12 (**P2ry12**), which belongs to the family of P2 purinergic receptors, is a specific marker for microglial cells in the human brain (Sasaki et al., 2003). Microglial chemotaxis and the extension of microglial foot processes are significantly inhibited by P2ry12 deficiency and thus perform unique functions in microglia development (Haynes et al., 2006). Notably, a highly expressed pattern of P2ry12 contribute to the identification of brain microglia in our decision rules.

Another protein coding gene, **Ctsd** (Cathepsin D), produces a member of the A1 family of peptidases. Cathepsin is a marker of gastric differentiation, and its expression is significantly correlated with the originated histological type of gastric cancer cell line (Konno-Shimizu et al., 2013). This finding supports the potential role of Ctsd in gastric-related tissue specificity.

P53 apoptosis effector related to PMP22 (**Perp**) is a component of intercellular desmosome junctions. It plays a role in stratified epithelial integrity and cell–cell adhesion by promoting desmosome assembly (Ihrie et al., 2005; Kiseljak-Vassiliades et al., 2017). Perp plays an antiapoptotic role, and the loss of Perp function leads to strong apoptosis in the skin, indicating that this gene is required for the survival of specific

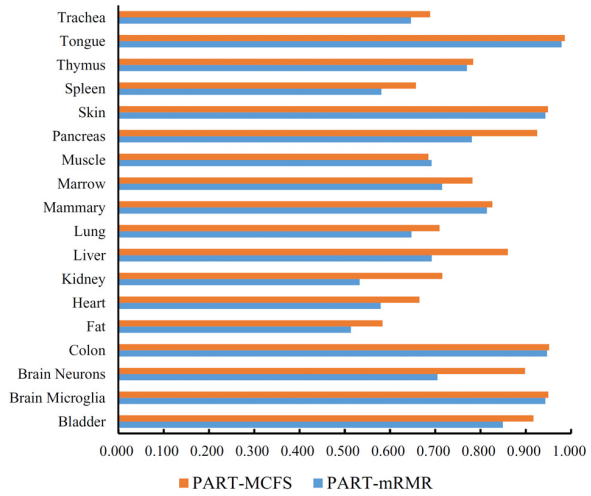


FIGURE 6 | Bar chart to show accuracies on 18 tissues yielded by the optimum PART classifiers on the feature lists of mRMR and MCFS methods.

cell types during development (Nowak et al., 2005). Notably, in the decision rules identifying heart tissues, several criteria that involve Perp, which require a relatively high expression of this gene, have experimental support. According to the immunohistochemical analysis, the Perp message is present in the intercalated discs of the cardiac muscle during embryogenesis but not in tissues containing simple epithelia, such as the lung. These results highlight the crucial role of Perp and the potential tissue-specific marker in stratified epithelia (Marques et al., 2006).

Ptprcap, also called Cd45-AP, is a transmembrane phosphoprotein that is associated with tyrosine phosphatase PRPRC/CD45, which can regulate T- and B-lymphocyte activation. It is overexpressed in PBMCs, which can enhance the phosphate activity of CD45 and increase tumor progression (Kitamura et al., 1995; Mao et al., 2008). It confirmed our predicted rules that the highly expressed pattern of Ptprcap is the indicator of marrow and thymus cell origin.

Legumain, also known as asparaginyl endopeptidase, which is encoded by the **Lgmn** gene, plays a role in the regulation of cell proliferation via its role in EGFR degradation and may be involved in the processing of proteins for MHC class II antigen presentation in the endosomal system (Manoury et al., 1998; Chen et al., 2001; Clerin et al., 2008). Legumain acts by regulating the differentiation fate of human bone marrow stromal cells, thereby regulating bone formation, which is independent of its enzymatic activity (Jafari et al., 2017). Legumain is overexpressed in bone marrow adipocytes, thereby supporting our decision rules regarding the classification of marrow, which require a highly expressed level of Lgmn, thus confirming the reliability of our predictor.

Analysis of Top Gene Features and Decision Rules Identified Using MCFS

7413 decision rules, involving 130 crucial features, were identified by MCFS and PART methods. Among the top features with the most relevance in terms of classification, some features had biological evidence of their potential tissue-specific expression patterns, which can thus be applied as biomarkers for distinguishing cell origins.

Notably, many of the features mentioned previously, including **P2ry12**, **Krt5**, **Lgals7**, **Lgals4**, and **Hexb**, were identified by mRMR and MCFS methods and have a remarkable relevance to our classifiers. These results strongly suggest that these genes have significant tissue-specific patterns and exert an important effect on the classification of different tissue cells.

DSC3 (Desmocollin 3), which ranks third among the relevant features identified by MCFS, may contribute to epidermal cell positioning by mediating the differential adhesiveness between cells that express different isoforms (Yue et al., 1995). In the decision rules for identifying lung and trachea tissues, Dsc3 should have a high expression level. RT-PCR results constantly showed that Dsc3 is expressed in the epithelium of the trachea and upregulated in the squamous cell in the lung (Nuber et al., 1996; Kettunen et al., 2004). Furthermore, desmosomal proteins are markers of epithelial differentiation (Moll et al., 1986). The expression pattern of Dsc3 changes with epidermal organization during skin development (Chidgey et al., 1997). Hence, Dsc3 may display specific expression patterns during cell differentiation and may thus support the process of distinguishing diverse stages of tissue development.

Cdx1 is a member of the caudal-related homeobox transcription factor gene family. The encoded DNA-binding protein regulates intestine-specific gene expression and enterocyte differentiation (Park et al., 2009). Homeobox genes are essential in the control of normal embryonic development. Recent publications on *Cdx1* suggested that early intestinal development, differentiation, and phenotype modulation are precisely regulated by effective transcription factors (Silberg et al., 2000). In addition, *Cdx1* is an important molecular mediator, which induces intestinal metaplasia in mouse stomach (Mutoh et al., 2004). These findings confirmed that in the criteria involving the decision rules for identifying colon tissues, highly expressed *Cdx1* indicates that the tissue may derived from colon associated tissues. In the same rules for identifying colon tissues, ***Gpx2***, which encodes the protein of the glutathione peroxidase family, requires a high expression like that of *Cdx1*. This gene is predominantly expressed in the gastrointestinal tract, and the overexpression of *Gpx2* is associated with increased differentiation and proliferation in colorectal cancer (Komatsu et al., 2001), thus contributing to colon development.

G protein-coupled receptors, such as **Gpr34**, mediate signals to the interior of the cell by activating heterotrimeric G proteins. Ubiquitous expression of Gpr34 is detectable in almost all human tissues; however the activity of promoters shows tissue-specific preference, which leads to different transcription patterns and various expression levels (Schöneberg et al., 1999). This special characteristic of Gpr34 allows its role in distinguishing different tissues and confirms that Gpr34 occurs in many decision rules with different criteria. Similarly, protein coding gene ***Cx3cr1***, which encodes fractalkine receptor, has diverse expression patterns in different cell types. The expression of

Cx3cr1 has been investigated in the mouse central nervous system, and its expression is elevated on microglia during chronic inflammation (Hughes et al., 2002). TGF-β1 plays an important role in regulating Cx3cr1 expression in rat microglia and inhibits fractalkine-stimulated signaling (Chen et al., 2002). The specific expression pattern of Cx3cr1 is consistent with our decision rules in which a high expression level indicates the category of brain microglia, although the criteria for identifying brain neurons require a low expression or absence of *Cx3cr1*.

Paired-like homeodomain 1 (**Pitx1**) encodes a member of the PITX homeobox family, which is involved in organ development and left-right asymmetry. This protein may act in the development of anterior structures and in specifying the identity or structure of hindlimbs (Logan and Tabin, 1999; Klopocki et al., 2012). Pitx1 exhibits the preferential expression in the hindlimb, and it critically modulates the potential patterning of specific hindlimb regions (Szeto et al., 1999). Pitx1 is expressed in lung epithelia cells, but its expression level varies during cancer development and progression, indicating that homeobox genes are associated with differentiation and show unique expression patterns at different development stages (Chen et al., 2007). It provides the basis for the use of Pitx1 as a potential biomarker.

Considering our single-cell profiling datasets, we carefully selected the crucial and characteristically expressed genes by using mRMR and MCFS, respectively, and their expression rules by using PART. These relevant gene features and decision rules may play essential roles in tissue development and organogenesis corresponding to 18 tissue types. Many biological studies about these may clarify the detailed mechanism of tissue development. Thus, our identified feature genes can be used as biomarkers to identify the origin of some unknown subgroups of cells, which can also be applied as potential therapy targets for developmental disorders.

CONCLUSION

This study gave an investigation on single-cell expression profiles of mouse cells in 18 tissues using several machine learning algorithms. Some essential genes that can be biomarkers for distinguishing cells of different tissues were extracted by feature selection methods and two RF classifiers were built to classify cells with high performance. In addition, two rule groups yielded by PART were reported to reveal specific expression patterns of cells in different tissues. The findings reported in this study can give a clear overview on the expression levels of different tissues.

DATA AVAILABILITY STATEMENT

The datasets for this study can be found in the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109774).

AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. FY, XP, and LC performed the experiments. FY, TZ, Y-HZ, and ZG analyzed the results. FY and XP wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This study was supported by the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Key R&D Program of China (2018YFC0910403), National Natural Science Foundation of China (31701151 and 61701298), Natural Science Foundation of Shanghai (17ZR1412500), Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), the fund of the Key Laboratory of Stem Cell Biology of Chinese Academy of Sciences (201703), Science and Technology Commission of Shanghai Municipality (STCSM) (18dz2271000), the science and technology project of Binzhou Medical University (BY2016KYQD22), and the Medicine and Health Science Technology Development Program of Shandong Province (2018WS541).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00350/full#supplementary-material

REFERENCES

Alam, H., Sehgal, L., Kundu, S. T., Dalal, S. N., and Vaidya, M. M. (2011). Novel function of keratins 5 and 14 in proliferation and differentiation of stratified epithelial cells. *Mol. Biol. Cell* 22, 4068–4078. doi: 10.1091/mbc.E10-08-0703

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., et al. (2013). An estimation of the number of cells in the human body. *Ann. Hum. Biol.* 40, 463–471. doi: 10.3109/03014460.2013.807878

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Cai, Y.-D., Zhang, S., Zhang, Y.-H., Pan, X., Feng, K., Chen, L., et al. (2018). Identification of the gene expression rules that define the subtypes in glioma. *J. Clin. Med.* 7:350. doi: 10.3390/jcm7100350

Che, J., Chen, L., Guo, Z.-H., Wang, S., and Aorigele. (2019). Drug target group prediction with multiple drug networks. *Comb. Chem. High Throughput Screen.* doi: 10.2174/1386207322666190702103927 [Epub ahead of print].

Chen, J.-M., Fortunato, M., Stevens, R. A., and Barrett, A. J. J. B. C. (2001). Activation of progelatinase A by mammalian legumain, a recently discovered cysteine proteinase. *Biol. Chem.* 382, 777–784.

Chen, L., Pan, X., Hu, X., Zhang, Y.-H., Wang, S., Huang, T., et al. (2018). Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* 143, 1731–1740. doi: 10.1002/ijc.31554

Chen, L., Pan, X., Zhang, Y.-H., Kong, X., Huang, T., and Cai, Y.-D. (2019). Tissue differences revealed by gene expression profiles of various cell lines. *J. Cell. Biochem.* 120, 7068–7081. doi: 10.1002/jcb.27977

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., et al. (2017). Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* 5, 26582–26590. doi: 10.1109/access.2017.2775703

Chen, S., Luo, D., Streit, W. J., and Harrison, J. K. (2002). TGF-β1 upregulates CX3CR1 expression and inhibits fractalkine-stimulated signaling in rat microglia. *J. Neuroimmunol.* 133, 46–55. doi: 10.1016/s0165-5728(02)00354-5

Chen, Y., Knösel, T., Ye, F., Pacyna-Gengelbach, M., Deutschmann, N., and Petersen, I. (2007). Decreased PITX1 homeobox gene expression in human lung cancer. *Lung Cancer* 55, 287–294. doi: 10.1016/j.lungcan.2006.11.001

Chidgey, M. A., Yue, K. K., Gould, S., Byrne, C., and Garrod, D. R. (1997). Changing pattern of desmocollin 3 expression accompanies epidermal organisation during skin development. *Dev. Dyn.* 210, 315–327. doi: 10.1002/(sici)1097-0177(199711)210:3<315::aid-aja11>3.0.co;2-9

Clerin, V., Shih, H. H., Deng, N., Hebert, G., Resmini, C., Shields, K. M., et al. (2008). Expression of the cysteine protease legumain in vascular lesions and functional implications in atherogenesis. *Atherosclerosis* 201, 53–66. doi: 10.1016/j.atherosclerosis.2008.01.016

Cui, H., and Chen, L. (2019). A binary classifier for the prediction of EC numbers of enzymes. *Curr. Proteomics* 16, 381–389.

Demers, M., Magnaldo, T., and St-Pierre, Y. (2005). A novel function for galectin-7: promoting tumorigenesis by up-regulating MMP-9 gene expression. *Cancer Res.* 65, 5205–5210. doi: 10.1158/0008-5472.can-05-0134

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2008). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Frank, E., and Witten, I. H. (1998). "Generating accurate rule sets without global optimization," in *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

Grün, D., and van Oudenaarden, A. (2015). Design and analysis of single-cell sequencing experiments. *Cell* 163, 799–810. doi: 10.1016/j.cell.2015.10.039

Haynes, S. E., Hollopeter, G., Yang, G., Kurpius, D., Dailey, M. E., Gan, W.-B., et al. (2006). The P2Y 12 receptor regulates microglial activation by extracellular nucleotides. *Nat. Neurosci.* 9, 1512–1519. doi: 10.1038/nn1805

Huflejt, M. E., Jordan, E. T., Gitt, M. A., Barondes, S. H., and Leffler, H. (1997). Strikingly different localization of galectin-3 and galectin-4 in human colon adenocarcinoma T84 cells Galectin-4 is localized at sites of cell adhesion. *J. Biol. Chem.* 272, 14294–14303. doi: 10.1074/jbc.272.22.14294

Huflejt, M. E., and Leffler, H. (2003). Galectin-4 in normal tissues and cancer. *Glycoconj. J.* 20, 247–255. doi: 10.1023/b:glyc.0000025819.54723.a0

Hughes, P. M., Botham, M. S., Frentzel, S., Mir, A., and Perry, V. H. (2002). Expression of fractalkine (CX3CL1) and its receptor, CX3CR1, during acute and chronic inflammation in the rodent CNS. *Glia* 37, 314–327. doi: 10.1002/glia.10037

Ihrie, R. A., Marques, M. R., Nguyen, B. T., Horner, J. S., Papazoglu, C., Bronson, R. T., et al. (2005). Perp is a p63-regulated gene essential for epithelial integrity. *Cell* 120, 843–856. doi: 10.1016/j.cell.2005.01.008

Jafari, A., Qanie, D., Andersen, T. L., Zhang, Y., Chen, L., Postert, B., et al. (2017). Legumain regulates differentiation fate of human bone marrow stromal cells and is altered in postmenopausal osteoporosis. *Stem Cell Rep.* 8, 373–386. doi: 10.1016/j.stemcr.2017.01.003

Kelley, C., Blumberg, H., Zon, L. I., and Evans, T. (1993). GATA-4 is a novel transcription factor expressed in endocardium of the developing heart. *Development* 118, 817–827.

Kettunen, E., Anttila, S., Seppänen, J. K., Karjalainen, A., Edgren, H., Lindström, I., et al. (2004). Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer. *Cancer Genet. Cytogenet.* 149, 98–106. doi: 10.1016/s0165-4608(03)00300-5

Kiseljak-Vassiliades, K., Mills, T. S., Zhang, Y., Xu, M., Lillehei, K. O., Kleinschmidt-Demasters, B., et al. (2017). Elucidating the role of the desmosome protein p53 apoptosis effector related to PMP-22 in growth hormone tumors. *Endocrinology* 158, 1450–1460. doi: 10.1210/en.2016-1841

Kitamura, K., Maiti, A., Ng, D. H., Johnson, P., Maizel, A. L., and Takeda, A. (1995). Characterization of the interaction between CD45 and CD45-AP. *J. Biol. Chem.* 270, 21151–21157. doi: 10.1074/jbc.270.36.21151

Klopocki, E., Kähler, C., Foulds, N., Shah, H., Joseph, B., Vogel, H., et al. (2012). Deletions in PITX1 cause a spectrum of lower-limb malformations including mirror-image polydactyly. *Eur. J. Hum. Genet.* 20, 705–708. doi: 10.1038/ejhg.2011.264

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the International Joint Conference on Artificial Intelligence*, (Burlington, MA: Morgan Kaufmann Publishers), 1137–1145.

Komatsu, H., Okayasu, I., Mitomi, H., Imai, H., Nakagawa, Y., and Obata, F. (2001). Immunohistochemical detection of human gastrointestinal glutathione peroxidase in normal tissues and cultured cells with novel mouse monoclonal antibodies. *J. Histochem. Cytochem.* 49, 759–766. doi: 10.1177/002215540104900609

Konno-Shimizu, M., Yamamichi, N., Inada, K.-I., Kageyama-Yahara, N., Shiogama, K., Takahashi, Y., et al. (2013). Cathepsin E is a marker of gastric differentiation and signet-ring cell carcinoma of stomach: a novel suggestion on gastric tumorigenesis. *PLoS One* 8:e56766. doi: 10.1371/journal.pone.0056766

Kursa, M. B., and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J. Stat. Softw.* 36, 1–13.

Laverriere, A. C., Macneill, C., Mueller, C., Poelmann, R. E., Burch, J., and Evans, T. (1994). GATA-4/5/6, a subfamily of three transcription factors transcribed in developing heart and gut. *J. Biol. Chem.* 269, 23177–23184.

Li, J., Chen, L., Zhang, Y. H., Kong, X., Huang, T., and Cai, Y. D. (2018). A computational method for classifying different human tissues with quantitatively tissue-specific expressed genes. *Genes* 9:449. doi: 10.3390/genes9090449

Li, J., Lu, L., Zhang, Y. H., Liu, M., Chen, L., Huang, T., et al. (2019). Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell. Biochem.* 120, 405–416. doi: 10.1002/jcb.27395

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230.

Logan, M., and Tabin, C. J. (1999). Role of Pitx1 upstream of Tbx4 in specification of hindlimb identity. *Science* 283, 1736–1739. doi: 10.1126/science.283.5408.1736

Magnaldo, T., Fowlis, D., and Darmon, M. (1998). Galectin-7, a marker of all types of stratified epithelia. *Differentiation* 63, 159–168. doi: 10.1046/j.1432-0436.1998.6330159.x

Manoury, B., Hewitt, E. W., Morrice, N., Dando, P. M., Barrett, A. J., and Watts, C. (1998). An asparaginyl endopeptidase processes a microbial antigen for class II MHC presentation. *Nature* 396, 695–699. doi: 10.1038/25379

Mao, X., Orchard, G., Mitchell, T. J., Oyama, N., Russell-Jones, R., Vermeer, M. H., et al. (2008). A genomic and expression study of AP-1 in primary cutaneous T-cell lymphoma: evidence for dysregulated expression of JUNB and JUND in MF and SS. *J. Cutan. Pathol.* 35, 899–910. doi: 10.1111/j.1600-0560.2007.00924.x

Marques, M. R., Ihrie, R. A., Horner, J. S., and Attardi, L. D. (2006). The requirement for perp in postnatal viability and epithelial integrity reflects an intrinsic role in stratified epithelia. *J. Investig. Dermatol.* 126, 69–73. doi: 10.1038/sj.jid.5700032

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* 7, 29–59. doi: 10.1146/annurev.genom.7.080505.115623

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Meacham, C. E., and Morrison, S. J. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature* 501, 328–337. doi: 10.1038/nature12624

Moll, R., Cowin, P., Kapprell, H., and Franke, W. W. (1986). Biology of disease. Desmosomal proteins: new markers for identification and classification of tumors. *Lab. Investig.* 54, 4–25.

Mutoh, H., Sakurai, S., Satoh, K., Osawa, H., Hakamata, Y., Takeuchi, T., et al. (2004). Cdx1 induced intestinal metaplasia in the transgenic mouse stomach: comparative study with Cdx2 transgenic mice. *Gut* 53, 1416–1423. doi: 10.1136/gut.2003.032482

Nishimura, M., and Naito, S. (2005). Tissue-specific mRNA expression profiles of human toll-like receptors and related genes. *Biol. Pharm. Bull.* 28, 886–892. doi: 10.1248/bpb.28.886

Nowak, M., Köster, C., and Hammerschmidt, M. (2005). Perp is required for tissue-specific cell survival during zebrafish development. *Cell Death Differ.* 12:52. doi: 10.1038/sj.cdd.4401519

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science* 194, 23–28. doi: 10.1126/science.959840

Nuber, U., Schäfer, S., Stehr, S., Rackwitz, H., and Franke, W. (1996). Patterns of desmocollin synthesis in human epithelia: immunolocalization of desmocollins 1 and 3 in special epithelia and in cultured cells. *Eur. J. Cell Biol.* 71, 1–13.

Orkin, S. H. (2004). Embryonic stem cells and transgenic mice in the study of hematopoiesis. *Int. J. Dev. Biol.* 42, 927–934.

Pan, X.-Y., Zhang, Y.-N., and Shen, H.-B. (2010). Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *J. Proteome Res.* 9, 4992–5001. doi: 10.1021/pr100618t

Park, M. J., Kim, H. Y., Kim, K., and Cheong, J. (2009). Homeodomain transcription factor CDX1 is required for the transcriptional induction of PPARγ in intestinal cell differentiation. *FEBS Lett.* 583, 29–35. doi: 10.1016/j.febslet.2008.11.030

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/tpami.2005.159

Sasaki, Y., Hoshi, M., Akazawa, C., Nakamura, Y., Tsuzuki, H., Inoue, K., et al. (2003). Selective expression of Gi/o-coupled ATP receptor P2Y12 in microglia in rat brain. *Glia* 44, 242–250. doi: 10.1002/glia.10293

Saussez, S., and Kiss, R. (2006). Galectin-7. *Cell. Mol. Life Sci.* 63, 686–697.

Schöneberg, T., Schulz, A., Grosse, R., Schade, R., Henklein, P., Schultz, G., et al. (1999). A novel subgroup of class I G-protein-coupled receptors. *Biochim. Biophys. Acta* 1446, 57–70.

Schuilenga-Hut, P. H., Vlies, P. V., Jonkman, M. F., Waanders, E., Buys, C. H., and Scheffer, H. (2003). Mutation analysis of the entire keratin 5 and 14 genes in patients with epidermolysis bullosa simplex and identification of novel mutations. *Hum. Mutat.* 21, 447–447. doi: 10.1002/humu.9124

Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18. doi: 10.1038/nmeth1156

Serewko, M. M., Popa, C., Dahler, A. L., Smith, L., Strutton, G. M., Coman, W., et al. (2002). Alterations in gene expression and activity during squamous cell carcinoma development. *Cancer Res.* 62, 3759–3765.

Shadeo, A., Chari, R., Vatcher, G., Campbell, J., Lonergan, K. M., Matisic, J., et al. (2007). Comprehensive serial analysis of gene expression of the cervical transcriptome. *BMC Genomics* 8:142. doi: 10.1186/1471-2164-8-142

Silberg, D. G., Swain, G. P., Suh, E. R., and Traber, P. G. (2000). Cdx1 and cdx2 expression during intestinal development. *Gastroenterology* 119, 961–971. doi: 10.1053/gast.2000.18142

Streuli, C. H., Schmidhauser, C., Bailey, N., Yurchenco, P., Skubitz, A. P., Roskelley, C., et al. (1995). Laminin mediates tissue-specific gene expression in mammary epithelia. *J. Cell Biol.* 129, 591–603. doi: 10.1083/jcb.129.3.591

Szeto, D. P., Rodriguez-Esteban, C., Ryan, A. K., O'connell, S. M., Liu, F., Kioussi, C., et al. (1999). Role of the Bicoid-related homeodomain factor Pitx1 in specifying hindlimb morphogenesis and pituitary development. *Genes Dev.* 13, 484–494. doi: 10.1101/gad.13.4.484

Tabula Muris Consortium. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562, 367–372. doi: 10.1038/s41586-018-0590-4

Triggs-Raine, B. L., Benoit, G., Salo, T. J., Trasler, J. M., and Gravel, R. A. (1994). Characterization of the murine β-hexosaminidase (HEXB) gene. *Biochim. Biophys. Acta* 1227, 79–86. doi: 10.1016/0925-4439(94)90110-4

Yamanaka, S., Johnson, O. N., Norflus, F., Boles, D. J., and Proia, R. L. (1994). Structure and expression of the mouse β-hexosaminidase genes. *Hexa Hexb. Genomics* 21, 588–596. doi: 10.1006/geno.1994.1318

Yue, K., Holton, J., Clarke, J., Hyam, J., Hashimoto, T., Chidgey, M., et al. (1995). Characterisation of a desmocollin isoform (bovine DSC3) exclusively expressed in lower layers of stratified epithelia. *J. Cell Science* 108, 2163–2173.

Zhang, X., Chen, L., Guo, Z.-H., and Liang, H. (2019). Identification of human membrane protein types by incorporating network embedding methods. *IEEE Access* 7, 140794–140805. doi: 10.1109/access.2019.2944177

Zhao, R., Chen, L., Zhou, B., Guo, Z.-H., Wang, S., and Aorigele. (2019). Recognizing novel tumor suppressor genes using a network machine learning strategy. *IEEE Access* 7, 155002–155013. doi: 10.1109/access.2019.2949415

Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinformatics* 14, 709–720. doi: 10.2174/1574893614666190220114644

Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* 306, 136–144. doi: 10.1016/j.mbs.2018.09.010

Zhou, J.-P., Chen, L., and Guo, Z.-H. (2019). iATC-NRAKEL: An efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. *Bioinformatics* 36, 1391–1396. doi: 10.1093/bioinformatics/btz757

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 05 May 2020
doi: 10.3389/fbioe.2020.00285

A Method for Prediction of Thermophilic Protein Based on Reduced Amino Acids and Mixed Features

Changli Feng[1], Zhaogui Ma[1], Deyun Yang[1], Xin Li[1], Jun Zhang[2] and Yanjuan Li[3]**

[1] College of Information Science and Technology, Taishan University, Tai'an, China, [2] Department of Rehabilitation, General Hospital of Heilongjiang Province Land Reclamation Bureau, Harbin, China, [3] Information and Computer Engineering College, Northeast Forestry University, Harbin, China

OPEN ACCESS

Edited by:
Yungang Xu,
The University of Texas Health Science Center at Houston, United States

Reviewed by:
Hifzur Rahman Ansari,
King Abdullah International Medical Research Center KAIMRC, Saudi Arabia
Leyi Wei,
Tianjin University, China
Bin Liu,
Beijing Institute of Technology, China

***Correspondence:**
Jun Zhang
zhangjun13902003@163.com
Yanjuan Li
liyanjuan@nefu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 15 January 2020
Accepted: 18 March 2020
Published: 05 May 2020

Citation:
Feng C, Ma Z, Yang D, Li X, Zhang J and Li Y (2020) A Method for Prediction of Thermophilic Protein Based on Reduced Amino Acids and Mixed Features. Front. Bioeng. Biotechnol. 8:285. doi: 10.3389/fbioe.2020.00285

The thermostability of proteins is a key factor considered during enzyme engineering, and finding a method that can identify thermophilic and non-thermophilic proteins will be helpful for enzyme design. In this study, we established a novel method combining mixed features and machine learning to achieve this recognition task. In this method, an amino acid reduction scheme was adopted to recode the amino acid sequence. Then, the physicochemical characteristics, auto-cross covariance (ACC), and reduced dipeptides were calculated and integrated to form a mixed feature set, which was processed using correlation analysis, feature selection, and principal component analysis (PCA) to remove redundant information. Finally, four machine learning methods and a dataset containing 500 random observations out of 915 thermophilic proteins and 500 random samples out of 793 non-thermophilic proteins were used to train and predict the data. The experimental results showed that 98.2% of thermophilic and non-thermophilic proteins were correctly identified using 10-fold cross-validation. Moreover, our analysis of the final reserved features and removed features yielded information about the crucial, unimportant and insensitive elements, it also provided essential information for enzyme design.

Keywords: thermophilic protein, reduced amino acids, mixed features, machine learning methods, non-thermophilic protein

INTRODUCTION

Proteins denature when the environmental temperature increases dramatically (Tang et al., 2017). However, thermophiles can survive in temperatures ranging from 41°C to 122°C (Takai et al., 2008; Fan et al., 2016) and produce enzymes that react well at higher environmental temperatures, such as 120°C (Fan et al., 2016). In enzyme engineering, identifying the functional mechanisms of these proteins will provide insights into the design and optimization of enzymes (Tang et al., 2017).

Protein thermostability has been shown to be related to hydrophobicity (Gromiha et al., 2013), hydrogen bonding (Bleicher et al., 2011), hydrophobic free energy (Gromiha et al., 1999; Saraboji et al., 2005), and residue (Meruelo et al., 2012) and inter-residue contacts (Gromiha, 2001). Moreover, Das and Gerstein (2000) found that salt bridges are essential for maintaining protein thermostability in thermophilic bacteria. The distribution of amino acids in proteins

(Fukuchi and Nishikawa, 2001; Zhou et al., 2008) and the presence of dipeptide (Ding et al., 2004; Zhang and Fang, 2006a,b) also affect protein thermostability. In a study by Vieille, the composition of Arg is greater in thermophiles than in mesophiles (Vieille and Zeikus, 2001). Guo also showed that expurgation of water-accessible thermo-labile residues, such as Gln and Met, affects the thermostability of enzymes expressed by thermophiles (Guo et al., 2014). Besides, Chen et al. (2016) found the pseudo amino acid composition had a big effect on the protein identification task, and constructed a web server to give a free way to use their algorithm[1].

Sequence-based protein identification provides an alternative method for studies of protein thermostability (Zhang and Fang, 2007; Wu et al., 2009; Li and Fang, 2010; Liu et al., 2011, 2019; Zuo et al., 2013; Fu et al., 2018; Wang et al., 2018; Zhang et al., 2018; Cheng et al., 2019; Yu et al., 2019b). Wang et al. (2011) introduced a feature selection method to identify vital features from the pseudo amino acid composition, amino acid composition, physicochemical features, composition transition, and distribution features using a support vector machine (SVM) to detect thermophilic proteins. Additionally, Tang proposed a two-step discrimination method with 94.44% accuracy using 5-fold cross-validation. Lin et al. constructed a dataset containing 915 thermophilic proteins and 793 non-thermophilic proteins, and predicted 93.8% thermophilic proteins and 92.7% non-thermophilic proteins using SVM. The same conclusion was also reached by Nakariyakul et al. (2012), who obtained 93.3% identification accuracy in the same database used by Lin. In another study, Fan et al. (2016) integrated information on the amino acid composition, evolution information, and acid dissociation constant to identify thermophiles by SVM, yielding an overall accuracy of 93.53%. Modarres et al. (2018) proposed a new thermophilic protein database, which contained 14 million protein sequences. In this database, all sequences were categorized according to the thermal stability and protein family property. Not only the sequences but also structures of thermophilic proteins were contained in the database. This online database gave the developers a powerful tool in the thermophilic protein prediction task.

In this study, we integrated 188 physicochemical characteristic features, auto-cross covariance (ACC) information, and dipeptide compositions of reduced amino acids to obtain a mixed feature set. Redundant features were then removed using correlation analysis, and dimensions were reduced using the max-relevance-max-distance (MRMD) method and principal component analysis (PCA). Finally, the SVM and other three machine learning methods were used to identify thermostability.

MATERIALS AND METHODS

The main framework of the method used in this study could be divided into the following four parts: (a) transforming thermophilic protein sequences to a reduced amino acid form; (b) extracting useful features; (c) using the SVM to train the extracted features; (d) predicting the test data by machine learning (Yu et al., 2017a,b; Zou et al., 2017a,b; Zhang et al., 2019a). The framework is shown in **Figure 1**.

Datasets

We used the dataset constructed by Lin et al. (Lin and Chen, 2011), whose data were chosen from the Universal Protein Resource (UniProt). The temperature of thermophilic proteins in this dataset was set to above 60°C and the temperature of non-thermophilic proteins was set to be less than 30°C. After removing redundancy and homology bias, there were 915 thermophilic and 793 non-thermophilic proteins. These data can be downloaded from http://www.labio.info/index-1therm.html.

Reduced Amino Acid Composition (RAAC)

In order to improve phylogenetic estimates, it is possible to recode the amino acids in the protein sequence (Susko and Roger, 2007). Furthermore, some reduced amino acid schemes, including the "Dayhoff classes" (AGPST, DENQ, HKR, ILMV, FWY, and C), have attracted attention (Susko and Roger, 2007).

In order to maximize the ratio of the expected number of substitutions within bins under the JTT model, Susko et al. proposed their reduced amino acid alphabet, which contains 30 schemes. In this study, we chose the final scheme as follows: A, C, D, E, F, G, H, IV, K, L, M, N, P, Q, R, S, T, W, Y. Thus, the 20 amino acids were classified into 19 types in the above scheme (Susko and Roger, 2007), in which Ile (I) and Val (V) were viewed as a single type, while every one of other categories had only one amino acid. Under this reduced scheme, we use the webserver of Zuo (Zheng et al., 2019) to calculate the RAAC of the thermophilic and non-thermophilic proteins.

Furthermore, dipeptides of proteins, like AA, A*A ($\lambda_{gap} = 1$), and A**A ($\lambda_{gap} = 2$), AK, A*K, A**K, etc., were also obtained using this webserver (Chen et al., 2016; Yang et al., 2019). The following formula was used to calculate the values of those features:

$$f_{361}^{\lambda}(j) = \frac{y_{361}^{\lambda}(j)}{\sum_j y_{361}^{\lambda}(j)} \qquad \lambda = 0, 1, 2, \cdots, 361,$$

where $y_{361}^{\lambda}(j)$ denotes the number of λ-gap dipeptides of type j in a protein sequence.

Feature Extraction

Physicochemical Characteristics

To quantitatively identify proteins, the physicochemical characteristics were obtained using a method (temporarily called 188d), which could extract sequence information and amino acid properties (Song et al., 2014; Xu et al., 2014, 2018; Fu et al., 2019; Liu, 2019; Zhu et al., 2019). The first 20 elements in the results of this method denoted the frequency of the 20 original amino acids (Zhu et al., 2019); the next 24 features reflected the group proportion corresponding to three groups (Qu et al., 2019); the following 120 dimensions were the distributions of three groups in five local positions (Cai et al., 2003); the last 24 features were the numbers of three types of dipeptides.

[1]http://lin-group.cn/server/Lypred/

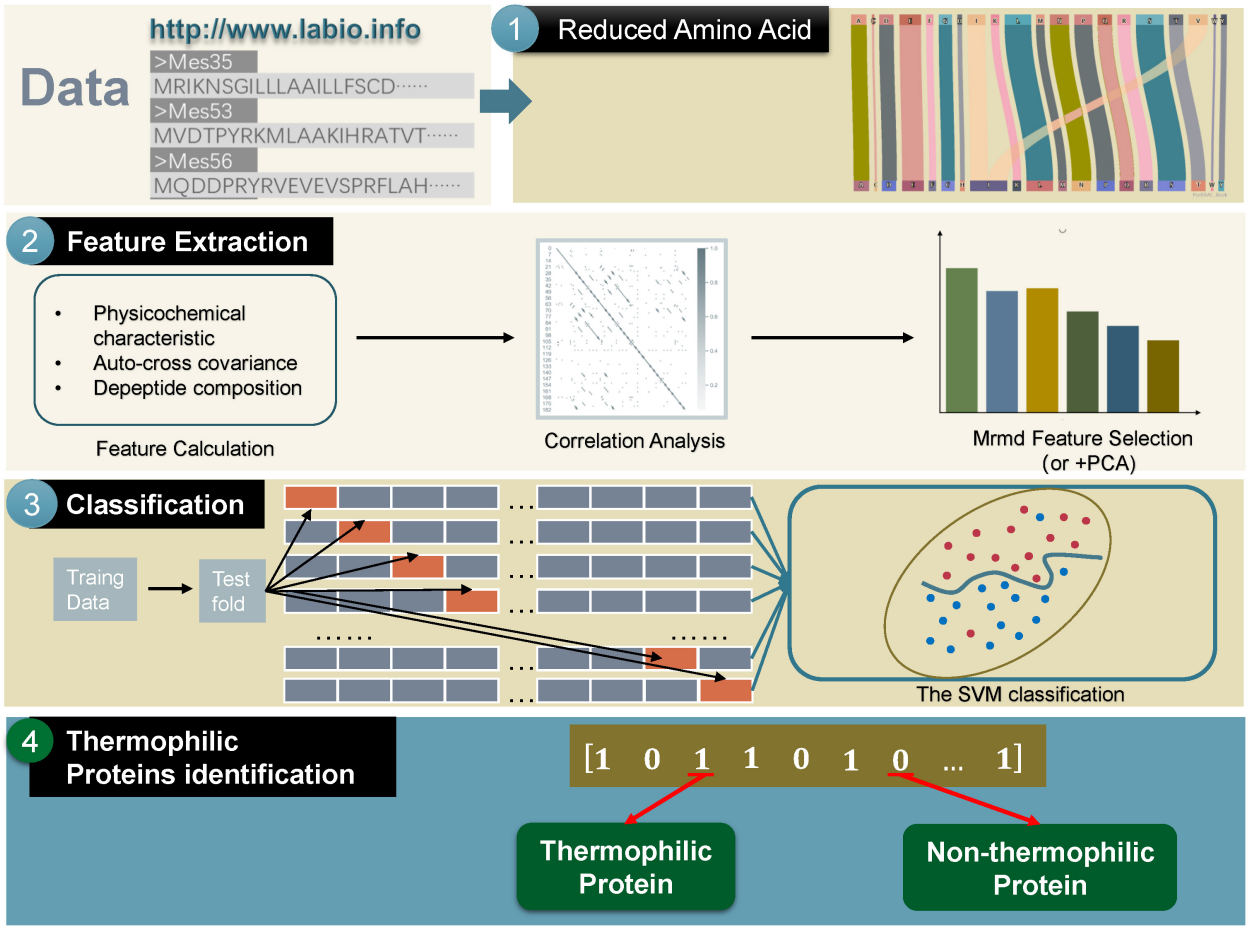


FIGURE 1 | The whole framework of the proposed method in this manuscript.

ACC

Auto covariance (AC) and cross-covariance (CC) calledACC, can reflect the relationship between amino acids with certain length features and contains AC and CC (Dong et al., 2009; Liu et al., 2015). The formula of CC transforms a protein sequence to a vector form Liu et al. (2016):

$$P' = [\varphi_1, \varphi_2, \varphi_3, \cdots, \varphi\, N * (N-1) * lg)]^T,$$

where N denotes the number of properties. φ_i can be calculated as:(Guo et al., 2008)

$$\varphi_n = AC\,(i, lg) = \frac{1}{N - lg} \sum_{j=1}^{L-lg} (S_{i,j} - \overline{S_i})(S_{i,j+lg} - \overline{S_i}),$$

where i is a residue, Ldenotes the length of the whole protein sequence, $S_{i,j}$ represents the i-th property of the j-th amino acid, and S_i reflects the mean value of the i-th property (Qu et al., 2019). In our experiment, the value of lg was set to 2.

Correlation Analysis

Some pairs in our feature set were found to be highly correlative, indicating that the effects of these two features were similar. Furthermore, this phenomenon denotes redundant and repeated information were present in the feature set. However, without the preprocess of discarding redundant information, machine learning models are associated with a risk of overfitting (Hua et al., 2009; Mwangi et al., 2014; Zeng et al., 2019b).

Thus, a correlation analysis-based redundant information expurgate method was proposed to discard one feature from each of the highly relevant feature pairs. As a prepare step, all feature values need to be normalized to [0,1] using the following equation:

$$x_i^n = \frac{x_i - \bar{x}}{x_{max} - x_{min}},$$

where $x_i (i = 1, 2, 3, \cdots)$ denotes the i-th value in the feature set, $\bar{x}$represents the mean value of the current feature vector, andx_{max}, x_{min} correspondingly reflect the maximum and minimum values of the feature vector.

Then, Pearson's correlation was used to evaluate the correlations between any two features. Its value was written as

follows:(Thibeault and Srinivasa, 2013; Jin et al., 2019)

$$\rho(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma(X)}\right)\left(\frac{Y_i - \bar{Y}}{\sigma(Y)}\right),$$

where X and Y are two given feature vectors, $\bar{X}$and $\bar{Y}$ represent the mean value of X and Y, respectively, and $\sigma(X)$ and $\sigma(Y)$ denote the standard deviations of X and Y, respectively.

In our experiment, for any feature pair X and Y, if the value of $\rho(X, Y)$ was larger than the threshold T, then X and Y were considered a highly correlated feature pair. In the next step, we decided whether to remove one of the features from the feature set while retaining the other in the feature set. Thus, for the first feature pair, a removed feature set D and a reserved feature set R were created and set as an empty set. Then, one feature was set to belong to D, while the other was set to belong to R randomly. In the following computation, the rule for assigning features could be expressed as follows: assuming that X-Y is a highly correlative feature pair,

- If $X \notin D$ and $X \notin R : Y \notin D$ and $Y \notin R \rightarrow Y \in D, X \in R$
- If $X \notin D$ and $X \notin R : Y \in D \rightarrow X \in R$
- If $X \notin D$ and $X \notin R : Y \in R \rightarrow X \in D$
- Elseif $X \in R : Y \notin D$ and $Y \notin R \rightarrow Y \in D$
- Elseif $X \in D : Y \notin D$ and $Y \notin R \rightarrow Y \in D$

Let $D = \left\{f_1^{'}, f_2^{'}, f_3^{'}, \cdots, f_M^{'}\right\}$ denote the final removed feature set. After all M features in D were removed from the feature set, the correlation between feature pairs was decreased dramatically. The threshold T used in our experiment was set as 0.85.

MRMD Feature Selection

Dimensionality reduction is a key process in machine learning research and application (Bhola and Singh, 2018). The MRMD method, as presented by Zou et al. (2016), was used to rank features in descending order and reduce the feature number. There were two object functions; the first reflected the relationship between the current feature and the target class, which could be written as follows (Zou et al., 2016):

$$PPC\left(\overrightarrow{F_i}, \overrightarrow{C_i}\right) = \frac{\frac{1}{N-1}\sum_{k=1}^{N}\left(f_{i,k} - \overline{f_i}\right)\left(C_{i,k} - \overline{C_i}\right)}{\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(C_{i,k} - \overline{C_i}\right)^2}\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(f_{i,k} - \overline{f_i}\right)^2}},$$

$$max\ MR_i = \left|PPC\left(\overrightarrow{F_i}, \overrightarrow{C_i}\right)\right|,$$

where $f_{i,k}$ and $C_{i,k}$ represent the k-th element in the feature vector F_i and C_i, respectively. The other object function was expressed in the following form Zou et al. (2016):

$$ED\left(\vec{X}, \vec{Y}\right) = \sqrt{\sum_{k=1}^{N}\left(x_k - y_k\right)^2},$$

$$\max MD_i = ED_i = \frac{1}{M-1}\sum ED\left(\overrightarrow{F_i}, \overrightarrow{F_k}\right).$$

Integrating the above two functions, we obtained the final objective function, which was written as follows:

$$max(MR_i + MD_i)$$

Solving this function, when the function reached the maximum ACC value, the iteration was stopped automatically, giving a feature dimension reduced set.

PCA

Principal component analysis (Price et al., 2006) is a widely used tool that can transform the features of observation into an uncorrelated feature set (Zeng et al., 2017, 2019a; Xiao et al., 2018; Zhang et al., 2019b). The main steps of PCA are as follows: (1) normalize the feature vector value; (2) calculate the covariance matrix by $\sum = \frac{1}{m}X \cdot X^T$; (3) use the singular value decomposition method $(U, S, V^T); = SVD(\Sigma)$; (4) extract the first k singular vectors from U and (5) calculate the i-th eigenvalue $\lambda_i,\ i = 1, 2, 3, \cdots$

We used ρ to evaluate the cumulative contribution value of the singular vectors; this value was defined as $\rho = \frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{m}\lambda_i} \geq T'$, where m denotes the dimension of the transformed features. The above function denotes there is enough information to serve as the optimal feature set for the identification 0task when the cumulative contribution value of singular vectors from the first one to the λ-th one reaches a value, namely, the threshold T'. Thus, through the threshold T', only a part of features were selected and then formed an optimal feature set, which made the model simple and fast to run.

Machine Learning Methods

In order to distinguish between thermophilic and non-thermophilic proteins, SVM (Ding et al., 2016a,b; He et al., 2018; Qiao et al., 2018; Wei et al., 2018; Fu et al., 2019; Wang et al., 2019b), random forest [RF, (Ding et al., 2017; Wang et al., 2019a)], decision tree (Mohasseb et al., 2018; Li et al., 2019), and naïve Bayes [NB, (Rajaraman and Chokkalingam, 2014)] methods were used in our experiment. The first two methods were implemented and optimized in the python 3.7 environment with our edited code. All four methods were also tested in the Weka environment, yielding similar results.

Evaluation of Performance

In order to evaluate the model performance, we used a 10-fold cross-validation scheme in our experiment and adopted three commonly used accuracy indicators for quantification (Jiang et al., 2013, 2018; Zeng et al., 2016; Wei et al., 2017a,b; Lu et al., 2018, 2019; Xiong et al., 2018; Chen et al., 2019; Ding et al., 2019; Lin et al., 2019; Shan et al., 2019; Shen et al., 2019; Xu et al., 2019; Yu and Gao, 2019; Yu et al., 2019a). The first indicator was sensitivity (Sn), which represents the ratio of the correctly identified thermophilic proteins and could be calculated as follows:

$$Sn = \frac{TP}{TP + FN} \times 100\%,$$

where TP, TN, FP, and FN represent the number of the correctly identified thermophilic proteins, the number of the correctly

indemnified non-thermophilic proteins, the number of non-thermophilic proteins predicted as thermophilic proteins, and the number of the thermophilic proteins predicted as non-thermophilic proteins, respectively (Lin and Chen, 2011).

The second indicator was specificity (Sp), which denotes the percentage of the correctly identified non-thermophilic proteins among all non-thermophilic observations. Sp was defined as follows:

$$Sp = \frac{TN}{TN + FP} \times 100\%.$$

The last indicator was accuracy (ACC), which reflected the percentage of correctly recognized thermophilic and non-thermophilic proteins among all observations, written as follows:

$$ACC = \frac{TN + TP}{TN + FP + FN + FP} \times 100\%.$$

RESULTS

Our experiments were performed on the basis of qualitative evaluation, quantitative analysis, and comparison with other counterparts, as shown in **Figure 2**. The data were calculated using 500 randomly selected thermophilic proteins and 500 randomly selected non-thermophilic proteins, and experiments were evaluated in 10-fold cross-validation format.

First, we evaluated the proposed method using qualitative analysis. In this analysis, all feature data were reduced to 12 dimensions through the PCA method. Furthermore, the t-SNE method (van der Maaten and Hinton, 2012; van der Maaten, 2014) is one of the powerful visualization tools for showing the structure of high-dimension data. Thus, we used the t-SNE method (van der Maaten and Hinton, 2012; van der Maaten, 2014) to differentiate thermophilic and non-thermophilic proteins in the figure. Additionally, the t-SNE method used here was not a part of the proposed model, but was a display tool of the experiment data. The first two features of the results using the t-SNE method are plotted in **Figure 2A**; from these data, a distinct boundary was observed for separating thermophilic and non-thermophilic observations. Moreover, it was easy to distinguish thermophilic proteins from non-thermophilic proteins.

In order to verify these findings, SVM was used to train and test the 12-dimensional data, and the results are shown in **Figure 2B**. Both types of proteins were separated successfully using this method. This phenomenon directly demonstrated that our proposed data had good separation quality and the SVM method had strong recognition ability for thermophilic proteins and non-thermophilic data.

Second, the processed data were tested using the other three machine learning methods, as detailed in **Figure 2C**. For every method, we also calculated three accuracy indicators: Sn, Sp, and ACC. The results showed that the SVM yielded the highest values for all three indicators, and all values reached at least 98.2%. NB also showed higher accuracy, with values of 96.25%, 97.56%, and 96.89%, respectively. The accuracy of the random forest model was higher than that of J48, for which the average value was only 91.48%.

Our method was also compared with the results of Lin (Lin and Chen, 2011) and the method of using the same dataset (Fan et al., 2016). The results are shown in **Figure 2D**. Notably, our method got the highest accuracy values based on the results of the MRMD methods, which denotes our proposed method outperformed the method described by Lin (Lin and Chen, 2011). Additionally, the performance of the proposed method was better than the effects described by Fan et al. (2016) too, suggesting that the proposed method could be a state-of-the-art model in current research.

Features using the original dipeptides were also tested in our study. All reduced features in our feature set were replaced with the original dipeptides. From the accuracy data shown in **Figure 2D**, the ability to distinguish thermophilic proteins from non-thermophilic ones was lower than that using the reduced amino acid dipeptides. Additionally, the receiver operating characteristic (ROC) curve was also plotted, which could be seen in **Figure 3A**. It is easy to found that the results of the ROC curve verified the identification efficiency of the proposed method too.

Finally, the newly released thermophilic protein database (Mohasseb et al., 2018) is also tested through the proposed method in this manuscript. In the experiment, we selected 106 thermophilic proteins and 101 psychrophilic proteins from the database. All those data can be downloaded on the website: http://www.labio.info/index-1therm.html. In the experiment, we did three experiments using three different thresholds in the correlation analysis step. The experiments are given in **Figure 3B**, from which it was easy to find that the identification accuracy was bigger than 0.97 in most cases when using the threshold of 0.95 and 0.90. It also showed that the classification efficiency was not ideal when using the threshold 0.85. The reason for this phenomenon may be the calculated features of the current data have a stronger correlation between each other than the previous thermophilic protein database. Thus, in this condition, a big value than 0.85 is needed to identify the thermophilic proteins accurately. It is worth noting that the results in this figure verified the perfect identification ability of the proposed method.

DISCUSSION

Many features are removed from the original feature set during correlation analysis and MRMD feature selection. Moreover, these removed features are typically not crucial or redundant for performing thermophilic protein recognition. However, the selection of features to remove and retain is essential, and further studies are needed to evaluate such approaches. Thus, in this study, we evaluated the removed features, as depicted in **Figure 4**.

The 10 most critical original features are shown in **Figure 4A**, and under our proposed model framework, the feature values of K*H, KR, TF, P*M, F**N, I**Y/V**Y, MW, and WQ (where * represents a gap in the residues) showed significant contributions to the recognition of thermophilic proteins. Additionally, residue K also plays a vital role in enhancing thermostability. Interestingly, our conclusions regarding residue K were consistent with the results of Lin (Lin and Chen, 2011).

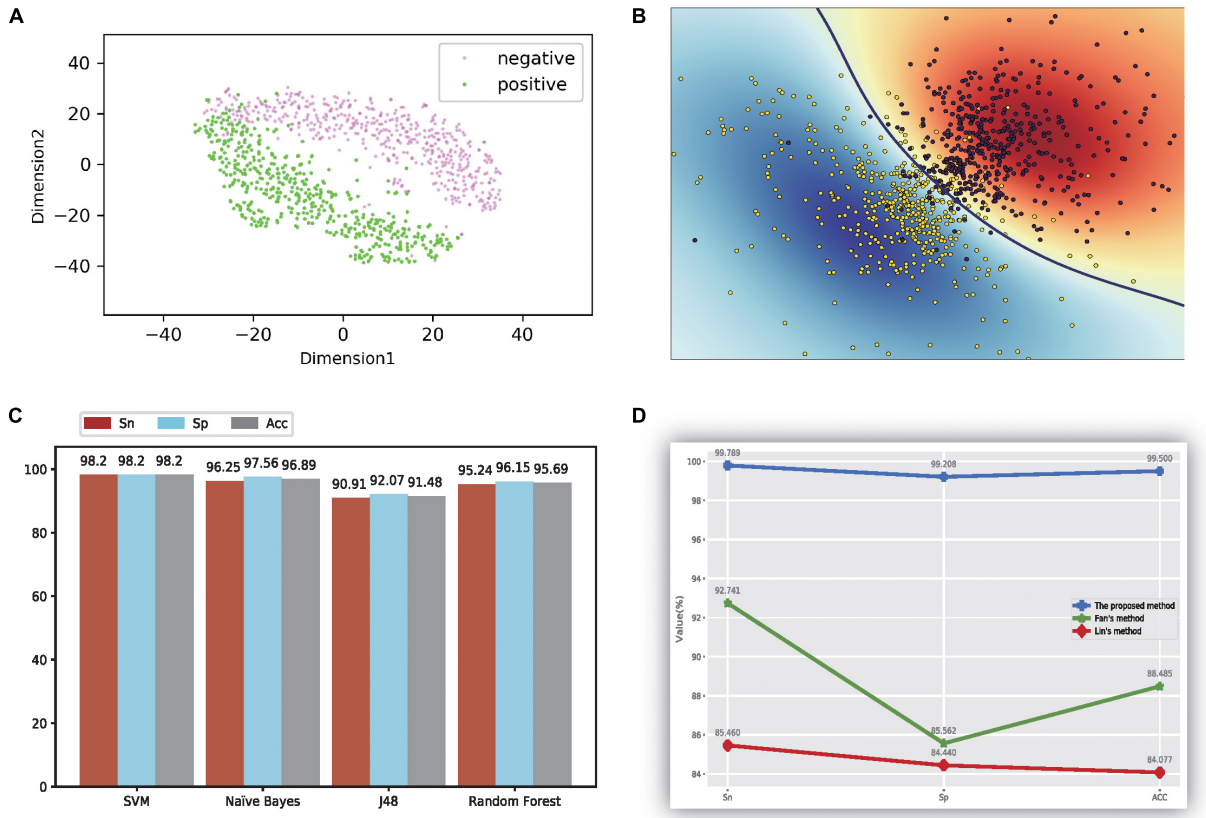


FIGURE 2 | The figure of model performance. **(A)** The first two dimensions of the result of compression characteristics of the TSNE method; **(B)** the figure of the ultra-classification surface of SVM method; **(C)** The accuracy values of four different models; **(D)** the comparison results with other methods.

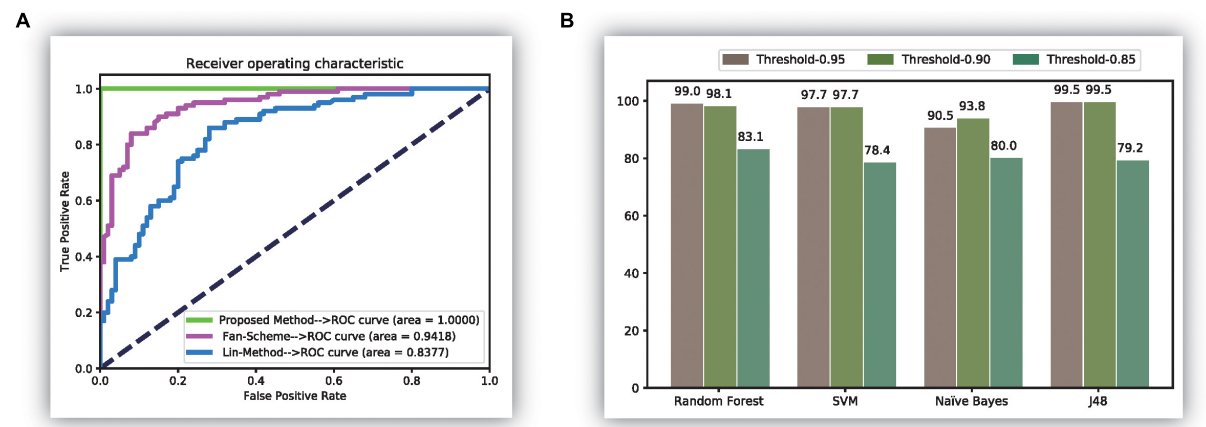


FIGURE 3 | The comparison results of experiments. **(A)** The receiver operation characteristic (ROC) curve of three methods; **(B)** the results of experiments over the database (Fan et al., 2016).

For the removed features, the results are shown in **Figures 4B–D**. There were four types of components in the final feature set: ACC features, physicochemical characteristics, amino acid frequencies (the first 20 features in the 188D feature), and reduced amino acid dipeptides. Approximately half of the physicochemical characteristics were deleted from the original feature set, and there were only a few reserved physicochemical characteristics in the first 50 crucial features. Thus, we concluded that the physicochemical characteristics were essential features, but not the most essential features,

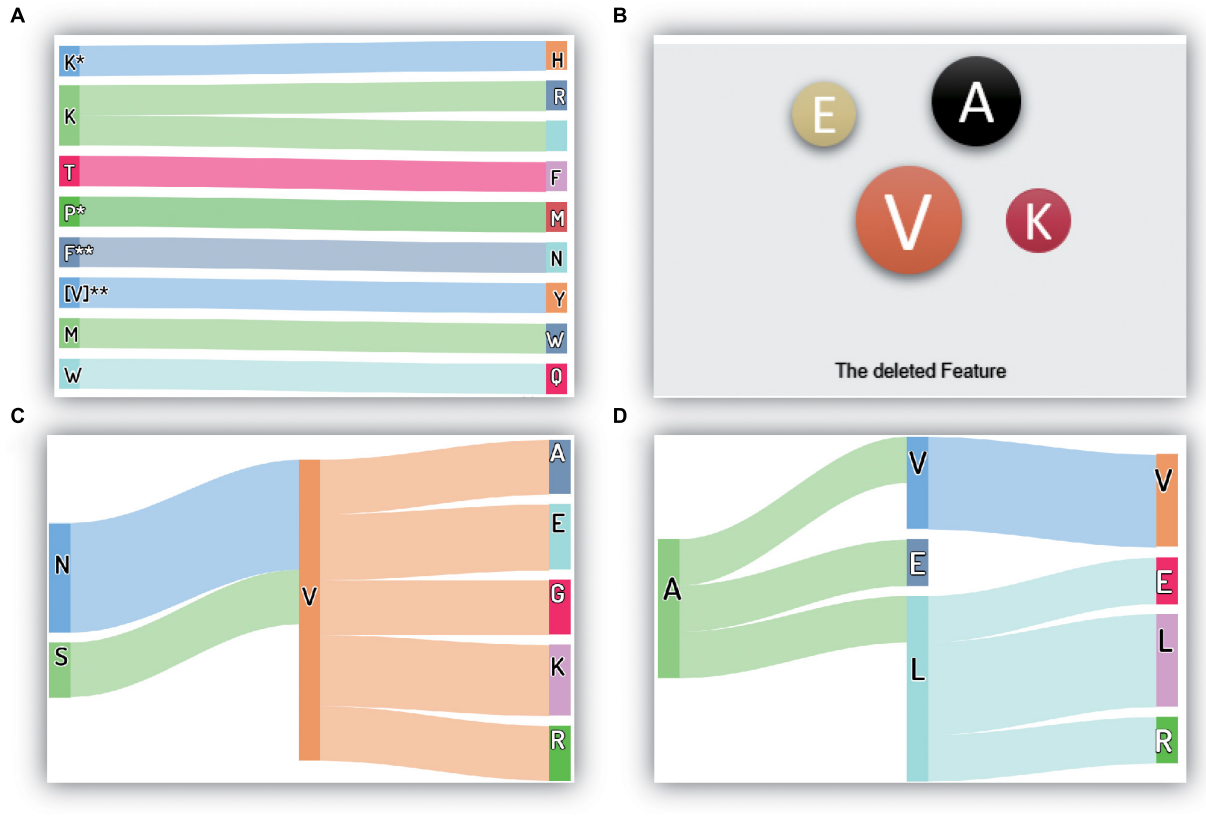


FIGURE 4 | The critical and removed features in the proposed method: **(A)** the most important features; **(B)** the deleted amino acid frequency features; **(C)** the deleted reduced-depiptides **(I)**; **(D)** the deleted reduced-depiptides **(II)**. The symbol "*" means any one of the 20 amino acids, it may be "A", "C", "P", or others. Besides, "**" has the same meaning; it represents a two-letter combination of 20 amino acids, "AA", "DC", "VP", for example.

for this recognition task. Accordingly, we did not analyze the details of the removed physicochemical characteristics. We also showed that only three ACC features were excluded from the final feature set, and the remaining 15 ACC features were retained, reflecting the crucial roles of the ACC features in this recognition task.

The amino acid frequency, which was one of the first 20 features in the 188D feature set, included only four residues removed from the feature set. These four residues were V (Ile and Val), A, E, and K, which had little contribution to recognizing thermophilic protein and non-thermophilic proteins. Interestingly, the reduced amino acid V, which included both Ile and Val, was also deleted. It is worth noting that the amino acid V appeared later in this manuscript denotes the reduced V, namely, both Ile and Val. This finding indicated that both Ile and Val were redundant and did not contribute to the identification task. If we used the original amino acid dipeptide features, additional useless features, including IA, I*A, and I**A, etc., would also be observed in the feature set. The number of additional redundant features in the original dipeptides could be as high as 39 if compared with the reduced amino acid dipeptides. As shown in **Figure 2D**, the smallest prediction accuracy was obtained, and represented those many additional useless features caused the classification model fail in the overfitting state when using the original dipeptides. Additionally, this observation could explain why the accuracy increased significantly when using the reduced amino acid dipeptides.

There were three types of dipeptides, expressed as AA ($\lambda = 0$), A*A ($\lambda = 1$), and A**A ($\lambda = 2$). The numbers of these types of removed dipeptides were 60, 61, and 71, respectively. To conveniently visualize these data, we counted the numbers of the same dipeptide (omitting the symbols * and **). If a dipeptide appeared more than twice, it was drawn in the figure. Thus, if the dipeptide NV was shown in the figure, there were at least two types of dipeptides, i.e., NV, N*V, or N**V, in the removed feature set.

All discovered dipeptides were classified into two parts, as shown in **Figures 4C,D**. The reduced dipeptides in **Figure 4C** were dipeptides having relationships with the reduced residue V, verifying the reduced power of the recognition task in the above analysis. Moreover, residue V enabled the discovery of seven related dipeptides in the removed features. This phenomenon demonstrated that residue V and some dipeptides containing V were insensitive to the recognition task under our proposed model framework. **Figure 4D** also shows another seven removed dipeptides, including VV, AV, AE, AL, LE, LL, and LR.

These results provide insights into the design of stable mutants to increase protein thermostability.

CONCLUSION

In this study, we aimed to develop an approach to distinguish thermophilic proteins from non-thermophilic proteins; to this end, a recognition method that combined mixed features of proteins and a machine learning method was established. First, an amino acid reduction method was introduced to reduce the categories of amino acids. Nest, we calculated the physicochemical characteristics, ACC, and reduced dipeptides of thermophilic and non-thermophilic proteins. After performing a dimension reduction step using correlation analysis, the MRMD method, and PCA, an optimal feature set was obtained. Finally, machine learning methods were used to train and predict feature data, and the results revealed that the proposed model could identify 98.2% of thermophilic proteins and non-thermophilic proteins if the data were operated in a 10-fold cross-validation mode. Furthermore, the feature values of K*H, KR, TF, P*M, F**N, V**Y, MW, and WQ were found to play vital roles in thermostability, and some residues and dipeptides, including V (Ile and Val), A, E, K, NV, VG, VA, AE, AL, and LE, were not important for identifying thermostability. As discussed in previous studies (Liu and Li, 2019; Liu and Zhu, 2019), the web-server is very important. In our future work, our research will focus on developing a free webserver that could provide a platform to test the currently proposed method using an easily accessible approach.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the http://www.labio.info/index-1therm.html.

AUTHOR CONTRIBUTIONS

CF and YL: conceptualization. CF: methodology, software, and writing – original draft preparation. YL, JZ, and XL: validation. DY: formal analysis. ZM and XL: investigation. ZM: resources. XL: data curation. DY, YL, and JZ: writing – review and editing, supervision. CF and XL: visualization. DY and ZM: project administration. YL and JZ: funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China, grant nos. 91935302 and 61971119.

REFERENCES

Bhola, A., and Singh, S. (2018). Gene selection using high dimensional gene expression data: an appraisal. *Curr. Bioinf.* 13, 225–233.

Bleicher, L., Prates, E. T., Gomes, T. C. F., Silveira, R. L., Nascimento, A. S., Rojas, A. L., et al. (2011). Molecular basis of the thermostability and thermophilicity of laminarinases: x-ray structure of the hyperthermostable laminarinase from rhodothermus marinus and molecular dynamics simulations. *J. Phys. Chem. B* 115, 7940–7949. doi: 10.1021/jp200330z

Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., and Chen, Y. Z. (2003). SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31, 3692–3697.

Chen, C., Zhang, Q. M., Ma, Q., and Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemometr. Intell. Labor. Syst.* 191, 54–64.

Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Biomed. Res. Int.* 2018:8.

Cheng, L., Wang, P. P., Tian, R., Wang, S., Guo, Q. H., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144.

Das, R., and Gerstein, M. (2000). The stability of thermophilic proteins: a study based on comprehensive genome comparison. *Funct. Integr. Genom.* 1, 76–88.

Ding, Y. J., Tang, J. J., and Guo, F. (2016a). Identification of protein-protein interactions via a novel matrix-based sequence representation model with amino acid contact information. *Int. J. Mol. Sci.* 17:14. doi: 10.3390/ijms17101623

Ding, Y. J., Tang, J. J., and Guo, F. (2016b). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinf.* 17:13. doi: 10.1186/s12859-016-1253-9

Ding, Y. J., Tang, J. J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inform. Sci.* 418, 546–560.

Ding, Y. J., Tang, J. J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224.

Ding, Y. R., Cai, Y. J., Zhang, G. X., and Xu, W. B. (2004). The influence of dipeptide composition on protein thermostability. *FEBS Lett.* 569, 284–288.

Dong, Q. W., Zhou, S. G., and Guan, J. H. (2009). A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics* 25, 2655–2662. doi: 10.1093/bioinformatics/btp500

Fan, G. L., Liu, Y. L., and Wang, H. (2016). Identification of thermophilic proteins by incorporating evolutionary and acid dissociation information into Chou's general pseudo amino acid composition. *J. Theor. Biol.* 407, 138–142. doi: 10.1016/j.jtbi.2016.07.010

Fu, X. Z., Ke, L. X., Cai, L. J., Chen, X. T., Ren, X. B., and Gao, M. Y. (2019). Improved prediction of cell-penetrating peptides via effective orchestrating amino acid composition feature representation. *IEEE Access.* 7, 163547–163555.

Fu, X. Z., Zhu, W., Liao, B., Cai, L. J., Peng, L. H., and Yang, J. L. (2018). Improved DNA-binding protein identification by incorporating evolutionary information into the chou's PseAAC. *IEEE Access.* 6, 66545–66556.

Fukuchi, S., and Nishikawa, K. (2001). Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J. Mol. Biol.* 309, 835–843.

Gromiha, M. M. (2001). Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. *Biophys. Chem.* 91, 71–77.

Gromiha, M. M., Oobatake, M., and Sarai, A. (1999). Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys. Chem.* 82, 51–67.

Gromiha, M. M., Pathak, M. C., Saraboji, K., Ortlund, E. A., and Gaucher, E. A. (2013). Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins-Struct. Funct. Bioinf.* 81, 715–721.

Guo, J. N., Luk, L. Y. P., Loveridge, E. J., and Allemann, R. K. (2014). Thermal adaptation of dihydrofolate reductase from the moderate thermophile *Geobacillus stearothermophilus*. *Biochemistry* 53, 2855–2863. doi: 10.1021/bi500238q

Guo, Y. Z., Yu, L. Z., Wen, Z. N., and Li, M. L. (2008). Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences. *Nucleic Acids Res.* 36, 3025–3030.

He, J. J., Fang, T., Zhang, Z. Z., Huang, B., Zhu, X. L., and Xiong, Y. (2018). PseUI: pseudouridine sites identification based on RNA sequence information. *BMC Bioinf.* 19:11. doi: 10.1186/s12859-018-2321-0

Hua, J. P., Tembe, W. D., and Dougherty, E. R. (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recogn.* 42, 409–424.

Jiang, L. M., Xiao, Y. K., Ding, Y. J., Tang, J. J., and Guo, F. (2018). FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association. *BMC Genomics* 19:15. doi: 10.1186/s12864-018-5273-x

Jiang, Q. H., Wang, G. H., Jin, S. L., Li, Y., and Wang, Y. D. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinf.* 8, 282–293.

Jin, J., Miao, Y. Y., Daly, I., Zuo, C. L., Hu, D. W., and Cichocki, A. (2019). Correlation-based channel selection and regularized feature optimization for MI-based BCI. *Neural Netw.* 118, 262–270. doi: 10.1016/j.neunet.2019.07.008

Li, M. J., Xu, H. H., and Deng, Y. (2019). Evidential decision tree based on belief entropy. *Entropy* 21, 14.

Li, Y. Q., and Fang, J. W. (2010). Distance-dependent statistical potentials for discriminating thermophilic and mesophilic proteins. *Biochem. Biophys. Res. Commun.* 396, 736–741. doi: 10.1016/j.bbrc.2010.05.005

Lin, H., and Chen, W. (2011). Prediction of thermophilic proteins using feature selection technique. *J. Microbiol. Methods* 84, 67–70. doi: 10.1016/j.mimet.2010.10.013

Lin, X., Quan, Z., Wang, Z.-J., Huang, H., and Zeng, X. (2019). A novel molecular representation with BiGRU neural networks for learning atom. *Brief. Bioinf.* doi: 10.1093/bib/bbz125

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., and Li, K. (2019). iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol. Ther. Nucleic Acids* 18, 80–87. doi: 10.1016/j.omtn.2019.08.008

Liu, B., Liu, F. L., Wang, X. L., Chen, J. J., Fang, L. Y., and Chou, K. C. (2015). Pse-in-one: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71.

Liu, B., Wang, S. Y., Dong, Q. W., Li, S. M., and Liu, X. (2016). Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning. *IEEE Trans. Nanobiosci.* 15, 328–334. doi: 10.1109/TNB.2016.2555951

Liu, B., and Zhu, Y. L. (2019). ProtDec-LTR3.0: protein remote homology detection by incorporating profile-based features into learning to rank. *IEEE Access.* 7, 102499–102507.

Liu, X. L., Lu, J. L., and Hu, X. H. (2011). Predicting thermophilic proteins with pseudo amino acid composition: approached from chaos game representation and principal component analysis. *Protein Peptide Lett.* 18, 1244–1250.

Liu, Y. M., Wang, X. L., and Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.* 20, 330–346. doi: 10.1093/bib/bbx126

Lu, X. G., Li, X., Liu, P., Qian, X., Miao, Q. M., and Peng, S. L. (2018). The integrative method based on the module-network for identifying driver genes in cancer subtypes. *Molecules* 23:15. doi: 10.3390/molecules23020183

Lu, X. G., Qian, X., Li, X., Miao, Q. M., and Peng, S. L. (2019). DMCM: a data-adaptive mutation clustering method to identify cancer-related mutation clusters. *Bioinformatics* 35, 389–397. doi: 10.1093/bioinformatics/bty624

Meruelo, A. D., Han, S. K., Kim, S., and Bowie, J. U. (2012). Structural differences between thermophilic and mesophilic membrane proteins. *Protein Sci.* 21, 1746–1753.

Modarres, H. P., Mofrad, M. R., and Sanati-Nezhad, A. (2018). ProtDataTherm: a database for thermostability analysis and engineering of proteins. *PLoS ONE* 13:9. doi: 10.1371/journal.pone.0191222

Mohasseb, A., Bader-El-Den, M., and Cocea, M. (2018). Question categorization and classification using grammar based approach. *Inform. Process. Manage.* 54, 1228–1243.

Mwangi, B., Tian, T. S., and Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics* 12, 229–244. doi: 10.1007/s12021-013-9204-3

Nakariyakul, S., Liu, Z. P., and Chen, L. N. (2012). Detecting thermophilic proteins through selecting amino acid and dipeptide composition features. *Amino Acids* 42, 1947–1953. doi: 10.1007/s00726-011-0923-1

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909.

Qiao, Y. H., Xiong, Y., Gao, H. Y., Zhu, X. L., and Chen, P. (2018). Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. *BMC Bioinf.* 19:16. doi: 10.1186/s12859-018-2009-5

Qu, K. Y., Wei, L. Y., Yu, J. T., and Wang, C. Y. (2019). Identifying plant pentatricopeptide repeat coding gene/protein using mixed feature extraction methods. *Front. Plant Sci.* 9:10. doi: 10.3389/fpls.2018.01961

Rajaraman, S., and Chokkalingam, A. (2014). Classification of denver system of chromosomes using similarity classifier guided by OWA operators. *Curr. Bioinf.* 9, 499–508.

Saraboji, K., Gromiha, M. M., and Ponnuswamy, M. N. (2005). Importance of main-chain hydrophobic free energy to the stability of thermophilic proteins. *Int. J. Biol. Macromol.* 35, 211–220.

Shan, X. Q., Wang, X. G., Li, C. D., Chu, Y. Y., Zhang, Y. F., Xiong, Y., et al. (2019). Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method. *J. Chem. Inform. Model.* 59, 4577–4586. doi: 10.1021/acs.jcim.9b00749

Shen, Y. N., Tang, J. J., and Guo, F. (2019). Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. *J. Theor. Biol.* 462, 230–239. doi: 10.1016/j.jtbi.2018.11.012

Song, L., Li, D. P., Zeng, X. X., Wu, Y. F., Guo, L., and Zou, Q. (2014). nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinform.* 15:10. doi: 10.1186/1471-2105-15-298

Susko, E., and Roger, A. J. (2007). On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* 24, 2139–2150.

Takai, K., Nakamura, K., Toki, T., Tsunogai, U., Miyazaki, M., Miyazaki, J., et al. (2008). Cell proliferation at 122 degrees C and isotopically heavy CH4 production by a hyperthermophilic methanogen under high-pressure cultivation. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10949–10954. doi: 10.1073/pnas.0712334105

Tang, H., Cao, R. Z., Wang, W., Liu, T. S., Wang, L. M., and He, C. M. (2017). A two-step discriminated method to identify thermophilic proteins. *Int. J. Biomathemat.* 10:8.

Thibeault, C. M., and Srinivasa, N. (2013). Using a hybrid neuron in physiologically inspired models of the basal ganglia. *Front. Comput. Neurosci.* 7:17. doi: 10.3389/fncom.2013.00088

van der Maaten, L. (2014). Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* 15, 3221–3245.

van der Maaten, L., and Hinton, G. (2012). Visualizing non-metric similarities in multiple maps. *Mach. Learn.* 87, 33–55. doi: 10.1186/s12859-018-2537-z

Vieille, C., and Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* 65, 1–43.

Wang, D., Yang, L., Fu, Z. Q., and Xia, J. B. (2011). Prediction of thermophilic protein with pseudo amino acid composition: an approach from combined feature selection and reduction. *Protein Peptide Lett.* 18, 684–689.

Wang, G. H., Luo, X. M., Wang, J. N., Wan, J., Xia, S. L., Zhu, H., et al. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151.

Wang, X. Y., Yu, B., Ma, A. J., Chen, C., Liu, B. Q., and Ma, Q. (2019a). Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* 35, 2395–2402. doi: 10.1093/bioinformatics/bty995

Wang, Y., Shi, F. Q., Cao, L. Y., Dey, N., Wu, Q., Ashour, A. S., et al. (2019b). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinf.* 14, 282–294.

Wei, L. Y., Wan, S. X., Guo, J. S., and Wong, K. K. L. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L. Y., Xing, P. W., Zeng, J. C., Chen, J. X., Su, R., and Guo, F. (2017b). Improved prediction of protein-protein interactions using novel negative

samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Wei, L. Y., Zhou, C., Chen, H. R., Song, J. N., and Su, R. (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Wu, L. C., Lee, J. X., Huang, H. D., Liu, B. J., and Horng, J. T. (2009). An expert system to predict protein thermostability using decision tree. *Exp. Syst. Appl.* 36, 9007–9014.

Xiao, J., Liu, S. D., Hu, L., and Wang, Y. (2018). Filtering method of rock points based on BP neural network and principal component analysis. *Front. Comput. Sci.* 12:1149–1159. doi: 10.1007/s11704-016-6170-6

Xiong, Y., Wang, Q. K., Yang, J. C., Zhu, X. L., and Weil, D. Q. (2018). PredT4SE-Stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol.* 9:9.

Xu, H., Zeng, W. H., Zhang, D. F., and Zeng, X. X. (2019). MOEA/HD: a multiobjective evolutionary algorithm based on hierarchical decomposition. *IEEE Trans. Cybernet.* 49, 517–526. doi: 10.1109/TCYB.2017.2779450

Xu, L., Liang, G. M., Shi, S. H., and Liao, C. R. (2018). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *Int. J. Mol. Sci.* 19:11. doi: 10.3390/ijms19061773

Xu, R. F., Zhou, J. Y., Liu, B., Yao, L., He, Y. L., Zou, Q., et al. (2014). enDNA-Prot: identification of DNA-binding proteins by applying ensemble learning. *Biomed. Res. Int.* 2014:10.

Yang, W., Zhu, X. J., Huang, J., Ding, H., and Lin, H. (2019). A brief survey of machine learning methods in protein sub-golgi localization. *Curr. Bioinf.* 14, 234–240.

Yu, B., Qiu, W., Chen, C., Ma, A., Jiang, J., Zhou, H., et al. (2019a). SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics* 36, 1074–1081. doi: 10.1093/bioinformatics/btz734

Yu, L., Yao, S. Y., Gao, L., and Zha, Y. H. (2019b). Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments. *Front. Genet.* 9:13. doi: 10.3389/fgene.2018.00745

Yu, L., and Gao, L. (2019). Human pathway-based disease network. *IEEE-ACM Trans. Comput. Biol. Bioinf.* 16, 1240–1249.

Yu, L., Su, R. D., Wang, B. B., Zhang, L., Zou, Y. P., Zhang, J., et al. (2017a). Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. *IEEE-ACM Trans. Comput. Biol. Bioinf.* 14, 966–977.

Yu, L., Zhao, J., and Gao, L. (2017b). Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome. *Artif. Intell. Med.* 77, 53–63. doi: 10.1016/j.artmed.2017.03.009

Zeng, X., Lin, Y., He, Y., Lv, L., Min, X., and Rodriguez-Paton, A. (2019a). Deep collaborative filtering for prediction of disease genes. *IEEE-ACM Trans. Comput. Biol. Bioinf.* doi: 10.1109/TCBB.2019.2907536

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2019b). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief. Bioinf.* doi: 10.1093/bib/bbz080

Zeng, X. X., Liao, Y. L., Liu, Y. S., and Zou, Q. (2017). Prediction and validation of disease genes using hetesim scores. *IEEE-ACM Trans. Comput. Biol. Bioinf.* 14, 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X. X., Zhang, X., and Zou, Q. (2016). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinform.* 17, 193–203. doi: 10.1093/bib/bbv033

Zhang, F., Ma, A. J., Wang, Z., Ma, Q., Liu, B. Q., Huang, L., et al. (2018). A central edge selection based overlapping community detection algorithm for the detection of overlapping structures in protein-protein interaction networks. *Molecules* 23:16. doi: 10.3390/molecules23102633

Zhang, G. Y., and Fang, B. S. (2006a). Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. *Process Biochem.* 41, 1792–1798.

Zhang, G. Y., and Fang, B. S. (2006b). Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. *Process Biochem.* 41, 552–556.

Zhang, G. Y., and Fang, B. S. (2007). LogitBoost classifier for discriminating thermophilic and mesophilic proteins. *J. Biotechnol.* 127, 417–424.

Zhang, M., Li, F. Y., Marquez-Lago, T. T., Leier, A., Fan, C., Kwoh, C. K., et al. (2019). MULTiPly: a novel multi-layer predictor for discovering general, and specific types of promoters. *Bioinformatics* 35, 2957–2965.

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. X. (2019). Meta-path methods for prioritizing candidate disease miRNAs. *IEEE-ACM Trans. Comput. Biol. Bioinf.* 16, 283–291. doi: 10.1109/TCBB.2017.2776280

Zheng, L., Huang, S., Mu, N., Zhang, H., Zhang, J., Chang, Y., et al. (2019). RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. *Database* 2019, 1–12.

Zhou, X. X., Wang, Y. B., Pan, Y. J., and Li, W. F. (2008). Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids.* 34, 25–33.

Zhu, X. J., Feng, C. Q., Lai, H. Y., Chen, W., and Hao, L. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.* 163, 787–793.

Zou, Q., Chen, L., Huang, T., Zhang, Z. G., and Xu, Y. G. (2017a). Machine learning and graph analytics in computational biomedicine. *Artif. Intell. Med.* 83, 1–1.

Zou, Q., Mrozek, D., Ma, Q., and Xu, Y. G. (2017b). Scalable data mining algorithms in computational biology and biomedicine. *Biomed. Res. Int.* 2017:3.

Zou, Q., Zeng, J. C., Cao, L. J., and Ji, R. R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354.

Zuo, Y. C., Chen, W., Fan, G. L., and Li, Q. Z. (2013). A similarity distance of diversity measure for discriminating mesophilic and thermophilic proteins. *Amino Acids* 44, 573–580. doi: 10.1007/s00726-012-1374-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 28 May 2020
doi: 10.3389/fbioe.2020.00496

WERFE: A Gene Selection Algorithm Based on Recursive Feature Elimination and Ensemble Strategy

Qi Chen[1,2], *Zhaopeng Meng*[1,3] *and Ran Su*[1,4]*

[1] School of Computer Software, College of Intelligence and Computing, Tianjin University, Tianjin, China, [2] Military Transportation Command Department, Army Military Transportation University, Tianjin, China, [3] Tianjin University of Traditional Chinese Medicine, Tianjin, China, [4] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China

Gene selection algorithm in micro-array data classification problem finds a small set of genes which are most informative and distinctive. A well-performed gene selection algorithm should pick a set of genes that achieve high performance and the size of this gene set should be as small as possible. Many of the existing gene selection algorithms suffer from either low performance or large size. In this study, we propose a wrapper gene selection approach, named WERFE, within a recursive feature elimination (RFE) framework to make the classification more efficient. This WERFE employs an ensemble strategy, takes advantages of a variety of gene selection methods and assembles the top selected genes in each approach as the final gene subset. By integrating multiple gene selection algorithms, the optimal gene subset is determined through prioritizing the more important genes selected by each gene selection method and a more discriminative and compact gene subset can be selected. Experimental results show that the proposed method can achieve state-of-the-art performance.

Keywords: WERFE, gene selection, RFE, ensemble, wrapper

OPEN ACCESS

Edited by:
Fengfeng Zhou,
Jilin University, China

Reviewed by:
Wen Zhang,
Huazhong Agricultural University, China
Xiuting Li,
*Singapore Bioimaging Consortium (A*STAR), Singapore*
Lin Gu,
National Institute of Informatics, Japan

***Correspondence:**
Ran Su
ran.su@tju.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *08 February 2020*
Accepted: *28 April 2020*
Published: *28 May 2020*

Citation:
Chen Q, Meng Z and Su R (2020) WERFE: A Gene Selection Algorithm Based on Recursive Feature Elimination and Ensemble Strategy. Front. Bioeng. Biotechnol. 8:496. doi: 10.3389/fbioe.2020.00496

1. INTRODUCTION

Gene expression data contains gene activity information, and it reflects the current physiological state of the cell, for example, whether the drug is effective on the cell, etc. It plays important roles in clinical diagnosis and drug efficacy judgment, such as assisting diagnosis and revealing disease occurrence mechanism (Lambrou et al., 2019). Gene expression data is rather complex, large in volume and grows fast. Since the dimensionality of gene expression data is often up to tens of thousands, it often consumes huge amount of time for analysis and it is difficult to make full use of it. The performance is not satisfied without proper processing. Although the dimensionality of gene expression data is extremely high, sometimes only a handful of the genes are informative and discriminative. Therefore, before the analysis of gene expression data, gene selection, which aims to reduce the dimensionality, is always carried out as the first step.

Gene selection is one special type of feature selection algorithm. It is a method to find the optimal gene subset from the original data set according to the actual needs (Su et al., 2019c). Over the years, many have studied the feature selection from different aspects. Kira et al. proposed a relief algorithm and defined the feature selection as a way to find the minimum feature subset that is necessary and sufficient to identify the target in ideal situations (Kira and Rendell, 1992). From the perspective of improving prediction accuracy, John et al. viewed the feature selection as a calculation procedure,

which could increase classification accuracy or reduce the feature dimension without reducing the classification accuracy (John et al., 1994). In the definition of Koller et al.'s study, feature selection aims to select the smallest feature subset, and ensure that the predicted class distribution is similar to the original data class distribution (Koller and Sahami, 1996). In Dash et al.'s study, they considered the feature selection as a method to select a feature subset as small as possible, and meet conditions that not reduce the classification accuracy significantly and not change the class distribution significantly (Dash and Liu, 1997). Although the definition varied from study to study, they had the same goal, that is, to find a smallest feature subset to identify the target effectively and achieve an accuracy as high as possible. Their definition of feature selection takes into account both classification accuracy and class distribution. Based on algorithm model structure, feature selection method has been divided into three categories: filter, wrapper, and embedded method. The gene selection can also be divided into these three categories.

Filter method is an early feature selection method, which selects the optimal feature subset at the first place and then using this feature subset to train the model. The two steps are independent. Another way to think about it is that it measures the importance of each feature, ranks the features, selects the top ranked features, or the top ranked percentage of all the features as the final feature subset. This method has often been used to pre-process the raw data. Phuong et al. (2005) proposed an effective method filter-based method for finding tagging SNPs. In the study of Zhang et al.'s, the filter method is used to pre-process 3D image data (Zhang et al., 2015). Roffo et al. (2016) proposed a new filter-based feature selection method which achieved state-of-the-art performance.

Unlike filter method, wrapper method uses the output of the learning model as the evaluation criterion of each feature subset. In wrapper method, feature selection algorithm plays as an integral part of the learning algorithm, and the classification output is used to evaluate the importance of the feature subsets (here we focus on classification issues). By generating different combinations of genes, evaluating each combination, and then comparing between combinations, this type of approach eventually becomes an optimization problem in terms of determination of the finally selected subset. The wrapper algorithm has been studied extensively. Zhang et al. (2014) built a spam detection model and used a wrapper-based feature selection method to extract crucial features. Li Yeh et al. used the idea of wrapper algorithm, combined the tabu search and binary particle swarm optimization for feature selection, and successfully classified the micro-array data (Li Yeh et al., 2009). Shah et al. developed a new approach for predicting drug effect, and decision-tree based wrapper method was used in a global searching mechanism to select significant genes (Shah and Kusiak, 2004).

Wrapper method integrates feature selection process and model training process into one entirety (Su et al., 2019b). That is, the feature selection is carried out automatically during the learning process. This method is often coupled with well-performed classification methods such as support vector machine (SVM) or random forests (RF) in order to improve the classification accuracy and efficiency. Wrapper method has shown impressive performance in gene studies. Su et al. proposed a MinE-RFE gene selection method which conducted the gene selection inside the RF classification algorithm and achieved good performance (Su et al., 2019b). They also proposed a gene selection algorithm combing GeneRank and gene importance to select gene signatures for Non-small cell lung cancer subtype classification (Su et al., 2019f). The third class, embedded method, is similar to wrapper methods. Different from the wrapper method, an intrinsic model building metric is used during learning in embedded approach. Duval et al. (2009) presented a memetic algorithm which was an embedded approach dealing with gene selection for supervised classification of micro-array data. Hernandez and Hao (2007) tried a genetic embedded approach which performed the selection task combining a SVM classifier and it gave highly competitive results.

Ensemble strategy has been used widely to deal with diverse types of issues (Wei et al., 2017a,b, 2018a; Wang et al., 2018; Zhang W. et al., 2018; Su et al., 2019d; Zhang et al., 2019a). It takes advantages of different algorithms and the optimal outcome is obtained based on the optimization of the multiple algorithms. In this study, we propose an wrapper approach for gene selection, named WERFE, to deal with classification issues within a recursive feature elimination (RFE) framework. This WERFE employs an ensemble strategy, takes advantages of a variety of gene selection methods and assembles the top selected genes in each approach as the final gene subset. By integrating multiple gene selection algorithms, the optimal gene subset is determined through prioritizing the more important genes of each gene selection method. A more compact and discriminative gene subset is then selected.

2. METHODOLOGY

2.1. Data Sets and Preprocessing

In our study, we used five data sets to validate the proposed method, RatinvitroH, Nki70, ZQ_188D, Prostate and Regicor. RatinvitroH was retrieved from Open TG-GATEs database, which is a large-scale toxicogenomics database (https://toxico.nibiohn.go.jp/english/index.html). It stores gene expression profiles and toxicological data derived from *in vivo* (rat) and *in vitro* (primary rat hepatocytes and primary human hepatocytes) exposed to 170 compounds at multiple dosages and time points (Yoshinobu et al., 2015; Su et al., 2018). Here we identified hepatotoxic compounds based on the toxicogenomics data. We used the liver toxicogenomics data of rat *in vitro* and we selected the data at 24 h as at this time point the gene expression is higher in the single-dose study (Otava et al., 2014; Su et al., 2019e). All 31,042 genes of 116 compounds in the database were picked to build and estimate the gene selection method. Gene expression levels at three concentrations, low, middle, and high were recorded and we employed the response at the high concentration to represent the potency of the drugs. The gene expression was profiled with Affymetrix GeneChip.

Nki70 is a data set assembling expression of 70 breast cancer-related genes of 144 samples. CPPsite (http://crdd.osdd.net/raghava/cppsite/) is a manually curated

TABLE 1 | The details of the five data sets.

Dataset	Gene number	Sample number
RatinvitroH	31,042	116
Nki70	70	144
ZQ_188D	188	9,024
Prostate	100	50
Regicor	22	300

database of experimentally validated 843 cell-penetrating peptides (CPPs) (Gautam et al., 2012), and CPPsite3.0 is the updated version of CPPsite2.0 (Piyush et al., 2015). ZQ_188D is derived from CPPsite3.0. It picks 188 CPPs of 9,024 samples. The Prostate data set contained 100 genes and 50 samples and it was used for cancer classification based on gene expression (Torrente et al., 2013). Regicor data set contained 22 genes and 300 samples (Subirana et al., 2014). It was used to identify death using cardiovascular risk factors. **Table 1** shows the details of the five data sets we used in this study.

2.1.1. Support Vector Machine (SVM)

SVM is a widely used classification and regression analysis method in machine learning. It maps the raw data into high dimensional space through kernel functions to make the data linearly separable (Wang et al., 2019; Wei et al., 2019a,b). It was developed in Vapnik et al.'s study of statistical learning theory (Cortes and Vapnik, 1995), with the core idea to find the hyperplane between different categories, so that samples in different categories can be grouped into different sides of the separating hyperplane as far as possible. The early SVM was flat and limited. Then using more complicated kernel function, the application scope of SVM was greatly enlarged (Zhang N. et al., 2018).

SVM has the cost function as follows (Su et al., 2019a):

$$J(\theta) = C\sum_{i=1}^{M}[\, y^i cost_1(\theta^T x^i) + (1-y^i) cost_0(\theta^T x^i)\,] + \frac{1}{2}\sum_{j=1}^{\gamma}\theta_j{}^2 \tag{1}$$

where θ is the adjustable parameter of the model and γ is the number of θ; M is the number of the samples. y^i represents the category of the i-th sample. Here we considered binary classification with label 0 and 1. $cost_1$ and $cost_0$ are the objective function when y^i is equal to 1 and 0, respectively. C is the degree of penalty for controlling mis-classified training samples. It can only be set as a positive value. Here we used the SVM with linear kernel.

2.1.2. Random Forest (RF)

Random forest (RF) is another classifier we used to train the model and obtain the importance of genes. RF is a method of discriminating and classifying data through voting of different classification trees (Ho, 1995; Gong et al., 2019; Lv et al., 2019). It is an ensemble learning method composed of multiple tree classifiers. It takes a random sample from the sample set with replacement, and then the samples are fed into the tree classifiers. Finally the class of the sample is determined by voting with the principle of majority rule. As it classifies the data, it can also provide the importance score of each variable (gene) and evaluate the role of each variable in the classification. In the process of applying RF, two parameters need to be determined. One is the number of samples selected each time and the other one is the number of decision trees in the random forest. The two parameters are determined according to the size of the data set.

2.2. Gene Selection Based on Recursive Feature Elimination

Gene selection was widely used in a number of fields (Fajila, 2019; Shahjaman et al., 2019). The most popular methods include Fisher-based methods (Gu et al., 2011), Relief-based methods (Robnik-Sikonja and Kononenko, 1997), FSNM methods (Nie et al., 2010), and mRMR (Peng et al., 2005) etc. All of these methods firstly rank the genes based on an evaluation criteria. Then based on the rank of genes, an appropriate gene subset is determined. However, the relationship between the number of selected genes and the classification precision cannot be fully reflected using these gene selection methods. Recently, Su et al. developed an algorithm balancing performance and gene number under the framework of recursive feature elimination (RFE) (Su et al., 2019b). Inspirited by their work, we designed the WERFE inside the RFE framework.

The RFE is a greedy algorithm which iteratively builds gene sets and the optimal subset is chosen from them. It was proposed by Guyon et al. with the intention to detect cancer (Guyon et al., 2002). The RFE iteratively eliminates the least important genes and conducts classification based on the new gene subsets. All the gene subsets are evaluated based on their classification performance. In our study, the finally selected subset is the one with the highest accuracy.

2.3. The Proposed Gene Selection Algorithm WERFE

2.3.1. Gene Ranking Algorithm

In this study, we developed a gene selection algorithm, named WERFE. Its main idea is to integrate two or more independent gene selection algorithms and the final decision is made based on all of these algorithms. The WERFE can be divided into two parts, the first is the gene ranking algorithm, and the second part is the determination of the optimal gene subset. **Figure 1** illustrates the entire process of the gene ranking algorithm. Cross validation is widely used to evaluate the model (Liu et al., 2017; Zeng et al., 2017a, 2018). Therefore, the WERFE was performed inside a ten-fold cross validation procedure. In each fold, different gene selection algorithms used the training and test data to pick gene subsets. Then we put all the selected genes which were obtained from different algorithms into a voting pool (Chen et al., 2018). We counted the votes of each gene in the voting pool and ranked the genes based on the votes. In this way, we obtained a list of genes, G_R, ranking from high to low. This ranking would be used for further gene selection. The pseudo code in Algorithm 1 shows the process of gene ranking. Here ten-Fold cross validation was

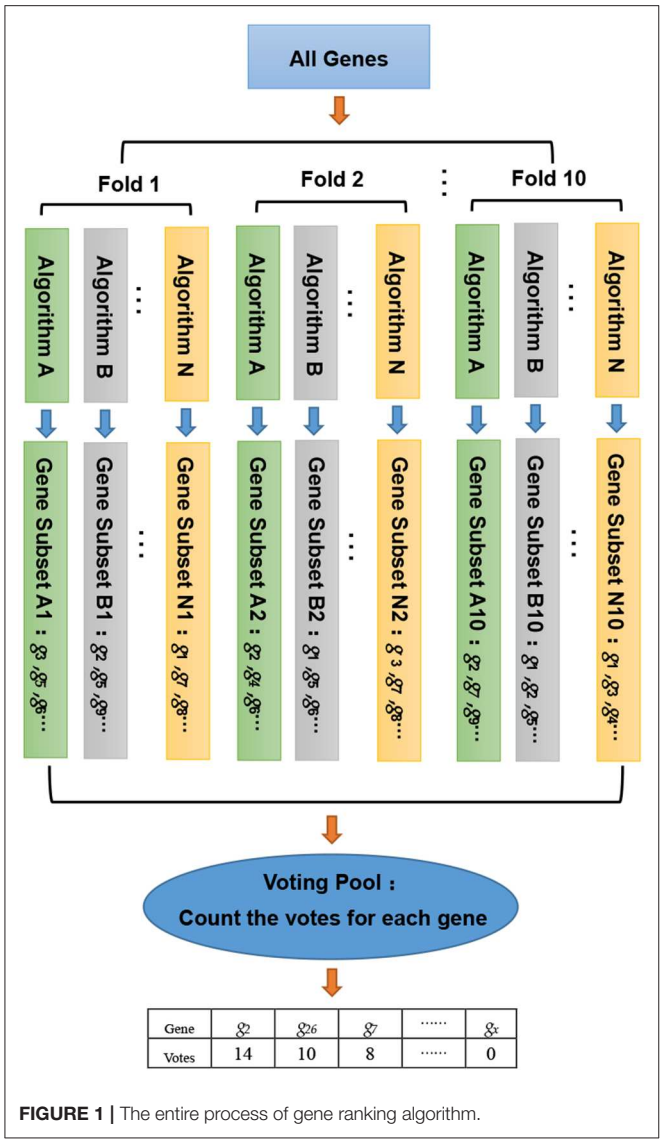


FIGURE 1 | The entire process of gene ranking algorithm.

used in WERFE, and two gene selection algorithms RF and SVM are integrated.

2.3.2. Determination of the Optimal Gene Subset

In our study, we generated different gene subsets, gathered all the genes selected through different gene selection algorithms, and chose an optimal gene subset according to the votes for each gene. We assume that G_{final} is the gene subset eventually selected, and there are p genes in G_{final}. According to the votes we obtained for each gene, G_{final} is acquired as follows:

$$G_{final} = G_r : \{G_{r1}, \cdots, G_{rl}\} \mid max(Acc(G_r, t_0)), \quad t_f > t_0, t_f \in [1, 10N], t_0 \in [0, 10N - 1]. \tag{2}$$

where G_r is the top ranked l genes of G_R; Each of these l genes present vote value t_f larger than a threshold t_0. $Acc()$ means the accuracy values of G_r. Assuming we integrated N gene selection algorithms, and thus we would have N ten-fold cross validation, respectively. Since all the selected subsets would be put into the voting pool, it made that the number of votes for each gene ranged from 0 to $10 \times N$. Therefore, the t_f ranges from 1 to $10 \times N$ and the threshold t_0 ranged from 0 to $10 \times N - 1$. Each time, we selected genes with t_f larger than t_0 and tested the performance for the selected genes. As we set various t_0 values and each t_0 corresponded to a gene subset with l genes, the performance using this subset could be calculated. Thus, we obtained a list of accuracy values corresponding to each t_0. Then the subset with the highest accuracy was selected as the final gene subset.

Algorithm 1: Gene ranking of Wrapper Embedded Recursive Feature Elimination (WERFE)

Input: Input data $\mathbf{X} : \mathbf{x_1}, \mathbf{x_1} \ldots \mathbf{x_m}$ and labels $\mathbf{Y} : \mathbf{y_1}, \mathbf{y_1} \ldots \mathbf{y_m}$, where m is the number of samples. $\mathbf{x}$ is n-dimensional gene vector.s is the step size of RFE.

Output: Ranked genes G_R of all the genes.

for $k = 1 : 10$ **do**
  The data set was randomly divided into ten equal parts;
  Keep one part as a test data; The remaining nine parts are used as training data;
  while X is not empty **do**
    Train a model based on training data of **X** using SVM;
    Calculate the prediction accuracy of the model using the test data;
    Obtain the weight of each gene produced from SVM;
    Remove s least weighted genes and update **X**;
  end while
  Obtain the gene subset G_1 with the highest prediction accuracy;
  while X is not empty **do**
    Train a model based on training data of **X** using RF;
    Calculate the prediction accuracy of the model using the test data;
    Obtain the importance of each gene produced from RF;
    Remove s least weighted genes and update **X**;
  end while
  Obtain the gene subset G_2 with the highest prediction accuracy;
  Count the votes for all the genes contained in both G_1 and G_2;
end for
Rank genes based on votes and obtain G_R.

2.4. Performance Measurements

Classification sensitivity, specificity and accuracy are important indicators for performance evaluation, which are widely used in diverse applications (Zeng et al., 2017b; Wei et al., 2018b, 2019c; Jin et al., 2019; Zhang et al., 2019b). In this study, we used these three measurements to estimate the performance of the gene subset. They are formulated as follows:

TABLE 2 | Voting and predicted results on RatinvitroH data set using WERFE.

t_f	t_f	GN[a]	Acc.RF[b]	Sen.RF	Spe.RF	Acc.SVM	Sen.SVM	Spe.SVM
19	20	0	–	–	–	–	–	–
18	19, 20	2	75.79	74.58	56.19	60.45	100	0
17	18–20	17	77.30	81.10	47.26	57.80	95.42	3.33
16	17–20	685	77.15	81.46	48.10	76.67	90.69	60.48
15	16–20	1,092	77.43	85.82	53.10	75.00	82.27	69.76
14	15–20	6,142	75.70	80.17	43.10	65.53	69.57	65.48
0	1–20	31,042	76.84	81.74	66.62	60.23	49.52	50.71

[a]*GN, gene number.*
[b]*Acc.RF, Acc using RF as classifier. Other abbreviations in the first row mean in the same way.*

$$\begin{aligned} \text{Sensitivity(Sen)} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\%, \\ \text{Specificity(Spe)} &= \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100\%, \\ \text{Accuracy(Acc)} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \times 100\%. \end{aligned} \tag{3}$$

The receive operating characteristic (ROC) curves as well as the area under the ROC, named AUC, were also implemented to measure the performance.

3. EXPERIMENTAL RESULTS

3.1. Performance Using Different Voting Threshold

Theoretically, the proposed WERFE can ensemble any number of gene selection algorithms. Here in order to made the calculation efficient, we integrated two of the most popular wrapper gene selection algorithms, the RFRFE and SVMRFE, and performed the ten-fold cross validation to pick the most informative genes. In each fold, using the same data splitting strategy, RFRFE and SVMRFE selected their gene subsets respectively. Then we obtained 20 gene subsets considering the ten-fold cross validation. These gene subsets were gathered and put into the voting pool. Based on votes of each gene, we obtained gene rank G_R, which is in descending order. Then we re-generated gene subsets by setting different threshold t_0. We evaluated the classification performance of each new gene subset and made the final decision. Here we used RF and SVM as the classifier respectively after obtaining the final gene subset. We used RatinvitroH to validate the WERFE as it is high in dimension. **Table 2** shows part of the intermediate outcome of applying WERFE method to RatinvitroH data set. Here as the vote of each gene ranges from 1 to 20, we set the threshold t_0 from 0 to 19.

From **Table 2**, it shows that no gene has 20 votes. It can also be seen that RF performs significantly better than SVM. Two genes obtain 19 votes, and the classification using gene subset composed of these two genes has reached 75.95% of accuracy, 74.58% of sensitivity, and 56.19% of specificity, based on RF. With the increase of the number of genes in the gene subset, the classification accuracy ranges from 75.70 to 77.43%, sensitivity ranges from 74.58 to 85.82%, and specificity ranges from 43.10 to 66.62%, using RF evaluation method. The accuracy achieves the highest when the t_0 is set to 15. However, a huge number of genes are obtained, which makes the computation slow down. In order to balance the gene number and the accuracy, we selected 17 genes as the final gene subset when t_0 equals to 17 and t_f ranges from 18 to 20, and obtained an accuracy of 77.30%, sensitivity of 81.10%, and specificity of 47.26%. That means we can obtain a relatively high classification result with a small number of genes.

TABLE 3 | Comparison with RFRFE.

Dataset	WERFE				RFRFE			
	GN[a]	Acc	Sen	Spe	GN[a]	Acc	Sen	Spe
RatinvitroH	17	77.30	81.10	47.26	11	72.27	68.71	34.95
Nki70	5	82.27	49.75	86.13	43	80.15	35.36	83.92
ZQ_188D	1	93.81	98.43	100.00	41	95.80	17.29	99.98
Prostate	4	98.00	95.00	100.00	3	95.31	90.00	100.00
Regicor	4	76.54	65.34	62.71	5	77.76	68.95	64.70

[a]*GN, gene number.*

TABLE 4 | Comparison with SVMRFE.

Dataset	WERFE				SVMRFE			
	GN[a]	Acc	Sen	Spe	GN[a]	Acc	Sen	Spe
RatinvitroH	17	77.30	81.10	47.26	51	70.30	80.86	53.79
Nki70	5	82.27	49.75	86.13	25	77.10	57.42	88.17
ZQ_188D	1	93.81	98.43	100.00	1	93.81	0	100.00
Prostate	4	98.00	95.00	100.00	42	98.00	96.67	100.00
Regicor	4	76.54	65.34	62.71	3	65.33	62.21	72.24

[a]*GN, gene number.*

3.2. Comparison and Analysis With Non-ensemble Algorithms

In theory, our ensemble strategy assumes that integrating more gene selection algorithms is able to give better performance, yet will lead to large calculation cost. Here we only integrated two wrapper algorithms, RFRFE and SVMRFE in the proposed WERFE. We compared WERFE with RFRFE and SVMRFE, respectively and show the results in **Tables 3**, **4**. The comparison was made based on the five data sets.

In **Table 3**, for RatinvitroH, Nki70 and Prostate, it can be clearly seen that the classification accuracy of WERFE is similar or higher than the RFRFE method and the gene subset number is similar or less; while for ZQ_188D and Regicor, although the performance is slightly lower, the gene number is also smaller. The overall performance of WERFE is better than the RFRFE.

From **Table 4**, we can find that the WERFE performs better on all the five data set than SVMRFE. The accuracy is higher or similar and gene number is smaller or similar.

Comparing across tables, we find WERFE outperforms the other two methods. For example, Nki70's classification accuracy

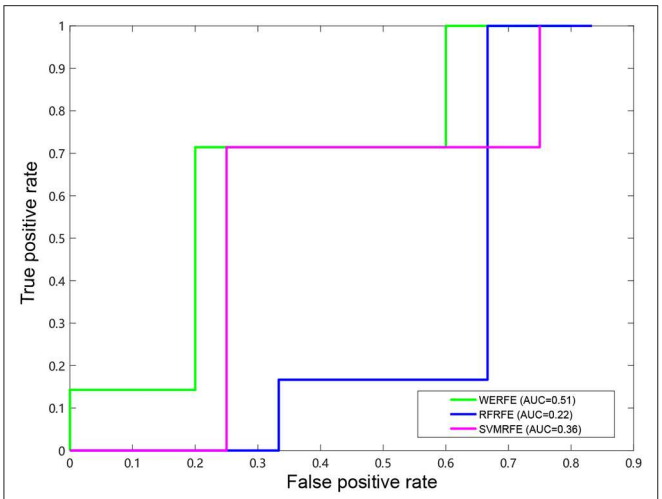


FIGURE 2 | ROC curve on RatinvitroH dataset.

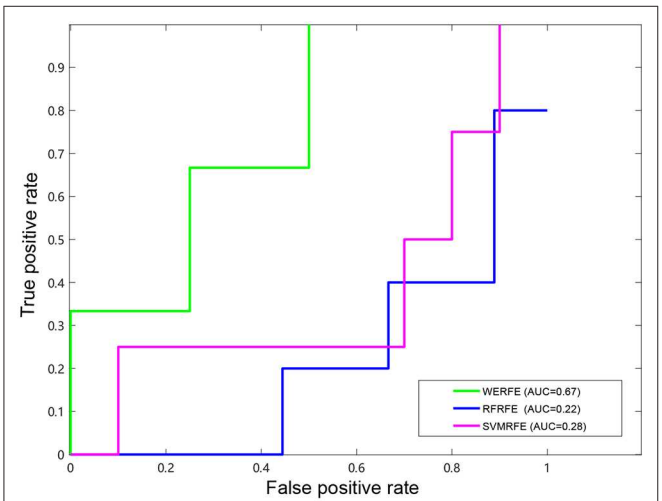


FIGURE 3 | ROC curve on Nki70 dataset.

reaches 82.27% using WERFE algorithm. While using RFRFE, the accuracy is 80.15% (**Table 3**) and using SVMRFE, the classification accuracy is 77.10% (**Table 4**). The number of selected genes is 5, 43, and 25, respectively. WERFE achieves the highest accuracy using the least number of genes. It is obvious to see the similar trend for the other data sets. Even the accuracy is lower using WERFE, e.g., for data ZD_188D, the accuracy is 2% lower, the much smaller number of gene subset can compensate the slight decrease of accuracy.

Figures 2, **3** show the ROC curves of the three methods on RatinvtroH and Nki70 data set. WERFE stays on the top left of RFRFE and SVMRFE, which shows it performs better on RatinvtroH and Nki70 data sets than the other two methods.

TABLE 5 | Performance between lightGBM with WERFE and without WERFE.

Dataset	With WERFE				Without WERFE			
	GN[a]	Acc	Sen	Spe	GN[a]	Acc	Sen	Spe
RatinvitroH	17	77.30	81.10	47.26	31042	59.13	73.90	36.93
Nki70	5	82.27	49.75	86.13	70	63.60	31.25	80.00
ZQ_188D	1	93.81	98.43	100.00	188	96.80	61.50	98.90
Prostate	4	98.00	95.00	100.00	100	89.80	88.00	91.70
Regicor	4	76.54	65.34	62.71	22	59.90	64.00	55.70

[a] *GN, gene number.*

3.3. Validation Using Other Classifiers

We have shown the results of WERFE using both RF and SVM as the classifiers in section 3.1. Besides classification, RF and SVM also provide gene ranking criteria for WERFE. In order to provide a fair evaluation of WERFE, we used another algorithm, LightGBM algorithm to classify the five data sets and we compared the results with or without WERFE gene selection. LightGBM, a gradient Boosting framework proposed in recent years (Ke et al., 2017), is a distributed and efficient machine learning algorithm based on Gradient Boosting Decision Tree (GBDT) with two key techniques, Gradient-based One-Side Sampling (GOSS), and Exclusive Feature Bundling (EFB). It has been used in gene studies and shown impressive performance (Su et al., 2019e). We show the results using lightGBM with WERFE and lighGBM without WERFE in **Table 5**.

Table 5 shows that, with the exception of the ZQ_188D data set, the classification accuracy and sensitivity of lightGBM plus WERFE is much higher than that of using LightGBM alone. And the WERFE greatly reduces the gene number. This shows that WERFE algorithm performs well in gene selection of most data sets and achieves the purpose of using fewer genes to reach higher classification accuracy.

3.4. Comparison With Other Gene Selection Algorithms

We also compared the WERFE with some widely used gene selection approaches including Nie et al.'s method (Nie et al., 2010), Fisher score-based approach and ReliefF approach (Kononenko et al., 1997). We denoted them with FSNM, Fisher, and ReliefF, respectively. These three gene algorithms were conducted combining an incremental search method (ISM). Firstly, the genes were ranked (descending order) using FSNM, Fisher score, and ReliefF, respectively. Then according to the rank, we assumed the basic gene subset include the top ranked θ genes. Next, by adding step size genes each time on top of the basic gene subset, we constructed a group of gene subsets. In order to be consistent with the evaluation method of WERFE algorithm, we also used RF and SVM as the classification methods, and took the subset with the highest accuracy as the result of gene selection. In our study, we set θ to 10 and the step size to 10. The results are shown in **Tables 6**, **7** for data RatinvitroH and Nki70, respectively.

Table 6 shows that, in the RF column, FSNM algorithm uses the gene subset composed of 60 genes to obtain the classification

TABLE 6 | Comparison with other gene selection algorithms on RatinvitroH.

Algorithms	RF				SVM			
	GN[a]	Acc	Sen	Spe	GN[a]	Acc	Sen	Spe
WERFE	17	77.30	81.10	47.26	685	76.67	90.69	60.48
FSNM	60	77.50	83.65	43.52	100	74.85	83.95	60.02
Fisher	20	73.39	69.60	34.02	10	59.85	93.02	14.83
ReliefF	40	73.21	74.60	40.45	80	62.20	97.46	8.17

[a]*GN, gene number.*

TABLE 7 | Comparison with other gene selection algorithms on Nki70.

Algorithms	RF				SVM			
	GN[a]	Acc	Sen	Spe	GN[a]	Acc	Sen	Spe
WERFE	5	82.27	49.75	86.13	5	72.33	33.00	92.17
FSNM	63	80.85	22.93	88.06	28	81.33	61.79	90.86
Fisher	35	81.46	35.33	92.94	35	74.24	46.12	89.14
ReliefF	21	80.31	39.36	82.11	35	75.76	50.62	87.86

[a]*GN, gene number.*

accuracy of 77.50%, which is the highest among the four algorithms, and the classification accuracy obtained by WERFE algorithm by using the gene subset composed of 17 genes is 77.30%. Through the comparison of FSNM and WERFE, we find that, although the classification accuracy is similar, the number of genes selected by WERFE algorithm is 20, while the number of genes selected by FSNM is 60, which is 40 more than that of WERFE. Therefore, it is reasonable to choose the WERFE in real applications considering both performance and computation consumption. In the SVM column, the WERFE selects more genes than FSNM but achieved an increase of 2% of accuracy.

Similarly, we applied these gene selection algorithms on the Nki70 dataset. **Table 7** shows a comparison of the results of these methods. For the RF column, it is easy to find that WERFE method has the highest classification accuracy 82.27%, when 5 genes were selected as the gene subset. But in the SVM column the WERFE has the worst performance. This indicates that it is better to combine WERFE with RF to perform the gene selection and classification.

4. CONCLUSION

A good gene selection can improve the performance of the classification and play an important role in further analysis. It should take both gene number and classification accuracy into account. In this paper, we proposed an ensemble gene selection algorithm, WERFE, which belongs to a wrapper method within a RFE framework, and conducts the gene selection combining cross validation. The WERFE takes good advantages of multiple gene selection algorithms. Through evaluating each gene with different gene selection algorithms, a small set of genes are selected and the classification accuracy is also improved.

It is expected that better performance can be achieved if integrating more gene selection algorithms. Our study integrates two gene selection algorithms in order to reduce the computation cost. Some of our operations are inspired by the non-ensemble embedded algorithm that we proposed in previous studies (Chen et al., 2018). For instance, we also completed the integration of the algorithm within ten-fold cross-validation. In each fold, under the same training set and test set, different gene selection algorithms were used to obtain the optimal gene subsets, respectively. Then we put the genes contained in each subset of each fold into a voting pool to obtain the votes for each gene. The number of votes of each gene in the voting pool is an important indicator for us to evaluate the gene's importance and based on the votes, we obtained a gene ranking. We constructed new gene subsets according to the ranking and a pre-set threshold was set. Eventually each gene subset was evaluated and a final gene subset was selected.

We used five data sets (RatinvitroH, Nki70, ZQ_180D, Prostate, and Regicor) to validate the proposed method. In order to verify the effectiveness of the gene selection algorithm, we designed three groups of comparative experiments. Firstly, we chose two wrapper algorithms, which are also the two basic algorithms integrated into our proposed algorithm, to compare with the WERFE. The results show that the proposed method outperforms the other two wrapper algorithms. Secondly, we used another classification algorithm, lightGBM, to evaluate the proposed method. We compared the performance between methods using WERFE and not using WERFE. And the results show that lightGBM performs better when using WERFE. Finally, we compared the WERFE with three other gene selection algorithms. It shows from the results that WERFE is best in both improving classification accuracy and reducing gene number. However, there are some limitations of the proposed method. For instance, this method needs to consume more computing resources if more gene selection algorithms are integrated. When the number of genes is large, the operation time will be relatively long.

In the future, we will test this algorithm on more types of data sets to further improve the algorithm. At the same time, we will also try to integrate more gene selection methods, aiming to evaluate the importance of genes in a more objective way, and meanwhile reduce the calculation time. We target to solve this through deep learning method.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://toxico.nibio.go.jp/english/index.html; http://crdd.osdd.net/raghava/cppsite/.

AUTHOR CONTRIBUTIONS

RS conceived and designed the experiments and revised the manuscript. QC collected the data, performed the analysis, and wrote the paper. ZM contributed the analysis tools and participated in revising the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 61702361), Natural Science Foundation of Tianjin (No. 18JCQNJC00800), the Science and Technology Program of Tianjin, China (Grant No. 16ZXHLGX00170), the National Key Technology R&D Program of China (Grant No. 2015BAH52F00), and the National Key Technology R&D Program of China (Grant No. 2018YFB1701700).

REFERENCES

Chen, Q., Meng, Z., Liu, X., Jin, Q., and Su, R. (2018). Decision variants for the automatic determination of optimal feature subset in RF-RFE. *Genes* 9:301. doi: 10.3390/genes9060301

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn*. 20, 273–297. doi: 10.1007/BF00994018

Dash, M., and Liu, H. (1997). Feature selection for classification. *Intell. Data Anal.* 1, 131–156. doi: 10.3233/IDA-1997-1302

Duval, B., Hao, J. K., and Hernandez, J. C. H. (2009). "A memetic algorithm for gene selection and molecular classification of cancer," in *Genetic & Evolutionary Computation Conference* (Montreal, CA), 201–208. doi: 10.1145/1569901.1569930

Fajila, M. N. F. (2019). Gene subset selection for leukemia classification using microarray data. *Curr. Bioinformatics* 14, 353–358. doi: 10.2174/1574893613666181031141717

Gautam, A., Singh, H., Tyagi, A., Chaudhary, K., Kumar, R., Kapoor, P., et al. (2012). CPPsite: a curated database of cell penetrating peptides. *Database* 2012:bas015. doi: 10.1093/database/bas015

Gong, Y., Niu, Y., Zhang, W., and Li, X. (2019). A network embedding-based multiple information integration method for the MiRNA-disease association prediction. *BMC Bioinformatics* 20:468. doi: 10.1186/s12859-019-3063-3

Gu, Q., Li, Z., and Han, J. (2011). "Generalized fisher score for feature selection," in *Twenty-seventh Conference on Uncertainty in Artificial Intelligence* (Barcelona), 266–273.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn*. 46, 389–422. doi: 10.1023/A:1012487302797

Hernandez, J. C. H., and Hao, J. K. (2007). "A genetic embedded approach for gene selection and classification of microarray data," in *European Conference on Evolutionary Computation* (Valencia), 90–101. doi: 10.1007/978-3-540-71783-6_9

Ho, T. K. (1995). "Random decision forests," in *International Conference on Document Analysis & Recognition* (Montreal, CA), 278–282.

Jin, Q., Meng, Z., Pham, T. D., Chen, Q., Wei, L., and Su, R. (2019). DUNet: A deformable network for retinal vessel segmentation. *Knowl. Based Syst*. 178, 149–162. doi: 10.1016/j.knosys.2019.04.025

John, G., Kohavi, R., and Pfleger, K. (1994). "Irrelevant features and the subset selection problem," in *Machine Learning Proceedings* (New Brunswick; New Jersey, NJ), 121–129. doi: 10.1016/B978-1-55860-335-6.50023-4

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree," in *31st Conference on Neural Information Processing Systems* (Long Beach, CA), 3149–3157.

Kira, K., and Rendell, L. A. (1992). "The feature selection problem: traditional methods and a new algorithm," in *Tenth National Conference on Artificial Intelligence* (San Jose, CA), 129–134.

Koller, D., and Sahami, M. (1996). "Toward optimal feature selection," in *Thirteenth International Conference on International Conference on Machine Learning* (Bari), 284–292.

Kononenko, I., Simec, E., and Robnik-Sikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with RELIEFF. *Appl. Intell.* 7, 39–55. doi: 10.1023/A:1008280620621

Lambrou, G. I., Sdraka, M., and Koutsouris, D. (2019). The "gene cube": A novel approach to three-dimensional clustering of gene expression data. *Curr. Bioinformatics* 14, 721–727. doi: 10.2174/1574893614666190116170406

Li Yeh, C., Cheng-Huei, Y., and Cheng Hong, Y. (2009). Tabu search and binary particle swarm optimization for feature selection using microarray data. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 16, 1689–1703. doi: 10.1089/cmb.2007.0211

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2017). Inferring MicroRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 14, 905–915. doi: 10.1109/TCBB.2016.2550432

Lv, Z., Jin, S., Ding, H., and Zou, Q. (2019). A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.* 7:215. doi: 10.3389/fbioe.2019.00215

Nie, F., Huang, H., Cai, X., and Ding, C. (2010). "Efficient and robust feature selection via joint ℓ_21-norms minimization," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems, Vol. 2* (Kyoto), 1813–1821.

Otava, M., Shkedy, Z., and Kasim, A. (2014). Prediction of gene expression in human using rat in *vivo* gene expression in Japanese toxicogenomics project. *Syst. Biomed.* 2, 8–15. doi: 10.4161/sysb.29412

Peng, H. C., Long, F. H., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Phuong, T. M., Lin, Z., and Altman, R. B. (2005). "Choosing SNPs using feature selection," in *Computational Systems Bioinformatics Conference* (Stanford, CA), 301–309. doi: 10.1109/CSB.2005.22

Piyush, A., Sherry, B., Sadullah, U. S., Sandeep, S., Kumardeep, C., S., and Ankur, G. (2015). CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic Acids Res.* 44, D1098–D1103. doi: 10.1093/nar/gkv1266

Robnik-Sikonja, M., and Kononenko, I. (1997). "An adaptation of relief for attribute estimation in regression," in *Fourteenth International Conference on Machine Learning* (Nashville, TN), 296–304.

Roffo, G., Melzi, S., and Cristani, M. (2016). "Infinite feature selection," in *IEEE International Conference on Computer Vision* (Santiago), 4202–4210. doi: 10.1109/ICCV.2015.478

Shah, S. C., and Kusiak, A. (2004). Data mining and genetic algorithm based gene/SNP selection. *Artif. Intell. Med.* 31, 183–196. doi: 10.1016/j.artmed.2004.04.002

Shahjaman, M., Kumar, N., and Mollah, N. H. (2019). Performance improvement of gene selection methods using outlier modification rule. *Curr. Bioinformatics* 14, 491–503. doi: 10.2174/1574893614666181126110008

Su, R., Liu, T., Sun, C., Jin, Q., Jennane, R., and Wei, L. (2019a). Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. *Neurocomputing*. 385, 300–309. doi: 10.1016/j.neucom.2019.12.083

Su, R., Liu, X., and Wei, L. (2019b). MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy. *Brief. Bioinformatics*. doi: 10.1093/bib/bbz021

Su, R., Liu, X., Wei, L., and Zou, Q. (2019c). Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods* 166, 91–102. doi: 10.1016/j.ymeth.2019.02.009

Su, R., Liu, X., Xiao, G., and Wei, L. (2019d). Meta-GDBP: a high-level stacked regression model to improve anti-cancer drug response prediction. *Brief. Bioinformatics*. doi: 10.1093/bib/bbz022

Su, R., Wu, H., Liu, X., and Wei, L. (2019e). Predicting drug-induced hepatotoxicity based on biological feature maps and diverse classification strategies. *Brief. Bioinformatics*. doi: 10.1093/bib/bbz165

Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2018). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 16, 1231–1239. doi: 10.1109/TCBB.2018.2858756

Su, R., Zhang, J., Liu, X., and Wei, L. (2019f). Identification of expression signatures for Non-Small-Cell Lung Carcinoma subtype classification. *Bioinformatics*. 36, 339–346. doi: 10.1093/bioinformatics/btz557

Subirana, I., Sanz, H., and Vila, J. (2014). Building bivariate tables: the comparegroups package for R. *J. Stat. Softw.* 57, 1–16. doi: 10.18637/jss.v057.i12

Torrente, A., López-Pintado, S., and Romo, J. (2013). DepthTools: an R package for a robust analysis of gene expression data. *BMC Bioinformatics* 14:237. doi: 10.1186/1471-2105-14-237

Wang, B., Lu, K., Long, H., Zhou, Y., Zheng, C.-H., Zhang, J., et al. (2018). Early stage identification of Alzheimer's disease using a two-stage ensemble classifier. *Curr. Bioinformatics* 13, 529–535. doi: 10.2174/1574893613666180328093114

Wang, Y., Shi, F., Cao, L., Dey, N., Wu, Q., Ashour, A. S., et al. (2019). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinformatics* 14, 282–294. doi: 10.2174/1574893614666190304125221

Wei, L., Chen, H., and Su, R. (2018a). M6APred-EL: a sequence-based predictor for identifying n6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids* 12, 635–644. doi: 10.1016/j.omtn.2018.07.004

Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018b). Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* 117, 212–217. doi: 10.1016/j.jpdc.2017.08.009

Wei, L., Su, R., Wang, B., Li, X., Zou, Q., and Gao, X. (2019a). Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites. *Neurocomputing* 324, 3–9. doi: 10.1016/j.neucom.2018.04.082

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Shi, G., Ji, Z. L., and Zou, Q. (2019b). Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 16, 1264–1273. doi: 10.1109/TCBB.2017.2670558

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z. S., and Zou, Q. (2019c). CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J. Proteome Res.* 16, 2044–2053. doi: 10.1021/acs.jproteome.7b00019

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017b). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Yoshinobu, I., Noriyuki, N., Tomoya, Y., Atsushi, O., Yasuo, O., Tetsuro, U., et al. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43:D921. doi: 10.1093/nar/gku955

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017a). Prediction and validation of disease genes using HeteSim scores. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 14, 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017b). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* 13:e1005420. doi: 10.1371/journal.pcbi.1005420

Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 34, 2425–2432. doi: 10.1093/bioinformatics/bty112

Zhang, N., Sa, Y., Guo, Y., Lin, W., Wang, P., and Feng, Y. (2018). Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinformatics* 13, 50–56. doi: 10.2174/1574893611666160608102537

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019a). SFLLN: A sparse feature learning ensemble method with linear neighborhood regularization for predicting drug-drug interactions. *Inform. Sci.* 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019b). A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 1–1. doi: 10.1109/TCBB.2019.2931546

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018). SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting lncRNA-protein interactions. *PLoS Comput. Biol.* 14:e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, Y., Dong, Z., Phillips, P., Wang, S., Ji, G., Yang, J., et al. (2015). Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. *Front. Comput. Neurosci.* 9:66. doi: 10.3389/fncom.2015.00066

Zhang, Y., Wang, S., Phillips, P., and Ji, G. (2014). Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. *Knowl. Based Syst.* 64, 22–31. doi: 10.1016/j.knosys.2014.03.015

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 May 2020
doi: 10.3389/fgene.2020.00403

Identification of CircRNA–miRNA–mRNA Regulatory Network in Gastrointestinal Stromal Tumor

*Fang-wen Zou[1], Ding Cao[2], Yi-fang Tang[3], Long Shu[1], Zhongkun Zuo[2] and Lei-yi Zhang[2]**

[1] Department of Oncology, The Second Xiangya Hospital of Central South University, Changsha, China, [2] Department of Minimally Invasive Surgery, The Second Xiangya Hospital of Central South University, Changsha, China, [3] Department of Anesthesiology, The Second Xiangya Hospital of Central South University, Changsha, China

Circular RNA (circRNA) abnormal expression and regulation are involved in the occurrence and development of a variety of tumors. However, the role of circRNAs still remains unknown in gastrointestinal stromal tumors (GISTs). In the present study, the differential circRNA expression profile of GISTs was screened by human circRNAs chip and verified by qRT-PCR. The circRNA–miRNA–mRNA regulatory network was constructed using the cytoHubba plugin based on the Cytoscape software. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were performed to explore circRNA functions. Six significantly differential circRNAs were also verified in 20 pairs of GISTs and adjacent tissues by qRT-PCR. The result showed that a total of 543 differentially expressed circRNAs were identified in GISTs, of which 242 were up-regulated and 301 were down-regulated. Additionally, the circRNA–miRNA–mRNA network contained six circRNAs, 30 miRNAs, and 308 mRNAs, and the targeted mRNAs were associated with "regulation of biological process," "intracellular organelle," "protein binding," and enriched in Wnt signaling pathway. Furthermore, qRT-PCR demonstrated that hsa_circRNA_061346, hsa_circRNA_103114, and hsa_circRNA_103870 were significantly up-regulated in GISTs (n = 20), and hsa_circRNA_405324, hsa_circRNA_406821, and hsa_circRNA_000361 were dramatically down-regulated in GISTs (n = 20). In addition, all of these circRNAs were shown to have high diagnostic values, and most of them were significantly associated with tumor size, mitotic figure, and malignant degrees in GISTs ($P < 0.05$). Therefore, we concluded that circRNAs were abnormally expressed in GISTs, and the circRNA–miRNA–mRNA regulatory network plays an important role in the occurrence and development of GISTs. Also, the identified six candidate circRNAs might be critical circRNAs and may present as potential diagnostic biomarkers for GISTs.

Keywords: gastrointestinal stromal tumors, circular RNA, miRNA, mRNA, biomarker

OPEN ACCESS

Edited by:
Yungang Xu,
The University of Texas Health Science Center at Houston, United States

Reviewed by:
Vincenzo Bonnici,
University of Verona, Italy
Pavel Loskot,
Swansea University, United Kingdom

***Correspondence:**
Lei-yi Zhang
zhangleiyixy@sina.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *13 January 2020*
Accepted: *30 March 2020*
Published: *28 May 2020*

Citation:
Zou F, Cao D, Tang Y, Shu L, Zuo Z and Zhang L (2020) Identification of CircRNA–miRNA–mRNA Regulatory Network in Gastrointestinal Stromal Tumor. Front. Genet. 11:403. doi: 10.3389/fgene.2020.00403

Abbreviations: CircRNA, circular RNA; DAVID, Database for Annotation, Visualization and Integrated Discovery; GISTs, gastrointestinal stromal tumors; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; MiRNAs, microRNAs; MREs, microRNA response element; ROC curve, receiver operating characteristic curve.

INTRODUCTION

Gastrointestinal stromal tumors are rarely one of the gastrointestinal carcinomas that originate from mesenchymal tissue. GISTs are characterized by expression of CD117 receptor in cells and have variable biological phenotypes ranging from benign to highly malignant (Gautam, 2020). As one of the most common non-epithelial neoplasms, they are mainly located in the stomach (55.6%) and small intestine (31.8%) (Tao et al., 2020). Radical surgery is the preferred treatment, and molecular target therapy, such as imatinib, can improve the survival of advanced patients with c-kit and/or PDGFRα mutations (Gupta and Rateria, 2020). However, a few effective tumor biomarkers are used for GIST diagnosis and prediction (Etherington and DeMatteo, 2019).

Circular RNA is a novel class of endogenous non-coding RNA characterized with 3′- and 5′-ends covalently linked in a closed-loop structure (Mahmoudi et al., 2020), which makes circRNAs resistant to exonucleases and more stable than traditional linear RNA, such as lncRNA and miRNA (Ding et al., 2020). Accordingly, the circRNAs can be divided into four types according to the source (Meng et al., 2017): exonic circRNAs (ecircRNA), intronic circRNA (ciRNA), exonic–intronic circRNA (EIciRNA), and intergenic circRNAs. Among them, 80% of circRNAs are ecircRNA. circRNAs may act as microRNA (miRNA/miR) sponges by competitively binding to miRNA response elements to influence downstream target gene expression, as well as affecting gene function at a post-translational level, and the same circRNA can regulate multiple miRNAs. Also, the same miRNAs can regulate multiple mRNA genes, thereby forming a large circRNA–miRNA–mRNA competitive network to affect the development of tumors (Xu S. et al., 2018). Also, some circRNAs can exert their activities via interaction with some proteins. Even then, some ecircRNAs may participate in the assembly and protein ribosomes translation. It was reported that circRNA is involved in various biological processes, including signal transduction and transcription, cell cycle regulation, RNA-binding protein, responses to stress, protein metabolism, cellular immunity, and cell structure (Jiang et al., 2018). Recent studies (Lu et al., 2018; Chaichian et al., 2020) have also demonstrated that circRNA abnormal expression and regulation are involved in the occurrence and development of a variety of tumors. Therefore, circRNAs are of great importance as a biomarker for cancer diagnosis, cancer prediction, and treatment feedback, and may even serve as targets for cancer treatment.

In this study, we first analyzed the circRNA differential expression profile in GISTs using circRNA chip and identified six potential key circRNAs by qRT-PCR. Also, the circRNA–miRNA–mRNA network was constructed, and the GO and KEGG pathway were performed via bioinformatics analysis. Our study provides a novel insight into the molecular mechanisms of GISTs from the circRNA–miRNA–mRNA network view, and these circRNAs gave new direction for diagnosis and treatment of GISTs.

TABLE 1 | The clinicopathological features of GISTs patients.

Variables	Cases (*n*)
Total	20
Age (years)	
≥60	12
<60	8
Gender	
Male	15
Female	5
Tumor size (cm)	
≤5	9
>5	11
Mitotic figure (HPF)	
≤5/50	13
>5/50	7
Malignant degrees	
Low/Moderate risk	12
High risk	8

MATERIALS AND METHODS

Patients and Samples

Twenty pairs of GISTs and adjacent tissues were collected from The Second Xiangya Hospital of Central South University. All pathological specimens were experienced pathologists confirmed, did not accept the pre-operative radiotherapy, chemotherapy, and imatinib targeted therapy. The clinicopathological features are shown in **Table 1**. All tissues were collected during surgical operation and instantly stored in liquid nitrogen. The present project was permitted by the ethics committee of The Second Xiangya Hospital of Central South University, and informed consents were obtained from all the participants.

CircRNA Chip Detection

Total RNAs were extracted by RNeasy Mini Kit (Qiagen, Hilden, Germany). Total RNA from each sample was quantified using the NanoDrop ND-1000. The sample preparation and microarray hybridization were performed based on the Arraystar's standard protocols. Total RNAs were digested with Rnase R to remove linear RNAs and enrich CircRNAs. Then, the enriched CircRNAs were amplified and transcribed into fluorescent cRNA (Arraystar Super RNA Labeling Kit; Arraystar). The labeled cRNAs were hybridized onto the Arraystar Human circRNA Array V2 (8 × 15 K; Arraystar). Agilent Feature Extraction software (version 11.0.1.1) was used to analyze acquired array images. Quantile normalization and subsequent data processing were performed using the R software limma package. Differentially expressed circRNAs with statistical significance between two groups were identified through volcano plot filtering. Differentially expressed circRNAs between two samples were identified through fold change filtering. Hierarchical clustering was performed to show the distinguishable circRNA expression pattern among samples.

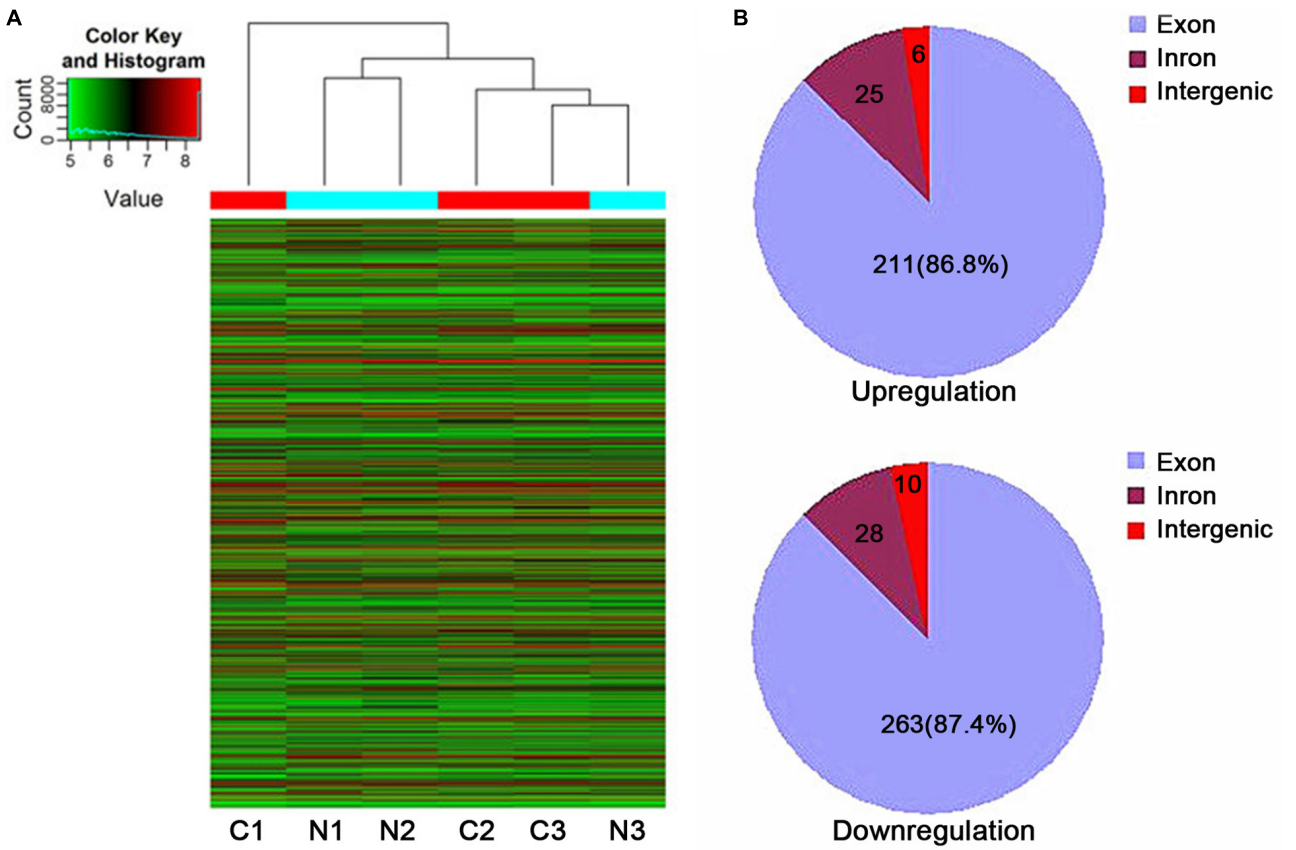


FIGURE 1 | Characterization of circRNA expression in GISTs. **(A)** Hierarchical cluster map. **(B)** The origin of upregulated and downregulated circRNAs. N, adjacent normal tissues; C, GIST tissues.

TABLE 2 | Top 20 significantly up-regulated circRNAs.

circRNA	*P*-value	FDR	FC (abs)	Regulation	source	chrom	strand	type	GeneSymbol
hsa_circRNA_061346	0.049017552	0.541205763	6.9877315	up	circBase	chr21	–	exonic	APP
hsa_circRNA_103114	0.027813714	0.541205763	6.8848564	up	circBase	chr21	–	exonic	APP
hsa_circRNA_103223	0.040961154	0.541205763	6.5339969	up	circBase	chr22	–	exonic	DDX17
hsa_circRNA_100582	0.02164493	0.541205763	4.5762691	up	circBase	chr10	+	exonic	ZEB1
hsa_circRNA_001481	0.047719242	0.541205763	4.5058956	up	circBase	chr5	–	sense overlapping	EMB
hsa_circRNA_103128	0.03858124	0.541205763	4.1057473	up	circBase	chr21	+	exonic	DYRK1A
hsa_circRNA_103870	0.039001555	0.541205763	4.0573338	up	circBase	chr5	+	exonic	SMA4
hsa_circRNA_404643	0.041732555	0.541205763	4.0082859	up	25070500	chr1	–	exonic	PIK3C2B
hsa_circRNA_004182	0.039090826	0.541205763	3.428553	up	circBase	chr2	+	intronic	CRIM1
hsa_circRNA_103977	0.03760255	0.541205763	3.426974	up	circBase	chr5	+	exonic	ARHGAP26
hsa_circRNA_104076	0.047784401	0.541205763	3.2920026	up	circBase	chr6	–	exonic	KIF13A
hsa_circRNA_101504	0.043980697	0.541205763	3.215775	up	circBase	chr15	+	exonic	PDIA3
hsa_circRNA_102510	0.029244589	0.541205763	3.1147657	up	circBase	chr19	+	exonic	LSM14A
hsa_circRNA_002164	0.048545781	0.541205763	3.082576	up	circBase	chr18	–	exonic	SS18
hsa_circRNA_103224	0.006590126	0.541205763	3.0780956	up	circBase	chr22	–	exonic	DDX17
hsa_circRNA_100915	0.030601002	0.541205763	3.0777302	up	circBase	chr11	–	exonic	PICALM
hsa_circRNA_403471	0.003668259	0.541205763	3.060178	up	25242744	chr5	+	exonic	ARHGAP26
hsa_circRNA_102248	0.047545686	0.541205763	3.056733	up	circBase	chr17	+	exonic	TBCD
hsa_circRNA_404446	0.005776341	0.541205763	3.0361211	up	25070500	chr1	–	exonic	CAPZB
hsa_circRNA_102378	0.045317666	0.541205763	2.9997955	up	circBase	chr18	+	exonic	ZNF532

TABLE 3 | Top 20 significantly down-regulated circRNAs.

circRNA	*P*-value	FDR	FC (abs)	Regulation	source	chrom	strand	type	GeneSymbol
hsa_circRNA_405324	0.013412911	0.541205763	4.3231649	down	25070500	chr15	+	sense overlapping	STARD9
hsa_circRNA_405443	0.028911656	0.541205763	3.3381633	down	25070500	chr16	+	intronic	NDE1
hsa_circRNA_406309	0.01797211	0.541205763	3.1653952	down	25070500	chr3	+	intronic	CMSS1
hsa_circRNA_100075	0.029545308	0.541205763	3.1382274	down	circBase	chr1	–	exonic	EMC1
hsa_circRNA_406821	0.020949363	0.541205763	3.0380172	down	25070500	chr6	+	exonic	ARMC2
hsa_circRNA_000361	0.027240076	0.541205763	3.0118108	down	circBase	chr3	–	antisense	PLCL2
hsa_circRNA_082335	0.035559522	0.541205763	3.008493	down	circBase	chr7	+	exonic	KLHDC10
hsa_circRNA_102207	0.013654637	0.541205763	3.0013356	down	circBase	chr17	+	exonic	AFMID
hsa_circRNA_405825	0.015220854	0.541205763	2.522762	down	25070500	chr2	+	exonic	KLF11
hsa_circRNA_074660	0.017052344	0.541205763	2.4763773	down	circBase	chr5	–	exonic	ATOX1
hsa_circRNA_104924	0.01941688	0.541205763	2.4549169	down	circBase	chr9	+	exonic	MVB12B
hsa_circRNA_056037	0.027008503	0.541205763	2.4478589	down	circBase	chr2	–	exonic	BUB1
hsa_circRNA_406780	0.028354789	0.541205763	2.3707217	down	25070500	chr6	–	sense overlapping	DNPH1
hsa_circRNA_024371	0.025742814	0.541205763	2.36132	down	circBase	chr11	+	exonic	PAFAH1B2
hsa_circRNA_061284	0.026417014	0.541205763	2.30245	down	circBase	chr21	+	exonic	USP25
hsa_circRNA_100456	0.009103973	0.541205763	2.2394652	down	circBase	chr1	+	exonic	KCNK2
hsa_circRNA_083919	0.037676327	0.541205763	2.213904	down	circBase	chr8	+	exonic	UNC5D
hsa_circRNA_406295	0.017374416	0.541205763	2.2027287	down	25070500	chr3	+	sense overlapping	SUCLG2-AS1
hsa_circRNA_035426	0.040284453	0.541205763	2.2016635	down	circBase	chr15	+	exonic	TCF12
hsa_circRNA_405296	0.023543643	0.541205763	2.1972163	down	25070500	chr15	+	sense overlapping	TUBGCP5

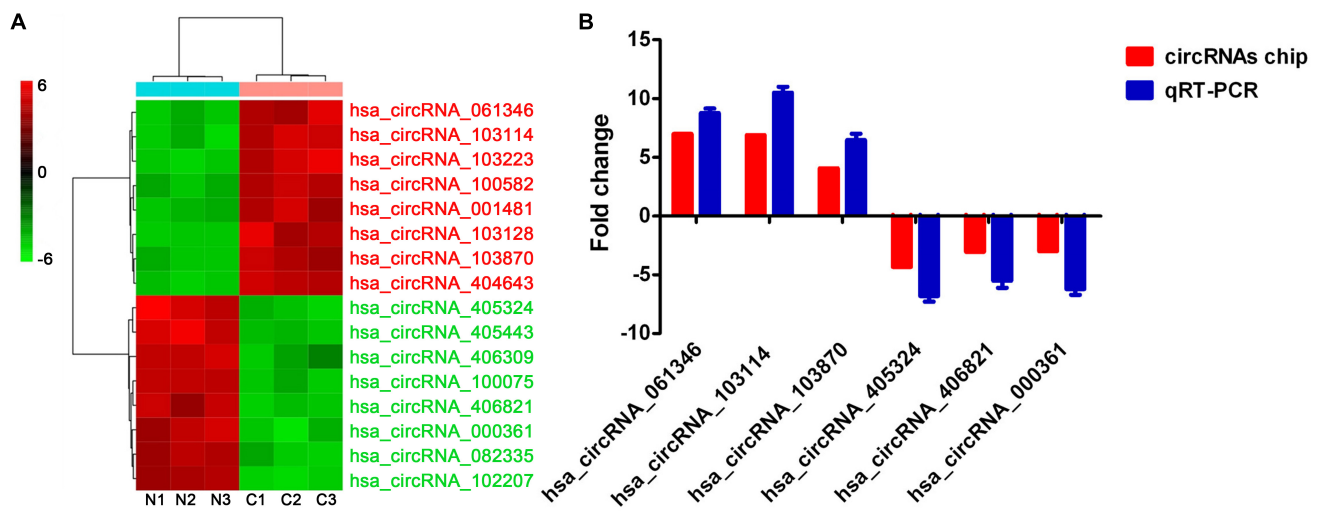


FIGURE 2 | qRT-PCR validation. **(A)** The cluster heat map of the top eight most up-regulated and down-regulated circRNAs. **(B)** Comparison of Arraystar human circRNAs chip data and qRT-PCR results. circRNAs expression levels were normalized to GAPDH. N, adjacent normal tissues; C, GIST tissues.

Quantitative Real-Time PCR

Total RNAs were extracted using RNeasy Mini kit (Qiagen, Hilden, Germany). Then, RNA was reversed into complementary DNA (cDNA) by SuperScript III Reverse Transcriptase (Invitrogen). qRT-PCR was performed with 95.0°C for 3 min, and 39 circles of 95.0°C for 10 s and 60°C for 30 s using SYBR Green PCR Master Mix system. The relative expression levels were calculated using the $22^{-\Delta\Delta Ct}$ method. RNA levels were normalized to GAPDH expression. The forward (F) and reverse (R) primer sequences for qRT-PCR were designed and synthesized by Shanghai Kangcheng Co., Ltd. (Chinese).

CircRNA–miRNA–mRNA Interaction Prediction

The fundamental structure of circRNAs was predicted using Cancer-Specific circRNA (CSCD[1]). circRNA–miRNA interactions were predicted using TargetScan and miRanda databases, and miRNA target gene was predicted using TargetScan, miRanda v5, and miBase prediction databases. Candidate miRNAs and mRNAs should be overlapped in at least two databases. Arraystar's miRNA target prediction software

[1]http://gb.whu.edu.cn/CSCD

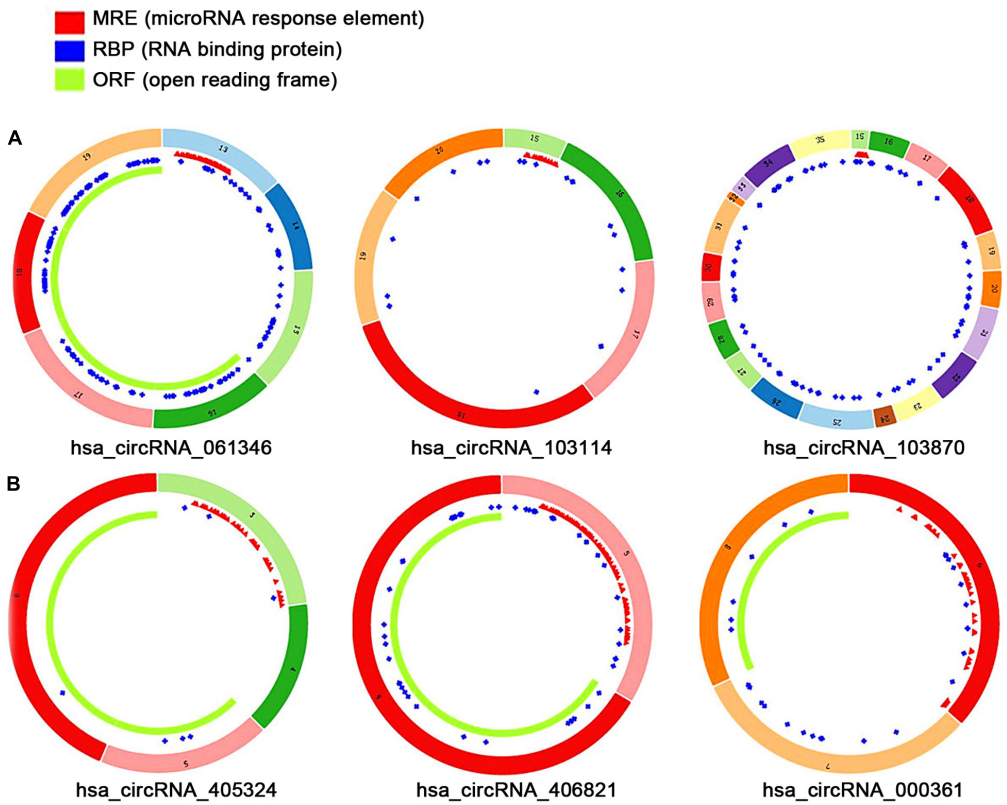


FIGURE 3 | The fundamental structure modes of the six candidate circRNAs predicted by CSCD. **(A)** Up-regulated circRNAs: hsa_circRNA_061346, hsa_circRNA_103114, hsa_circRNA_103870. **(B)** Down-regulated circRNAs: hsa_circRNA_405324, hsa_ circRNA _406821, hsa_circRNA_000361.

site: miRanda v5[2], TargetScan[3], and mibase[4]. The circRNA–miRNA–mRNA competitive network (cirCeNET) was visualized by Cytoscape software (version 3.6.1[5]).

Gene Ontology (GO) and KEGG Pathway Analysis

Gene ontology and KEGG pathways analysis was used to determine the function of candidate mRNAs in circRNA–miRNA–mRNA competitive network. DAVID[6] was used to predict the enriched functional categories and enriched signaling pathways. GO term, including BP, CC, MF, and KEGG pathway with $P < 0.05$ and FDR < 0.05 were considered as statistically significant.

Statistical Analysis

All data were analyzed by SPSS17.0 statistics software. Paired *t*-test was employed for the comparison of two groups. Chi-square test was used to investigate the relationship between circRNA expression and clinicopathologic features of GISTs patients. $P < 0.05$ was considered as statistically significant.

RESULTS

Differential CicrRNA Expression Profiles Were Established Successfully

The box plot showed similar distributions of tissues. In the volcano plots, differentially expressed circRNAs were categorized using fold change and *P* values. The scatter plots demonstrated the variation of differentially expressed circRNAs. Hierarchical cluster analysis showed differentially expressed circRNAs in GISTs with fold change >1.5 and $P < 0.05$ (**Figure 1A**). After normalization and data analysis, compared with adjacent tissues, a total of 543 differentially expressed circRNAs were identified, including 242 up-regulated circRNAs and 301 down-regulated circRNAs, of which, exonic circRNAs accounted for 86.8% in up-regulated circRNAs and 87.4% in down-regulated circRNAs (**Figure 1B**). The top 20 significantly up- and down-regulated circRNAs are listed in **Tables 2** and **3**.

qRT-PCR Validation

The top eight most upregulated circRNAs with fold change >4 and $P < 0.05$ and the top eight most downregulated circRNAs

[2]http://www.ebi.ac.uk/enright-srv/microcosiTi/htdo/targets/v5
[3]http://www.targetscan.org
[4]http://pictar.bio.nyu.edu
[5]http://cytoscape.org/
[6]http://www.david.abcc.ncifcrf.gov/

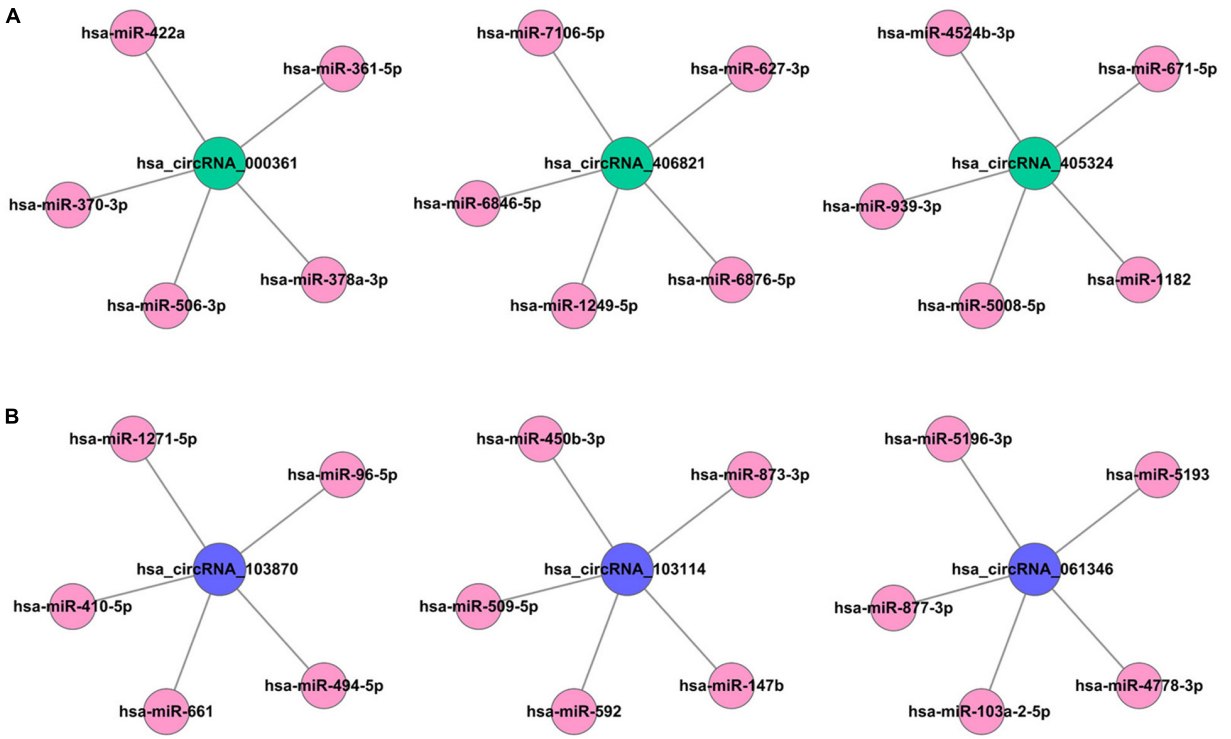


FIGURE 4 | circRNA–miRNA regulatory network. **(A)** The top five targeted miRNAs of up-regulated circRNAs. **(B)** The top five targeted miRNAs of down-regulated circRNAs.

with fold change >3 and $P < 0.05$ are shown in the cluster heat map (**Figure 2A**). qRT-PCR assay was used to assess the accuracy of circRNAs chip data. After filtering circRNAs with low raw intensity, six candidate circRNAs, including three up-regulated circRNAs (hsa_ circRNA_061346, hsa_circRNA_103114, hsa_circRNA_103870) and three down-regulated circRNAs (hsa_circRNA_405324, hsa_circRNA_406821, hsa_circRNA_ 000361) were selected for qRT-PCR analysis. The results showed that qRT-PCR results were consistent with the circRNAs chip data (**Figure 2B**), indicating the reliability of circRNAs chip data.

CircRNA–miRNA–mRNA Network Construction

The fundamental structure modes of the six candidate circRNAs predicted by CSCD are shown in **Figure 3**. To estimate the function of six candidate circRNAs, circRNA–miRNA interactions were constructed with TargetScan and miRanda databases. The top five targeted miRNAs of six candidate circRNAs are exhibited in **Figure 4**, and the detailed potential circRNA–miRNA interaction sites of targeted miRNAs with the highest context score percentile are shown in **Figure 5**. Then, the circRNA–miRNA–mRNA competitive network (cirCeNET) was visualized by Cytoscape software (version 3.6.1) based on circRNA–miRNA interactions and miRNA–mRNA interactions (**Figure 6**). This network contained six circRNAs, 30 miRNAs, and 308 mRNAs, which provided a comprehensive perspective into the links between circRNA, miRNA, and mRNAs in GISTs.

GO and KEGG Pathway Analysis

The GO analysis demonstrated that the term with the highest enrichment score was regulation of biological process (GO:0050789) for biological process terms (BP), intracellular organelle (GO:0043229) for cellular component terms (CC), and protein binding (GO:0005515) for molecular function terms (MF), respectively. The top 10 enrichment scores are shown in **Figures 7A–C**. The KEGG pathway with the highest enrichment score was the Wnt signaling pathway. The top 10 enriched KEGG pathways are shown in **Figure 7D**.

qRT-PCR Validation

Six significantly differential circRNAs were also verified in 20 pairs of GISTs and adjacent tissues by qRT-PCR. The results showed that circRNA_061346, circRNA_103114, and circRNA_103870 were significantly up-regulated in GIST tissues (**Figure 8**) ($P < 0.05$), and circRNA_405324, circRNA_ 406821, and circRNA_000361 were dramatically down-regulated in GIST tissues (**Figure 9**) $P < 0.05$), compared with corresponding adjacent tissues.

Diagnosis Values of CircRNA

In order to determine the diagnostic value of six candidate circRNAs in GISTs, the ROC curve was employed. Statistical analysis demonstrated that all six candidate circRNAs had high diagnostic efficiency with AUC = 0.9925, AUC = 0.9824, AUC = 0.9231, AUC = 0.9300, AUC = 0.9463, AUC = 0.9138

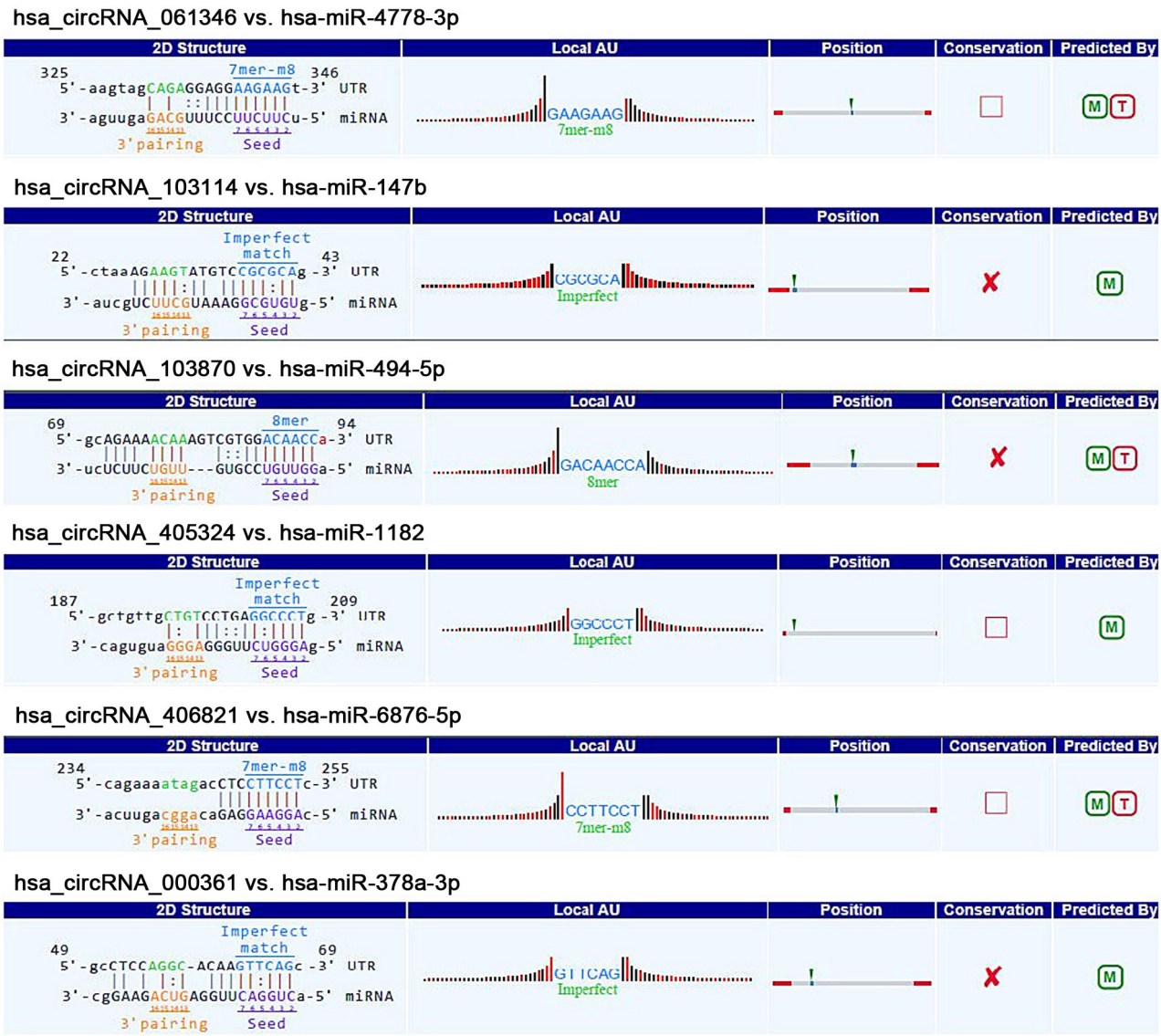


FIGURE 5 | The detailed potential circRNA–miRNA interaction sites of targeted miRNAs with highest context score percentile based on TargetScan and miRanda data.

for circRNA_061346, circRNA_103114, circRNA_103870 and circRNA_405324, circRNA_406821, circRNA_000361, respectively (**Figure 10**) ($P < 0.05$).

Correlation of CircRNA Expressions With Clinical Pathologic Features

In order to investigate the correlation of six candidate circRNA expressions with clinical–pathologic features, the median circRNA expression was used to divide the 20 pairs of GISTs tissues into higher and lower circRNA expression groups. Chi-square assay was employed for statistical analysis. The results suggested that circRNA_061346 and circRNA_103114 expressions were positively associated with tumor size, mitotic figure, malignant degrees, and circRNA_103870 expression was positively associated with tumor size, mitotic figure, but without relation to malignant degrees (**Figure 11**). On the contrary, circRNA_405324 expression was negatively associated with tumor size, mitotic figure, malignant degrees, and circRNA_406821 was negatively correlated with mitotic figure, malignant degrees, but not with tumor size; nevertheless, circRNA_000361 expression was only negatively related with mitotic figure (**Figure 11B**). However, there was no correlation with age, gender, and tumor location.

DISCUSSION

Gastrointestinal stromal tumors are a rare malignant tumor that occurs principally in the stomach, small intestine, and

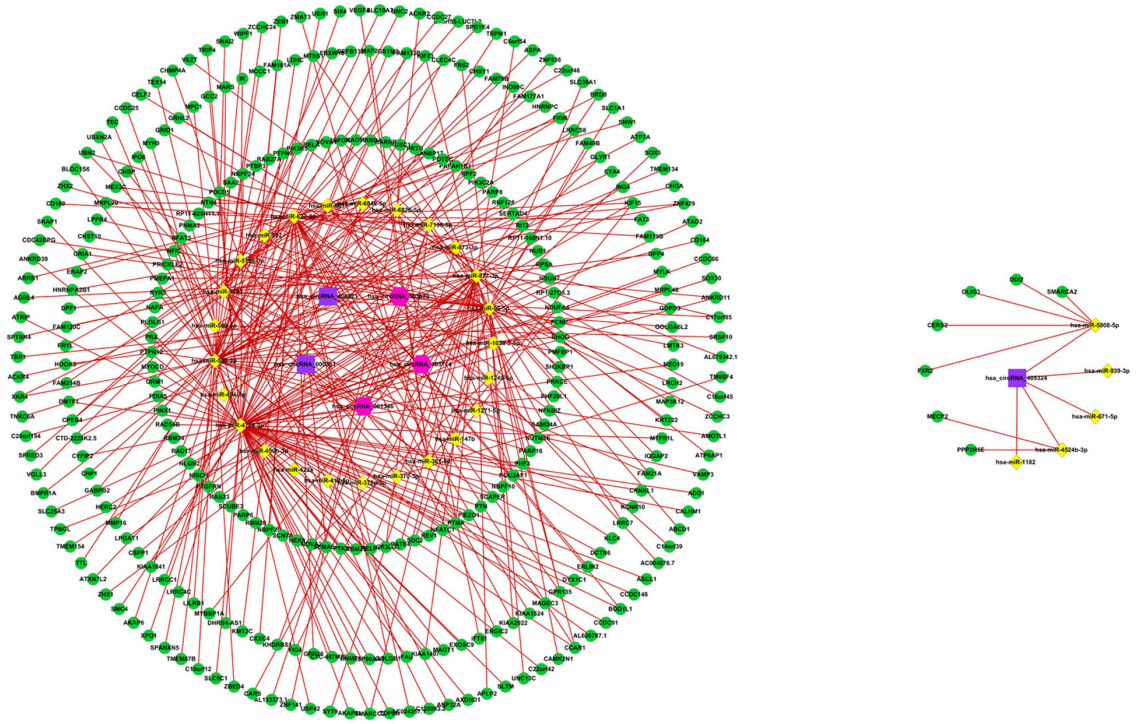


FIGURE 6 | circRNA–miRNA–mRNA regulatory network. Up-regulated circRNAs, down-regulated circRNAs, miRNA, and mRNA are presented as square, hexagon, diamond, and circle, respectively.

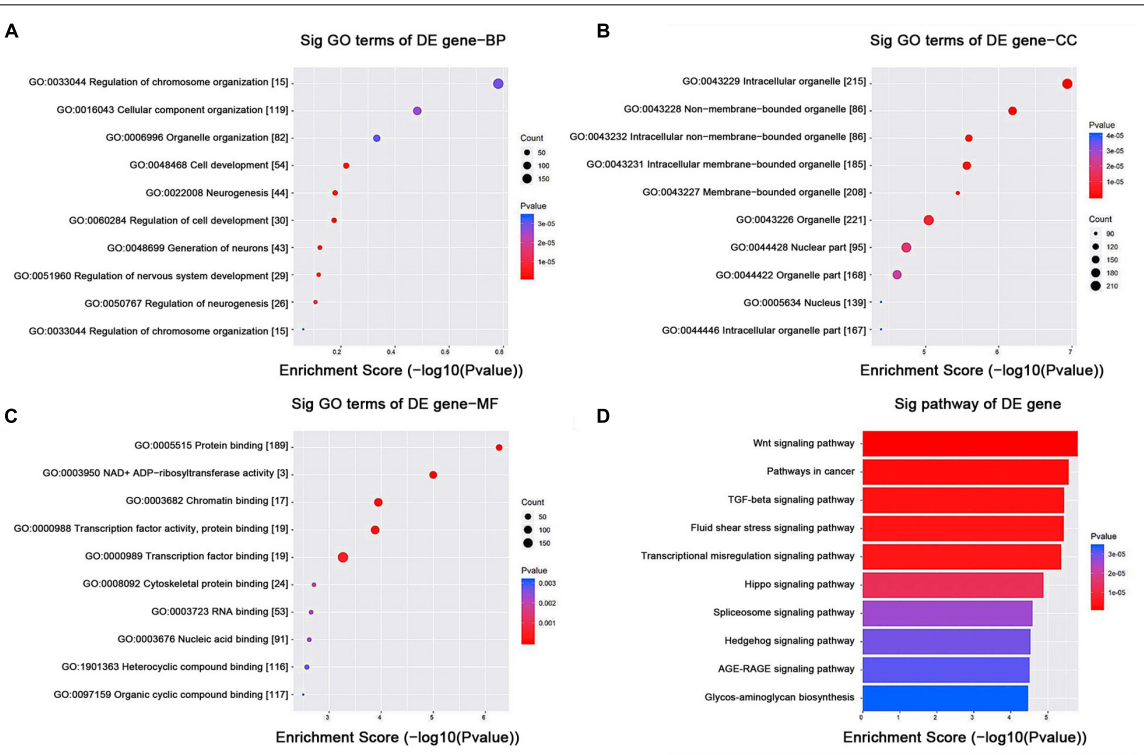


FIGURE 7 | GO and KEGG pathway analysis. **(A–C)** GO annotation of targeted mRNAs with the top 10 enrichment scores for biological process, cellular component, and molecular function, respectively. **(D)** The top 10 enriched KEGG pathways. GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

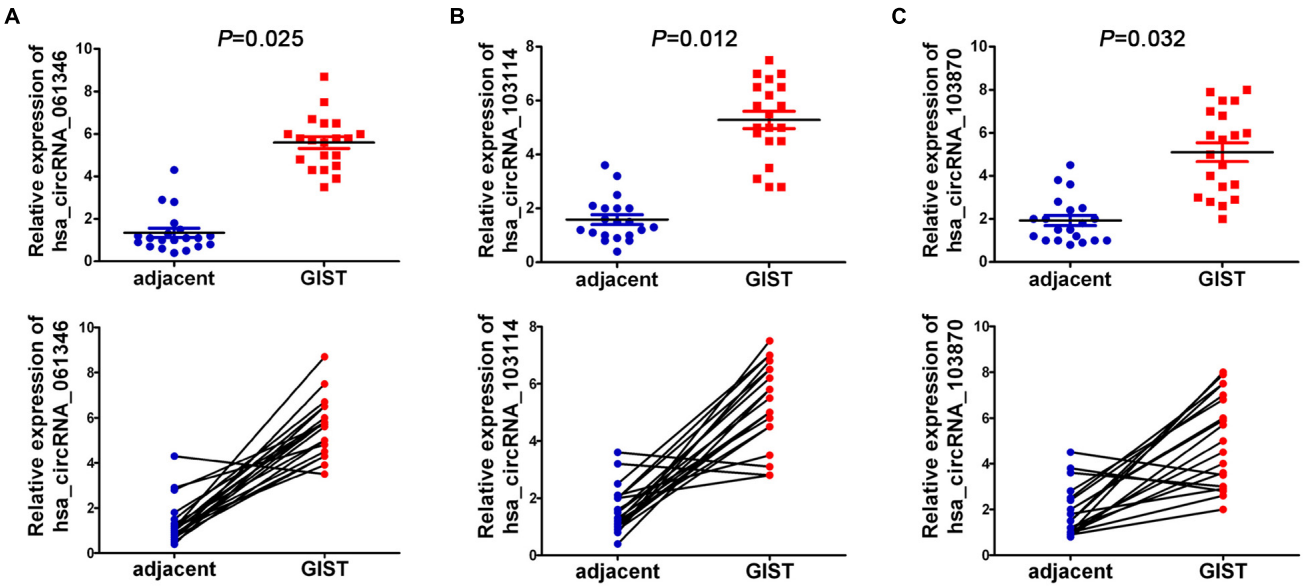


FIGURE 8 | The relative expression of up-regulated circRNAs of 20 GIST tissues were detected by qRT-PCR, **(A)** hsa_circRNA_061346; **(B)** hsa_circRNA_103114; **(C)** hsa_ circRNA_103870. circRNA expression levels were normalized to GAPDH.

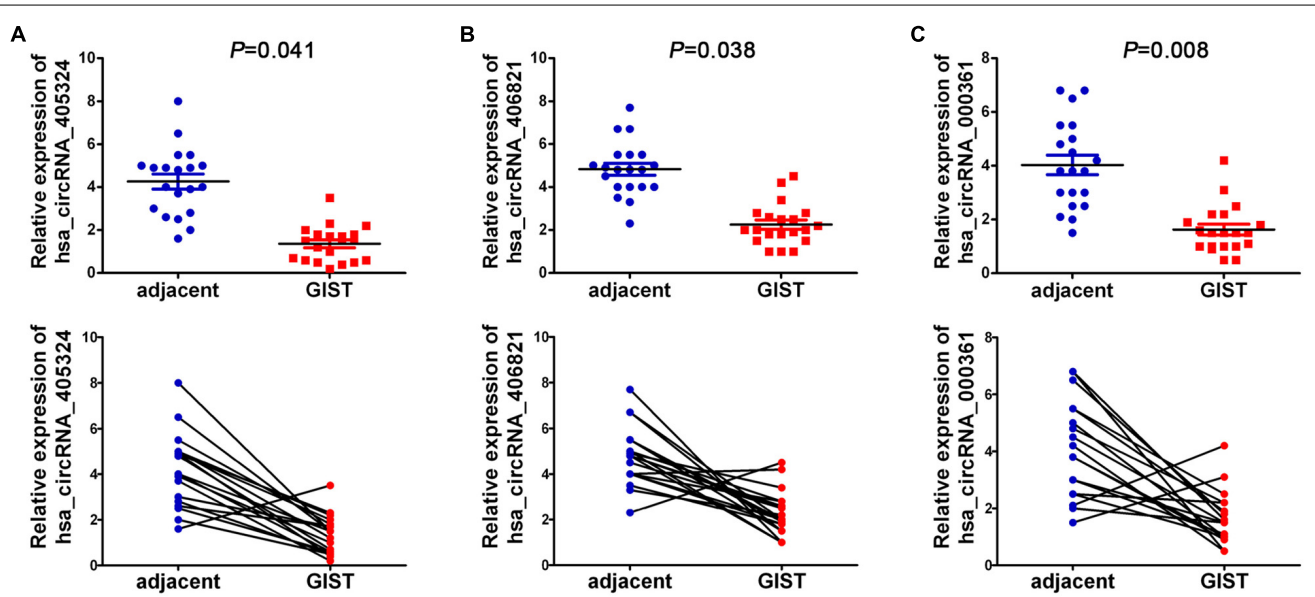


FIGURE 9 | The relative expression of down-regulated circRNAs of 20 GIST tissues was detected by qRT-PCR, **(A)** hsa_circRNA_405324; **(B)** hsa_circRNA _406821; **(C)** hsa_circRNA_000361. circRNA expression levels were normalized to GAPDH.

colon–rectum (Flavahan et al., 2019). K-ras gene mutation might be correlated with the mechanism of development and infiltration of GISTs, but the pathogenesis of GISTs is inadequately understood (Lasota et al., 2019). A growing body of research (Wang et al., 2018; Cai et al., 2019; Zaghlool et al., 2020) demonstrates that the development of cancer is often accompanied by abnormal expression of circRNA. Also, circRNA is characterized by inherent stability, highly conservative, and universality. Therefore, circRNA is of great importance as a biomarker for cancer screening, cancer diagnosis, cancer prediction, feedback of treatment, and prognosis. Additionally, circRNA's abnormal expression and the circRNA–miRNA–mRNA regulatory network regulation have been increasingly demonstrated in a variety of tumors, such as circRNA_CAMK2A–miR-615-5p–fibronectin 1 network in lung adenocarcinoma

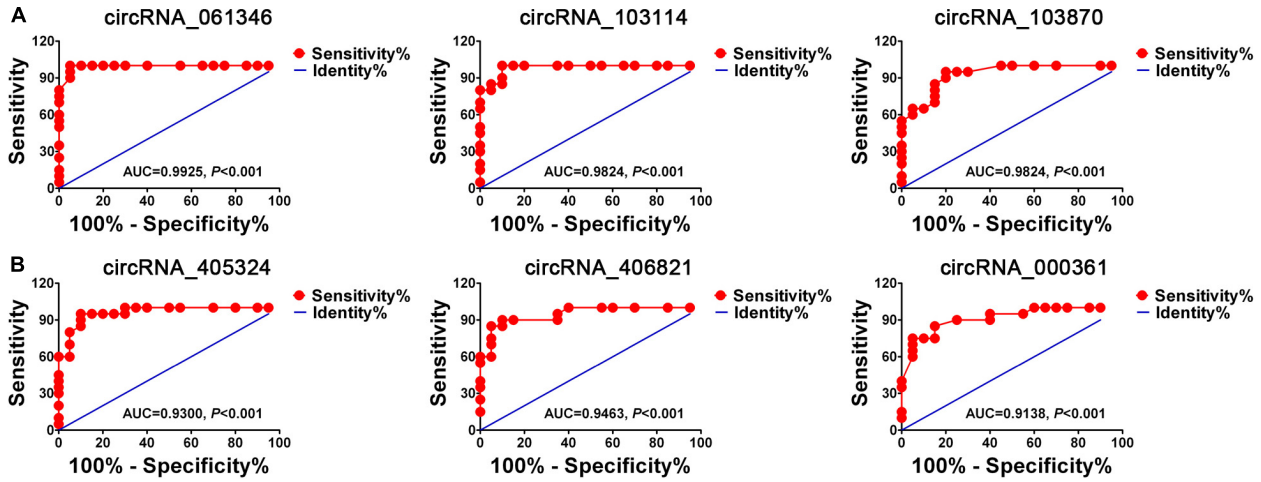


FIGURE 10 | The diagnostic value of six candidate circRNAs in GISTs. **(A)** ROC curve for three up-regulated circRNAs. **(B)** ROC curve for three down- regulated circRNAs.

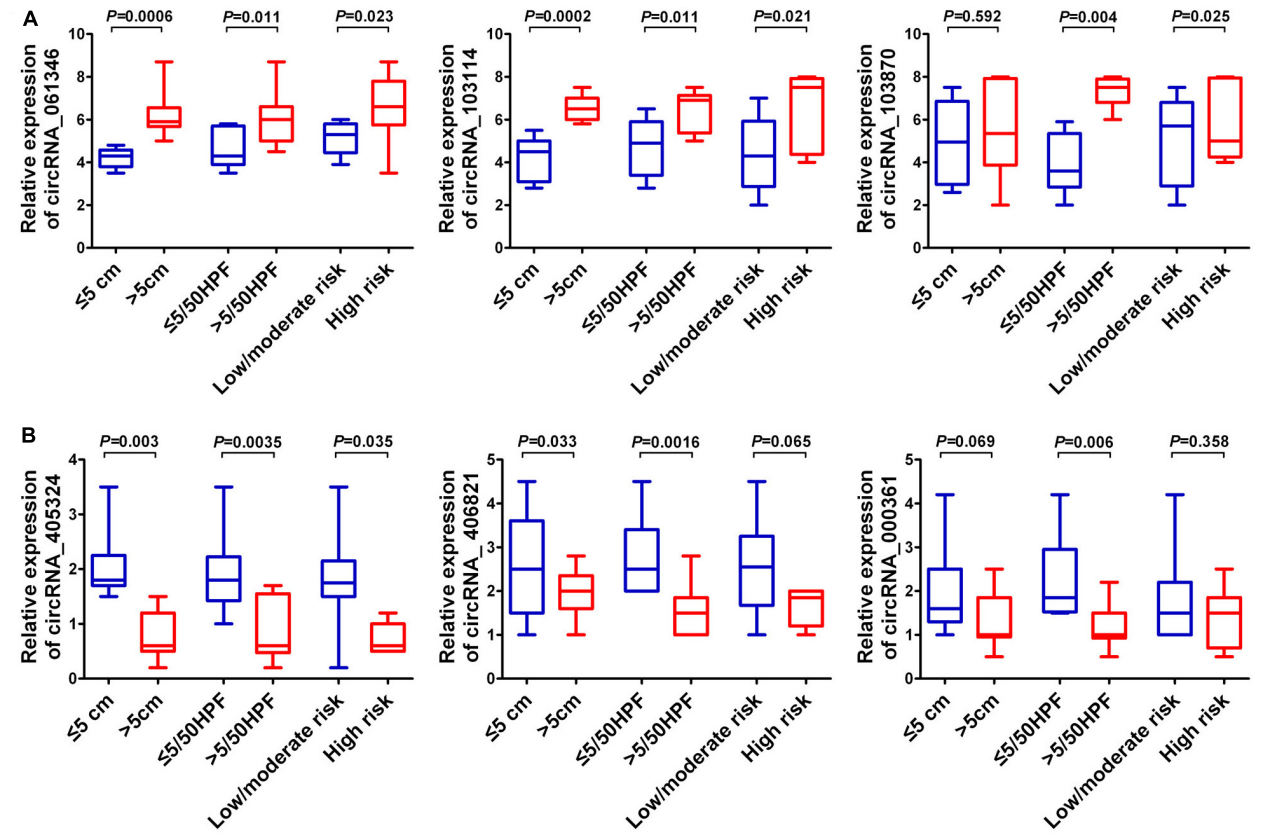


FIGURE 11 | Correlation of six candidate circRNAs with clinical pathologic features. **(A)** Box plots for three up-regulated circRNAs. **(B)** Box plots for three down-regulated circRNAs. Tumor size: ≤5 cm, >5 cm; mitotic figure: ≤5/50 HPF, <5/50 HPF; malignant degrees: low risk, moderate risk, high risk.

metastasis (Du et al., 2019), circRNA_0006948–miR-490-3p–HMGA2 network in esophageal squamous cell carcinoma (Pan et al., 2019), circRNA_ACAP2–miR-29a/b-3p–COL5A1 network in breast cancer (Zhao et al., 2019), circRNA_51217–miRNA-646–TGFβ1/p-Smad2/3 network in prostate cancer (Xu et al., 2019), etc. At present, there are few reports about the circRNA–miRNA–mRNA regulatory network in GISTs.

In the present study, we illuminate the molecular mechanisms of circRNAs in the occurrence and development of GISTs for the first time. We first performed circRNA chip analysis to assess differential cicrRNA expression profiles in GIST tissues and corresponding non-cancer tissues. A totally of 543 differentially expressed circRNAs were identified, of which 242 were significantly upregulated and 301 were significantly downregulated in GISTs tissues. Additionally, in order to fully elucidated the function of the circRNA-related ceRNA in GISTs, six candidate circRNAs including three up-regulated circRNAs (hsa_circRNA_061346, hsa_circRNA_103114, hsa_ circRNA_103870) and three down-regulated circRNAs (hsa_circRNA_405324, hsa_ circRNA_406821, hsa_circRNA_000361) were identified to be involved in the ceRNA network. The ceRNA network consists of six circRNAs, 30 miRNAs, and 308 mRNAs. In this network, previous studies (Xu Z.H. et al., 2018; Zhang et al., 2019) show that many miRNAs, such as miR-4778-3p, miR-147b, miR-1182, and miR-378a-3p, were involved in tumor cell growth, invasion, and metastasis. Also, many targeted genes, such as ZEB1, SOX5, AKAP1, CHP1, CNBP, VEGFR, and MAGT1, play a vital function in the cell shape, movement, invasion, adhesion, and polarity formation, so as to involve in many kinds of diseases such as malignant tumors, wound healing, and so on (Rinaldi et al., 2017; Caramel et al., 2018; Hu et al., 2018).

Furthermore, 20 GIST tissues and adjacent tissues were collected to verify the expression of identified six candidate circRNAs. qRT-PCR results showed that hsa_circRNA_061346, hsa_circRNA_103114, and hsa_circRNA_103870 were significantly up-regulated in GISTs, and hsa_ circRNA_405324, hsa_circRNA _406821, hsa_circRNA_ 000361 were dramatically down-regulated in GISTs. In addition, all of these circRNAs were shown to have high diagnostic values, and most of them were significantly associated with tumor size, mitotic figure, and malignant degrees in GISTs. The six candidate circRNAs might be critical circRNAs participating in the occurrence and development of GISTs and can serve as novel potential diagnostic biomarkers for GISTs patients. Through a large literature review, very limited data are available about these circRNAs's functions and their deregulation in cancer.

However, there are several limitations to our study. First, only 20 patients were enrolled in our study, the sample size is relatively small, and the result showed an association, rather than a definite, causal relationship. Also, the relation analysis of clinical factors and circRNAs needs to be supported by large samples. Second, in our study, we only conducted a network based on identified six critical circRNAs, miRNA, and target mRNA, but a total of 543 circRNAs were identified in GISTs. Other circRNAs may contribute as well. Third, our paper starts with a general analysis of circRNAs in GISTs, but the mechanism is not discussed in detail. Therefore, in our future work, further studies with larger groups of patients, a network based on 543 circRNAs are needed to confirm these findings, and the concrete mechanism of circRNAs in GISTs also needs to be further explored.

CONCLUSION

In the present study, the differential circRNA expression profile of GISTs was established, and a total of 543 differentially expressed circRNAs were screened. In addition, the circRNA–miRNA–mRNA regulatory network was constructed. hsa_circRNA_ 061346, hsa_circRNA_103114, hsa_circRNA_103870 and hsa_circRNA_405324, hsa_ circRNA_406821, hsa_circRNA_000361 were identified as critical circRNAs in the occurrence and development of GISTs and may present as potential diagnostic biomarkers for GISTs. In brief, our study provides a new insight into the pathogenesis of GISTs from the circRNA–miRNA–mRNA regulatory network view.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in NCBI GEO accession GSE147303.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of The Second Xiangya Hospital of Central South University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LZ coordinated all aspects of the research. FZ and DC were responsible for the clinical sample collection. FZ participated in most molecular and cellular experiments and manuscript preparation. LS and ZZ were responsible for data analysis. All authors read and approved the final manuscript.

FUNDING

The project received support from the Hunan Provincial Science and Technology Department (2017SK1032, LZ), fundamental research funds for the Central Universities of Central South University (2019zzts356, FZ), Changsha Science and Technology Plan Project (kq1907080, YT), and Changsha Science and Technology Plan Project (kq1907078, FZ).

ACKNOWLEDGMENTS

We highly appreciate the kind help and support from the Shanghai Kangcheng Co., Ltd. (Chinese). Personally, for my parents and wife and daughter, I really appreciate all your love and support up to my work. Thank you for all you did to us, really thanks, I love you all. Patient clinicopathological data used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

Cai, H., Li, Y., Niringiyumukiza, J. D., Su, P., and Xiang, W. (2019). Circular RNA involvement in aging: an emerging player with great potential. *Mech. Age. Dev.* 178, 16–24.

Caramel, J., Ligier, M., and Puisieux, A. (2018). Pleiotropic roles for ZEB1 in cancer. *Cancer Res.* 78, 30–35.

Chaichian, S., Shafabakhsh, R., Mirhashemi, S. M., Moazzami, B., and Asemi, Z. (2020). Circular RNAs: a novel biomarker for cervical cancer. *J. Cell. Physiol.* 235, 718–724.

Ding, H., Xu, Q., Wang, B., Lv, Z., and Yuan, Y. (2020). MetaDE-based analysis of circRNA expression profiles involved in gastric cancer. *Digest. Dis. Sci.* 12, 1–12.

Du, J., Zhang, G., Qiu, H., Yu, H., and Yuan, W. (2019). The novel circular RNA circ-CAMK2A enhances lung adenocarcinoma metastasis by regulating the miR-615-5p/ fibronectin 1 pathway. *Cell Mol. Biol. Lett.* 24, 72–78.

Etherington, M. S., and DeMatteo, R. P. (2019). Tailored management of primary gastrointestinal stromal tumors. *Cancer* 10, 9–15.

Flavahan, W. A., Drier, Y., Johnstone, S. E., Hemming, M. L., Tarjan, D. R., Hegazi, E., et al. (2019). Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs. *Nature* 575, 229–233.

Gautam, A. (2020). Gastrointestinal stromal tumor of the stomach infiltrating the hilum of spleen: case report. *Clin. Surg. J.* 3, 17–21.

Gupta, S. K., and Rateria, N. (2020). Gastrointestinal stromal tumors (GIST): an overview. *Indian J. Surg.* 24, 1–7.

Hu, J., Tian, J., Zhu, S., Sun, S., Yu, J., Tian, H., et al. (2018). Sox5 contributes to prostate cancer metastasis and is a master regulator of TGF-β-induced epithelial mesenchymal transition through controlling Twist1 expression. *Br. J. Cancer* 118, 88–97.

Jiang, L., Sun, D., Hou, J., Hou, J. C., and Ji, Z. L. (2018). CircRNA: a novel type of biomarker for cancer. *Breast Cancer* 25, 1–7.

Lasota, J., Kowalik, A., Felisiak-Golabek, A., Ziêba, S., and Wang, Z.-F. (2019). New mechanisms of mTOR pathway activation in KIT-mutant malignant GISTs. *Appl. Immunohist.* 27, 54–58.

Lu, C. X., Sun, X. M., Li, N., Wang, W. G., Kuang, D. X., Tong, P., et al. (2018). CircRNAs in the tree shrew (Tupaia belangeri) brain during postnatal development and aging. *Aging* 1, 833–840.

Mahmoudi, E., Kiltschewskij, D., Fitzsimmons, C., and Cairns, M. J. (2020). Depolarization-associated CircRNA regulate neural gene expression and in some cases may function as templates for translation. *Cells* 9, 25–30.

Meng, S., Zhou, H., Feng, Z., Xu, Z., Tang, Y., Li, P., et al. (2017). CircRNA: functions and properties of a novel potential biomarker for cancer. *Mol. Cancer* 16, 94–99.

Pan, Z., Lin, J., Wu, D., He, X., Wang, W., Hu, X., et al. (2019). Hsa_circ_0006948 enhances cancer progression and epithelial-mesenchymal transition through the miR-490-3p/HMGA2 axis in esophageal squamous cell carcinoma. *Aging* 11, 11937–11954.

Rinaldi, L., Sepe, M., Delle, D. R., Conte, K., Arcella, A., Borzacchiello, D., et al. (2017). Mitochondrial AKAP1 supports mTOR pathway and tumor growth. *Cell Death Dis.* 8, e2842.

Tao, K., Zeng, X., Liu, W., Wang, S., Gao, J., Shuai, X., et al. (2020). Primary gastrointestinal stromal tumor mimicking as gynecologic mass: characteristics, management, and prognosis. *J. Surg. Res.* 246, 584–590.

Wang, H., Xiao, Y., Wu, L., and Ma, D. (2018). Comprehensive circular RNA profiling reveals the regulatory role of the circRNA-000911/miR-449a pathway in breast carcinogenesis. *Intern. J. Oncol.* 52, 743–754.

Xu, H., Sun, Y., You, B., Huang, C. P., Ye, C., and Chang, C. (2019). Androgen receptor reverses the oncometabolite R-2-hydroxyglutarate-induced prostate cancer cell invasion via suppressing the circRNA-51217/miRNA-646/TGFβ1/p-Smad2/3 signaling. *Cancer Lett.* 472, 151–164.

Xu, S., Zhou, L. Y., Ponnusamy, M., Zhang, L., Dong, Y. H., Zhang, Y. H., et al. (2018). A comprehensive review of circRNA: from purification and identification to disease marker potentia. *PeerJ* 6:e5503.

Xu, Z. H., Yao, T. Z., and Liu, W. (2018). miR-378a-3p sensitizes ovarian cancer cells to cisplatin through targeting MAPK1/GRB2. *Biomed. Pharmacother.* 107, 1410–1417.

Zaghlool, S. B., Kühnel, B., Elhadad, M. A., Kader, S., Halama, A., Thareja, G., et al. (2020). Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits. *Nat. Commun.* 11, 1–12.

Zhang, Y., Li, P., Hu, J., Jhao, J. P., Ma, R., Li, W., et al. (2019). Role and mechanism of miR-4778-3p and its targets NR2C2 and Med19 in cervical cancer radioresistance. *Biochem. Biophys. Res. Commun.* 508, 210–216.

Zhao, B., Song, X., and Guan, H. (2019). CircACAP2 promotes breast cancer proliferation and metastasis by targeting miR-29a/b-3p-COL5A1 axis. *Life Sci.* 14:117179.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 04 June 2020
doi: 10.3389/fbioe.2020.00502

A Bioinformatics Tool for the Prediction of DNA N6-Methyladenine Modifications Based on Feature Fusion and Optimization Protocol

Jianhua Cai[1,2†], Donghua Wang[3†], Riqing Chen[4], Yuzhen Niu[1], Xiucai Ye[5], Ran Su[6‡], Guobao Xiao[1] and Leyi Wei[1,7*]*

[1] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, College of Computer and Control Engineering, Minjiang University, Fuzhou, China, [2] College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China, [3] Department of General Surgery, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China, [4] College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou, China, [5] Department of Computer Science, University of Tsukuba, Tsukuba, Japan, [6] College of Intelligence and Computing, Tianjin University, Tianjin, China, [7] School of Software, Shandong University, Jinan, China

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Dariusz Mrozek,
Silesian University of Technology, Poland
Renzhi Cao,
Pacific Lutheran University, United States

***Correspondence:**
Guobao Xiao
gbx@mju.edu.cn
Leyi Wei
weileyi@tju.edu.cn

[†] These authors have contributed equally to this work

[‡]Present address:
Ran Su,
School of Computer Software at Tianjin University, Tianjin, China

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 17 February 2020
Accepted: 29 April 2020
Published: 04 June 2020

Citation:
Cai J, Wang D, Chen R, Niu Y, Ye X, Su R, Xiao G and Wei L (2020) A Bioinformatics Tool for the Prediction of DNA N6-Methyladenine Modifications Based on Feature Fusion and Optimization Protocol. Front. Bioeng. Biotechnol. 8:502. doi: 10.3389/fbioe.2020.00502

DNA N^6-methyladenine (6mA) is closely involved with various biological processes. Identifying the distributions of 6mA modifications in genome-scale is of great significance to in-depth understand the functions. In recent years, various experimental and computational methods have been proposed for this purpose. Unfortunately, existing methods cannot provide accurate and fast 6mA prediction. In this study, we present 6mAPred-FO, a bioinformatics tool that enables researchers to make predictions based on sequences only. To sufficiently capture the characteristics of 6mA sites, we integrate the sequence-order information with nucleotide positional specificity information for feature encoding, and further improve the feature representation capacity by analysis of variance-based feature optimization protocol. The experimental results show that using this feature protocol, we can significantly improve the predictive performance. Via further feature analysis, we found that the sequence-order information and positional specificity information are complementary to each other, contributing to the performance improvement. On the other hand, the improvement is also due to the use of the feature optimization protocol, which is capable of effectively capturing the most informative features from the original feature space. Moreover, benchmarking comparison results demonstrate that our 6mAPred-FO outperforms several existing predictors. Finally, we establish a web-server that implements the proposed method for convenience of researchers' use, which is currently available at http://server.malab.cn/6mAPred-FO.

Keywords: DNA N6-methyladenine site, machine learning, feature representation, sequence-based predictor, feature fusion

KEYPOINTS

- In this study, we present 6mAPred-FO, a powerful bioinformatics tool for the prediction of 6mA sites.
- In 6mAPred-FO, we integrate the sequence-order information with nucleotide positional specificity information for feature encoding, and further improve the feature representation capacity by feature optimization.

- Comparative results showed that the proposed 6mAPred-FO significantly outperforms several existing predictors.
- We have established a webserver implementing the proposed 6mAPred-FO. It is publicly accessible at http://server.malab.cn/6mAPred-FO.

INTRODUCTION

N^6-methyladenine (6mA), as a dynamic DNA epigenetic modification, has been extensively discovered in the following three species: bacteria, archaea and eukaryotes (O'Brown and Greer, 2016). The newly studies have indicated that 6mA modification participates in a wide spectrum of important biological processes. In prokaryotes, for example, 6mA has been found to be closely correlated with a series of DNA activities, such as replication (Campbell and Kleckner, 1990; Li et al., 2019), repair (Pukkila et al., 1983), transcription (Robbins-Manke et al., 2005), and cellular defense (Luria and Human, 1952; Linn and Arber, 1968; Meselson and Yuan, 1968). In addition, some studies have demonstrated that 6mA can act as an epigenetic mark in Phytophthora genomes and there may be a relationship between patterns of 6mA methylation and adaptive evolution in these important plant pathogens (Chen H. et al., 2018). Besides, recent study demonstrated that DNA 6mA modification plays a significant role in cell fate transition of mammalian cells as well (Liang et al., 2016; Liao et al., 2016). Therefore, it is very indispensable to determine the distribution of 6mA modification sites in genome-scale to systematically interpret its biological functions.

To solve this problem, experimental efforts have been proposed, such as ultra-high performance liquid chromatography coupled with mass spectrometry (UHPLC-ms/ms) (Greer et al., 2015), capillary electrophoresis and laser-induced fluorescence (CE-LIF) (Krais et al., 2010), methylated DNA immunoprecipitation sequencing (MeDIP-seq) (Pomraning et al., 2009), and single-molecule real-time sequencing (SMRT-seq) (Flusberg et al., 2010). Notably, using mass spectrometry together with SMRT-seq, Zhou et al. obtained the first 6mA profile in rice genome (Zhou et al., 2018). Currently, there is a publicly available database namely "MethSMRT" that integrates multiple 6mA datasets derived from SMRT-seq (Ye et al., 2017). Although considerable progress has been made, the use of the high-throughput sequencing techniques is very limited as it is laborious and expensive.

Recently, as the rapid increase of the experimentally validated 6mA sites, more research efforts have been focused on the development of data-driven computational methods, especially machine learning based prediction methods. For instance, Chen et al., proposed the first machine learning based 6mA site predictor, named "i6mA-Pred," to predict 6mA sites in rice genome (Chen et al., 2019). The i6mA-Pred used nucleotide chemical properties and nucleotide frequency as features to formulate DNA sequences (Chen et al., 2017) and utilized support vector machine (SVM) to train the predictive model (Chen et al., 2019). The i6mA-Pred model achieved 83.13% in terms of the overall accuracy for identifying 6mA sites (Chen et al., 2019). More recently, researchers have proposed to use deep learning to identify 6mA sites, like iDNA6mA (5-step rule) (Tahir et al., 2019). This model can automatically extract features from DNA sequences by convolution neural network (CNN). Although these models have been proven to be effective and efficient in identifying DNA 6mA sites, the accuracy was not high enough to perform the genome-wide prediction.

In this study, we propose a new bioinformatics predictor, namely "6mAPred-FO." In this predictor, we aim to capture the discriminative characteristics of 6mA sites by different-view information integration and optimization. Based on the sequential features we extracted, we trained an SVM-based prediction model. Benchmarking comparative results have shown that under the 10-fold cross-validation, our model improves the exiting performance to 87.44% in the overall accuracy. Via further experimental analysis, we found that our performance improvement contributes mainly to our feature integration and optimization strategy. In particular, the nucleotide positional specificity information is complementary to sequence-order information to effectively distinguish 6mA sites from non-6mA sites. We anticipate this tool can be useful to discover new 6mA sites in other species, at least complementary to the high-throughput techniques.

MATERIALS AND METHODS

Benchmark Dataset

A high-quality benchmark dataset is essential for building an effective and unbiased supervised learning model. In this study, we used the same stringent benchmark dataset, which is originally proposed in Chen's study (Chen et al., 2019). In the dataset, the positive samples (sequences with 6mA sites) were obtained from NCBI Gene Expression Omnibus and the single-molecule real-time sequencing (Zhou et al., 2018). Afterwards, they separated out the sites with a modification score of <30 according to the Methylome Analysis Technical Note, and used the CD-HIT (Fu et al., 2012) software to eliminate sequences with the similarity of more than 60% (Chen et al., 2019). The negative samples (sequences without 6mA sites) were obtained from sub-sequences containing GAGG motifs in coding sequences (CDSs) of the rice genome (Zhou et al., 2018). Ultimately, 880 6mA sequences (positive samples) and 880 non-6mA sequences (negative samples) were retained in the dataset.

Framework of the Proposed 6mAPred-FO

Figure 1 illustrates the overall framework of the 6mAPred-FO method for DNA 6mA site prediction. The predictive procedure can be concluded as two phases: model training and prediction. In the training phase, the training samples are encoded and integrated by two feature representation algorithms: NPS (Nucleotide Positional Specificity) and PseDNC (Pseudo Dinucleotide Composition). Afterwards, the features are optimized to obtain the best feature subset for the training set. The resulting feature vectors are then fed into the SVM algorithm to train predictive model. In prediction phase, given the query sequences that are not characterized, we followed the similar procedure to encode the sequences, and used the trained model to predict whether the query sequences are 6mA sites or not.

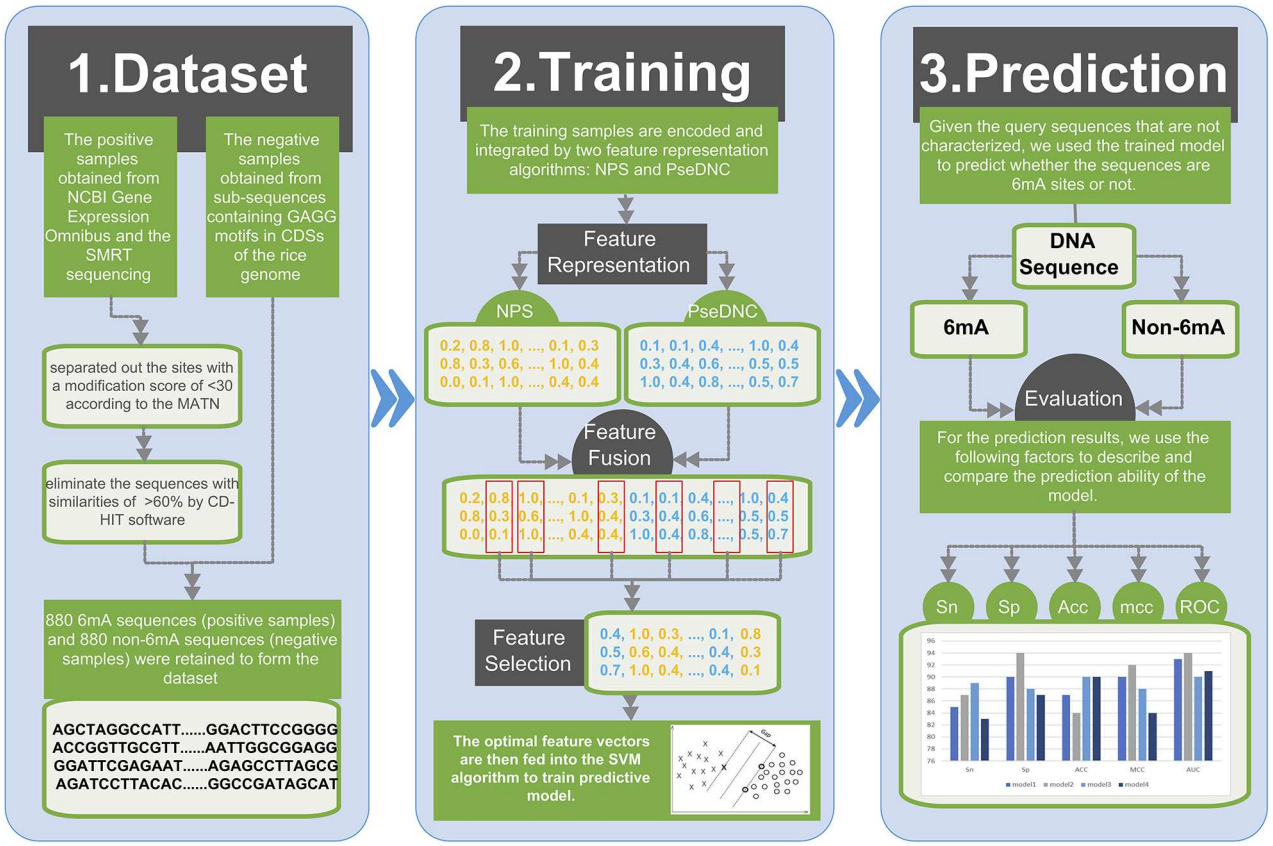


FIGURE 1 | Framework of the proposed 6mAPred-FO. The overall framework is divided into two phases: (1) model training phase and (2) prediction phase. In the model training phase, each DNA sequence is fed to two feature representation algorithms NPS and PseDNC to generate the corresponding features. Afterwards, through the feature fusion and optimization protocol, we yielded the optimal features for each sequence sample. To this end, we train an SVM model with the resulting feature set. In the prediction phase, the model predicts whether the sequence in the test dataset contains 6mA site. And then, compare our model with previous models in terms of ACC, etc. Finally, build a user-friendly web server to provide convenient 6mA site identification and prediction.

Feature Representation Algorithms

To convert DNA sequences into feature vectors that machine learning methods can handle, two feature representation algorithms, Nucleotide Positional Specificity (NPS) and Pseudo Dinucleotide Composition (PseDNC), are introduced for feature representation. Here is a brief introduction to the two algorithms.

Nucleotide Positional Specificity (NPS)

In this algorithm, two feature representation descriptors are used to encode the sequences.

The first feature is the positional binary encoding of flanking nucleotide sequence. We adopt the traditional method of flanking window to represent the 6mA site. On the premise that the minimum length 41 can perform well, if the 6mA site is located at both ends of the sequence, we fill the end of the sequence with the gap character "N." Therefore, in the orthogonal binary coding scheme, we transform nucleotide sequences into numeric vectors by the following rules: the codes of "A (adenine)," "T (thymine)," "C (cytosine)," "G (guanine)" and "N" are "(0, 0, 0, 1)," "(0, 0, 1, 0)," "(0, 1, 0, 0)," "(1, 0, 0, 0)," and "(0, 0, 0, 0)," respectively.

The second feature descriptor of NPS was the position-independent k-mer frequency. We calculated the frequencies of all possible k-mer nucleotides in a site-centered nearby flanking window. However, the vector dimension increases rapidly with the increase of k value, which leads to over-fitting. Thus, we set k to 2, 3, and 4. Finally, the 41-length DNA sequence is transformed into a 500-dimensional vector. More details about this method are available in the Xiang et al. (2016).

Pseudo Dinucleotide Composition (PseDNC)

PseDNC combines local and global pattern information of sequences. We use a vector to represent the DNA sequence as given below,

$$R = [d_1 \ d_2 \ \cdots \ d_{16} \ d_{16+1} \ \cdots \ d_{16+\lambda}]^T$$

where

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 16) \\ \dfrac{w\theta_{u-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (16 < u \le 16+\lambda) \end{cases}$$

In the formulation above, $f_u(u = 1, 2, \cdots, 16)$ is the normalized occurrence frequency of the u-th non-overlapping dinucleotides in the sequence. The w is the weight factor for balancing the component action of pseudo nucleotides. The θ_j is the j-th tier correlation factor that reflects the sequence order correlation between all the j-th most contiguous dinucleotides. What's more,

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1} C_{i,\ i+j} \qquad (j = 1, 2, \cdots, \lambda;\ \ \lambda < L)$$

where

$$C_{i,i+j} = \frac{1}{\mu}\sum_{g=1}^{\mu}\left[P_g\left(D_i\right) - P_g(D_{i+j})\right]^2$$

In the above two formulations, L is the length of DNA sequence and the number λ is an integer to reflect the correlation rank which is smaller than L. The $C_{i,i+j}$ is correlation function which is given above, where $P_g\left(D_i\right)$ is the numerical value of the g-th physicochemical property for the dinucleotide sequence D_i in the DNA, and so as $P_g(D_{i+j})$. The μ is the total number of correlation functions counted. It should be noticed that these values of physicochemical property were all subjected into a standard conversion by the formula below before substituting into the $P_g\left(D_i\right)$,

$$P_g\left(D_i\right) = \frac{P_g^0\left(D_i\right) - ave(P_g^0\left(D_i\right))}{SD\{ave\left(P_g^0\left(D_i\right)\right)\}}$$

where the symbol $ave()$ means getting the average of the values over the 16 different dinucleotides and $SD\{\ \}$ means the corresponding standard deviation. In the above equation, $P_g^0\left(D_i\right)$ is the original physicochemical property value for the dinucleotide. In this study, the following three physicochemical properties, namely enthalpy, entropy and free energy, are used to calculate the global or long-range sequence-order effects of the DNA. And their original values are given in **Table S1** of Supplementary material.

Ultimately, using this feature descriptor, we obtained 22 features. More details about these formulas can be found in the references Chen et al. (2014, 2015a,b), Liu (2019), Liu et al. (2019b).

Feature Fusion and Optimization Protocol

Feature fusion has been successfully applied into bio-sequence analysis (Zhang et al., 2017; Tang et al., 2018; Wei et al., 2018a,b; Liu et al., 2019d) and other bioinformatics tasks (Liang et al., 2018; Zhang et al., 2018, 2019a,b; Gong et al., 2019; Wang et al., 2019). It refers to merge different types of feature representations to more comprehensively capture the characteristics of samples from different perspectives. In this study, to make better use of different information, we fused the following two feature representations. One is 500-dimensional feature vector via NPS and the other is 22-dimensional feature vector via PseDNC. Accordingly, we yielded 522-dimensional features.

Generally, the fused feature space probably contains irrelevant or mutual information, impacting the predictive performance. Therefore, feature optimization is a necessary step forwards capturing the most discriminative features from the original feature space, building the optimal predictive model. It can help to eliminate irrelevant or redundant features, so as to reduce feature dimension, improve model accuracy as well as reduce computational cost. On the other hand, selecting relevant features can simplify the model and make it easier to understand the process of data generation. So far, in order to solve these problems, various effective feature optimization methods have been proposed, such as analysis of variance (Feng et al., 2019), binomial distribution (Su et al., 2018), minimal redundancy maximal relevance (Peng et al., 2005), and maximum relevance maximum distance (MRMD) (Zou et al., 2016; Chen W. et al., 2018).

To improve the feature representation ability, we used variance analysis in the filter method for feature selection. Its main idea is to calculate the variance of each feature by function f_classif in sklearn package. By doing so, we obtained the predictive contribution of each feature according to the corresponding f-value. The higher the f-value, the stronger the prediction ability. Afterwards, we selected the features one by one from high to low according to their f-values, and trained the SVM model for each feature subset. Different feature subsets of different dimensions can produce different models, and thus different prediction results can be obtained. The feature subset with the highest accuracy is yielded as the optimal feature subset. The analysis of feature optimization results is discussed in section "RESULTS AND DISCUSSION".

Support Vector Machine (SVM)

SVM is a powerful machine learning method for classification, regression and other machine learning tasks. It has been successfully applied in various fields to deal with a series of supervised learning problems (Zhang et al., 2016; Bu et al., 2018; Liu and Li, 2019; Manavalan et al., 2019a,b). The main principle of SVM is to transform the import data into high-dimensional feature space, and then determine the most suitable hyperplane for separating the samples in one class from another. After that, the trained hyperplane can be used to predict the unknown data. Based on this idea, a package namely LibSVM (Chih-chung and Chih-jen, 2011) was established to make the SVM more convenient to use. In this study, we implemented the SVM algorithm by using the LibSVM package. We chose the radial basis kernel (RBF) as a learning function, and optimized the parameters like cost and gamma by grid search to determine the optimal classification hyperplane of SVM. Given a sequence sample, the SVM model can calculate its probability score to be true 6mA sequence. If the probability is more than 50%, it is considered to be the 6mA sequence; otherwise, it is not the 6mA sequence.

Assessment of Predictive Ability

There are three cross-validation methods namely independent dataset test, n-fold cross-validation test and jackknife test

TABLE 1 | Comparison of single feature and fused features.

Features	Sn (%)	Sp (%)	ACC (%)	MCC	AUC
NPS	84.09	83.86	83.98	0.68	0.908
PseDNC	55.91	72.39	64.15	0.29	0.673
Fused Features	84.43	85.45	84.94	0.70	0.917

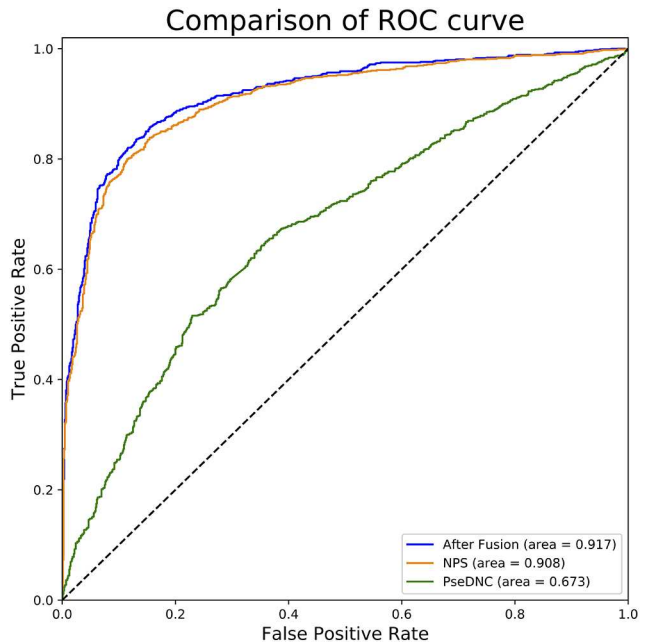


FIGURE 2 | ROC curves of single and fused features. This figure shows the ROC curve results of different features. The green curve characterizes ROC result of PseDNC and the orange one for NPS. The ROC curve of fused features is represented by blue.

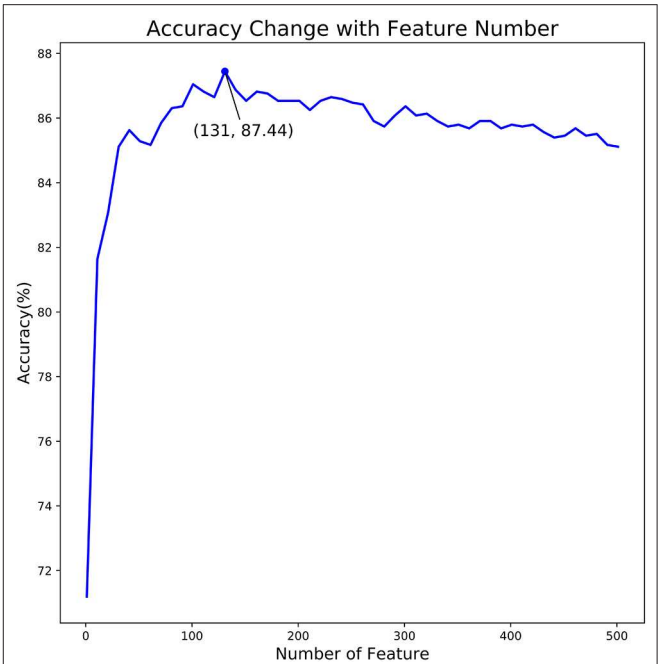


FIGURE 3 | The relationship curve of prediction accuracy and dimension of feature subset. The curve in this figure reflects the change of predictor accuracy with dimension of feature subset.

in statistical prediction to evaluate expected success rate of predictors (Manavalan and Lee, 2017; Wei et al., 2017a,d; He et al., 2018; Manavalan et al., 2018; Liu and Zhu, 2019; Liu et al., 2019a). In this study, we used n-fold cross-validation to examine the quality of the model. In the n-fold cross-validation, the dataset was randomly divided into n subsets, of which n-1 subsets were used as training data and the remaining one as testing data. This process would be repeated n times, each time using different testing data in turn. Corresponding accuracy and other evaluation metrics will be obtained in each test, and the average value of the evaluation index obtained from n-time results was used to evaluate the predictor. Generally, multiple n-fold cross-validation (such as 10 times n-fold cross-validation) is needed, and then its mean value is calculated to estimate the accuracy of the predictor.

Four metrics, sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthew's correlation coefficient (MCC), were used to evaluate the performance of the proposed method. The formulas of these metrics are given below:

$$\begin{cases} Sn = \frac{TP}{TP+FN} & 0 \le Sn \le 1 \\ Sp = \frac{TN}{TN+FP} & 0 \le Sp \le 1 \\ ACC = \frac{TP+TN}{TP+FP+TN+FN} & 0 \le ACC \le 1 \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN+FN)\times(TN+FP)\times(TP+FN)\times(TP+FP)}} & -1 \le MCC \le 1 \end{cases}$$

where, TP (True Positive) represents the number of positive samples correctly predicted; TN (True Negative) represents the number of negative samples correctly predicted; FP (False Positive) represents the number of negative samples incorrectly predicted to be the positives; FN (False Negative) represents the number of positive samples incorrectly predicted to be the negatives.

Moreover, we used the Receiver Operating Characteristic (ROC) curve to measure the overall performance of the predictive model. The area under the ROC curve (AUC) is to quantitively measure the quality of binary classifier. The closer the ROC curve is to the upper left corner, the better the performance of the predictor is. When the AUC value is closer to 0.5, it means that this is a random predictor (Hanley and Mcneil, 1982).

RESULTS AND DISCUSSION

Comparison of Single and Fused Features

In this section, we investigated the impact of the feature fusion protocol on the predictive performance. We compared two

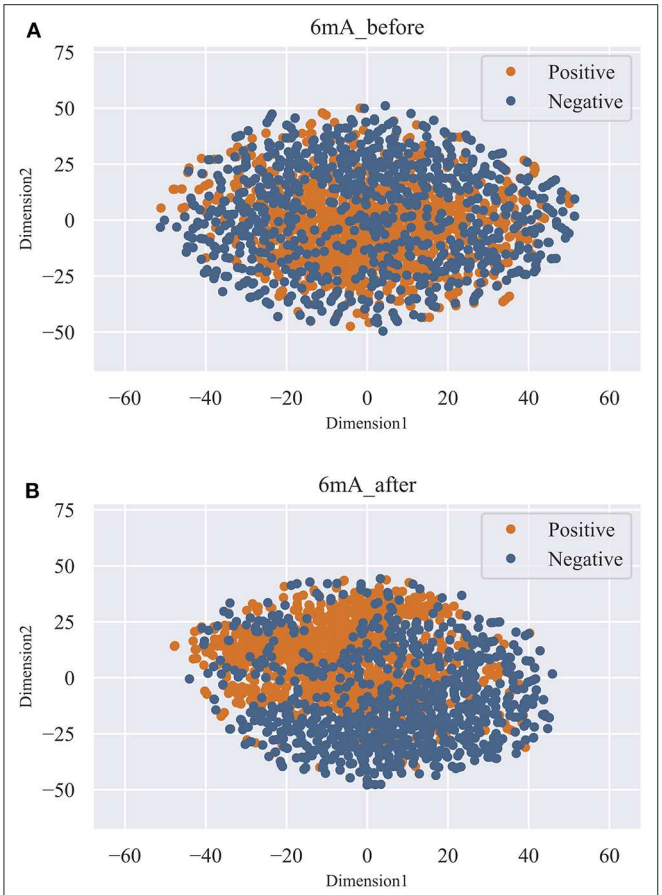


FIGURE 4 | T-SNE visualization of the feature space before and after feature optimization. **(A)** represents the distribution of the positives and negatives in original feature space. **(B)** represents the distribution of the positives and negatives in optimal feature space.

feature representations (NPS and PseDNC) with their fusion. They are evaluated with 10-fold cross validation on the same benchmark dataset used in this study. The comparison results are presented in **Table 1**. It can be seen that the fused features improve the performances in all the metrics. To be specific, the Sn, Sp, ACC, MCC, and AUC is enhanced by 0.34, 1.59, 1, 2, and 0.9%, as compared with the runner-up feature descriptor—NPS. For intuitive comparison, we further compared the ROC curves of different features in **Figure 2**. Similarly, the fused features show better performance than the single features. From the specific point of view in the **Figure 2**, the fused feature curve (the blue one) is closer to the upper left corner than the single feature curve. What's more, the AUC value of the fused feature is 0.917, which is higher than that of the single feature. This figure and accurate data can more intuitively support the conclusion above. Together, the results suggest that the information in different features is complementary to better capture the characteristics specificity of 6mA sites.

Feature Optimization Results

In the proposed feature optimization strategy, we firstly calculated the classification importance score of each feature in the feature set, and then the features are sorted from high to low according to their scores. Secondly, the feature in the sorted feature set is added to the feature subset one by one. Once a new feature is added to the feature subset, we obtained a new feature subset and train a new SVM model under its default parameters. We evaluated the performance of all feature subsets, respectively. The relationship between prediction accuracy and dimension of feature subset is illustrated in **Figure 3**.

As shown in **Figure 3**, we observed that the accuracy of the model increased rapidly as the feature number grows. Afterwards, the accuracy slightly declined as the feature number increases. When the feature number reached to 131, the model achieved the highest accuracy of 87.44%. Thus, the 131 features are considered as the optimal and used to train our predictive model. Moreover, the feature optimization results evaluated with other evaluation metrics, like MCC and ROC, can be found in **Table S2** of Supporting Information. To visually see how the feature space changes using feature optimization, we further compared the sample distribution between the original feature space and the optimal feature space, as depicted in **Figure 4**. It can be seen that the positives and negatives in the optimal feature space are more clearly distributed in two clear clusters as the original feature space. It demonstrates that using the feature optimization strategy, it helps to remove the irrelevant features and improve the feature representation ability.

Comparison of Different Kernel Functions

In this section, we compared the impact of RBF kernel function and other three kernel functions on the performance of our proposed model. They are Linear, Polynomial and Sigmoid. In this study, we used the same dataset to evaluate them. At the same time, they used the best feature subset after our fusion to show the performance. The 10-fold cross validation results can be found in **Table S3** of Supplementary material. According to the results in **Table S3**, we can find easily that SVM model using RBF kernel function achieves the highest prediction accuracy of 87.44% and performs better in other prediction factors. Moreover, with the help of RBF kernel function, AUC of the model is also the highest among several other kernel functions. In general, these results show that RBF kernel function is superior to other kernel functions in this study.

Comparison With Other Classifiers

To measure the superiority of SVM, we selected several other classifiers to compare with SVM. There are Gradient Boosting Decision Tree (GBDT) (Liao et al., 2017), K-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF) (Wei et al., 2017b,c; Lv et al., 2019; Ru et al., 2019). They are evaluated based on the same dataset used in this study with our fused feature set. The 10-flod cross validation results of prediction accuracy and AUC value are illustrated in **Figure 5**. In **Figure 5A** represents the comparison results of prediction accuracy of six classifiers, and **Figure 5B** represents the AUC value. As shown in **Figure 5**, we observed that the SVM got the highest score among the six classifiers not only in predictive accuracy but also AUC. The 10-fold cross validation results of other evaluation factors are illustrated in **Table S4** of

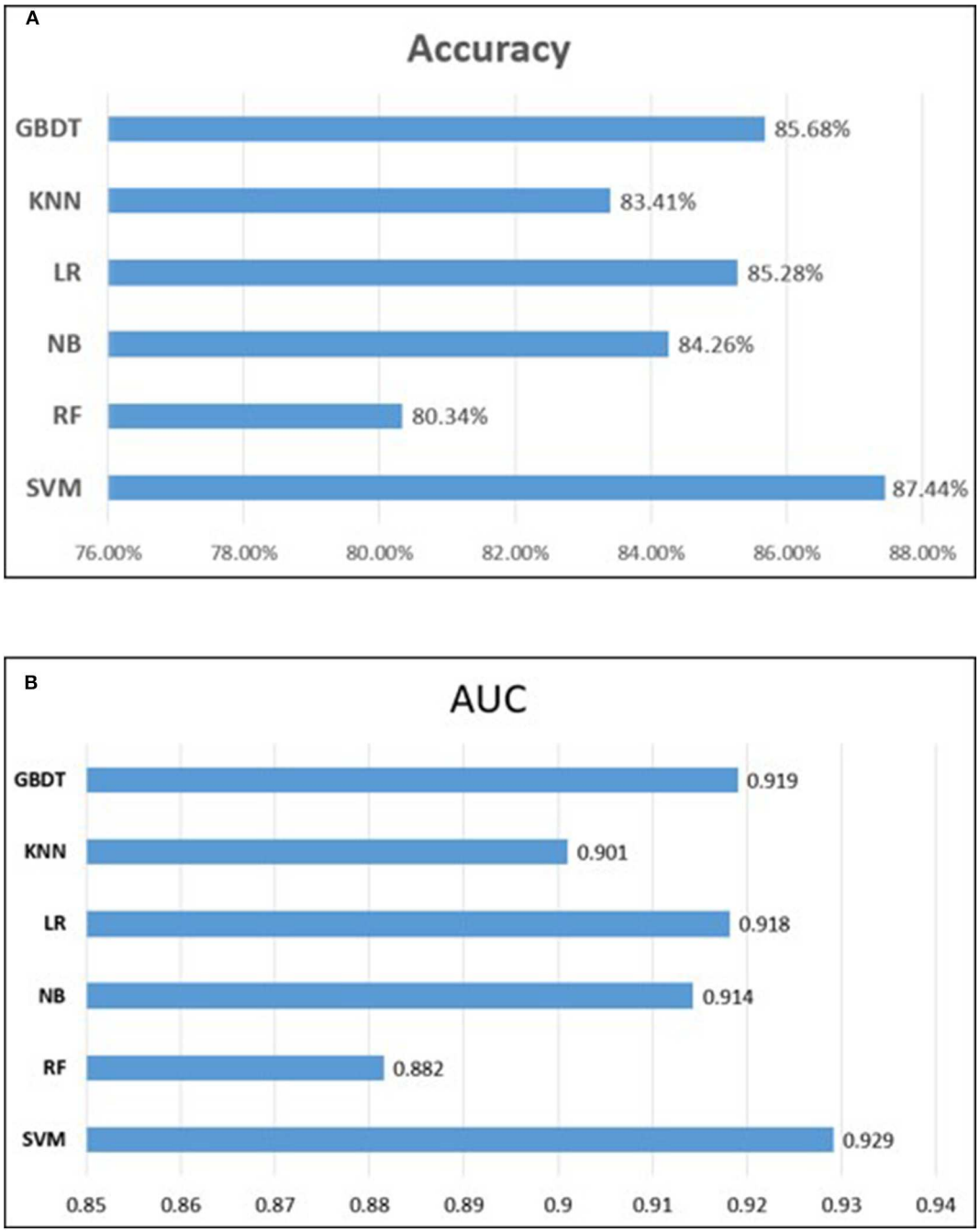


FIGURE 5 | Performance comparison of different classifiers. **(A)** represents the comparison results of prediction accuracy of six classifiers, and **(B)** represents the comparison results of auROC.

Supplementary material, which provide us with more specific classifier performance information. From **Table S4**, we can see that the SVM also performs better than other classifiers in other performance indicators. For intuitive comparison, we further compared their ROC curves as illustrated in **Figure 6**. As seen, SVM achieved 0.917 in terms of AUC, which is higher than GBDT and other classifiers. It can be seen from the figure that the ROC curve corresponding to SVM is at the top, which means that SVM has better classification performance than other classifiers. In general, these results demonstrate that SVM is better than other commonly used classifiers in this study.

Comparison With Existing Predictors

To measure the effectiveness of our predictive model—6mAPred-FO, we compared the model with i6mA-Pred

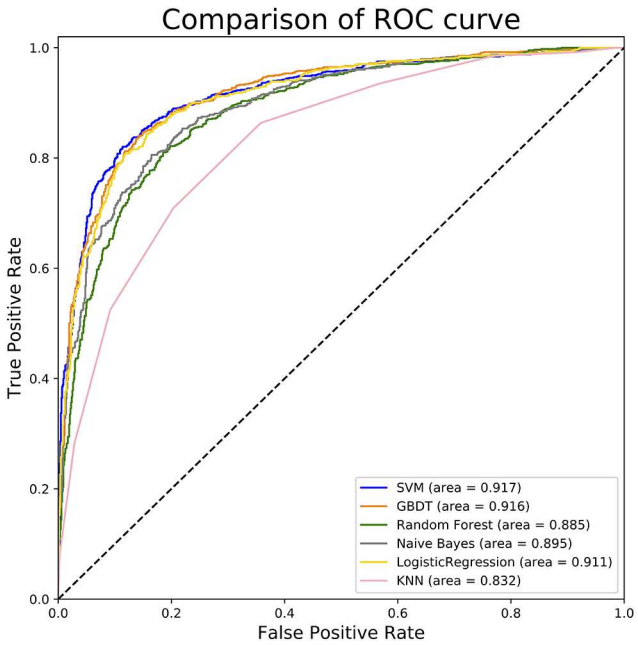


FIGURE 6 | ROC curves of different classifiers. This figure shows the ROC curve results of different classifiers. For example, according to legend, the blue curve characterizes ROC result of SVM and the orange one for GBDT. The value of auROC is also calculated after the name in the legend.

TABLE 2 | Comparison of the proposed 6mAPred-FO with existing predictors.

Method	Sn (%)	Sp (%)	ACC (%)	MCC	AUC
i6mA-Pred	82.95	83.30	83.13	0.66	0.886
iDNA6mA (5-step rule)	86.70	86.59	86.64	0.732	0.931
6mAPred-FO	86.93	87.95	87.44	0.75	0.929

(Chen et al., 2019) and iDNA6mA (5-step rule) on the same dataset, which are the best two among existing predictors to identify the 6mA site. The results are presented in **Table 2**. As shown in **Table 2**, i6mA-Pred obtains the accuracy of 83.13%, sensitivity of 82.95%, specificity of 83.30%, MCC of 0.66 and AUC of 0.886, while our prediction model obtains the accuracy of 87.44%, sensitivity of 86.93%, specificity of 87.95%, MCC of 0.75 and AUC of 0.929. Obviously, our method is superior to i6mA-Pred in all the metrics. Specifically, as compared to i6mA-Pred, our model achieved 4.31%, 3.98%, 4.65%, 0.09 and 0.043 higher in terms of ACC, Sn, Sp, MCC, and AUC, respectively. This demonstrated that our feature representations are more effective to capture the characteristic specificity of 6mA sites. In the **Table 2**, we also compared our predictor model with iDNA6mA (5-step rule). It can be seen that the accuracy of our 6mAPred-FO is 0.8% higher than iDNA6mA (5-step rule). All the other performance indicators except AUC value are slightly higher than those of iDNA6mA (5-step rule). Generally, it can be concluded that our 6mAPred-FO is better than existing predictors in distinguishing 6mA sites from non-6mA sites.

CONCLUSIONS

In this study, we have proposed a new machine learning based 6mA site predictor namely 6mAPred-FO. To sufficiently capture the characteristics of 6mA sites, we have combined the information from two feature representations NPS and PseDNC, and further optimized the features by feature selection. Feature analysis results showed that as compared with the single feature descriptor, the fused features perform better, demonstrating that different information are complementary to improve the predictive performance. Moreover, feature selection is an effective strategy to optimize the feature space and improve the feature representation ability. We have also compared our 6mAPred-FO with existing predictors on benchmark datasets. The comparative results showed that our approach improved the performance significantly in terms of multiple metrics like SN, SP, MCC, and AUC. This suggests that our feature fusion and selection scheme is more effective to represent 6mA sites in comparison with existing features. From our study results, we can make a reasonable inference that the recognition of 6mA site is closely related to the local and global pattern information represented by PseDNC. Then, the position specific information represented by NPS is fused to make our proposed algorithm more accurate for the recognition of 6mA sites. In general, our method provides a more accurate model for biological scientists to identify 6mA site in rice genome. In the future, we will pay more attention on deep learning (Liu et al., 2019c; Zou et al., 2019) for the accuracy improvement.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the http://server.malab.cn/6mAPred-FO/Download.html.

AUTHOR CONTRIBUTIONS

RC, YN, and DW: conceptualization. XY and RS: data curation. JC: writing—original draft preparation. GX and LW: writing—review and editing. JC: visualization. LW and GX: supervision.

FUNDING

This work is supported by National Natural Science Foundation of China (Grand Nos. 61701340, 61702431).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00502/full#supplementary-material

REFERENCES

Bu, H., Hao, J., Guan, J., and Zhou, S. (2018). Predicting enhancers from multiple cell lines and tissues across different developmental stages based on SVM method. *Curr. Bioinformatics* 13, 655–660. doi: 10.2174/1574893613666180726163429

Campbell, J. L., and Kleckner, N. (1990). *E. coli* oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. *Cell* 62, 967–979. doi: 10.1016/0092-8674(90)90271-F

Chen, H., Shu, H., Wang, L., Zhang, F., Li, X., Ochola, S. O., et al. (2018). Phytophthora methylomes are modulated by 6mA methyltransferases and associated with adaptive genome regions. *Genome Biol.* 19:181. doi: 10.1186/s13059-018-1564-4

Chen, W., Feng, P., Ding, H., and Lin, H. (2018). Classifying included and excluded exons in exon skipping event using histone modifications. *Front. Genet*. 9:433. doi: 10.3389/fgene.2018.00433

Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K. C. (2015a). iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem*. 490, 26–33. doi: 10.1016/j.ab.2015.08.021

Chen, W., Lei, T. Y., Jin, D. C., Lin, H., and Chou, K. C. (2014). PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem*. 456, 53–60. doi: 10.1016/j.ab.2014.04.001

Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 35, 2796–2800. doi: 10.1093/bioinformatics/btz015

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K. C. (2015b). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* 31, 119–120. doi: 10.1093/bioinformatics/btu602

Chih-chung, C., and Chih-jen, L. (2011). Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol*. 2, 1–27. doi: 10.1145/1961189.1961199

Feng, C. Q., Zhang, Z. Y., Zhu, X. J., Lin, Y., Chen, W., Tang, H., et al. (2019). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 35, 1469–1477. doi: 10.1093/bioinformatics/bty827

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465. doi: 10.1038/nmeth.1459

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gong, Y., Niu, Y., Zhang, W., and Li, X. (2019). A network embedding-based multiple information integration method for the MiRNA-disease association prediction. *BMC Bioinformatics* 20:468. doi: 10.1186/s12859-019-3063-3

Greer, E. L., Blanco, M. A., Gu, L., Sendinc, E., Liu, J., Aristizabal-Corrales, D., et al. (2015). DNA Methylation on N6-Adenine in C. elegans. *Cell* 161, 868–878. doi: 10.1016/j.cell.2015.04.005

Hanley, J. A., and Mcneil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

He, W., Jia, C., Duan, Y., and Zou, Q. (2018). 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol.* 12(Suppl. 4):44. doi: 10.1186/s12918-018-0570-1

Krais, A. M., Cornelius, M. G., and Schmeiser, H. H. (2010). Genomic N(6)-methyladenine determination by MEKC with LIF. *Electrophoresis* 31, 3548–3551. doi: 10.1002/elps.201000357

Li, B., Du, K., Gu, S., Xie, J., Liang, T., Xu, Z., et al. (2019). Epigenetic DNA modification N(6)-Methyladenine inhibits DNA replication by DNA polymerase of pseudomonas aeruginosa Phage PaP1. *Chem. Res. Toxicol*. 32, 840–849. doi: 10.1021/acs.chemrestox.8b00348

Liang, D., Wang, H., Song, W., Xiong, X., Zhang, X., Hu, Z., et al. (2016). The decreased N(6)-methyladenine DNA modification in cancer cells. *Biochem. Biophys. Res. Commun*. 480, 120–125. doi: 10.1016/j.bbrc.2016.09.136

Liang, S., Ma, A., Yang, S., Wang, Y., and Ma, Q. (2018). A review of matched-pairs feature selection methods for gene expression data analysis. *Comput. Struct. Biotechnol. J*. 16, 88–97. doi: 10.1016/j.csbj.2018.02.005

Liao, Z., Li, D., Wang, X., Li, L., and Zou, Q. (2016). Cancer diagnosis through isomiR expression with machine learning method. *Curr. Bioinform*. 13, 57–63. doi: 10.2174/1574893611666160609081155

Liao, Z., Wan, S., He, Y., and Zou, Q. (2017). Classification of small GTPases with hybrid protein features and advanced machine learning techniques. *Curr. Bioinform*. 12, 1–9. doi: 10.2174/1574893612666171121162552

Linn, S., and Arber, W. (1968). Host specificity of DNA produced by Escherichia coli, X. *In vitro* restriction of phage fd replicative form. *Proc. Natl. Acad. Sci. U.S.A.* 59, 1300–1306. doi: 10.1073/pnas.59.4.1300

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief Bioinform*. 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Chen, S., Yan, K., and Weng, F. (2019a). iRO-PsekGCC: Identify DNA Replication Origins Based on Pseudo k-Tuple GC Composition. *Front. Genet* 10:842. doi: 10.3389/fgene.2019.00842

Liu, B., Gao, X., and Zhang, H. (2019b). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res*. 47:e127. doi: 10.1093/nar/gkz740

Liu, B., Li, C. C., and Yan, K. (2019c). DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief Bioinform*. doi: 10.1093/bib/bbz098

Liu, B., and Li, K. (2019). iPromoter-2L2.0: Identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol. Ther. Nucleic Acids* 18, 80–87. doi: 10.1016/j.omtn.2019.08.008

Liu, B., and Zhu, Y. (2019). ProtDec-LTR3.0: protein remote homology detection by incorporating profile-based features into learning to rank. *IEEE Access* 7, 102499–102507. doi: 10.1109/ACCESS.2019.2929363

Liu, S., Zhao, X., Zhang, G., Li, W., Liu, F., Liu, S., et al. (2019d). PredLnc-GFStack: a global sequence feature based on a stacked ensemble learning method for predicting lncRNAs from transcripts. *Genes (Basel)* 10:672. doi: 10.3390/genes10090672

Luria, S. E., and Human, M. L. (1952). A nonhereditary, host-induced variation of bacterial viruses. *J. Bacteriol* 64, 557–569. doi: 10.1128/JB.64.4.557-569.1952

Lv, Z., Jin, S., Ding, H., and Zou, Q. (2019). A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.* 7:215. doi: 10.3389/fbioe.2019.00215

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019a). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35, 2757–2765. doi: 10.1093/bioinformatics/bty1047

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019b). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Manavalan, B., and Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* 33, 2496–2503. doi: 10.1093/bioinformatics/btx222

Manavalan, B., Shin, T. H., and Lee, G. (2018). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol*. 9:476. doi: 10.3389/fmicb.2018.00476

Meselson, M., and Yuan, R. (1968). DNA restriction enzyme from *E. coli*. *Nature* 217, 1110–1114. doi: 10.1038/2171110a0

O'Brown, Z. K., and Greer, E. L. (2016). N6-Methyladenine: a conserved and dynamic DNA mark. *Adv. Exp. Med. Biol*. 945, 213–246. doi: 10.1007/978-3-319-43624-1_10

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell*. 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Pomraning, K. R., Smith, K. M., and Freitag, M. (2009). Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods* 47, 142–150. doi: 10.1016/j.ymeth.2008.09.022

Pukkila, P. J., Peterson, J., Herman, G., Modrich, P., and Meselson, M. (1983). Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in *Escherichia coli*. *Genetics* 104, 571–582.

Robbins-Manke, J. L., Zdraveski, Z. Z., Marinus, M., and Essigmann, J. M. (2005). Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase- and mismatch repair-deficient Escherichia coli. *J. Bacteriol.* 187, 7027–7037. doi: 10.1128/JB.187.20.7027-7037.2005

Ru, X., Li, L., and Zou, Q. (2019). Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. *J. Proteome Res.* 18, 2931–2939. doi: 10.1021/acs.jproteome.9b00250

Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* 34, 4196–4204. doi: 10.1093/bioinformatics/bty508

Tahir, M., Tayara, H., and Chong, K. T. (2019). iDNA6mA (5-step rule): identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. *Chemometr. Intell. Lab. Systems* 189, 96–101. doi: 10.1016/j.chemolab.2019.04.007

Tang, G., Shi, J., Wu, W., Yue, X., and Zhang, W. (2018). Sequence-based bacterial small RNAs prediction using ensemble learning strategies. *BMC Bioinformatics* 19, 13–23. doi: 10.1186/s12859-018-2535-1

Wang, Y., Yang, S., Zhao, J., Du, W., Liang, Y., Wang, C., et al. (2019). Using machine learning to measure relatedness between genes: a multi-features model. *Sci. Rep.* 9, 1–15. doi: 10.1038/s41598-019-40780-7

Wei, L., Chen, H., and Su, R. (2018a). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids* 12, 635–644. doi: 10.1016/j.omtn.2018.07.004

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z. S., and Zou, Q. (2017b). CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J. Proteome Res.* 16, 2044–2053. doi: 10.1021/acs.jproteome.7b00019

Wei, L., Xing, P., Tang, J., and Zou, Q. (2017c). PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Trans. Nanobiosc.* 16, 240–247. doi: 10.1109/TNB.2017.2661756

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017d). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018b). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Xiang, S., Yan, Z., Liu, K., Zhang, Y., and Sun, Z. (2016). AthMethPre: a web server for the prediction and query of mRNA m(6)A sites in Arabidopsis thaliana. *Mol. Biosyst.* 12, 3333–3337. doi: 10.1039/C6MB00536E

Ye, P., Luan, Y., Chen, K., Liu, Y., Xiao, C., and Xie, Z. (2017). MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res.* 45, D85–D89. doi: 10.1093/nar/gkw950

Zhang, N., Sa, Y., Guo, Y., Lin, W., Wang, P., and Feng, Y. (2016). Discriminating ramos and jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine. *Curr. Bioinform.* 11, 50–56. doi: 10.2174/1574893611666160608102537

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019a). SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions. *Inf. Sci.* 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019b). A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2931546

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018). SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PLoS Comput. Biol.* 14:e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, W., Zhu, X., Fu, Y., Tsuji, J., and Weng, Z. (2017). Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods. *BMC Bioinformatics* 18(Suppl. 13):464. doi: 10.1186/s12859-017-1875-6

Zhou, C., Wang, C., Liu, H., Zhou, Q., Liu, Q., Guo, Y., et al. (2018). Identification and analysis of adenine N(6)-methylation sites in the rice genome. *Nat Plants* 4, 554–563. doi: 10.1038/s41477-018-0214-x

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 29 June 2020
doi: 10.3389/fgene.2020.00630

EHR2Vec: Representation Learning of Medical Concepts From Temporal Patterns of Clinical Notes Based on Self-Attention Mechanism

Li Wang[1,2], Qinghua Wang[1], Heming Bai[2], Cong Liu[3], Wei Liu[2], Yuanpeng Zhang[1,2], Lei Jiang[4], Huji Xu[4,5,6], Kai Wang[7,8] and Yunyun Zhou[7]**

[1] Department of Medical Informatics, Medical School, Nantong University, Nantong, China, [2] Research Center for Intelligence Information Technology, Nantong University, Nantong, China, [3] Department of Biomedical Informatics, Columbia University, New York, NY, United States, [4] Department of Rheumatology and Immunology, Changzheng Hospital, The Second Military Medical University, Shanghai, China, [5] Beijing Tsinghua Chang Gung Hospital, School of Clinical Medicine, Tsinghua University, Beijing, China, [6] Peking-Tsinghua Center for Life Sciences, Tsinghua University, Beijing, China, [7] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA, United States, [8] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Zhiguo Zhou,
University of Central Missouri, United States
Donghan Yang,
UT Southwestern Medical Center, United States

***Correspondence:**
Kai Wang
wangk@email.chop.edu
Yunyun Zhou
zhouy6@email.chop.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 03 March 2020
Accepted: 26 May 2020
Published: 29 June 2020

Citation:
Wang L, Wang Q, Bai H, Liu C, Liu W, Zhang Y, Jiang L, Xu H, Wang K and Zhou Y (2020) EHR2Vec: Representation Learning of Medical Concepts From Temporal Patterns of Clinical Notes Based on Self-Attention Mechanism. Front. Genet. 11:630. doi: 10.3389/fgene.2020.00630

Efficiently learning representations of clinical concepts (i. e., symptoms, lab test, etc.) from unstructured clinical notes of electronic health record (EHR) data remain significant challenges, since each patient may have multiple visits at different times and each visit may contain different sequential concepts. Therefore, learning distributed representations from temporal patterns of clinical notes is an essential step for downstream applications on EHR data. However, existing methods for EHR representation learning can not adequately capture either contextual information per-visit or temporal information at multiple visits. In this study, we developed a new vector embedding method called EHR2Vec that can learn semantically-meaningful representations of clinical concepts. EHR2Vec incorporated the self-attention structure and showed its utility in accurately identifying relevant clinical concept entities considering time sequence information from multiple visits. Using EHR data from systemic lupus erythematosus (SLE) patients as a case study, we showed EHR2Vec outperforms in identifying interpretable representations compared to other well-known methods including Word2Vec and Med2Vec, according to clinical experts' evaluations.

Keywords: natural language processing, representation learning, electronic health record, unstructured clinical notes, word vector

INTRODUCTION

In the field of clinical natural language processing (NLP), deep learning (DL) techniques outperform other NLP methods in many tasks, such as information extraction, named entity detection, and relationship assignment, etc. In DL-based NLP approaches, learning representative features is the critical step for the following analysis, such as classifications, clustering, and more. According to a recent review by Wu et al. (2020), among 1,737 clinical NLP articles, 74.1% of which used the DL-based Word2Vec method till 2018. Word2Vec is an unsupervised feature extraction method for representation learning, which converts word to numerical embedding by mapping

each of the word tokens into high-dimensional vector space (Mikolov et al., 2013). Words that are closely related in certain conditions will be clustered together and share short distance in high-dimensional semantic space.

There are two types of DL architectures in Word2Vec model, CBOW, and Skip-gram. However, Word2Vec has limitations in capturing contextual information globally when analyzing clinical notes of electronic health records (EHR) (Mikolov et al., 2013). First, it scans the nearby relationship of each center word by a sliding window with a fixed size, yet considers each word in the window with equal importance; second, it does not incorporate temporal relationships of events on the same patients over time. In real-world EHR data, identifying representative medical entities is more complicated since the sequential relationship among entities per visit and temporal relationship among multiple visits need to be considered. For example, each patient has multiple visits for different reasons mapping to different visiting events at different time points; and the intervals of the patient's multiple medical events may vary from a few days to several months. More issues complicated these problems are that entities from different medical categories (i.e., diagnosis, medication, and procedures) often constitute disordered sequential collections and ignored long-range semantic dependencies.

To overcome the challenges of handling temporal issues, Med2Vec was proposed by Choi et al. (2016) to learn medical entity representations for EHR data at the visit level (Choi et al., 2016). Med2Vec essentially adopted the Word2Vec structure but has two layers; the first layer is to capture the relations between medical entities within a visit, and the second layer is to capture the relations between medical visit sequences. However, Med2Vec does not overcome the limitations of Word2Vec that considers the surrounding words within a window to be equally important, which makes the representative learning less efficient and inaccurate. In real-world EHR data, the impact of its nearby terms differs relative to the center word.

In this project, we proposed a new representation learning method for embedding medical entities for EHR data, called EHR2Vec. The EHR2Vec model incorporated a self-attention mechanism to learn important representations by updating values of the context words as a whole per visiting event. The self-attention algorithm, which is a DL-based method, was initially used in image processing (Vaswani et al., 2017). Instead of considering every part of the entire image equally important, it focuses on a specific portion while down weighting other parts of an image, which greatly improves the learning accuracy for the point of interest. Similar to NLP task in general documents, each patient's EHR data from multiple visits can be considered as a document composed of many sentences, while each visit can be considered as a sentence composed of many medical entity tokens. Because of the information heterogeneity of medical entities, we grouped them into four categories: medication, diagnosis, symptom, and lab test. Since time sequence information is an important factor in finding the most relevant representations at a certain time point, we sorted the medical events of each patient in the order of time to improve learning accuracy.

Compared to existing methods, EHR2Vec method has following characteristics: (1) it applies self-attention algorithm with multi-headed design to identify important global representations at visit level, which greatly improves embedding accuracy compared to previous word embedding methods; (2) it enables more accurate symptom detections in a temporal order, which can be used to facilitate predictions of disease progression trajectories. In the current study, we applied EHR2Vec on Systemic lupus erythematosus (SLE) data and compared the results with clinicians' manual interpretations. The results of the experiment indicated that EHR2Vec's high-dimensional embedding features are interpretable and consistent with clinician's opinions.

MATERIALS AND METHODS

Description of Experimental Data

Since SLE is a chronic autoimmune disease which can gradually develop to multiple comorbidities from mild to serious as time elapses, we used SLE as a apt disease on which to test our approach. SLE involves multiple organs and systems throughout the body, often accompanied by various comorbidities, of which lupus nephropathy (LN) is one and can cause visceral organ failures (Almaani et al., 2017). The EHR data used in the study were from the hospitalization and discharge records of SLE patients from 13 Grade III Level A hospitals in China (**Supplementary Table 1**). The study was performed according to a protocol approved by the Institutional Committee on Ethics of Biomedicine. There are 14,439 de-identified SLE patients with a total of 57,367 Chinese clinical notes, enrolled from October 28, 2001 to May 31st, 2016. Privacy information, such as patients' name, photo ids, home addresses, and diagnostic dates has been de-identified and anonymized. These SLE patients' averaged diagnosis age is 33.4 years old, including 13,062 females (90.46%). We performed rigorous data quality control by recruiting patients whose EHR recorded the related diagnosis and treatment information corresponding to the time of each admission and discharge. Patients who missed first diagnosis date and only had one visit were excluded from further study. Ultimately, the final data set contains 14,219 patients with 49,752 notes. Among these patients, 14,039 (98.7%) patients have <10 visits.

Pre-processing to Extract Medical Entities From Unstructured Clinical Notes

The overview of the EHR2Vec workflow is shown in **Figure 1**. Medical entities extracted from Chinese clinical notes follow standard pipeline, including word segmentation, part-of-speech tagging, named entity recognition (NER), annotation and normalization using our in-house customized scripts. We collected standardized Chinese medical vocabulary knowledgebase, including thesaurus, grammar, and semantic database, as the prior information for NER task. Word segmentation used a string matching method to identify thesaurus and then used a bi-directional matching method to parse grammar rules. Part-of-speech tagging used a dynamic viterbi algorithm for semantic analysis (Klein and Manning, 2003; Schmid, 2004). Finally, NER task used bi-LSTM

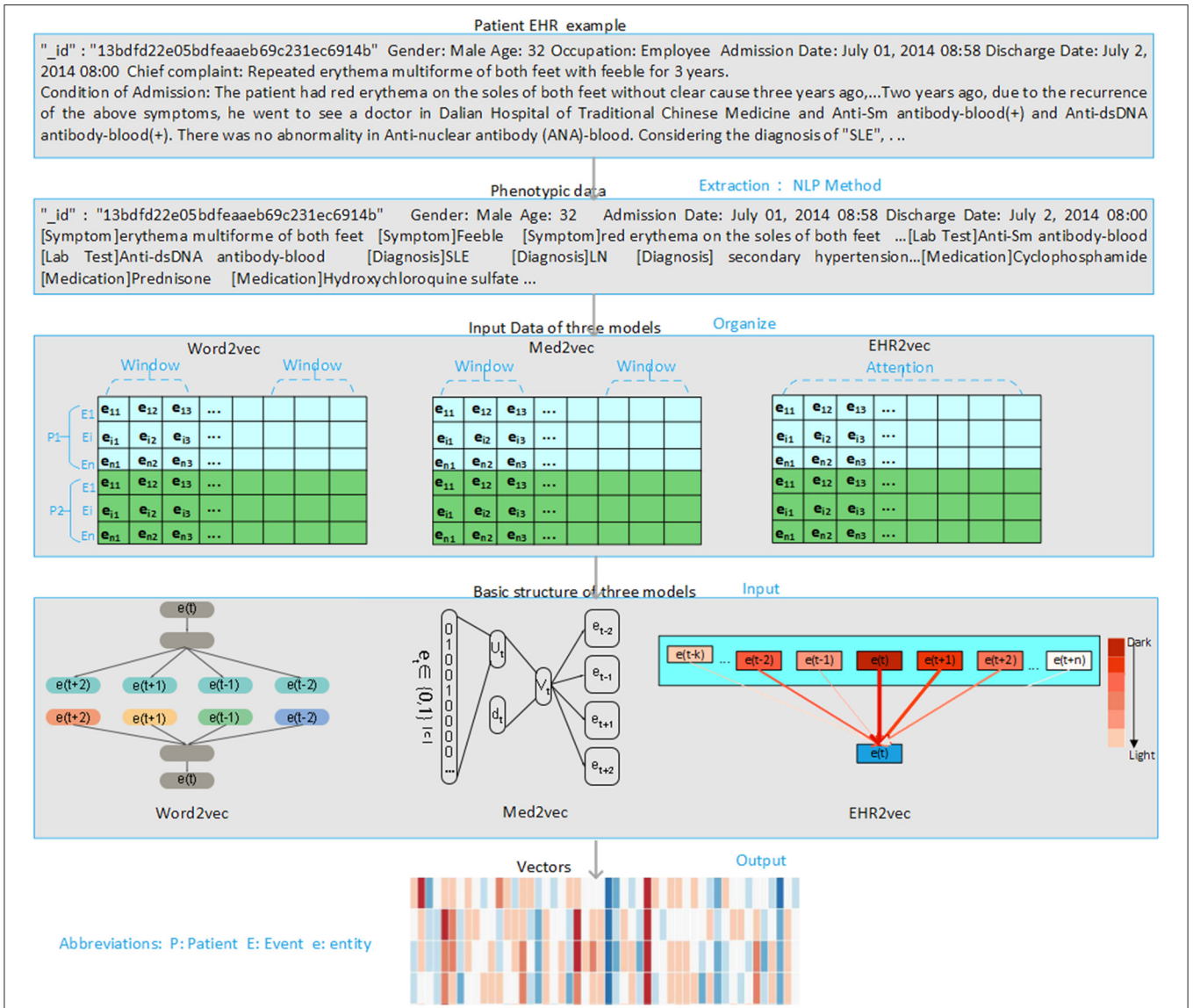


FIGURE 1 | Overview of project design. Our project has four steps. Firstly, we extract medical concepts (i.e., symptoms, lab test, medication, and diagnosis) from free clinical notes of SLE patients' EHR data. Next, we align these medical concepts into structured data format for each note per visit. Then, we sort notes in time series order for each patient. Finally, we comparatively study the performance of three embedding methods: Word2Vec, Med2Vec, and EHR2Vec.

algorithm to improve the accuracy and recall continuously. All the medical entities of the patients were tokenized as the input of Word2Vec, Med2Vec, and EHR2Vec for performance comparison.

EHR2Vec Method for Representations Learning

EHR2Vec has two layers, of which the first is to capture the relations between medical entities within each of the patient's medical event, and the second is to capture the global relations among the different medical visit events. In EHR2Vec, each medical entity within each visiting event undergoes attention computation with the goal of learning the dependencies between medical entity vectors. Assuming each patient P had multiple visiting events $E = \{E_1, E_2, \ldots E_n\}$. In a particular visiting event E_i, medical entity j can be represented as e_{ij}. The medical entities were grouped into four categories for further analysis including symptom, medicine, lab test, and diagnosis. **Supplementary Figure 1** illustrated an example of information extraction at different time points and intervals for a patient.

The initialized vector-matrix W, is in vector space R^{h*c}, where c is the dimension of each entity vector, h is the number of entities in all visits. Here, we used default value $c = 512$, which means each entity maps to 512-dimensional vector space. We first input the patient's initialized vector matrix to the first sublayer (attention mechanism). Equation (1) is the core formula of the attention mechanism that is used, in which Q, K, and V represent

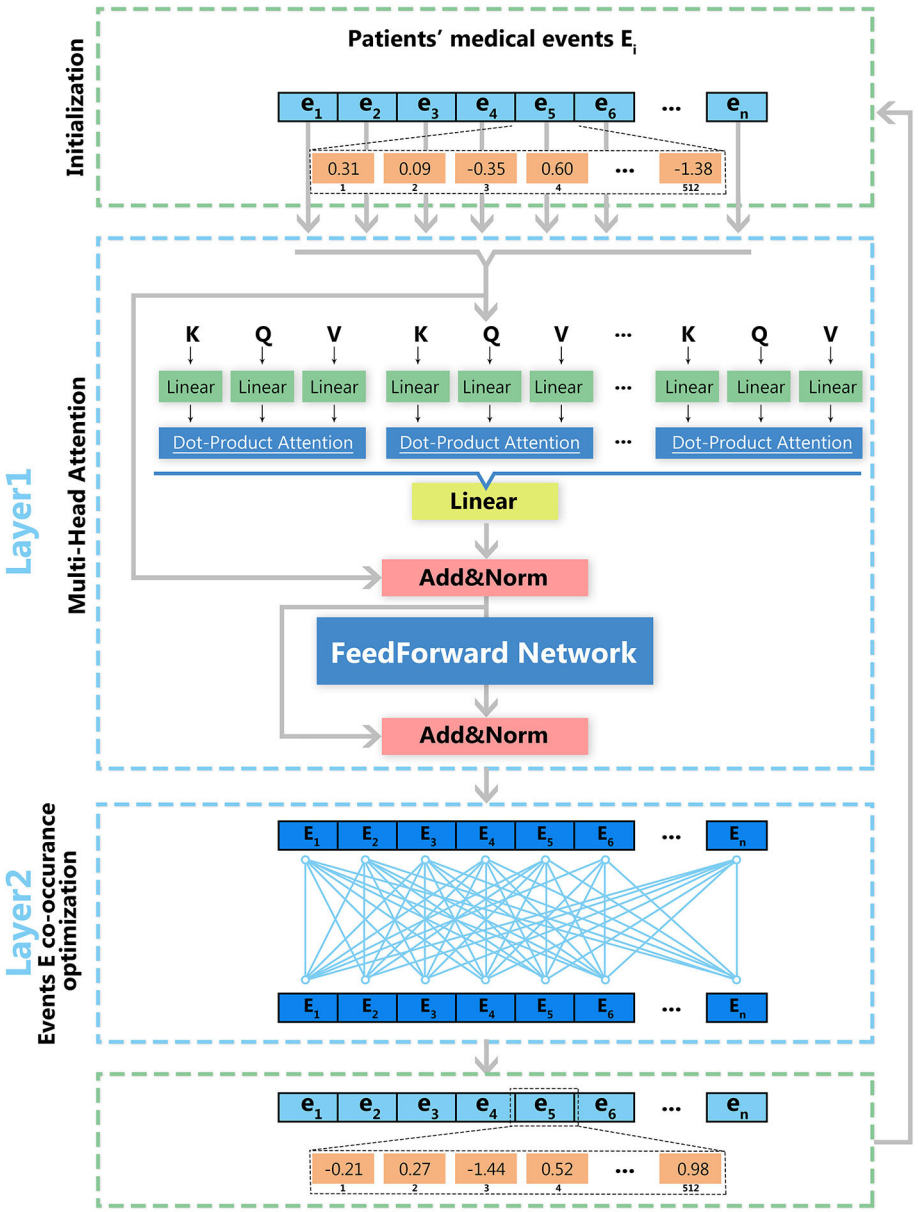


FIGURE 2 | Deep learning architecture of EHR2Vec. EHR2Vec is developed under deep learning framework including two layers of optimizations. The first layer is based on self-attention structure with multi heads to capture the relationship of different medical concepts within each visit event. The second layer is based on co-occurrence of visits to capture the relationships among visits of patients.

the query vector, key vector, and value vector, respectively, and d_k represents the dimension of Q, K, or V. The reason for the division by the square root of d_k is to prevent the product of QK^T from being too large, which may cause the softmax function to enter the saturation region so that the gradient would be too small (Vaswani et al., 2017).

In the model, to extract more features, a multi-head attention structure is adopted, in which a total of eight attention heads are used. Each head can capture different layers of dependency relationships. The eight attention heads are equivalent to eight subtasks, each subtask generating its own attention. The attention calculation of the eight attention heads can be performed through parallel computing to speed up the calculation.

$$\text{Attention}\,(\text{Q},\text{K},\text{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\text{V} \tag{1}$$

Model Optimization

As shown in **Figure 2**, model optimization has two steps. The first step is the optimization for a multi-head attention structure within each visit event. The second step is the optimization of deep feedforward networks for multiple visit events. Vector

matrix W is obtained through iterative training, and each row of W represents the vector in the medial entity set. Therefore, we obtain the final medical entity vectors by continuously optimizing the vector-matrix W. The log-likelihood function is used to optimize the obtained medical event vectors as shown in Equation (2), in which e_i and e_j represent medical entities in each medical event and T represents the number of medical events. By maximizing this function's value, we obtain the optimized vector-matrix W.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i\,:\,e_i\in E_t}\sum_{j\,:\,e_j\in E_t,\,j\neq i}\log p\left(e_j|e_i\right) \qquad \text{where,}$$

$$p\left(e_j|e_i\right)=\frac{\exp\left(W\left[i,:\right]^T W\left[j,:\right]\right)}{\sum_{k=1}^{all}\exp\left(W\left[k,:\right]^T W\left[i,:\right]\right)} \tag{2}$$

Intrusion Analysis for Representations Evaluation

Intrusion analysis is used to test whether the identified representations agree with human judgment. To evaluate the accuracy of EHR2Vec, we performed two types of intrusion analysis (Chang et al., 2009; Murphy et al., 2012; Luo et al., 2015). The first intrusion experiment is to compare with clinical experts' opinions. We calculated the cosine correlation values of a given medical entity's vector against all other medical entities e_{ij} and then ranked them. We picked the top five medical entities and randomly chosen one medical entity from the last 50% of the ranks, consisting of six medical entities. Clinicians were asked to pick the correct entity set, and the accuracy of the correct choice was calculated using Equation (3), in which a_k^m represents the kth given medical entity in the mth model, i_k^m is the kth intrusive medical entity in the mth model chosen by the expert, and S is the total number of medical entities in the mth model.

$$MP_k^m=\sum_s 1(i_k^m=a_k^m)/S \tag{3}$$

The second intrusive experiment is based on the assumption that if a vector accurately captures the relations between the medical entities and patient's medical events, then a certain dimension of the vector will have a certain meaning (Murphy et al., 2012). In order to verify this assumption, we randomly chose several dimensions from the vector result of EHR2Vec, ranked their vector values in descending order and obtained the medical entities corresponding to the first k values, as indicated in Equation (4), in which i represents the i-th dimension, and rank the indices of a vector.

$$\text{argsort}\left(\text{W}\left[:,\text{i}\right]\right)\left[-\text{k}:\right] \tag{4}$$

Implementation and Training Details

EHR2Vec and Med2Vec were implemented and trained using the python TensorFlow 1.8.0 deep learning framework (Abadi et al., 2016). All models were performed on a CentOS server equipped with two 16G NVIDIA TESLA P100 graphics cards. EHR2Vec used the Adadelta optimizer to optimize the target function with a drop rate of 0.1 to achieve model convergence. EHR2Vec used eight attention heads in the self-attention mechanism, and 512 vector dimensions for each entity. To be consistent, the numbers of word vector dimensions of Med2Vec and Word2Vec were also set to 512. The Word2Vec model was implemented by python genism 3.6.0 package, with a window size of 5 and a minimum word frequency of 5. Both EHR2Vec and Med2Vec have trained 20 epochs for the best result.

EXPERIMENTAL RESULTS AND DISCUSSION

Illustration of Extracted Medical Entities

The statistics of the number of identified NERs can be found in **Supplementary Table 2**. In details, a total of 10,469 Chinese medical entities, including 1,106 diagnosis entities, 963 medication entities, 8,365 symptom entities, and 35 lab test entities extracted from 49,752 notes, have been translated into English standardized medical vocabularies for results delivery. As the data are shown in **Supplementary Table 6**, the first column are the de-identified patient IDs, the second column are the de-identified patients' visiting ids, and the rest columns are the example extracted entities.

Experimental Results Comparing Three Models at a Fixed 512 Vector Dimension

We used LN as the target word in SLE, and calculated its cosine correlation to other medical entity vectors. Cosine distance measures the cosine of the angle between two vectors projected in a high-dimensional space. It is advantageous because even if the two similar medical entities are far apart by the Euclidean disease due to the size of the terms, they may still be oriented closer together. Top 20 medical entities correlated with LN in SLE patients were calculated with medical entities from diagnosis, medication, lab test, and symptom, respectively.

Table 1 showed the comparison of the top 20 medicines associated with LN using EHR2Vec, Word2Vec, and Med2Vec. The top 20 drugs using the EHR2Vec method match the three doctors' medication preference order. For example, the top three drugs, such as hydroxychloroquine sulfate, are the most commonly used hormone prescription for LN treatment. Other drugs, such as Calcium carbonate and vitamin D3 are all auxiliary drugs for the treatment of the related target organ damage and complications, with less relevance. Similarly, **Supplementary Tables 3–5** showed the top 20 entities from diagnosis, lab test and symptom. Results from the three methods were also consistent with clinicians' experiences. For example, in the diagnosis results from EHR2Vec, SLE ranked on the top followed by hypertension, lung infection and diabetes, which match with clinician's cognitions. As for the association of pregnancy, one explanation might be that lupus could lead to abnormal birth during pregnancy (Mok et al., 2014).

However, top rankings from Med2Vec are not consistent with clinical cognitions. For example, the top three ranks in diagnosis from Med2Vec, such as chronic viral hepatitis B, typhoid and fever of unknown origin, were diagnosed poorly

TABLE 1 | Top 20 medication entities with the highest correlation to LN in the vector results obtained using four models.

Rank	Word2Vec		Med2Vec		EHR2Vec	
	Correlation	Medication	Correlation	Medication	Correlation	Medication
1	0.68	Albumen	0.31	Tabellae rhei ET natrii bicarbonatis	0.93	Hydroxychloroquine sulfate
2	0.67	Lamivudine	0.31	Ranitidine hydrochloride	0.92	Prednisone acetate
3	0.66	Felodipine	0.27	Iron sucrose	0.90	Methylprednisolone
4	0.66	Cefotaxime	0.25	Terazosin hydrochloride	0.89	Cyclophosphamide
5	0.65	Dexamethasone	0.24	Arotinolol hydrochloride	0.86	Calcium carbonate and vitamin D3
6	0.65	Metoclopramide	0.24	Enalapril maleate	0.82	Omperazole
7	0.65	Dengzhanxixin	0.24	Diammonium glycyrrhizinate	0.79	Calcitriol
8	0.65	Colquhounia root	0.24	Clopidogrel hydrogen sulfate	0.75	Alfacalcidol
9	0.64	Fasudil hydrochloride	0.23	Rabeprazole	0.72	Leflunomide
10	0.63	Salvianolate	0.23	Haloperidol	0.71	Total glucosides of paeony
11	0.63	Thiamazole	0.23	Prednisone	0.70	Aspirin
12	0.62	Cefoperazone Sodium and Tazobactam Sodium	0.23	Levothyroxine sodium	0.65	Prednisolone acetate
13	0.62	Leigongteng	0.23	Lithium carbonate	0.63	Folic acid
14	0.62	Thyroid	0.23	Urokinase	0.62	Levothyroxine sodium
15	0.62	Prednisone	0.22	Penicillins	0.61	Warfarin Sodium
16	0.62	Fluvoxamine maleate	0.22	Carvedilol	0.60	Mycophenolate mofetil
17	0.62	Sodium valproate	0.22	Mecobalamin	0.60	Pantoprazole
18	0.61	Salvianolate	0.21	Furosemide and spironolactone	0.60	Valsartan
19	0.61	Tacrolimus	0.21	Deslanoside	0.58	Spironolactone
20	0.61	Sanqi Panax Notoginseng	0.21	Cefradine	0.57	Low Molecular Weight Heparin Calcium

related to LN. The top two results of Word2Vec were similar to Med2Vec, but not for the rest. For example, hyperlipemia, hypothyroidism, and fatty liver were considered poorly related to LN. **Figure 3** showed a summary comparison of the three models for the four categories. Particularly, EHR2Vec showed over 40% improvement in detecting LN relevant medications in the medicine category.

Interpretability Evaluation of EHR2Vec Representations at Three Random Dimensions

The results in **Table 2** showed the top 10 medical entities at three arbitrarily selected vector dimensions: 180, 274, and 480. Identified entities in the dimension of 180 represented a class of women with a history of pregnancy who suffered from LN, tested positive on the anti-nuclear antibody (ANA)-B and abnormalities on Complement 3-B and Complement 4-B, and used hormone drugs, such as prednisone acetate and methylprednisolone sodium succinate, as well as hydroxychloroquine sulfate. Correlations between these diagnoses, lab tests, and medications are highly consistent with clinical observations. Results from dimension 274 indicated patients who suffered from LN, tested positive on the anti-nuclear antibody (ANA)-B lab test and took commonly used drugs for patients with SLE, such as prednisone acetate and prednisolone sodium succinate while using calcium-supplementary drugs, such as Calcium carbonate and vitamin D3 and gastrointestinal agents, such as omeprazole and aspirin. Dimension 480 represents a number of highly relevant symptoms that often manifest in patients with SLE, including rash, telangiectasia, muscle pain, etc.

To show the robustness of our interpretable model, we performed another independent experiment by arbitrarily selecting the extra three sets of dimensions: 360, 440, 457. We can see the results in **Supplementary Table 7**, dimension 360 is more related to manifestations of the mucous membrane of the skin, such as subcutaneous bleeding, facial rash, palm erythema, and oral ulcer; dimension 457 is more related to appearance symptoms, such as phenotypes in face, skin, ulcers, etc.; dimension 440 is more associated with vasculitis, such as hypertension, edema, rash, and erythema. All the above evidence indicated the interpretability of our model by showing that the top medical entities in each dimension are highly relevant.

SLE Disease Comorbidity Prevalence and Progression Over Time

Figure 4 showed a summary of the prevalence of SLE comorbidity changes over time diagnosed by clinicians. The comorbidities of SLE diagnosed by clinicians are correlated with medical entities, such as symptoms and diagnosis ranked by EHR2Vec. For example, skin mucous membrane lesions (SMML) is the most prevalent comorbidity than any others. The number

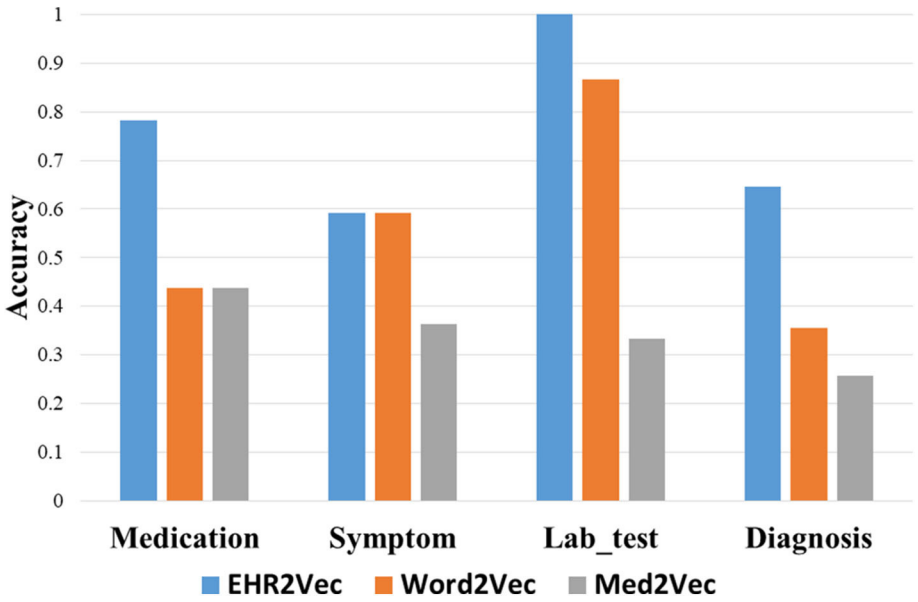


FIGURE 3 | Performance comparison by Intrusion analysis. We perform intrusion analysis to evaluate model performance by comparing clinicians' opinion with our identified medical concepts from four groups. The EHR2Vec shows higher accuracy than the other two models, Word2vec and Med2vec.

TABLE 2 | Top 10 medical entities in terms of vector value rank in three different dimensions.

Dimension 180	Dimension 274	Dimension 480
[Lab Test] Complement 3-B	[Medication] Omperazole	[Symptom] Widespread facial red rash
[Diagnosis] Pregnancy	[Lab Test] Urine protein qualitative test-U	[Symptom]Migratory double joint and shoulder pain
[Lab Test] Complement 4-B	[Medication] Calcium carbonate and vitamin D3	[Symptom] Systemic diffusive and red rash
[Drug] Calcium carbonate and vitamin D3	[Medication]Aspirin	[Symptom]Slightly swollen left-hand fingers
[Medication] Methylprednisolone	[Symptom] Cough	[Symptom]Scattered bleeding points on hands
[Medication] Methylprednisolone sodium succinate	[Medication] Methylprednisolone sodium succinate	[Symptom] Facial rash relief
[Lab Test]Anti-nuclear antibody (ANA)-B	[Medication]Prednisone acetate	[Symptom]Capillary and facial capillary expansion
[Medication] Prednisone acetate	[Lab Test]Anti-nuclear antibody (ANA)-B	[Symptom] Muscle and body tenderness
[Diagnosis] LN	[Diagnosis] LN	[Symptom] Scattered red rash
[Medication] Hydroxychloroquine	[Medication] Hydroxychloroquine	[Symptom] Left chest pain

of this group of patients is gradually progression year by year. Some skin related symptoms, such as rash are also ranked by top candidate entities related to LN from EHR2Vec analysis, which has a high agreement with clinician's diagnosis. While some of the comorbidities, such as blood system, are more complex in disease progression. For example, no or little progression in the first 5 years, but dramatically increased after 5 years. These types of comorbidities' symptoms may have challenges to be detected by EHR2Vec, which requires further validation by clinicians. Nevertheless, existing evidence showed that EHR2Vec is able to rank the most relevant disease-related phenotypic information from raw EHR data automatically through key word query.

DISCUSSION

EHR data with sequential visiting records provides opportunities for disease early detection. Automatically extracting medical concepts from EHR data and converting them into embedding vectors can contribute significantly to monitor the conditions change for chronic disease. In this project, we developed a new embedding method, EHR2Vec, to learn representative medical concepts from EHR data. EHR2Vec incorporated self-attention algorithm with multi-layer deep learning optimizations at both medical codes and visits level. EHR2Vec overcomes existing embedding methods that ignored contextual information at visit level or missed multi-visit information to capture temporal patterns from clinical notes. In the experiments of SLE data, one of a chronic disease, EHR2Vec has displayed its significant improvement in key medical applications, while providing clinically meaningful interpretations. Using EHR2Vec to learn representations will improve the accuracy for detecting disease prevalence and progression in precision medicine.

Traditional NLP word embedding approaches, such as Word2Vec published in 2013, are unable to capture long-distant dependency relations for words in a sentence, here medical

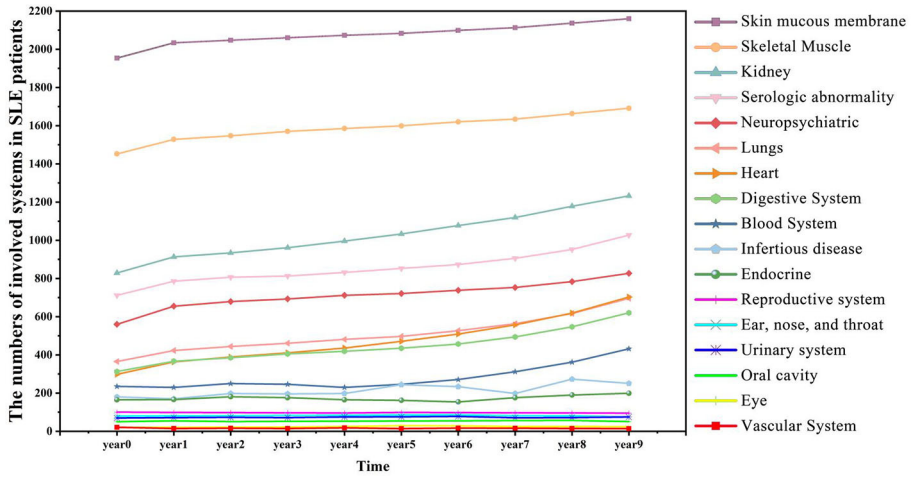


FIGURE 4 | SLE affects more patients' body organs and systems over time. SLE is a chronic disease with many comorbid conditions. This figure shows that more SLE patients are affected by comorbidities as the time accumulated. For example, in the initial year (Year 0), 829 patients manifested kidney diseases, and in Year 9, the number increased to 1,232.

entities in a visiting event. Later, improved architectures RNN-based Bi-LSTM method, such as ElMo (Peters et al., 2018), were developed to capture contextualized information in time order, however its computational efficiency is less optimal in identifying long-range dependency since it has to remember all the state of words sequence by sequence and disallowance of the parallel computing in a long sequence. Recently, self-attention based algorithms overcome those limitations and has become the leading architecture in NLP tasks since 2018. It is the key component utilized in many state-of-art transformer learning methods, such as BERT (Devlin et al., 2018), GTP-2 (Radford et al., 2019), and XLnet (Yang et al., 2019), allowing to identify long-range relationships that are far away through parallel computation. By applying self-attention structure in real-world EHR data, our EHR2Vec optimized both sequential information from different visits and different types of medical entity relations within each visit, which greatly improves traditional methods in representation learning for relevant comorbidity detection in chronic disease.

EHR2Vec showed great performance in comparison with the other two most popular representation learning methods in EHR data. We queried a medical term, LN, which is a common comorbidity in SLE patients. The experimental results of EHR2Vec from medication category indicated that the medication preference has the highest correlation with LN. While the logical sequence was chaotic and the frequencies of the features were small with no representativeness in Word2Vec or Med2Vec method. In the Word2Vec method, the vector of the center word is obtained by simply summing the vector values of the context words, so the generated vector has an indistinctive boundary, causing heterogeneous medical concept ambiguities. The Med2Vec model also has a fixed window size, leading to poor word vector discriminations. EHR2Vec solves the problem by self-attention structure to capture global information, thus the concept representation vector calculated is more accurate while allowing further optimization on the medical event sequence of each patient, thereby leading to higher discrimination of the final medical entity vector.

We also recognize that there are some areas that can be further improved in EHR2Vec in the future. First, although we optimized the visiting events at the time orders for each patient's EHR data, some of the time sequence information within each visit might be poorly extracted and captured. Second, a more comprehensive propagation networking algorithm could be combined with self-attention structure to quickly target information of importance for each event. Nevertheless, these tasks are the most challenging part for the clinical NLP in every EHR data and more accurate models can be developed according to specific scenarios in the real-world situation. Finally, we only evaluated a large-scale SLE dataset in the current study, but we plan to expand to other disease areas where clinical symptoms and comorbidity change over time, such as autism spectrum disorders, to assess the generalizability of the proposed approach in the future.

CONCLUSION

In this study, we proposed the EHR2Vec model, a new deep learning model that generates medical entity vectors based on the attention mechanism. Compared with other widely used word vector generation models (e.g., Word2Vec and Med2Vec), EHR2Vec can correct target medical entity more accurately using self-attention structure to capture relations from surrounding medical entities. We compared and tested the performance of EHR2Vec through clinical expert assessments and an intrusion experiment using the SLE dataset, an actual clinical disease dataset, and found that EHR2Vec could generate more accurate vectors. In the future, we will integrate more medical knowledge (e.g., doctors' prior knowledge and patient image data) into the EHR2Vec model and apply the resulting vector

to more scenarios, e.g., SLE complication and hospitalization length predictions.

DATA AVAILABILITY STATEMENT

EHR2Vec code can be found: https://github.com/jingsongs/EHR2Vec.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Committee on Ethics of Biomedicine, Second Military Medical University. The Ethics Committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

LW and QW contributed on the project design and code implementation. HB, WL, and YZha contributed on the EHR database management and data curation. LJ and HX performed the statistic analysis. YZho and KW contributed on the project design and supervised the project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Ministry of Science and Technology Key Research and Development Program of China (No. 2018YFC0116902) and National Science Foundation of China (No. 81873915).

ACKNOWLEDGMENTS

We thank the reviewers whose comments and suggestions helped improve this manuscript. We thank Hitales (Shanghai, China) for algorithm and software engineering support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00630/full#supplementary-material

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (Savannah, GA), 265–283.

Almaani, S., Meara, A., and Rovin, B. H. (2017). Update on lupus nephritis. *Clin. J. Am. Soc. Nephrol.* 12, 825–835. doi: 10.2215/CJN.05780616

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). "Reading tea leaves: how humans interpret topic models," in *Advances in Neural Information Processing Systems*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Vancouver, CA: Neural Information Processing Systems), 288–296.

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., et al. (2016). "Multi-layer representation learning for medical concepts," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA: ACM), 1495–1504. doi: 10.1145/2939672.2939823

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* 1810.04805.

Klein, D., and Manning, C. D. (2003). "A parsing: fast exact Viterbi parse selection," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (Stroudsburg, PA: Association for Computational Linguistics), 40–47. doi: 10.3115/1073445.1073461

Luo, H., Liu, Z., Luan, H., and Sun, M. (2015). "Online learning of interpretable word embeddings," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon), 1687–1692. doi: 10.18653/v1/D15-1196

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Lake Tahoe, NV: Neural Information Processing Systems), 3111–3119.

Mok, C. C., Yap, D. Y., Navarra, S. V., Liu, Z., Zhao, M., Lu, L., et al. (2014). Overview of lupus nephritis management guidelines and perspective from Asia. *Nephrology* 19, 11–20. doi: 10.1111/nep.12136

Murphy, B., Talukdar, P., and Mitchell, T. (2012). "Learning effective and interpretable semantic models using non-negative sparse embedding," in *Proceedings of COLING 2012* (Mumbai), 1933–1950.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. *arXiv* 1802.05365. doi: 10.18653/v1/N18-1202

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). *Language Models Are Unsupervised Multitask Learners*. San Francisco, CA: OpenAI Blog 1.

Schmid, H. (2004). "Efficient parsing of highly ambiguous context-free grammars with bit vectors," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics* (Geneva), 162–168. doi: 10.3115/1220355.1220379

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA: Neural Information Processing Systems), 5998–6008.

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., et al. (2020). Deep learning in clinical natural language processing: a methodical review. *J. Am. Med. Inform. Assoc.* 27, 457–470. doi: 10.1093/jamia/ocz200

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: generalized autoregressive pretraining for language understanding. *arXiv* 1906.08237.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 30 June 2020
doi: 10.3389/fbioe.2020.00607

High-Throughput Screen for Cell Wall Synthesis Network Module in *Mycobacterium tuberculosis* Based on Integrated Bioinformatics Strategy

Xizi Luo[1†], Jiahui Pan[1†], Qingyu Meng[1], Juanjuan Huang[1], Wenfang Wang[1], Nan Zhang[2] and Guoqing Wang[1]*

[1] Department of Pathogenobiology, The Key Laboratory of Zoonosis, Chinese Ministry of Education, College of Basic Medical Sciences, Jilin University, Changchun, China, [2] College of Mathematics, Jilin University, Changchun, China

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Shili Liu,
Shandong University, China
Yanna Shen,
Tianjin Medical University, China

***Correspondence:**
Guoqing Wang
qing@jlu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 06 April 2020
Accepted: 18 May 2020
Published: 30 June 2020

Citation:
Luo X, Pan J, Meng Q, Huang J, Wang W, Zhang N and Wang G (2020) High-Throughput Screen for Cell Wall Synthesis Network Module in Mycobacterium tuberculosis Based on Integrated Bioinformatics Strategy. Front. Bioeng. Biotechnol. 8:607. doi: 10.3389/fbioe.2020.00607

Background: *Mycobacterium tuberculosis* is one of the deadliest pathogens in humans. Co-infection of *M. tuberculosis* with HIV and the emergence of multi-drug-resistant tuberculosis (TB) constitute a serious global threat. However, no effective anti-TB drugs are available, with the exception of first-line drugs such as isoniazid. The cell wall of *M. tuberculosis*, which is primarily responsible for the lack of effective anti-TB drugs and the escape of the bacteria from host immunity, is an important drug target. The core components of the cell wall of *M. tuberculosis* are peptidoglycan, arabinogalactan, and mycotic acid. However, the functional genome and metabolic regulation pathways for the *M. tuberculosis* cell wall are still unknown. In this study, we used the biclustering algorithm integrated into cMonkey, sequence alignment, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and other bioinformatics methods to scan the whole genome of *M. tuberculosis* as well as to identify and statistically analyze the genes related to the synthesis of the *M. tuberculosis* cell wall.

Method: We performed high-throughput genome-wide screening for *M. tuberculosis* using Biocarta, KEGG, National Cancer Institute Pathway Interaction Database (NCI-PID), HumanCyc, and Reactome. We then used the Database of Origin and Registration (DOOR) established in our laboratory to classify the collection of operons for *M. tuberculosis* cell wall synthetic genes. We used the cMonkey double clustering algorithm to perform clustering analysis on the gene expression profile of *M. tuberculosis* for cell wall synthesis. Finally, we visualized the results using Cytoscape.

Result and Conclusion: Through bioinformatics and statistical analyses, we identified 893 *M. tuberculosis* H37Rv cell wall synthesis genes, distributed in 20 pathways, involved in 46 different functions related to cell wall synthesis, and clustered in 386 modules. We identified important pivotal genes and proteins in the cell wall synthesis

pathway such as *murA,* a class of operons containing genes involved in cell wall synthesis such as ID6951, and a class of operons indispensable for the survival of the bacteria. In addition, we found 41 co-regulatory modules for cell wall synthesis and five co-expression networks of molecular complexes involved in peptidoglycan biosynthesis, membrane transporter synthesis, and other cell wall processes.

Keywords: *Mycobacterium tuberculosis*, cell wall, module, regulatory networks, enrichment analysis

INTRODUCTION

Mycobacterium tuberculosis is considered one of the world's most successful pathogens. The disease caused by it has been a major global health challenge (Sher et al., 2020). Since the 1950s, the discovery of first-line anti-tuberculosis (TB) drugs such as isoniazid, rifampicin, and ethambutol has effectively improved the cure rate and survival rate of TB patients. However, the emergence of multiple forms of drug-resistant strains, including a single isoniazid-resistant strain, a multi-drug-resistant strain, and a widely drug-resistant strain, has again made *M. tuberculosis* one of the leading causes of death worldwide, with a mortality of 1.5 million people in 2018 (Merker et al., 2020). Co-infection of HIV and *M. tuberculosis* increases the burden of curing TB; therefore, the development of new and effective anti-TB drugs is critical (Turner et al., 2020).

The cell wall structure of *M. tuberculosis* is unique and is extremely important for the invasion, survival, and reproduction of the bacterium in a host. The main reason for the difficulty in developing drugs for *M. tuberculosis* is that the bacterium has a hard cell wall and very low permeability. The development of *M. tuberculosis* resistance is also associated with the cell wall. Howard et al. (2018) found that *M. tuberculosis* carrying a rifampicin-resistance mutation reprograms macrophage metabolism through cell wall lipid changes. Maitra et al. (2019) described *M. tuberculosis* cell wall peptidoglycan as its fatal weakness. Thus, the cell wall of *M. tuberculosis* is an important target for the development of new anti-TB drugs.

In this study, we performed high-throughput screening of *M. tuberculosis* cell wall synthesis genes and screened key genes using bioinformatics and statistical methods to obtain new key targets for the development of anti-TB drugs.

MATERIALS AND METHODS

Synthetic Gene Data for *M. tuberculosis* H37Rv Cell Wall

The relevant data for *M. tuberculosis* cell wall synthesis genes used in this study were obtained from the screening and integration of the following databases: TubercuList (Lew et al., 2011), TBDB (Galagan et al., 2010), PATRIC (Gillespie et al., 2011), MycoDB (Chaudhuri, 2009), GenoMycDB (Catanho et al., 2006), MyBASE (Zhu et al., 2009), MabsBase (Heydari et al., 2013), and MGDD (Vishnoi et al., 2008).

Sequence Alignment

We used online software[1] to compare the amino acid sequence of *M. tuberculosis* H37Rv with the amino acid sequence of *Mycobacterium smegmatis*, *Mycobacterium leprae*, *Mycobacterium bovis*, and *M. tuberculosis* H37Ra. Genes with homology greater than 60% were selected (Kuroda et al., 2001; Hellweger et al., 2014).

Screening Essential Genes

The whole genome information for *M. tuberculosis* H37Rv was obtained from the National Center for Biotechnology Information (NCBI) and annotated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database with KEGG Orthology (KO) in accordance with the "binary relationships" provided by the KEGG Brite database. The types and functions of cell wall synthesis genes were determined using Clusters of Orthologous Groups with KO (KO COG) and the P-Score and E-Score for each KO were calculated. The E-Score was calculated with KO using the same path annotation and the P-Score was determined from the e-score. The P-Score-KEGG and P-Score-COG were also calculated based on the KEGG and COG annotations (Kong et al., 2019). These two values were in the range of 0 to 1, with 0 indicating a lack of necessity and 1 indicating necessity.

Screening Operon Set

We applied the operon Database of Origin and Registration (DOOR) (Cao et al., 2019) established in our laboratory to classify the operon collection of cell wall genes. The DOOR database uses two prediction procedures. For operon genomes with a large number of experimental verifications, we used a non-linear classifier to train the known operon subsets based on the general characteristics of the genome and the characteristics of specific genomes. For genomes without experimental data, we used linear classification to predict operons for the general characteristics of the genome.

Screening Co-regulatory Gene Modules

We selected all *M. tuberculosis* H37Rv gene chips in NCBI after filtering out irrelevant chip data and performed min-max normalization on each chip. We used the cMonkey double clustering algorithm to establish seed clusters (Waltman et al., 2010). We calculated the *P*-values of three such model components based on the amount of co-expressed genes, upstream sequences, and association networks. We optimized

[1] www.ncbi.nlm.nih.gov/blast/

seed clusters by adding or removing related genes and proceeded to build new clusters. We used the Monte Carlo procedure to calculate the probability of each gene or condition sampled as a dual cluster gene with the conditional probability at each stage. Through these procedures, the genomic co-regulation network was identified.

Functional Enrichment Analysis

We performed a Gene Ontology (GO) analysis of the target genes using the comprehensive database Davide[2] for enrichment analysis, annotation, and visualization. We used the Biocarta, KEGG, National Cancer Institute Pathway Interaction Database (NCI-PID), HumanCyc, and Reactome pathway databases for pathway enrichment of the target genes. $P < 0.05$ was considered statistically significant when the threshold was $\geq$ two genes. We used R software and the Perl language to visualize the enrichment results. We also installed "Rcpp," "ggplot2," and other related software packages (Postma and Goedhart, 2019).

Construction of Gene Regulatory Network

The protein–protein interaction (PPI) network was constructed using a gene interaction search tool database (STRING) and Cytoscape 3.6.1 was used for visualization. The Minimal Common Oncology Data Elements (MCODE), a Cytoscape network analysis plug-in for molecular complex detection, was used to deeply mine the existing modules in the network structure to find the core gene clustering modules with the highest levels of interaction.

RESULTS

Statistical Analysis of Cell Wall-Related Genes in Mycobacteria

Through database annotation and sequence alignment, we screened the cell wall synthesis genes for mycobacteria. As shown in **Table 1**, there were 892 cell wall synthesis genes for *M. tuberculosis* H37Rv, 888 for *M. tuberculosis* H37Ra, 780 for *M. bovis*, 508 for *M. smegmatis*, and 454 for *M. leprae*.

We used the operon database DOOR to assess the module distribution of cell wall synthesis genes. In *M. tuberculosis* H37Rv, 893 genes related to cell wall synthesis were located in 684 operons and 37 operons contained three or more cell wall-related genes. Multiple genes located in an operon are usually regulated by the same control region and constitute a transcription unit. The 149 genes contained in these 37 operons may be key genes that play an important role in the synthesis of the *M. tuberculosis* cell wall. There are four sets of operons, which contain more than seven genes related to the cell wall, including operons with ID numbers 7375, 7760, 6927, and 7590 displayed in the DOOR database. The ID number of the operon with the largest number of genes is 7558, up to 9. The genes *yrbE1A* and *yrbE1B* encode cell wall membrane proteins (Pasricha et al., 2011). The proteins encoded by *mce3A* and *mce3B* are not only present in the cell wall, but are also important for the virulence of *M. tuberculosis* during host invasion (Ahmad et al., 2005); 37 pairs of operons in this pathway and their details are shown in **Supplementary Table S1**.

The main cause of infection of the host with *M. tuberculosis* is the virulence factor. We obtained all coding genes related to virulence of TB from the VFDB database, of which 115 genes are cell wall synthesis genes. The cell wall genes that belong to virulence included the *mmpl* family which encoded cell wall lipid transporters, the cell wall mycolic acid synthase *mmA4*, and *Rv2224c* with little research and unknown specific function. Genes related to cell walls and virulence factors are shown in **Supplementary Table S2**.

Function Analysis of Cell Wall-Related Genes

Essential genes are often critical for sustaining the activities of living organisms. As shown in **Table 1**, there are 236 essential genes related to cell wall synthesis in the whole genome of *M. tuberculosis* H37Rv. These genes are located in 161 operons, among which there are 10 operons containing more than three essential genes and five operons with more than four essential genes. Three or more operons have five or more essential genes. The six genes controlled by the operon ID6951 are all required genes that play key roles in cell wall synthesis. The six required genes include *eccA3-E3*, a member of the ESAT6 secretory system (ESX), and membrane-anchored mycosin *mycP3*. ESX secretion systems mediate various functions, participate in the metabolism of zinc and iron, and play an important role in cell wall integrity (Gaur et al., 2017).

We used KEGG, BioCyc, and Reactome pathway data to analyze cell wall synthetic genes (**Figures 1A,B**). The 892 cell wall synthesis genes in the H37Rv strain were distributed in 39 signaling pathways. The essential gene *murA* participates in the metabolic pathway (KEGG mtu01100), peptidoglycan biosynthesis (KEGG mtu00550), and UDP-*N*-acetylmuramoyl-pentapeptide biosynthesis I (BioCyc pwy6387). *AftB* is involved in the super pathways of mycolyl-arabinogalactan-peptidoglycan complex biosynthesis (BioCyc pwy6404) and Lipoarabinomannan biosynthesis (KEGG mtu00571). *MurA* and *aftB* are thought to be key node genes in the cell wall biosynthesis pathway. In addition, we identified some genes whose functions are currently unknown but which are located in important pathways such as the

[2]https://david.ncifcrf.gov/

TABLE 1 | Cell wall synthesis network module in mycobacteria.

Strain	Cell wall-related genes	Essential genes in cell wall	Operon	Pathway
H37Rv	892	236	684	20
H37Ra	888	323	689	15
M. leprae	454	149	455	6
M. bovis	780	160	636	7
M. smegmatis	508	92	394	11

mycolyl-arabinogalactan-peptidoglycan complex biosynthesis (BioCyc PWY-6397) pathway.

We annotated the gene functions using GO and identified 46 GO items in the 892 cell wall synthesis genes. As shown in **Figure 2**, there were 19 items related to biological processes (BPs), nine items related to cell components (CCs), and 18 items related to molecular function (MF). The most significant BP terms were related to cell wall organization (GO:0071555), regulation of cell shape (GO:0008360), and peptidoglycan biosynthetic process (GO:0009252), as shown in **Figure 2A**.

We also visualized and clustered the enriched GO and KEGG terms using the cluego in Cytoscape (**Figure 3**). We found

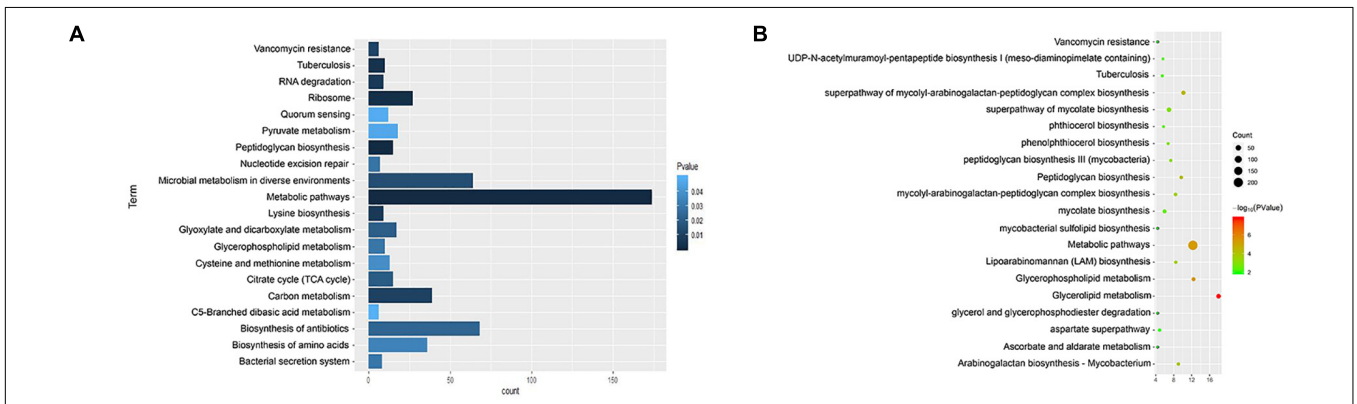


FIGURE 1 | Enrichment of cell wall-related gene pathways in the standard strain of *M. tuberculosis* H37Rv. **(A)** The top 20 pathways with the lowest *p*-values (<0.05) for KEGG were selected and a histogram was created. **(B)** The top 20 pathways with the lowest *p*-values (<0.05) for Biocarta, KEGG, NCI-PID, HumanCyc, and Reactome are shown in a bubble chart.

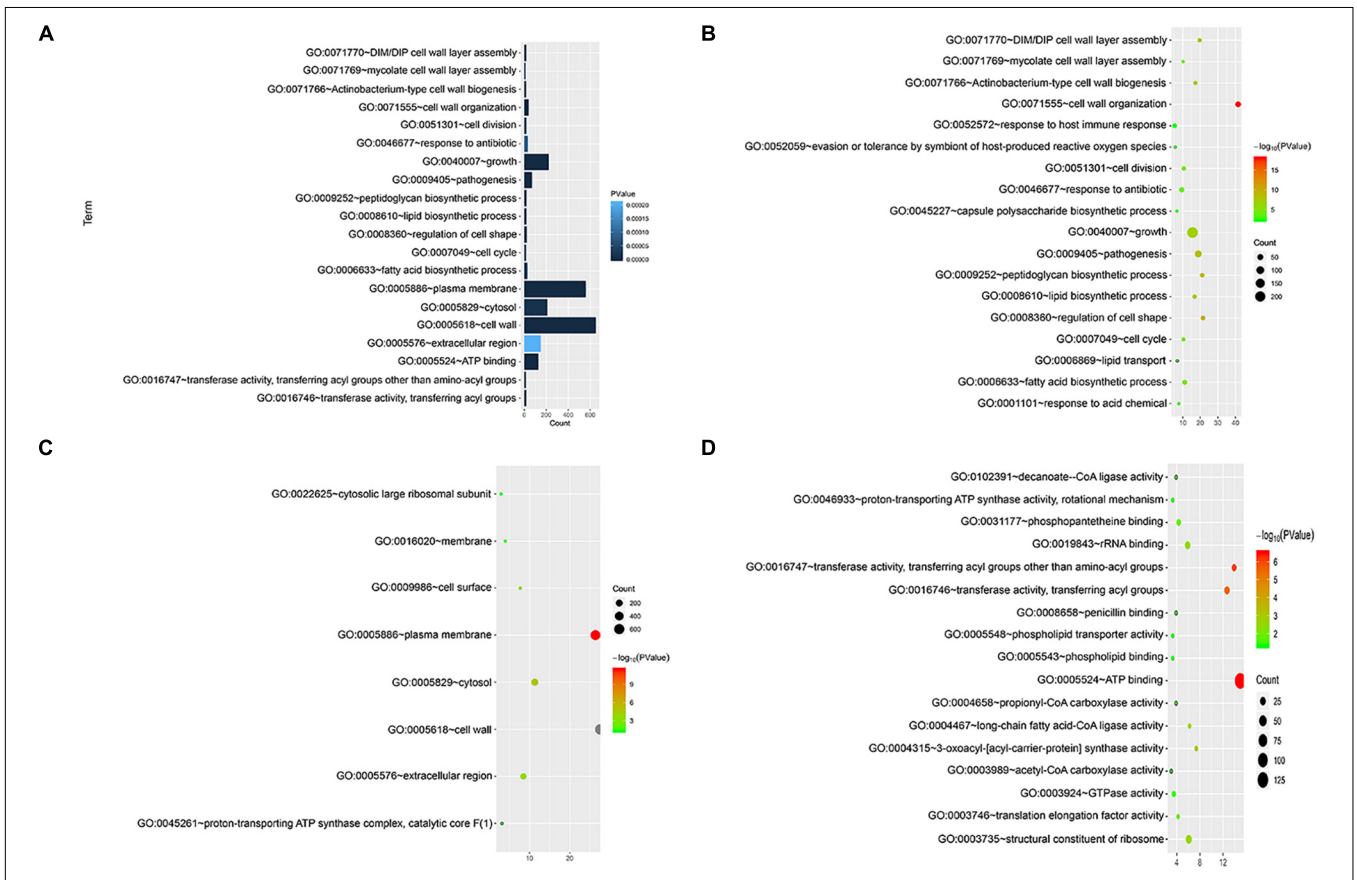


FIGURE 2 | Functional significance analysis of cell wall-related genes in the standard strain of *M. tuberculosis* H37Rv. **(A)** Histogram showing the top 20 pathways identified in function analysis of cell wall-related genes. **(B–D)** Bubble charts showing the top 20 pathways in BP, CC, and MF.

that most genes are enriched in important cell wall-related pathways, such as lipid biosynthetic process, peptidoglycan-base cell wall synthesis, lipid synthesis, and 3-oxoacyl-acyl-carrier-protein sythase activity. In addition, it is closely related to the pathogenicity of the host, symbiosis of the host, secretion, and pathogenesis.

Analysis of the *M. tuberculosis* Cell Wall-Related Modules

By screening the *M. tuberculosis* cell wall modules using gene chips and the cMonkey double clustering algorithm, we found that the total number of *M. tuberculosis* modules was 600, among which 386 contained the target genes for cell wall synthesis.

Among the modules containing the target genes, 41 modules contained more than four target genes. Among these 41 modules, 16 were related to the synthesis of sugar in the cell wall, such as bicluster_0098 for the mannosyl transfer process and bicluster_0329 for the peptidoglycosyl transfer process. Fifteen modules were related to the synthesis of lipids. The modules bicluster_0068 and bicluster_0012 were related to the synthesis of mycobacterial acid (Saelens et al., 2018). There were 10 modules related to cell wall surface proteins and virulence. Among them, bicluster_0384 contained the largest number of target genes for cell wall synthesis in a single module. The nine genes contained in this module are all involved in the biosynthetic process for arabinose. For example, the *Rv0129c* coding protein plays a role in the addition of mycosyl residues in the cell wall arabinose (Jiang et al., 2020) and *Rv3806c* plays a role in the synthesis of decenyl phosphate D-arabinose (Safi et al., 2013).

In the process of gene transcription, transcription factors complete the binding of proteins to DNA by identifying specific sequences of the double helix structure (motif). The motif is short and conservative, consisting of about 20 base pairs. Many key regulatory pathways in the cell are usually recruited by a motif (Ivarsson and Jemth, 2019). Genes located in a

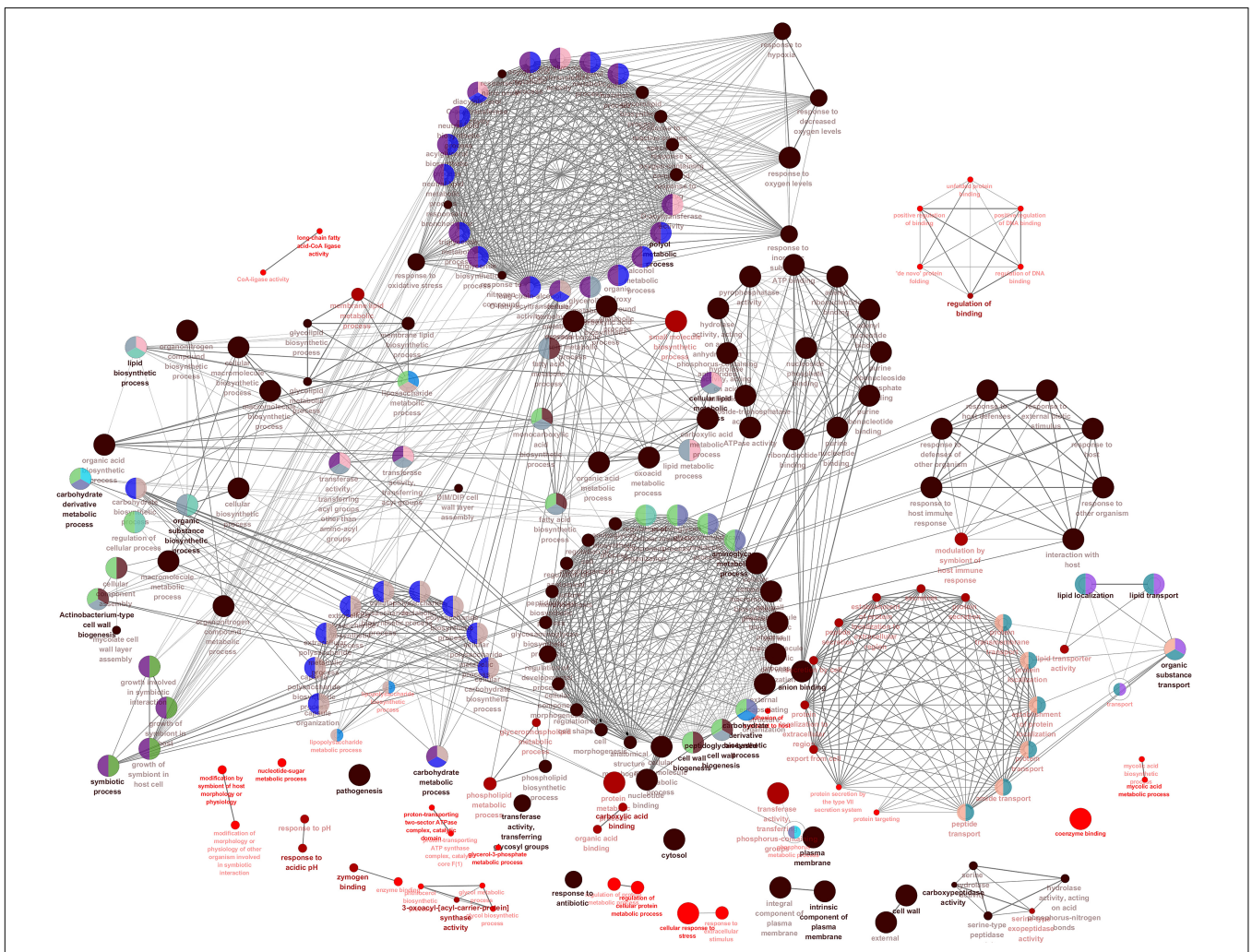


FIGURE 3 | The enrichment map of GO annotation and KEGG pathway. Node size represents the number of cell wall genes expressed in specific terms. The edge thickness represents the number of genes shared by the two items connected by the edge.

module are regulated by a transcription factor and have the same motif. We mapped the motif base distribution for the four modules with the largest number of cell wall genes, as shown in **Figures 4A–D**.

Establishment of PPI Network and Screening of Key Genes

We enriched the function of cell wall synthesis gene and constructed the network between cell wall synthesis gene and gene function. As shown in **Figure 5**, the cell wall genes screened are mainly related to 18 functions, including fatty acid biosynthesis process, DIM cell wall layer assembly, and plasma membrane.

Using the STRING database, we analyzed the interaction relationships between the cell wall synthesis genes of *M. tuberculosis* and constructed a PPI network of cell wall-related genes after deleting unconnected nodes. As shown in **Figure 6A**, in order to identify the key genes in the network diagram, we used MCODE to screen out five important subnets and several related genes under the condition of k-score = 2.

As shown in **Figure 6B**, Subnet 1 contains 14 key genes in the cell wall peptidoglycan synthesis process. Alr, the *ddla* coding protein, plays a role in the synthesis of alanine peptidoglycan (Bhat et al., 2017; Meng et al., 2019). *Ftsw*, *ftsz*, and *pbp3*-encoded proteins can form a ternary complex to potentially regulate peptidoglycan biogenesis. Roda glycosyltransferase is also involved in peptidoglycan synthesis (Wu et al., 2016). **Figure 6C** shows that Subnet 2 contains 13 ESX-1 secretory system-related genes. The ESX-1 secretory system is not only an important determinant of *M. tuberculosis* virulence, but is also closely related to cell wall synthesis (Wong, 2017). After elimination of the *espa* gene encoding the ESX-1 substrate, *M. tuberculosis* bacteria lose the ability to synthesize a complete cell wall structure (Chen et al., 2013). ESX-A is an early secreted antigen target that promotes the synthesis of the ESX-1 substrate and interacts with the cell membrane and cell wall of bacteria. Subnet 3 (**Figure 6D**) contains membrane lipid transporters. In Subnet 4 (**Figure 6E**), *ddrA-C* is not only the key gene in cell wall synthesis, but also the key gene for drug resistance in *M. tuberculosis* bacteria (Selvam et al., 2013). The other eight genes are related to the synthesis of lipid phthiocerol dimycocerosates (PDIM) in the cell wall. Among them, *ppsA-E* encodes the PDIM catecholic dipolyoleate (Gopal et al., 2016). All seven genes in Subnet 5 (**Figure 6F**) are regulated by the mymA operon and play a role in cell wall fatty acid modification (Singh et al., 2005).

Through PPI, we identified some node genes that are crucial in cell wall biosynthesis of important sugars and lipids. In **Figure 7**, the petal diagram shows the genes contained in the top five annotations in BP, MF, and CC and the genes contained in the first five paths in KEGG BioCyc and Reactome. We selected the

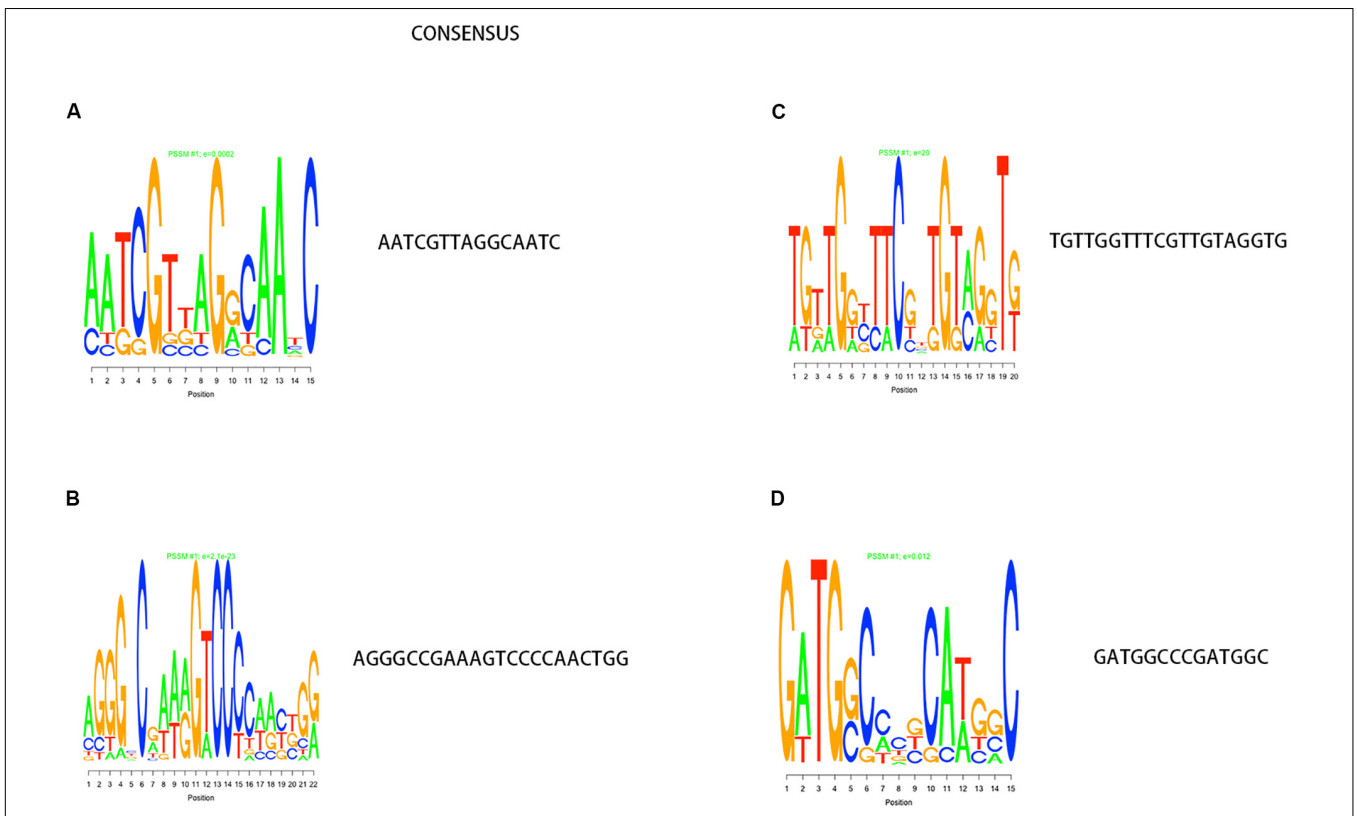


FIGURE 4 | Motif analysis results with the largest number of cell wall genes in a single module cluster. **(A–D)** The motif base distribution of the four modules with the largest number of cell wall genes. In each BP base distribution in the motif, the size of the base is proportional to the corresponding frequency.

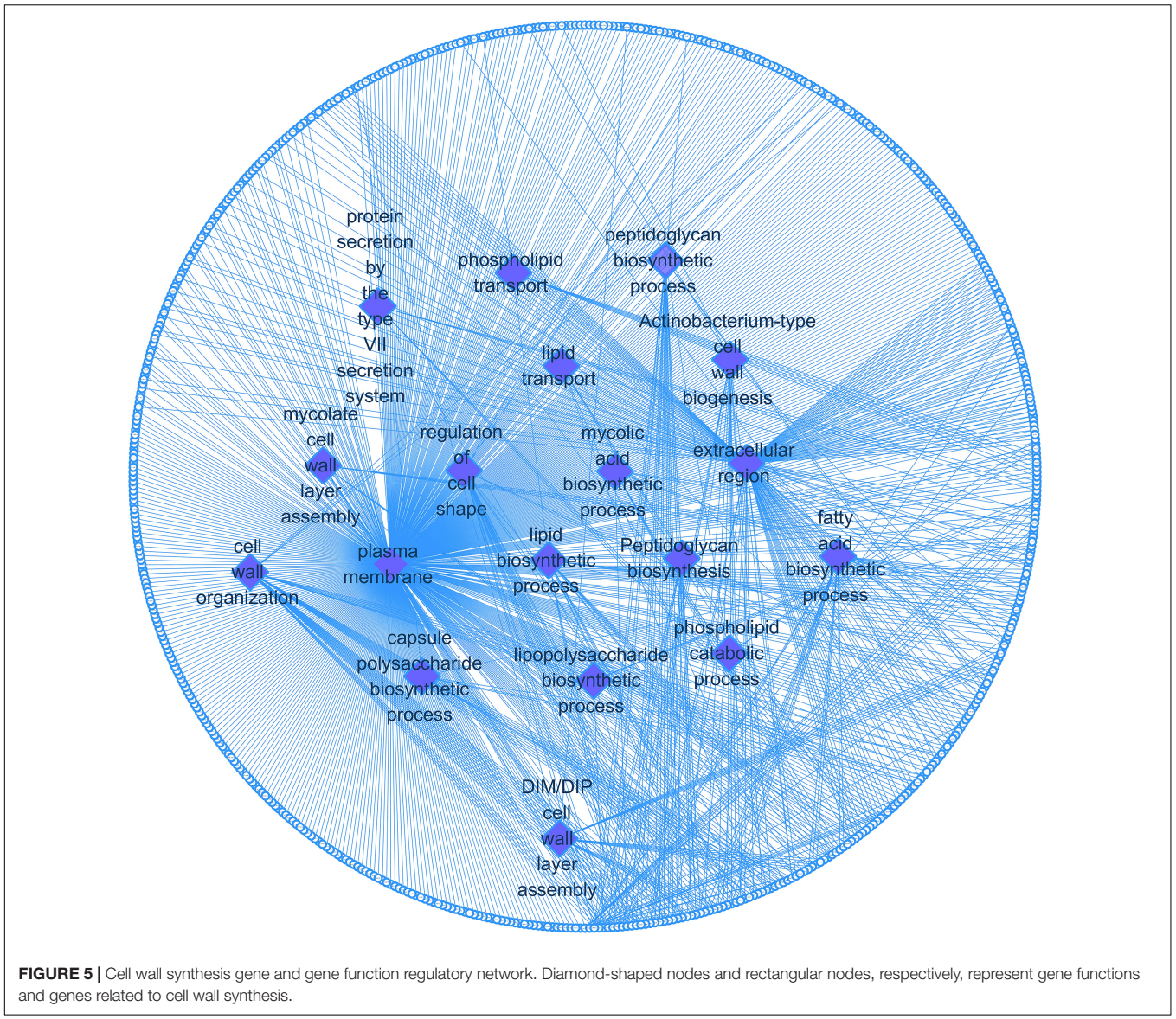


FIGURE 5 | Cell wall synthesis gene and gene function regulatory network. Diamond-shaped nodes and rectangular nodes, respectively, represent gene functions and genes related to cell wall synthesis.

top five groups with the lowest *P*-values in all enrichments. This is to intuitively demonstrate the functional enrichment pathway for the co-regulation of key genes. The key genes for this pathway are shown in **Supplementary Table S3**.

DISCUSSION

As an important target for the development of new anti-TB drugs, the *M. tuberculosis* cell wall has attracted increasing attention. Maan and Kaur (2019) discovered *Rv2223c* in the cell wall of *M. tuberculosis*, which is a carboxyl transferase. Bothra et al. knocked out *mmpl11* and the resulting mutant strain exhibited a change in the biological activity related to mycolate wax and long-chain triacylglycerol. The knockout strain was also damaged compared to the wild strain *in vitro* granuloma model, thus demonstrating the important role of *mmpl11* in cell wall and biofilm syntheses (Bothra et al., 2018). Quigley et al. (2017) found that the expression of lipid PDIM in the cell wall of *M. tuberculosis* was negatively regulated by a novel transcription repressor, *Rv3167c*. Although extensive *M. tuberculosis* cell wall-related research has been conducted, there is still no comprehensive summary of the key genes involved in the process of cell wall synthesis.

In this study, we first screened the genes related to cell wall anabolism using multiple *M. tuberculosis* gene annotation databases. Next, we screened the essential genes for cell wall synthesis by GO functional annotation. We then evaluated the distribution of cell wall synthesis genes in the whole genome using the DOOR database established in our laboratory. Using the above methods, we obtained a lot of valuable information. For example, we identified the entire operon containing genes involved in cell wall synthesis, which is necessary for the survival of the bacterium. We employed module analysis and

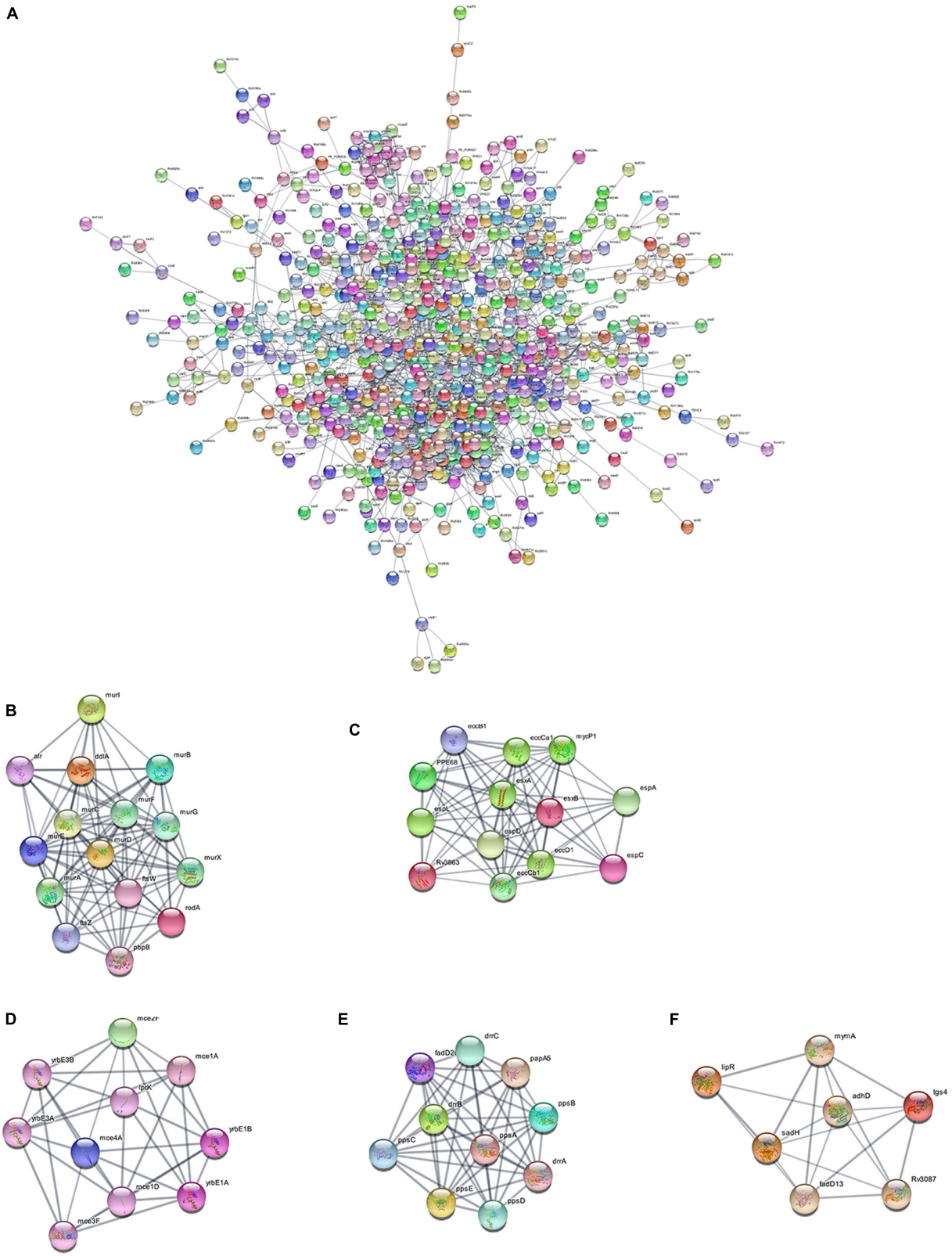


FIGURE 6 | PPI network of cell wall-related genes of *M. tuberculosis*. **(A)** Protein interaction networks visualized with Cytoscape. **(B)** Molecular complex detection (MCODE) with deep excavation of the core subnet. A modular gene involved in peptidoglycan synthesis in Subnet 1. **(C)** The gene cluster for the ESX-1 secretory system. **(D)** Gene clusters encoding membrane lipid transporters. **(E)** Key gene clusters for cell wall resistance. **(F)** Fatty acids modify gene clusters.

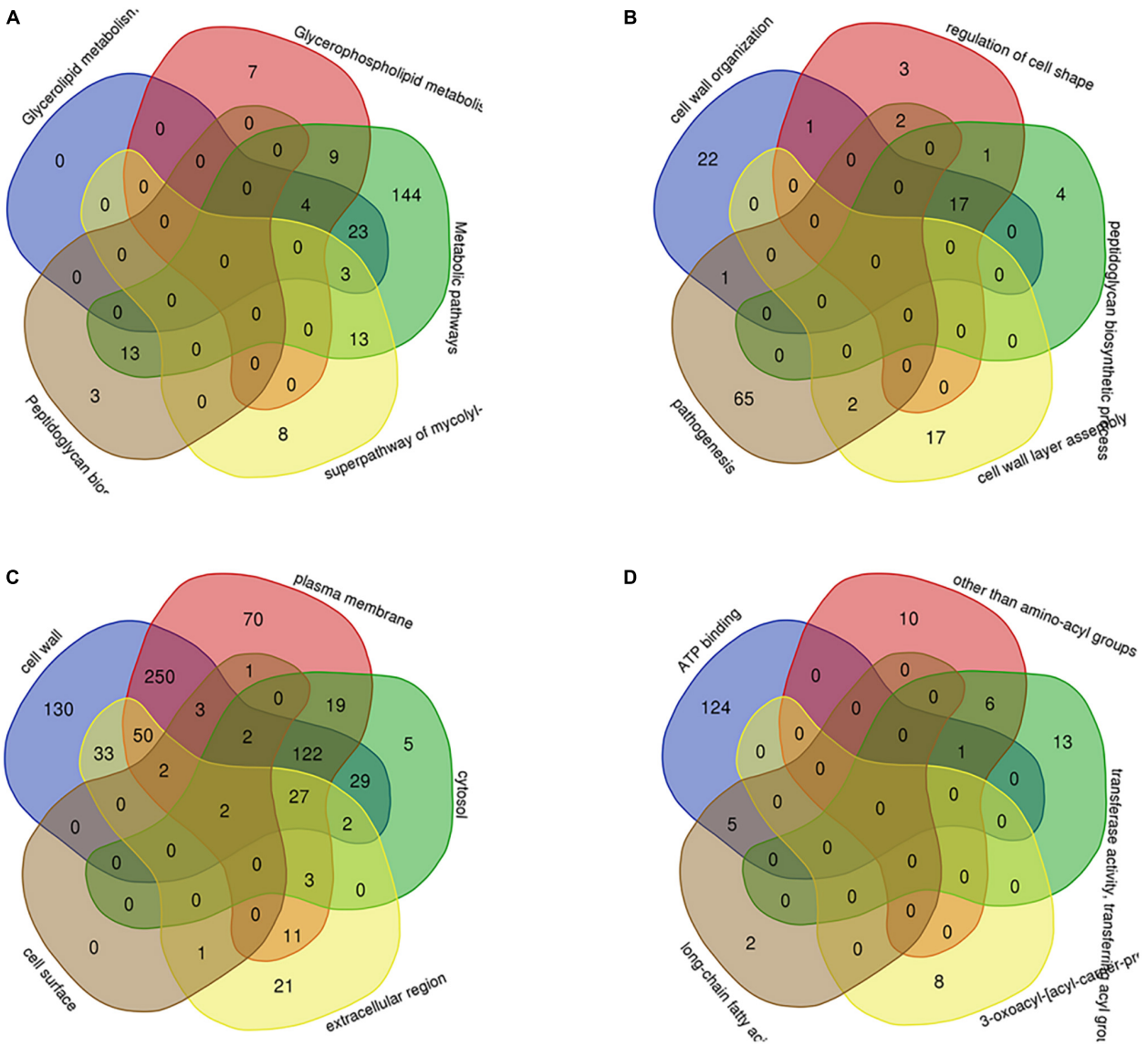


FIGURE 7 | The key genes with multifunctional correlation intermingle through multiple pathways screened for the cell wall-related genes of *M. tuberculosis*. Intersecting genes in the five pathways with the most significant differences in KEGG **(A)**, BP **(B)**, CC **(C)**, and MF **(D)** are displayed in a petal diagram.

the cMonkey double clustering algorithm to cluster the cell wall synthesis genes. We also identified key genes by screening co-regulatory clustering modules. Through functional analysis of cell wall synthesis genes by GO and KEGG, we screened the key genes for the synthesis of important components of the cell wall, such as mycotic acid and peptidoglycan, and the key hub genes involved in multi-pathway synthesis. Finally, we created a PPI network and identified five important subnets through MCODE analysis. The intrinsic relationship between proteins in the network was used to deeply explore the genes. Molecular complexes containing key genes were extracted based on closely related regions in the PPI. Finally, we obtained the five most valuable subnets. Using Subnet 3 as an example, all genes contained in this subnet are part of the mammalian cell entry (MCE) operon (Gioffre et al., 2005). The MCE operon is present in all genera of mycobacteria and actinomycetes. However, the number of MCE operons in different strains varies, with MCE 4 in *M. tuberculosis*, MCE 3 in *M. smegmatis*, and MCE 1, 2, and 4 in *M. bovis*. It is unknown why the MCE 3 operon is absent from *M. bovis* (Kumar et al., 2005). The MCE operons help *M. tuberculosis* ingest cholesterol in the host to keep the bacteria alive. Lack of the MCE operon causes a serious imbalance of lipid content in the *M. tuberculosis* cell wall. Sally et al. reported free mycolic acid accumulation in the cell wall of the MCE 1 operon mutant strain of *M. tuberculosis* (Singh et al., 2018). However, the genes contained in Subnet 4, such as *ppsa* and *ppsb*, were significantly altered in drug-resistant bacteria (Cantrell et al., 2013). We believe that *ppsa* changes the expression of PDIM in the cell wall by changing the approach

of the multi-subunit non-iterated polyketide synthase system (Vergnolle et al., 2015). This makes the bacterial cell wall thicker and causes bacterial drug efflux. We used bioinformatics and statistical methods to comprehensively scan all the genes synthesized in the *M. tuberculosis* cell wall and to screen out new targets that can be used as new anti-*M. tuberculous* cell wall targeting drugs.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

GW designed this study. XL and JP wrote the manuscript. XL analyzed the data. QM, WW, JH, and NZ contributed in the data collection. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (#81871699), Foundation of Jilin Province Science and Technology Department (#172408GH010234983), and the epidemiology, early warning, and response techniques of major infectious diseases in the Belt and Road Initiative (#2018ZX10101002).

ACKNOWLEDGMENTS

We thank Medjaden Bioscience Limited (Hong Kong, China) for editing and proofreading this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00607/full#supplementary-material

REFERENCES

Ahmad, S., El-Shazly, S., Mustafa, A. S., and Al-Attiyah, R. (2005). The six mammalian cell entry proteins (Mce3A-F) encoded by the mce3 operon are expressed during in vitro growth of *Mycobacterium tuberculosis*. *Scand. J. Immunol.* 62, 16–24. doi: 10.1111/j.1365-3083.2005.01639.x

Bhat, A. H., Pathak, D., and Rao, A. (2017). The alr-groEL1 operon in *Mycobacterium tuberculosis*: an interplay of multiple regulatory elements. *Sci. Rep.* 7:43772. doi: 10.1038/srep43772

Bothra, A., Arumugam, P., Panchal, V., Menon, D., Srivastava, S., Shankaran, D., et al. (2018). Phospholipid homeostasis, membrane tenacity and survival of Mtb in lipid rich conditions is determined by MmpL11 function. *Sci. Rep.* 8:8317. doi: 10.1038/s41598-018-26710-z

Cantrell, S. A., Leavell, M. D., Marjanovic, O., Iavarone, A. T., Leary, J. A., and Riley, L. W. (2013). Free mycolic acid accumulation in the cell wall of the mce1 operon mutant strain of *Mycobacterium tuberculosis*. *J. Microbiol.* 51, 619–626. doi: 10.1007/s12275-013-3092-y

Cao, H., Ma, Q., Chen, X., and Xu, Y. (2019). DOOR: a prokaryotic operon database for genome analyses and functional inference. *Brief. Bioinform.* 20, 1568–1577. doi: 10.1093/bib/bbx088

Catanho, M., Mascarenhas, D., Degrave, W., and Miranda, A. B. (2006). GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. *Genet. Mol. Res.* 5, 115–126.

Chaudhuri, R. R. (2009). MycoDB: an online database for comparative genomics of the mycobacteria and related organisms. *Methods Mol. Biol.* 465, 419–431. doi: 10.1007/978-1-59745-207-6_27

Chen, J. M., Zhang, M., Rybniker, J., Basterra, L., Dhar, N., Tischler, A. D., et al. (2013). Phenotypic profiling of *Mycobacterium tuberculosis* EspA point mutants reveals that blockage of ESAT-6 and CFP-10 secretion in vitro does not always correlate with attenuation of virulence. *J. Bacteriol.* 195, 5421–5430. doi: 10.1128/JB.00967-13

Galagan, J. E., Sisk, P., Stolte, C., Weiner, B., Koehrsen, M., Wymore, F., et al. (2010). TB database 2010: overview and update. *Tuberculosis* 90, 225–235. doi: 10.1016/j.tube.2010.03.010

Gaur, A., Sharma, V. K., Shree, S., Rai, N., and Ramachandran, R. (2017). Characterization of EccA3, a CbbX family ATPase from the ESX-3 secretion pathway of *M. tuberculosis*. *Biochim. Biophys. Acta Proteins Proteom.* 1865, 715–724. doi: 10.1016/j.bbapap.2017.04.001

Gillespie, J. J., Wattam, A. R., Cammer, S. A., Gabbard, J. L., Shukla, M. P., Dalay, O., et al. (2011). PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.* 79, 4286–4298. doi: 10.1128/IAI.00207-11

Gioffre, A., Infante, E., Aguilar, D., Santangelo, M. P., Klepp, L., Amadio, A., et al. (2005). Mutation in mce operons attenuates *Mycobacterium tuberculosis* virulence. *Microbes Infect.* 7, 325–334. doi: 10.1016/j.micinf.2004.11.007

Gopal, P., Yee, M., Sarathy, J., Low, J. L., Sarathy, J. P., Kaya, F., et al. (2016). Pyrazinamide resistance is caused by two distinct mechanisms: prevention of coenzyme a depletion and loss of virulence factor synthesis. *ACS Infect. Dis.* 2, 616–626. doi: 10.1021/acsinfecdis.6b00070

Hellweger, F. L., van Sebille, E., and Fredrick, N. D. (2014). Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science* 345, 1346–1349. doi: 10.1126/science.1254421

Heydari, H., Wee, W. Y., Lokanathan, N., Hari, R., Mohamed Yusoff, A., Beh, C. Y., et al. (2013). MabsBase: a *Mycobacterium abscessus* genome and annotation database. *PLoS One* 8:e62443. doi: 10.1371/journal.pone.0062443

Howard, N. C., Marin, N. D., Ahmed, M., Rosa, B. A., Martin, J., Bambouskova, M., et al. (2018). *Mycobacterium tuberculosis* carrying a rifampicin drug resistance mutation reprograms macrophage metabolism through cell wall lipid changes. *Nat. Microbiol.* 3, 1099–1108. doi: 10.1038/s41564-018-0245-0

Ivarsson, Y., and Jemth, P. (2019). Affinity and specificity of motif-based protein-protein interactions. *Curr. Opin. Struct. Biol.* 54, 26–33. doi: 10.1016/j.sbi.2018.09.009

Jiang, M. J., Liu, S. J., Su, L., Zhang, X., Li, Y. Y., Tang, T., et al. (2020). Intranasal vaccination with Listeria ivanovii as vector of *Mycobacterium tuberculosis* antigens promotes specific lung-localized cellular and humoral immune responses. *Sci. Rep.* 10:302. doi: 10.1038/s41598-019-57245-6

Kong, X., Zhu, B., Stone, V. N., Ge, X., El-Rami, F. E., Donghai, H., et al. (2019). ePath: an online database towards comprehensive essential gene annotation for prokaryotes. *Sci. Rep.* 9:12949. doi: 10.1038/s41598-019-49098-w

Kumar, A., Chandolia, A., Chaudhry, U., Brahmachari, V., and Bose, M. (2005). Comparison of mammalian cell entry operons of mycobacteria: in silico analysis and expression profiling. *FEMS Immunol. Med. Microbiol.* 43, 185–195. doi: 10.1016/j.femsim.2004.08.013

Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I., et al. (2001). Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet* 357, 1225–1240. doi: 10.1016/s0140-6736(00)04403-2

Lew, J. M., Kapopoulou, A., Jones, L. M., and Cole, S. T. (2011). TubercuList–10 years after. *Tuberculosis* 91, 1–7. doi: 10.1016/j.tube.2010.09.008

Maan, P., and Kaur, J. (2019). Rv2223c, an acid inducible carboxyl-esterase of *Mycobacterium tuberculosis* enhanced the growth and survival of

Mycobacterium smegmatis. *Future Microbiol.* 14, 1397–1415. doi: 10.2217/fmb-2019-0162

Maitra, A., Munshi, T., Healy, J., Martin, L. T., Vollmer, W., Keep, N. H., et al. (2019). Cell wall peptidoglycan in *Mycobacterium tuberculosis*: an Achilles' heel for the TB-causing pathogen. *FEMS Microbiol. Rev.* 43, 548–575. doi: 10.1093/femsre/fuz016

Meng, J., Gao, P., Wang, X., Guan, Y., Liu, Y., and Xiao, C. (2019). Digging deeper to save the old anti-tuberculosis target: D-Alanine-D-Alanine Ligase With a Novel Inhibitor, IMB-0283. *Front. Microbiol.* 10:3017. doi: 10.3389/fmicb.2019.03017

Merker, M., Kohl, T. A., Barilar, I., Andres, S., Fowler, P. W., Chryssanthou, E., et al. (2020). Phylogenetically informative mutations in genes implicated in antibiotic resistance in *Mycobacterium tuberculosis* complex. *Genome Med.* 12:27. doi: 10.1186/s13073-020-00726-5

Pasricha, R., Chandolia, A., Ponnan, P., Saini, N. K., Sharma, S., Chopra, M., et al. (2011). Single nucleotide polymorphism in the genes of mce1 and mce4 operons of *Mycobacterium tuberculosis*: analysis of clinical isolates and standard reference strains. *BMC Microbiol.* 11:41. doi: 10.1186/1471-2180-11-41

Postma, M., and Goedhart, J. (2019). PlotsOfData-A web app for visualizing data together with their summaries. *PLoS Biol.* 17:e3000202. doi: 10.1371/journal.pbio.3000202

Quigley, J., Hughitt, V. K., Velikovsky, C. A., Mariuzza, R. A., El-Sayed, N. M., and Briken, V. (2017). The cell wall lipid PDIM contributes to phagosomal escape and host cell exit of *Mycobacterium tuberculosis*. *mBio* 8:e00148-17. doi: 10.1128/mBio.00148-17

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* 9:1090. doi: 10.1038/s41467-018-03424-4

Safi, H., Lingaraju, S., Amin, A., Kim, S., Jones, M., Holmes, M., et al. (2013). Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-beta-D-arabinose biosynthetic and utilization pathway genes. *Nat. Genet.* 45, 1190–1197. doi: 10.1038/ng.2743

Selvam, K., Duncan, J. R., Tanaka, M., and Battista, J. R. (2013). DdrA, DdrD, and PprA: components of UV and mitomycin C resistance in Deinococcus radiodurans R1. *PLoS One* 8:e69007. doi: 10.1371/journal.pone.0069007

Sher, J. W., Lim, H. C., and Bernhardt, T. G. (2020). Global phenotypic profiling identifies a conserved actinobacterial cofactor for a bifunctional PBP-type cell wall synthase. *eLife* 9:54761. doi: 10.7554/eLife.54761

Singh, A., Gupta, R., Vishwakarma, R. A., Narayanan, P. R., Paramasivan, C. N., Ramanathan, V. D., et al. (2005). Requirement of the mymA operon for appropriate cell wall ultrastructure and persistence of *Mycobacterium tuberculosis* in the spleens of guinea pigs. *J. Bacteriol.* 187, 4173–4186. doi: 10.1128/JB.187.12.4173-4186.2005

Singh, P., Sinha, R., Tyagi, G., Sharma, N. K., Saini, N. K., Chandolia, A., et al. (2018). PDIM and SL1 accumulation in *Mycobacterium tuberculosis* is associated with mce4A expression. *Gene* 642, 178–187. doi: 10.1016/j.gene.2017.09.062

Turner, C. T., Gupta, R. K., Tsaliki, E., Roe, J. K., Mondal, P., Nyawo, G. R., et al. (2020). Blood transcriptional biomarkers for active pulmonary tuberculosis in a high-burden setting: a prospective, observational, diagnostic accuracy study. *Lancet Respir. Med.* 8, 407–419. doi: 10.1016/S2213-2600(19)30469-2

Vergnolle, O., Chavadi, S. S., Edupuganti, U. R., Mohandas, P., Chan, C., Zeng, J., et al. (2015). Biosynthesis of cell envelope-associated phenolic glycolipids in *Mycobacterium marinum*. *J. Bacteriol.* 197, 1040–1050. doi: 10.1128/JB.02546-14

Vishnoi, A., Srivastava, A., Roy, R., and Bhattacharya, A. (2008). MGDD: *Mycobacterium tuberculosis* genome divergence database. *BMC Genom.* 9:373. doi: 10.1186/1471-2164-9-373

Waltman, P., Kacmarczyk, T., Bate, A. R., Kearns, D. B., Reiss, D. J., Eichenberger, P., et al. (2010). Multi-species integrative biclustering. *Genome Biol.* 11:R96. doi: 10.1186/gb-2010-11-9-r96

Wong, K. W. (2017). The Role of ESX-1 in *Mycobacterium tuberculosis* pathogenesis. *Microbiol. Spectr.* 5, 627–634. doi: 10.1128/microbiolspec.TBTB2-0001-2015

Wu, M. L., Gengenbacher, M., Chung, J. C., Chen, S. L., Mollenkopf, H. J., Kaufmann, S. H., et al. (2016). Developmental transcriptome of resting cell formation in *Mycobacterium smegmatis*. *BMC Genomics* 17:837. doi: 10.1186/s12864-016-3190-4

Zhu, X., Chang, S., Fang, K., Cui, S., Liu, J., Wu, Z., et al. (2009). MyBASE: a database for genome polymorphism and gene function studies of Mycobacterium. *BMC Microbiol.* 9:40. doi: 10.1186/1471-2180-9-40

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 03 July 2020
doi: 10.3389/fbioe.2020.00670

SCAU-Net: Spatial-Channel Attention U-Net for Gland Segmentation

Peng Zhao[1], *Jindi Zhang*[2], *Weijia Fang*[1]* *and Shuiguang Deng*[1,2]*

[1] First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China

With the development of medical technology, image semantic segmentation is of great significance for morphological analysis, quantification, and diagnosis of human tissues. However, manual detection and segmentation is a time-consuming task. Especially for biomedical image, only experts are able to identify tissues and mark their contours. In recent years, the development of deep learning has greatly improved the accuracy of computer automatic segmentation. This paper proposes a deep learning image semantic segmentation network named Spatial-Channel Attention U-Net (SCAU-Net) based on current research status of medical image. SCAU-Net has an encoder-decoder-style symmetrical structure integrated with spatial and channel attention as plug-and-play modules. The main idea is to enhance local related features and restrain irrelevant features at the spatial and channel levels. Experiments on the gland dataset GlaS and CRAG show that the proposed SCAU-Net model is superior to the classic U-Net model in image segmentation task, with 1% improvement on Dice score and 1.5% improvement on Jaccard score.

Keywords: deep learning, semantic segmentation, attention mechanism, medical image, gland

OPEN ACCESS

Edited by:
Yucong Duan,
Hainan University, China

Reviewed by:
Qiang He,
Swinburne University of Technology, Australia
Shangguang Wang,
Beijing University of Posts and Telecommunications (BUPT), China
Jun Yu,
Hangzhou Dianzi University, China

***Correspondence:**
Weijia Fang
weijiafang@zju.edu.cn
Shuiguang Deng
dengsg@zju.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *23 March 2020*
Accepted: *28 May 2020*
Published: *03 July 2020*

Citation:
Zhao P, Zhang J, Fang W and Deng S (2020) SCAU-Net: Spatial-Channel Attention U-Net for Gland Segmentation. Front. Bioeng. Biotechnol. 8:670. doi: 10.3389/fbioe.2020.00670

1. INTRODUCTION

In clinical practice, biomedical image analysis (Litjens et al., 2017) provides doctors with digital and quantitative medical information, and helps doctors make objective and accurate diagnosis. Image segmentation is a basic problem in medical image analysis. In short, it is to identify the target area in an image and distinguish the research object from the background. For instance, glands are important tissues of the human body that secrete special proteins and hormones. Malignant tumors caused by glandular differentiation, i.e., adenocarcinoma, is a common form of cancer. Different grades of differentiated glands have various morphological structures. In pathological examination, pathologists usually use Hematoxylin and Eosin (H&E) to stain glandular tissues, then evaluate the malignancy of adenocarcinoma and determine the grade of cancer (Niazi et al., 2019). Early detection of glandular differentiation can greatly improve the cure rate of patients, and these treatment methods often require detailed gland information, such as the size, shape and location of the glands before and after treatment, in order to propose a suitable treatment plan. At present, this work is mainly performed by expert pathologists. However, the morphology of glands in different histological differentiation grads is quite complex, and the texture and size vary from patient to patient. It is still a very challenging task.

Manually detecting and segmenting medical images consumes a lot of energy and time of doctors. In recent years, with the deepening cooperation between artificial intelligence and medical image analysis, the research of computer-aided medical image segmentation have exploded.

Computer automatic segmentation enables doctors to quickly and easily obtain image markers related to the disease treatment process, detect malignant tumors early in time. Especially for the automatic segmentation of H&E gland images, pathologists can quickly extract important morphological features from massive histological images. This work helps pathologists to provide services to more patients while ensuring diagnostic accuracy. To some extent, it can solve the problem of imbalanced distribution of medical resources and lack of expert pathologists.

In this paper, we propose a deep learning network named Spatial-Channel Attention U-Net (SCAU-Net) for gland segmentation. The contributions of this paper are as follows:

1. Our model has a symmetrical structure. It exploits skip connections to concatenate outputs of encoder to the decoder in corresponding level. Multi-level features are fused to improve segmentation results.
2. We introduce spatial attention and channel attention as plug-and-play modules for the basic encoder-decoder structure. The module exploits hidden layer neural network to capture the non-linear relationship between spatial-wise and channel-wise feature, and essentially introduces a self-attention mechanism. The attention module performs feature recalibration to enhance local related features and restrain irrelevant features at the spatial and channel levels.

2. RELATED WORK

2.1. Biomedical Image Segmentation

Computer automatic image segmentation algorithms are categorized as traditional algorithms based on manual features and deep learning algorithms based on Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012).

The main idea of traditional image segmentation algorithms is to segment the image into regions with similar properties, such as color and texture (Sharma and Aggarwal, 2010). Divided in principle, including the following types of methods: (1) Edge based segmentation. Algorithm exploits discontinuity principle such as grayscale and color to detect boundaries between regions (Hancock and Kittler, 1990; Liow, 1991). Fuzzy boundaries and noise can easily affect the performance of the method. (2) Region based segmentation. Pixels with similar properties are aggregated to form a complete object regions. Wu et al. (2005) proposed a intestinal gland images segmentation based on iterative region growing. The segmentation results of this method are sensitive to the number of clusters and regions initialization. (3) Textural feature based segmentation. This method divides the image regions according to texture properties (Sirinukunwattana et al., 2015).

In recent years, deep learning has become the main research method in many fields, and CNN is widely used in many different computer vision tasks. Unlike previous traditional methods, CNN is a data-driven method that can automatically learn advanced features from image without the need for artificial feature design and prior knowledge. In the medical field, CNN has also achieved good results in the detection and segmentation of cells (Raza et al., 2017), pancreas (Roth et al., 2015), liver tumors (Dou et al., 2016; Christ et al., 2017), glands (Chen et al., 2016; Xu et al., 2016; Yang et al., 2017; Graham et al., 2019), and other human tissues.

The full convolutional network (FCN) (Long et al., 2015) is the first method for image semantic segmentation using end-to-end deep neural networks. The innovation is that the fully connected layer is replaced by fully convolutional layer. This important innovation enables the network to adapt to the input of any resolution.

Datasets containing large amounts of labeled images have been established in other fields, such as ImageNet, COCO, etc. However, in the field of medical images, due to the high annotation cost, it is almost impossible to provide such a large dataset. Therefore, how to train a good model in the case of small datasets is a difficult research point. U-Net (Ronneberger et al., 2015) is based on the FCN structure, and exploits skip connections to transfer and fuse the output of feature maps with different resolutions to obtain more accurate outputs. It is firstly used for segmentation of neuron and cell images and has excellent performance on many medical image datasets. In the last few years of medical image segmentation, many works have been developed and improved on the basis of the U-Net (Çiçek et al., 2016; Milletari et al., 2016; Gordienko et al., 2018; Zhou et al., 2018). Unlike many recent studies focus on instance segmentation (Xu et al., 2016; Graham et al., 2019; Yu et al., 2020), SCAU-Net proposed in this paper extends U-Net as basic model in order to improve the accuracy of segmentation while retaining the original advantages. In addition, our method can be easily extended to other medical image segmentation such as liver, cell, etc.

2.2. Vision Attention

When looking at a scene, we often firstly scan the whole scene quickly and focus on the region of interest (ROI). This selective attention mechanism that mimics the Human Visual System (HVS) has been widely used in computer vision (Itti and Koch, 2001; Wang and Shen, 2017). There is no strict mathematical definition of the attention mechanism. Oktay et al. (2018) proposed a network of encoder-decoder-style called Attention U-Net, which exploits a Attention Gates control. Another modular attention mechanism is called self-attention. The computation and parameter overhead of the feature map's attention generation process is much smaller, which can be used as a plug-and-play module of the existing basic CNN architecture. This method introduces additional neural network modules, which can assign different weights to spatial-wise or channel-wise.

Spatial attention learns to focus on spatial location (where), and weights are assigned to each pixel. Therefore, the form of weights is a $H \times W$ 2D matrix. Jaderberg et al. (2015) introduced a learnable Spatial Transformer module, which can learn the location of object regions by the input feature map.

Channel attention learns to select important feature dimensions (what), and weights are assigned to each channel. Therefore, the form of weights is a 1D vector. Hu et al. (2018) proposed the Squeeze-and-excitation (SE) module, which learns the non-linear relationship between channels and performs dynamic channel-wise feature recalibration.

In addition, spatial and channel attention modules can be combined in a parallel or sequential manner. e.g., Dual Attention Network (Fu et al., 2019) parallels spatial and channel attention and fuses output features of attention module. Woo et al. (2018) proposed Convolutional Block Attention Module (CBAM), which sequentially builds the channel and spatial attention modules. Non-Local attention (Wang et al., 2018) computes the response at a position by capturing long-range dependencies at all positions. Bottleneck attention module (Park et al., 2018) generates a 3D attention map in two streams, i.e., spatial stream and channel stream.

3. METHOD

Inspired by U-Net network structure and attention mechanism, we propose a deep learning network named SCAU-Net. The entire structure is shown in **Figure 1**.

We define "Block(x)" which executes a 3×3 convolution followed by a batch normalization and ReLU activation, two times. x refers to the output channel number. The role of the encoder part is to extract features from the image and obtain compressed expression of the image features at multi-level. Down-sampling is performed by 2 × 2 max-pooling operation. During each down-sampling, the image size is reduced and the number of feature channels is doubled. The role of the decoder part is to gradually restore the details and spatial dimensions of the image according to the image features, and obtain the result of image segmentation mask. Up-sampling is performed by bilinear interpolation. Finally, a 1 × 1 convolutional layer is applied to predict the class of each pixel, denoted as Conv(1 × 1, C), where C is the number of classes. For image semantic segmentation, C is set to 2. The decoder part has a symmetrical structure to the encoder part. The copy operation links the corresponding down-sampling and up-sampling feature maps. The feature map is a combination of high-level and low-level features, and multi-level features are fused.

The medical image structure is simpler and more fixed than other images. For gland slices, the shooting angle and position are fixed, and the glands of approximate differentiation degree are often similar in shape. Inspired by the work of SE (Hu et al., 2018) and CBAM (Woo et al., 2018), we propose spatial attention module and channel attention module, which are used as plug-and-play modules in the network. Attention will focus on the objects and ignore the cluttered background. Especially, model will pay more attention on the edges of the glands because the fuzzy edge is the most worthy of the segmentation task.

3.1. Spatial Attention

Attention in the spatial-wise ignores the information of the channel, and treats the features of different channels equally. We add the spatial attention module to the low-level feature map since the low-level feature map mainly extracts the spatial feature such as contour, edge, with fewer channels. The module self-learns the interaction of spatial points, enhance key areas, and restrain irrelevant areas. The structure of the spatial attention module is shown in **Figure 2**. Firstly we pass the feature map $\boldsymbol{U} \in \mathbb{R}^{C\times H\times W}$ to the aggregation operation, which generates a spatial descriptor $\boldsymbol{p} \in \mathbb{R}^{H\times W}$ by aggregating the feature map in its channel dimension (C). It generates a global distribution of spatial features:

$$\boldsymbol{p}_{hw} = F_{ac}(\boldsymbol{u}_{hw}) = \frac{1}{C}\sum_{i=1}^{C}\boldsymbol{u}_{hw}(i) \tag{1}$$

where $\boldsymbol{u}_{hw} \in \mathbb{R}^{C}$ refers to the local feature at spatial position (h, w). The aggregate function F_{ac} uses global average pooling for channel dimension.

This is followed by a weight self-learning operation. It is implemented by convolutional layers. The function $F_l(\boldsymbol{p},\boldsymbol{f})$ aims to fully capture the spatial correlation and adaptively generates the spatial weights map $\boldsymbol{t} \in \mathbb{R}^{H\times W}$. The calculation formula is as follows:

$$\boldsymbol{t} = F_l(\boldsymbol{p},\boldsymbol{f}) = \sigma(g(\boldsymbol{p},\boldsymbol{f})) = \sigma(\boldsymbol{f_2}\delta(\boldsymbol{f_1}\boldsymbol{p})) \tag{2}$$

where $\boldsymbol{f_1}$ refers to 3 × 3 convolution, denoted as Conv(3×3, m), and $\boldsymbol{f_2}$ refers to 3 × 3 convolution, denoted as Conv(3 × 3, 1). m refers to the channel number of hidden feature map. δ refers to activation function ReLU, and σ is a sigmoid activation function used to generate spatial weight $\boldsymbol{t}_{hw} \in (0, 1)$, at position (h, w). In essence, the convolution operation that takes the original spatial descriptor as input can be considered as a spatial-wise self-attention function, and it can capture the non-linear inter-spatial relationship.

The weights calculated in the previous step are applied to the feature map $\boldsymbol{U}$. By spatial-wise recalibration $F_{re}(\boldsymbol{u}_{hw}, \boldsymbol{t}_{hw})$, the feature values of different position in $\boldsymbol{U}$ are multiplied by different weights to generate the output $\boldsymbol{U}'$ of the SA module:

$$\boldsymbol{u}'_{hw} = F_{re}(\boldsymbol{u}_{hw}, \boldsymbol{t}_{hw}) = \boldsymbol{u}_{hw} \cdot \boldsymbol{t}_{hw} \tag{3}$$

3.2. Channel Attention

Similarly, we add the channel attention module at the last layer of the encoder, since the hight-level feature map mainly expresses complex feature with large receptive field and more channels. This mechanism allows the network to perform feature recalibration, through learning to exploit global information to selectively enhance useful features and restrain useless features. The structure of the channel attention module is shown in **Figure 3**. Firstly we pass the feature map $\boldsymbol{U} \in \mathbb{R}^{C\times H\times W}$ to the aggregation operation, which generates a channel descriptor $\boldsymbol{q} \in \mathbb{R}^{C}$ by aggregating the feature map in its spatial dimension $(H \times W)$. It generates a global distribution of channel features:

$$\boldsymbol{q}_c = F_{as}(\boldsymbol{u}_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{i=j}^{W}\boldsymbol{u}_c(i,j) \tag{4}$$

where $\boldsymbol{u}_c \in \mathbb{R}^{H\times W}$ refers to the local feature of channel c. The aggregate function F_{as} uses global average pooling for spatial dimension.

This is followed by a weight self-learning operation. It is implemented by fully connected layers. The function $F_l(\boldsymbol{q}, \boldsymbol{w})$

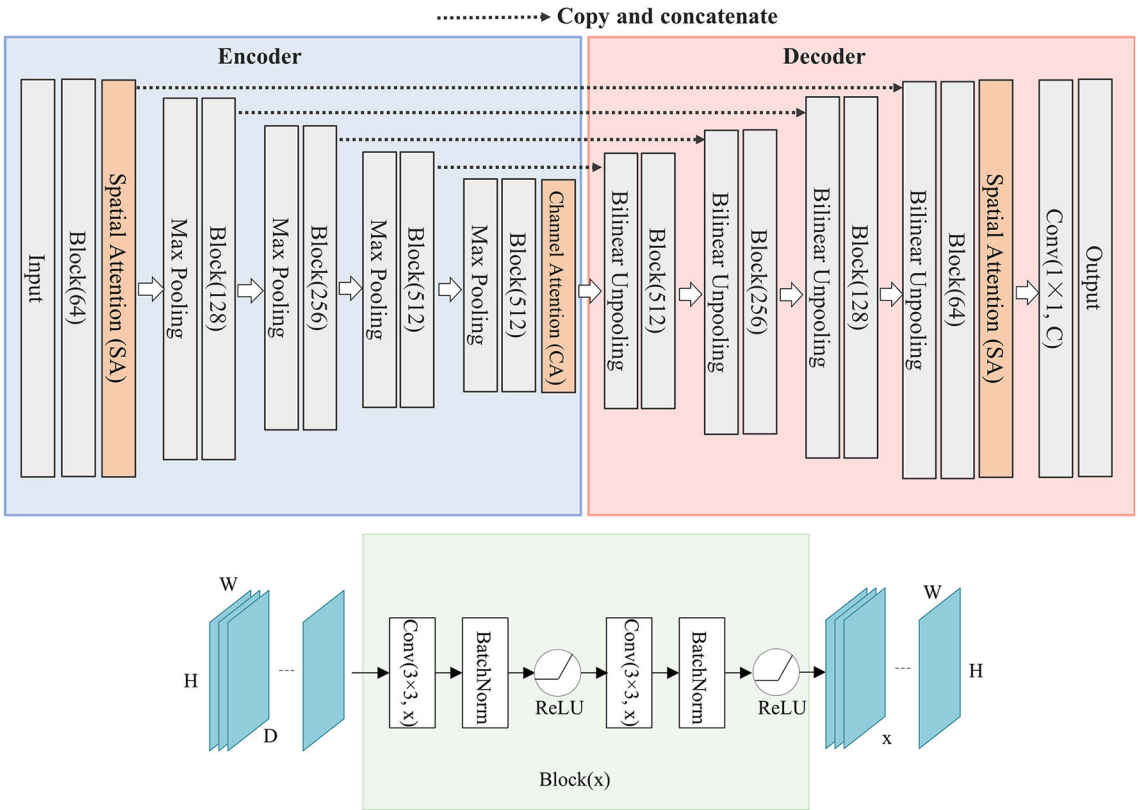


FIGURE 1 | The structure of SCAU-Net. The entire structure is divided into four parts: encoder, decoder, spatial attention module, channel attention module. Given an input feature map of size D×H×W, the output size of Block(x) is x××H×W. W, width; H, height of feature map; D, input channel number; x, output channel number.

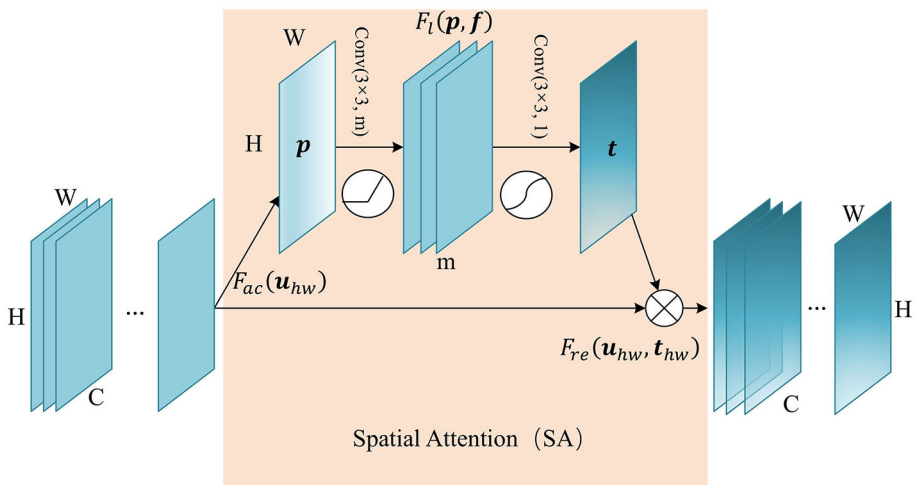


FIGURE 2 | The structure of Spatial Attention (SA) module. The aggregate function F_{ac} generates a spatial descriptor $\boldsymbol{p} \in \mathbb{R}^{H\times W}$. Self-learning function F_l implemented by two convolutional layers generates the spatial weights map $\boldsymbol{t} \in \mathbb{R}^{H\times W}$. Finally, function F_{re} uses $\boldsymbol{t}$ to generate the output of the SA module.

aims to fully capture the dependencies between channels and adaptively generates the channel weights map $\boldsymbol{v} \in \mathbb{R}^C$. The calculation formula is as follows:

$$\boldsymbol{v} = F_l(\boldsymbol{q}, \boldsymbol{w}) = \sigma(g(\boldsymbol{q}, \boldsymbol{w})) = \sigma(\boldsymbol{w}_2\delta(\boldsymbol{w}_1\boldsymbol{q})) \quad (5)$$

where $\boldsymbol{w_1} \in \mathbb{R}^{K\times C}$, $\boldsymbol{w_2} \in \mathbb{R}^{C\times K}$. K refers to number of hidden neurons. σ is a sigmoid activation function used to generate channel weights $\boldsymbol{v}_c \in (0, 1)$, at channel c. With fully-connected hidden layers, it can capture the non-linear interaction between channels.

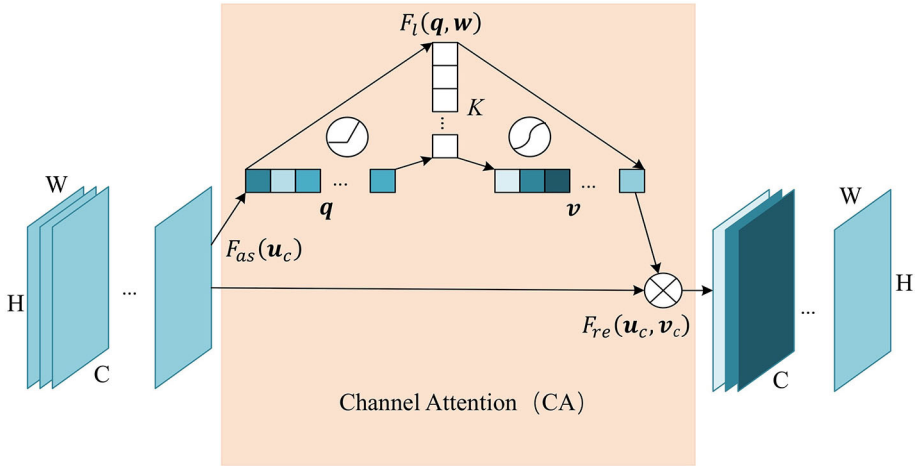


FIGURE 3 | The structure of Channel Attention (CA) module. The aggregate function F_{as} generates a channel descriptor $q \in \mathbb{R}^C$. Self-learning function F_l implemented by two fully connected layers generates the channel weights map $v \in \mathbb{R}^C$. Finally, function F_{re} uses v to generate the output of the CA module.

The weight calculated in the previous step is applied to the feature map U. By channel-wise recalibration $F_{re}(u_c, v_c)$, the feature values of different channels in U are multiplied by different weights to generate the output U' of the CA module:

$$u'_c = F_{re}(u_c, v_c) = u_c \cdot v_c \tag{6}$$

4. EXPERIMENTS AND RESULTS

4.1. Dataset

The two gland datasets used in the experiments are provided by a team of pathologists at the University Hospitals Coventry and Warwickshire, UK. (1) Gland Segmentation Challenge Contest (GlaS) (Sirinukunwattana et al., 2015) in MICCAI 2015. (2) The colorectal adenocarcinoma gland (CRAG) (Graham et al., 2019) dataset. The images are Haematoxylin and Eosin (H&E) stained slides of a variety of histologic grades. The GlaS dataset is split into 85 training images (benign/malignant = 37/48) and 80 testing images (benign/malignant = 37/43). We random split from 165 images using 80% images as the training set and the remaining 20% for testing. Images are mostly of size 780 × 520 pixels. The CRAG dataset is split into 173 training images and 40 test images. Images are mostly of size 1,510 × 1,510 pixels. And the ground truth annotations of the glands are provided by expert pathologists.

All the images processed by the network have fixed size of 512×512 pixels. Since the dataset is small, the training data is extended by using the data augmentation method in our experiments, i.e., a series of random changes such as rotation, scaling, cropping, etc., to increase the robustness and reduce overfitting.

4.2. Experimental Setting

The proposed network was implemented using Pytorch (Paszke et al., 2019) deep learning framework. Experiments are carried out on Ubuntu 16.04 operating system, NVIDIA Tesla K80 GPU, CUDA 10.1.

4.3. Training Process

The loss function defined in experiment is a combination of cross-entropy loss and dice loss:

$$CELoss = -\frac{1}{n}\sum y * log(y') + (1 - y) * log(1 - y') \tag{7}$$

$$DiceLoss = \frac{2\sum(y' * y)}{\sum y' + \sum y} \tag{8}$$

$$Loss = \lambda * CELoss + (1 - \lambda) * DiceLoss \tag{9}$$

where y is the ground truth of each pixel, and y' is model prediction. Dice loss function (Milletari et al., 2016) is based on dice coefficient and helps to establish the loss balance between foreground and background pixels. The loss function allocates the cross-entropy loss function and the dice loss function with λ. We set λ to 0.5 in the experiment. We use the Adam optimization (Kingma and Ba, 2014) and set initial learning rate to 0.0001. The input mini-batch size is 4. The total epoch is set to 100 with the learning rate decay strategy. Every 30 epochs, the learning rate is reduced to 1/10 of the previous value. For spatial attention module, we set the channel number of hidden feature map to 16. For channel attention module, we set the number of hidden neurons to 32.

4.4. Quality Measures

In order to evaluate the performance of the proposed method, we use the quality metrics commonly used in the field of medical image. Metric applies to the semantic segmentation of binary values which only considers glands as foreground, and everything else as background. Given A a set of pixels annotated as a ground truth object and B a set of pixels segmented as a gland object.

TABLE 1 | Our method's segmentation results compare with U-Net on dataset GlaS and CRAG.

	GlaS			CRAG		
Method	**Dice**	**Jaccard**	**RVD**	**Dice**	**Jaccard**	**RVD**
U-Net	0.8963	0.8175	0.0079	0.9003	0.8243	−0.0042
SCAU-Net(CA)	0.9004	0.8242	0.0190	0.9069	0.8333	−0.0072
SCAU-Net(SA)	0.9054	0.8322	−0.0166	0.9067	0.8330	−0.0033
SCAU-Net(SA+CA)	**0.9063**	**0.8332**	0.0197	**0.9100**	**0.8381**	−0.0074
DeepLabv3+	0.8866	0.7994	−0.0203	0.8672	0.7691	−0.0492
SegNet	0.7930	0.6643	−0.0582	0.8990	0.8209	−0.0030
U-Net++	0.8952	0.8166	0.0256	0.8870	0.8010	−0.0182

CA refers to channel attention module, SA refers to spatial attention module. We also compare with the network SegNet, U-Net++, DeepLabv3+. Significant results are highlighted in bold font.

Dice Similarity Coefficient (Dice):

$$\frac{2(A \cap B)}{A + B} \tag{10}$$

Jaccard Coefficient (Jaccard):

$$\frac{A \cap B}{A \cup B} \tag{11}$$

Relative Volume Difference (RVD):

$$\frac{|B| - |A|}{|A|} \tag{12}$$

In order to save the best model parameters during the training process, we use the Dice coefficient as the main evaluation metric. The larger the coefficient, the better the method performance. When the coefficient is 1, the predict result is consistent with the ground truth.

4.5. Results and Discussions

The experimental results are shown in **Table 1**. We compare our method with the baseline model U-Net. When our network using the channel attention (CA) alone, in the dataset GlaS, Dice score has a 0.4% improvement, and the dataset CRAG has a 0.6% improvement. When our network using the spatial attention (SA) alone, in the dataset GlaS, Dice score has a 0.9% improvement, and the dataset CRAG has a 0.6% improvement. Combining spatial and channel attention (SA+CA), there is 1% improvement on Dice score and 1.6% improvement on Jasccard score in the dataset GlaS. There is 1% improvement on Dice score and 1.4% improvement on Jasccard score in the dataset CRAG. Besides, compared with the network SegNet (Badrinarayanan et al., 2017), U-Net++ (Zhou et al., 2018), DeepLabv3+ (Chen et al., 2018), the overall performance of SCAU-Net is excellent, and it is more robust to different datasets.

As shown in **Figure 4**, we compare the training process between the U-Net and SCAU-Net. It can be observed that the SCAU-Net with spatial and channel attention (SA+CA) achieves the highest accuracy on validation sets. For the dataset GlaS, the SCAU-Net slightly over-fits after about the 60th epoch, while dataset CRAG doesn't. We analyze the results and believe that the added attention mechanism makes the model parameters increase, and the model is more likely to over-fit with less data amount.

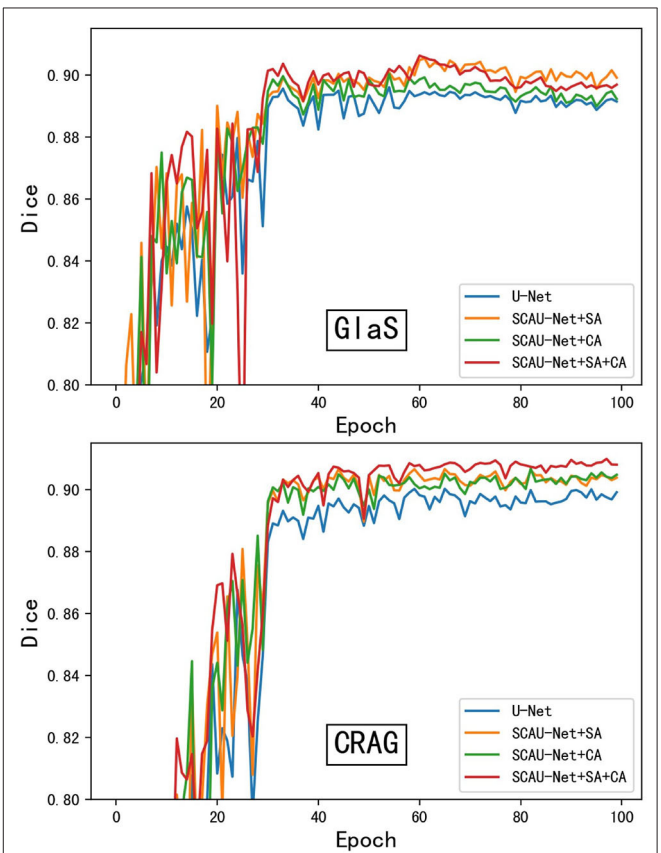


FIGURE 4 | The training process of different model. Figure shows dice curve of the U-Net and SCAU-Net with different settings (CA, SA, CA+SA) on validation sets. CA, channel attention module; SA, spatial attention module.

Figure 5 shows the visualization results of the method. As shown in **Figures 5A,B**, for some gland objects, the U-Net

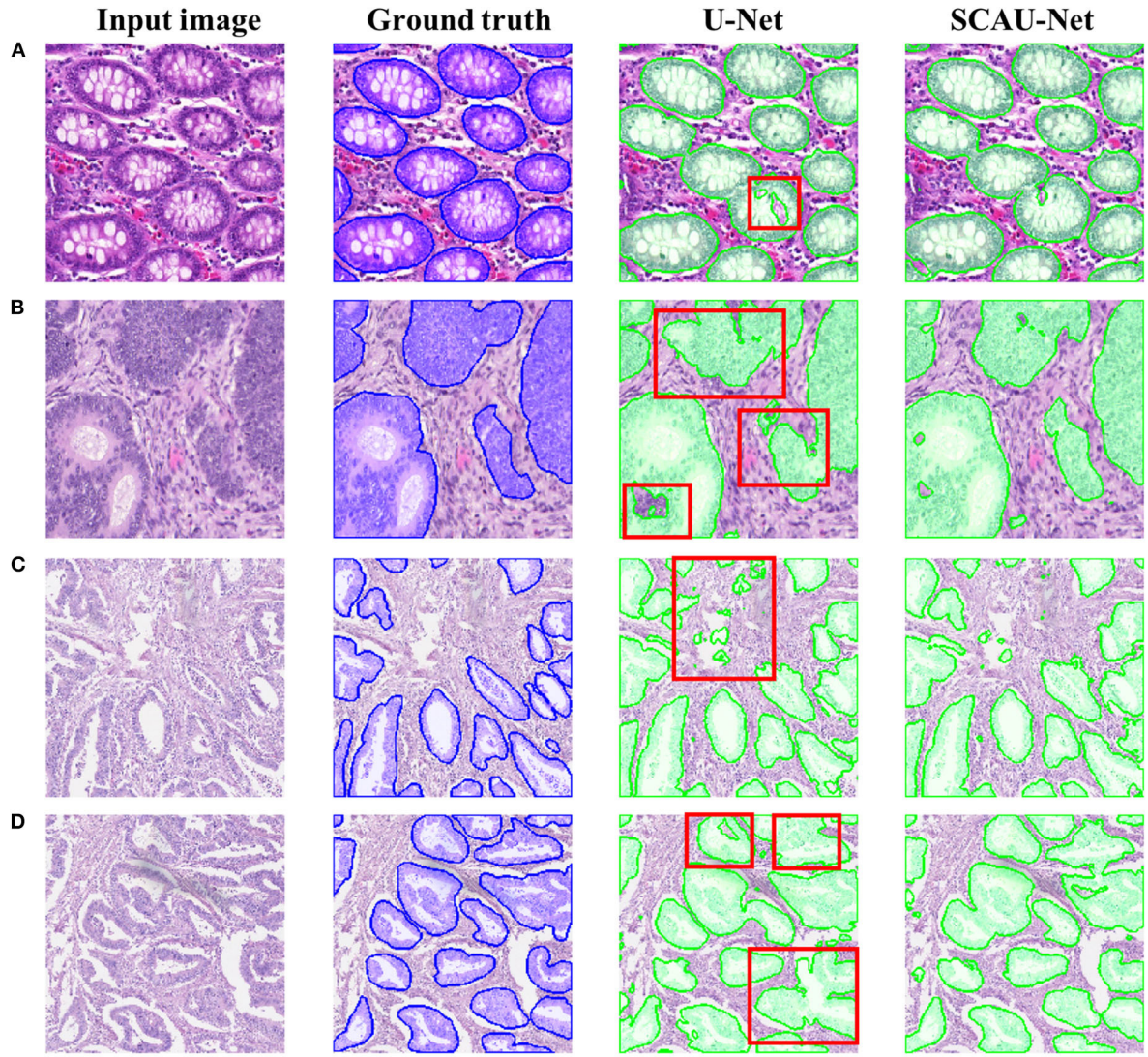


FIGURE 5 | Comparison of segmentation results. Examples **(A,B)** are from the GlaS dataset, and examples **(C,D)** are from the CRAG dataset. The red boxes indicate areas with poor segmentation results.

network misclassifies the white area inside the gland as the background, while SCAU-Net performs better. It shows that our method has better object connectivity. For some complex scenes, SCAU-Net can accurately distinguish background noise, as shown in **Figure 5C**, and can distinguish the edges of multiple gland objects well to prevent "sticking," as shown in **Figure 5D**. On the whole, SCAU-Net outperforms U-Net in the segmentation of glands.

In order to explore how the attention mechanism works, we visualize the effect of the model with the spatial attention mechanism added. For visual display, we extract the encoder output feature map of Block(64). Compared with the basic U-Net network, SCAU-Net exploits spatial attention weights to recalibrate the feature map. As shown in **Figure 6**, the feature maps extracted show the differences between the two methods. The contrast of the feature map by SCAU-Net is more prominent, indicating the wider range of values. The spatial attention weights map learned by SCAU-Net has different weight assignments in different regions, as shown in weights map. Spatial attention assigns lower weights on easily distinguishable backgrounds, non-glandular noise tissue areas, obvious contours, etc. The fuzzy boundaries of the indistinguishable contours are assigned higher weights, indicating that the network pays more attention to these difficult-to-classify regions.

5. CONCLUSION

In this paper, we extend the U-Net encoder-decoder framework, propose a new network named Spatial-Channel Attention U-Net (SCAU-Net) for image semantic segmentation. We perform the segmentation tasks on GlaS and CRAG gland dataset. The

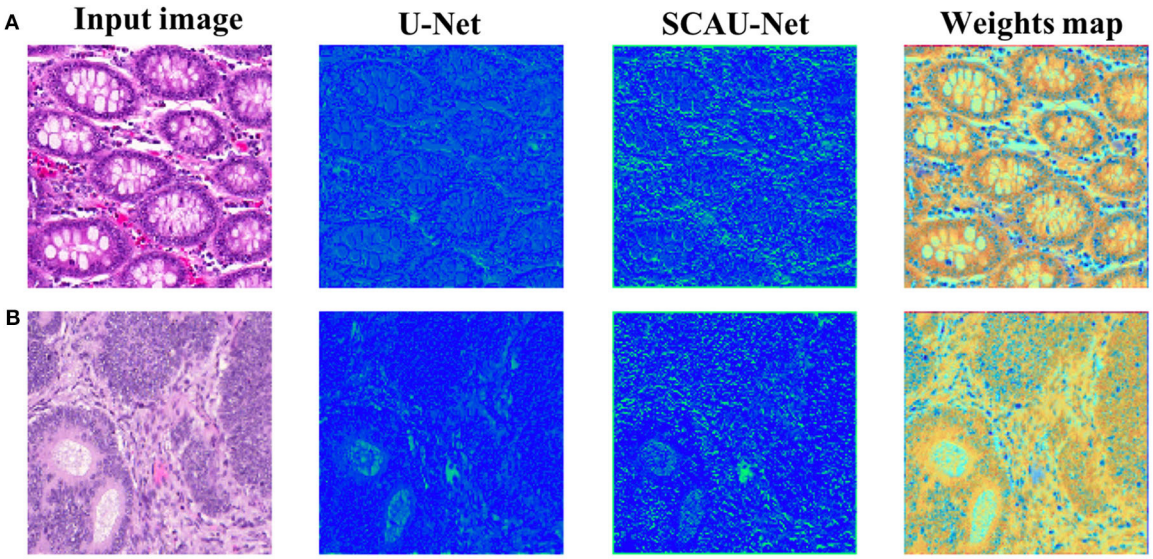


FIGURE 6 | Visualization the output feature activations after Block(64). Compared with the basic U-Net network, SCAU-Net exploits spatial attention weights to recalibrate the feature map. Weights map by SCAU-Net is shown in the last column. Examples **(A,B)** are from the GlaS dataset.

experiment results and comparisons with classic U-Net model demonstrate that our proposed model can achieve a better segmentation performance, with 1% improvement on Dice score and 1.5% improvement on Jaccard score. We also visualize the effect of attention mechanism on feature extraction to explain how the mechanism works.

In the future, the spatial and channel attention modules proposed in this paper need further exploration for the number of convolutional layers, the number of fully connected layers, and the location settings of the module embedding.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

PZ proposed the main idea. JZ implemented the experiments and wrote most of the manuscript. WF and SD wrote parts of the manuscript, read, and approved the final manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

Chen, H., Qi, X., Yu, L., and Heng, P.-A. (2016). "Dcan: deep contour-aware networks for accurate gland segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2487–2496. doi: 10.1109/CVPR.2016.273

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 801–818. doi: 10.1007/978-3-030-01234-2_49

Christ, P. F., Ettlinger, F., Grün, F., Elshaera, M. E. A., Lipkova, J., Schlecht, S., et al. (2017). Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. *arXiv preprint arXiv:1702.05970*.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 424–432. doi: 10.1007/978-3-319-46723-8_49

Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., and Heng, P.-A. (2016). "3D deeply supervised network for automatic liver segmentation from CT volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 149–157. doi: 10.1007/978-3-319-46723-8_18

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., et al. (2019). "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 3146–3154. doi: 10.1109/CVPR.2019.00326

Gordienko, Y., Gang, P., Hui, J., Zeng, W., Kochura, Y., Alienin, O., et al. (2018). "Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer," in *International Conference on Computer Science, Engineering and Education Applications* (Kiev: Springer), 638–647. doi: 10.1007/978-3-319-91008-6_63

Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.-A., Snead, D., et al. (2019). Mild-net: minimal information loss dilated network for gland instance segmentation in colon histology images. *Med. Image Anal.* 52, 199–211. doi: 10.1016/j.media.2018.12.001

Hancock, E. R., and Kittler, J. (1990). Edge-labeling using dictionary-based relaxation. *IEEE Trans. Pattern Anal. Mach. Intell.* 12, 165–181. doi: 10.1109/34.44403

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 7132–7141. doi: 10.1109/CVPR.2018.00745

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203. doi: 10.1038/35058500

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015). "Spatial transformer networks," in *Advances in Neural Information Processing Systems* (Montreal, QC), 2017–2025.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Lake Tahoe, NV), 1097–1105.

Liow, Y.-T. (1991). A contour tracing algorithm that preserves common boundaries between regions. *CVGIP* 53, 313–321. doi: 10.1016/1049-9660(91)90019-L

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3431–3440. doi: 10.1109/CVPR.2015.7298965

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA: IEEE), 565–571. doi: 10.1109/3DV.2016.79

Niazi, M. K. K., Parwani, A. V., and Gurcan, M. N. (2019). Digital pathology and artificial intelligence. *Lancet Oncol.* 20, e253–e261. doi: 10.1016/S1470-2045(19)30154-8

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Park, J., Woo, S., Lee, J.-Y., and Kweon, I. S. (2018). Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 8024–8035.

Raza, S. E. A., Cheung, L., Epstein, D., Pelengaris, S., Khan, M., and Rajpoot, N. M. (2017). "Mimo-net: A multi-input multi-output convolutional neural network for cell segmentation in fluorescence microscopy images," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (Melbourne, VIC), 337–340. doi: 10.1109/ISBI.2017.7950532

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Roth, H. R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E. B., et al. (2015). "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 556–564. doi: 10.1007/978-3-319-24553-9_68

Sharma, N., and Aggarwal, L. M. (2010). Automated medical image segmentation techniques. *J. Med. Phys.* 35:3. doi: 10.4103/0971-6203.58777

Sirinukunwattana, K., Snead, D. R., and Rajpoot, N. M. (2015). A stochastic polygons model for glandular structures in colon histology images. *IEEE Trans. Med. Imaging* 34, 2366–2378. doi: 10.1109/TMI.2015.2433900

Wang, W., and Shen, J. (2017). Deep visual attention prediction. *IEEE Trans. Image Process.* 27, 2368–2378. doi: 10.1109/TIP.2017.2787612

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 7794–7803. doi: 10.1109/CVPR.2018.00813

Woo, S., Park, J., Lee, J.-Y., and So Kweon, I. (2018). "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 3–19. doi: 10.1007/978-3-030-01234-2_1

Wu, H.-S., Xu, R., Harpaz, N., Burstein, D., and Gil, J. (2005). Segmentation of intestinal gland images with iterative region growing. *J. Microsc.* 220, 190–204. doi: 10.1111/j.1365-2818.2005.01531.x

Xu, Y., Li, Y., Liu, M., Wang, Y., Lai, M., Eric, I., et al. (2016). "Gland instance segmentation by deep multichannel side supervision," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 496–504. doi: 10.1007/978-3-319-46723-8_57

Yang, L., Zhang, Y., Chen, J., Zhang, S., and Chen, D. Z. (2017). "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Quebec City, QC: Springer), 399–407. doi: 10.1007/978-3-319-66179-7_46

Yu, J., Yao, J., Zhang, J., Yu, Z., and Tao, D. (2020). Sprnet: Single-pixel reconstruction for one-stage instance segmentation. *IEEE Trans. Cybern.* 1–12. doi: 10.1109/TCYB.2020.2969046

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Granada: Springer), 3–11. doi: 10.1007/978-3-030-00889-5_1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 July 2020
doi: 10.3389/fgene.2020.00659

Comprehensive Network Analysis Reveals Alternative Splicing-Related lncRNAs in Hepatocellular Carcinoma

Junqing Wang[1], Xiuquan Wang[2], Akshay Bhat[3], Yixin Chen[4], Keli Xu[5], Yin-yuan Mo[6], Song Stephen Yi[3] and Yunyun Zhou[7]**

[1] Department of General Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [2] Department of Mathematics and Computer Science, Tougaloo College, Jackson, MS, United States, [3] Department of Oncology and Department of Biomedical Engineering, University of Texas at Austin, Austin, TX, United States, [4] Department of Computer and Information Science, University of Mississippi, Oxford, MS, United States, [5] Department of Neurobiology and Anatomical Sciences, University of Mississippi Medical Center, Jackson, MS, United States, [6] Department of Pharmacology and Toxicology, University of Mississippi Medical Center, Jackson, MS, United States, [7] Department of Data Science, University of Mississippi Medical Center, Jackson, MS, United States

OPEN ACCESS

Edited by:
Yungang Xu,
The University of Texas Health Science Center at Houston, United States

Reviewed by:
Xue Xiao,
The University of Texas Southwestern Medical Center, United States
Renzhi Cao,
Pacific Lutheran University, United States

***Correspondence:**
Song Stephen Yi
stephen.yi@austin.utexas.edu
Yunyun Zhou
yzhou.umc@gmail.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 27 February 2020
Accepted: 29 May 2020
Published: 15 July 2020

Citation:
Wang J, Wang X, Bhat A, Chen Y, Xu K, Mo Y, Yi SS and Zhou Y (2020) Comprehensive Network Analysis Reveals Alternative Splicing-Related lncRNAs in Hepatocellular Carcinoma. Front. Genet. 11:659. doi: 10.3389/fgene.2020.00659

It is increasingly appreciated that long non-coding RNAs (lncRNAs) associated with alternative splicing (AS) could be involved in aggressive hepatocellular carcinoma. Although many recent studies show the alteration of RNA alternative splicing by deregulated lncRNAs in cancer, the extent to which and how lncRNAs impact alternative splicing at the genome scale remains largely elusive. We analyzed RNA-seq data obtained from 369 hepatocellular carcinomas (HCCs) and 160 normal liver tissues, quantified 198,619 isoform transcripts, and identified a total of 1,375 significant AS events in liver cancer. In order to predict novel AS-associated lncRNAs, we performed an integration of co-expression, protein-protein interaction (PPI) and epigenetic interaction networks that links lncRNA modulators (such as splicing factors, transcript factors, and miRNAs) along with their targeted AS genes in HCC. We developed a random walk-based multi-graphic (RWMG) model algorithm that prioritizes functional lncRNAs with their associated AS targets to computationally model the heterogeneous networks in HCC. RWMG shows a good performance evaluated by the ROC curve based on cross-validation and bootstrapping strategies. As a conclusion, our robust network-based framework has derived 31 AS-related lncRNAs that not only validates known cancer-associated cases MALAT1 and HOXA11-AS, but also reveals new players such as DNM1P35 and DLX6-AS1with potential functional implications. Survival analysis further provides insights into the clinical significance of identified lncRNAs.

Keywords: long non-coding RNAs (lncRNA), alternative splicing, multi-graphic random walk, gene-regulatory network analysis, random walk, hepatocellular carcinoma, integrative network analysis

INTRODUCTION

Alternative splicing (AS) events are frequently observed in tumorigenesis and serve as cancer-driving genes. AS can originate from somatic mutations that disrupt splicing regulatory mechanisms or influence the expression levels of splicing factors or transcription factors (Climente-Gonzalez et al., 2017). Hence, AS-associated genes are recognized as important signatures for

tumorigenesis and are of significance in developing therapeutic targets for cancer clinical trial. For example, the SF3B1-targeting compound spliceosome inhibitor E7107 has been implemented in advanced tumor treatment (Eskens et al., 2013).

Studies from Zhang et al. (2016) and Romero-Barrios et al. (2018) showed that long non-coding RNAs [generally more than >200 nucleotides (nt) in length] are associated with a variety of AS mechanisms. lncRNAs may interact with specific alternative splicing factors (ASF) or with other intermediate molecules that affect chromatin remodeling to fine tune the splicing of target genes (Romero-Barrios et al., 2018). For instance, our previous experimental study showed that MALAT1 regulated the ASF, SRSF1 (SF2) in gastric cancer cells (Wang et al., 2014; West et al., 2014). In addition, Ji et al. (2014) reported that MALAT1 promoted tumor growth and metastasis in colorectal cancer through the binding of SFPQ in order to release the oncogene PTBP2. On the other hand, LINC01133 has been reported to interact with splice factor SRSF6 in patients suffering from colorectal cancer (Kong et al., 2016) and non-small cell lung cancer (NSCLC) (Zang et al., 2016).

Proteins that have multiple splicing regulators and that promote the transformation of target genes generally get triggered by transcriptional factors (TFs). For example, the transcription regulator MYC, induces upregulation of hnRNP A1/2, that, in turn, regulates alternative splicing events in expressing the cancer-associated pyruvate kinase M2 (PKM2) isoform (David et al., 2010; Koh et al., 2015). Since lncRNAs occur specifically during pre-transcriptional or post-transcriptional modifications, effectors (such as miRNAs, TFs, or ASFs) that are away from their targets, act as cofactors or guides to alter TF-promoter interactions.

Although studies have identified the correlation of lncRNAs and AS to be important in cancer prognosis, there still remains gaps within current studies as only a few cancer-related AS events are known to be regulated by lncRNAs. In addition, it was not clear how the lncRNAs were linked to specific AS sites, hence, providing no evidence to correlate clinical outcomes. Next-generation sequencing technologies have helped identify ∼40K novel lncRNAs cancer, whose regulatory functions in AS remain unknown in tumorigenesis. Hence, computationally predicting novel lncRNAs and associated alternative splicing events may help in the comprehensive understanding of the HCC disease at a systems level.

In this study, we established an innovative technology for propagating molecular networks called the random walk-based multi-graphic (RWMG) model. The RWMG model simultaneously integrates sophisticated biological connections among lncRNA targets [such as transcription factors (TF), alternative splice factors (ASF), and microRNAs] based on both biophysical interaction networks and their co-expression profiles within a single analytical framework. When comparing conventional random walk algorithms that considers equal proportion of all input genes, our flexible and scalable method can be formulated to rank a subset of lncRNAs based on literature survey. In addition, the method we propose has better accuracy than other previously defined "shortest path" network-based algorithms, with advantages of overcoming "noise" and "incomplete" dimensional heterogenicity from the data.

In addition, previous published reports on comparing tumor and normal tissues are generally limited to normal adjacent tissues (NAT). However, these tissues are not truly "normal" as they are usually surrounded by tumor contaminations. Therefore, many potential cancer biomarkers involved in AS may be missed. Hence, to increase the performance of such analysis, we combined healthy liver tissue samples that were downloaded from GTEx along with expression data from TCGA.

MATERIALS AND METHODS

Data Description and Project Design

The framework of the underlying biological hypothesis and model assumption for this project is described in **Supplementary Figures S1A,B**. The analysis in this manuscript relied on using multi-omics data. We downloaded gene expression data for 110 normal liver samples from the GTEx and TCGA along with clinical information for 369 liver tumors and 50 normal samples from the UCSC Xena database[1]. The sequencing platform for obtaining gene expression was Illumina HiSeq 2000, and pre-processing of raw data was done following the UCSC's Xena Toil (Vivian et al., 2017) method in order to quantify gene and transcript isoform expression. Annotation of coding and non-coding genes was obtained using GENCODE v23 (Harrow et al., 2012).

Identification of HCC Tumor Non-coding Genes (lncRNAs or Pseudogenes) From TCGA and GTEx RNA-seq Data

We performed a method of trimmed mean of *M*-values (TMM) normalization for RNAseq data (Robinson and Oshlack, 2010) so that the expression level for lncRNAs and pseudogenes are comparable. The TMM normalized data was further transformed to log2-counts per million for our linear model. HCC differentially expressed (DE) lncRNAs and pseudogenes between tumor and normal samples (T/N) were analyzed by R package limma (Smyth, 2005) with a statistical cutoff ($p < 1.0E-04$ and fold-change > 2). The identification of DE miRNAs had been reported in our previous work (Wang et al., 2018). The identified HCC-specific expressed features (lncRNAs, pseudogenes, and miRNAs) are expected to represent potential key mechanisms in liver neoplasm.

Analysis of Alternative Splicing Isoforms and Functional Consequence

In order to analyze alternative splice isoforms, we first discarded those isoforms that contained nil values on abundance levels across all the samples. We used R package "IsoformSwitchAnalyzeR" to analyze individual isoform switches

[1]http://xena.ucsc.edu/

from T/N comparison and their biological processes (Vitting-Seerup and Sandelin, 2017, 2018). Differentially switched isoforms between T/N were determined by the following criteria: difference in isoform fraction (dIF) > 0.1 and FDR-corrected q-value < 0.05. The functional consequences of switched isoforms were further analyzed for protein-coding potential (CPAT) (Wang et al., 2013), Nonsense-mediated decay (NMD) status, protein domains (Pfam) (Finn et al., 2015; Potter et al., 2018), and open reading frames (ORF). We used the cutoff 0.364 as suggested to distinguish coding and non-coding isoforms in CPAT analysis. On the other hand, NMD is a process that recognizes mRNAs carrying a premature termination codon (PTC) and that triggers their degradation in order to prevent the synthesis of dysfunctional proteins. AS that controls expression of genes is an important process facilitating mRNA degradation in specific isoforms and could lead to NMD (Cuccurese et al., 2005). Since exon structure of all isoforms in a given gene are with isoform switching capabilities, we obtained their corresponding spliced nucleotide sequence and corresponding coding sequence from ORF positions (Weischenfeldt et al., 2012). The alternative splicing (AS) patterns of switching isoforms were predicted by spliceR (Vitting-Seerup et al., 2014) to include alternative 3′ acceptor sites (A3), alternative 5′ donor sites (A5), exon skipping (ES), mutually exclusive exons (MEE), AS at TF start sites (ATSS), AS at termination site (ATTS), and intron retention (IR). Gene enrichment analysis of features that compared normal vs. tumor samples were performed by following the statistical testing of Fisher's exact-test. P-values were corrected for multiple testing using the Benjamin–Hochberg scheme with an FDR < 0.05.

Construction of AS-Associated lncRNA Epigenetic Regulatory Interaction Subnetworks in HCC

We collected physical interaction information of lncRNAs and associated targeted genes through database searching and text mining. These interactions were evidenced from experimental validations, neighboring gene pairs, gene fusions, and co-occurrence of lncRNAs that connect with miRNA-, TF-, ASF-, and switched genes. Furthermore, HCC lncRNA-target networks were compiled from the following resources: Chiu et al. (2018), miWalker2.0 (Dweep and Gretz, 2015), STARBASE v2 (Li et al., 2013), and lncRNA-disease (Bao et al., 2018) that were analyzed from several high-throughput assays, including ENCODE enhanced version of the crosslinking and immunoprecipitation assay (eCLIP) and chromatin immunoprecipitation sequencing (ChIP-seq) data (Consortium, 2004). HCC-specific miRNA-target networks have been described in our previous published results (Wang et al., 2018); TF-target predicted interaction networks were manually curated from the following databases and publications: Chiu et al. (2018) (**Supplementary Table S5**), HTRIdb (Bovolenta et al., 2012), Whitfield (Whitfield et al., 2012), and TRANSFAC (Matys et al., 2006) that were based on combined evidence from ENCODE ChIP-Seq assays and positioned weighted matrix (PWM) for TF motif analysis.

Features that were enriched in AS regulatory pathways were collected from pathCards (Belinky et al., 2015), KEGG spliceosome (Kanehisa and Goto, 2000), NCBI Biosystems mRNA processing (Geer et al., 2009), REATOME mRNA splicing pathway, and processing of capped intron-containing pre-mRNA pathway (Croft et al., 2010). These features were involved in an essential component of splicing factors or non-snRNA spliceosome required for the second catalytic step of pre-mRNA splicing. Among these collected 335 splicing regulator genes, 86 were experimentally validated as alternative splicing factors (ASF). ASF and target gene interactions were manually confirmed from SpliceAid 2 (Giulietti et al., 2012), ASF motif analysis from SFmap (Paz et al., 2010), a subset of RNA-binding protein network by Chiu et al. (2018) (**Supplementary Table S6**), and STRING database (Franceschini et al., 2012).

Finally, identified HCC-DE lncRNAs, pseudogenes, and miRNAs were mapped to the global regulatory networks to construct HCC-specific sub-networks that contain switched genes as the targets or TF/ASF as the co-effectors of non-coding RNA regulators.

Construction of HCC lncRNA-AS Regulatory Networks at Isoform Level

Pearson correlation was used to estimate the lncRNA co-expression relationships at isoform level. We only included connections for the pairs of lncRNA and protein-coding genes, with absolute correlation coefficient greater than 0.75 and FDR p < 0.05. The types of protein were either TFs, ASFs, or genes with isoform switches. lncRNAs that were negatively correlated with their targeted protein-coding genes were predicted to be inhibitors, while positive correlation indicated activators.

Random Walk Multi-Graphic Model for the Integration of Heterogeneous Interaction Networks

Random walk multi-graphic (RWMG) model is an integrative application of random walk with restart (RWR) algorithm on multiple layers of heterogeneous network. Our framework is encoded with data sets for the same cohort of patients including:

(1) Co-expression network, which is a bipartite graph containing the association between n lncRNA and l AS genes.
(2) Epigenetic regulatory network, which is also a bipartite graph containing the association between n lncRNA and k AS genes ($p \neq$ l). Note: the Epigenetic regulatory network and Co-expression network share the same set of n lncRNA nodes, but the AS genes of the two networks are partially distinct.
(3) Splicing pathway PPI networks, which is an $m \times m$ AS gene–AS gene interaction network with m nodes. The node set is the union of distinct AS genes from the Co-expression and Epigenetic regulatory networks with size m. There is no information about interaction between lncRNA.

We first create an extended graph $G(V, E_k)$ with N *nodes* for each given network, where V is the union of n lncRNA and m AS gene nodes and $N = n + m$, $k = 1,2,3$, which represent co-expression, epigenetic, and splicing pathway PPI networks, respectively. In addition, these were merged into one undirected association network $MG(V, E)$, $E = UE_k$. Multiple edges are allowed to connect between any two nodes based on the relationship defined from networks. Merged network with the overlapped node features and the union of edges will augment each individual network with missing connections. We let A denote the adjacency matrix of a (weighted) molecular interaction multi-graph network $MG(V, E)$. Edge $(i, j) \in E, 1 \leq i, j \leq N$ is weighted by the connectivity score between these vertices. The connectivity score

$$E_{i,j} = \frac{\sum_{k=1}^{3} [E_k]_{ij}}{3}$$

is the average of all included edge scores connecting nodes *i,j*. It is the edge weight to shape the adjacent matrix A.

Each entry B_{ij} in the transition probability matrix B, which stores the probability of a transition from node j to node i, is computed as

$$B_{ij} = \frac{A_{ij}}{\sum_{k=1}^{N} A_{kj}}$$

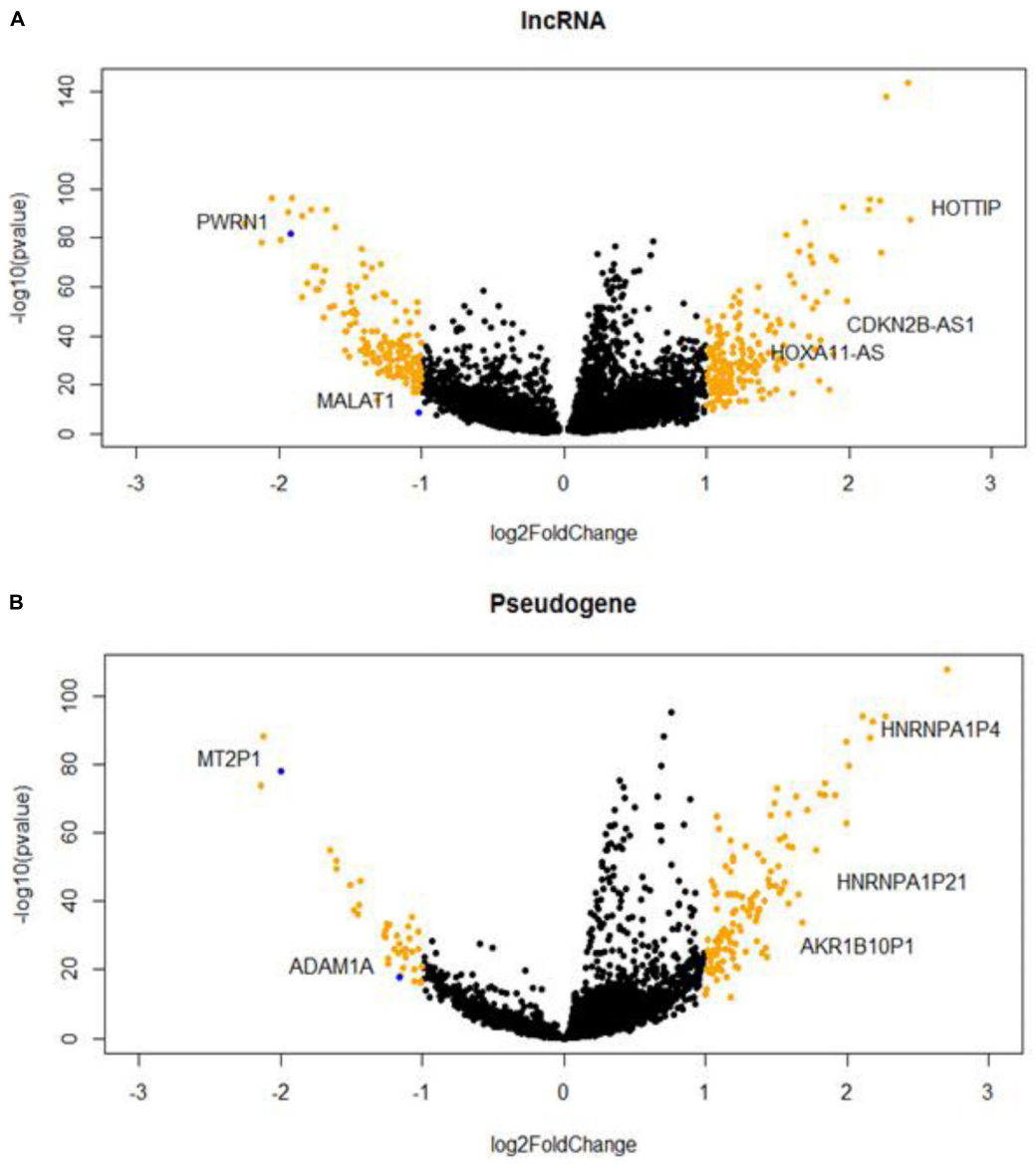


FIGURE 1 | Hepatocellular carcinoma (HCC)-specific long non-coding RNAs (lncRNAs) **(A)** and pseudogenes **(B)** that are differentially expressed in tumor and normal samples.

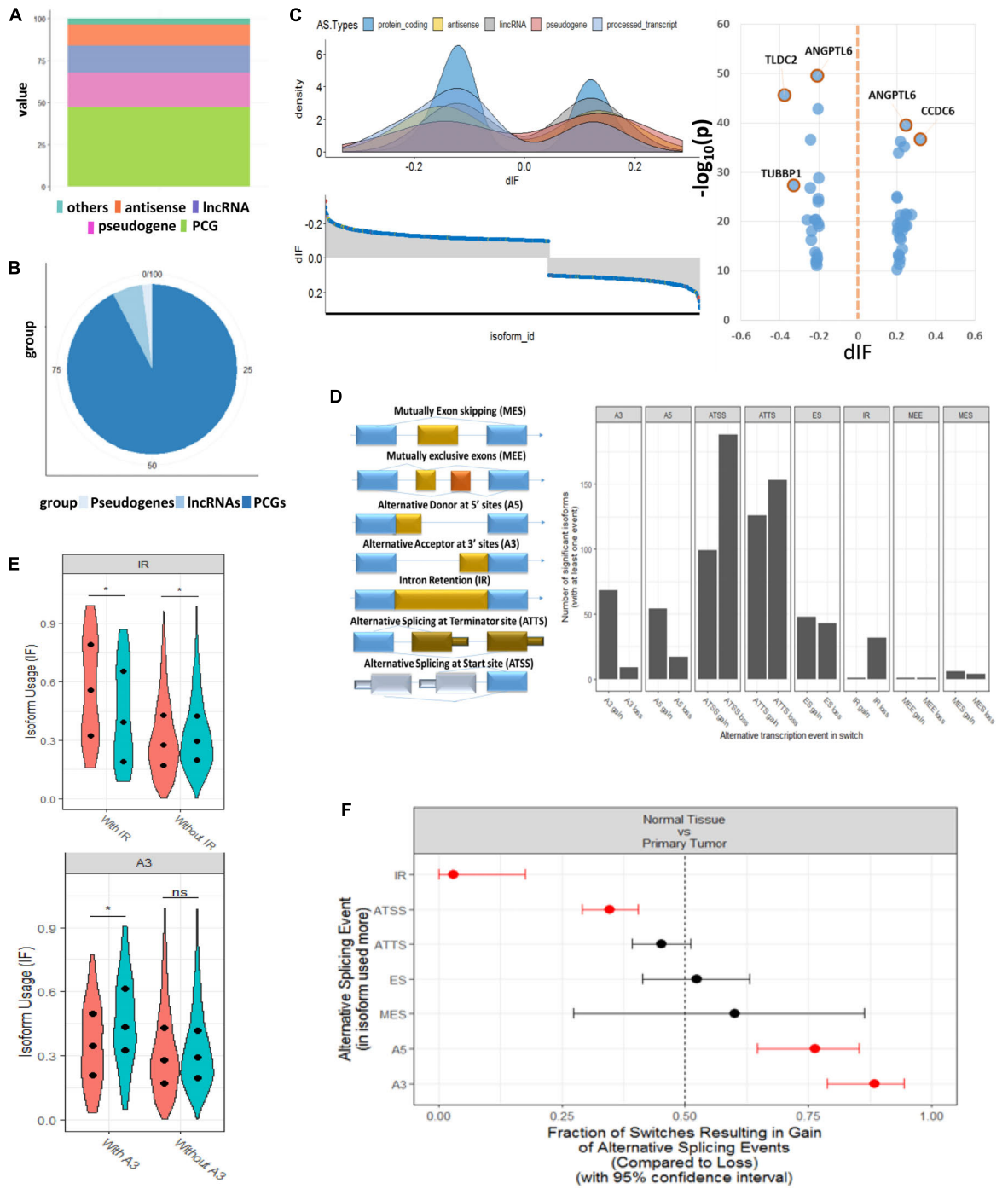


FIGURE 2 | Genome-wide transcript analysis for switched isoforms between tumor vs. normal comparison HCC. **(A)** Global distribution of whole genome transcriptions based on GENCODE annotation. The percentage of coding and non-coding genes is about half and half. **(B)** Distribution of the HCC-switched isoforms in coding and non-coding region. About 95% of switched isoforms are from protein-coding genes. **(C)** Distribution of differential isoform fraction (dIF) stratified by coding or non-coding isoform types. The most significantly switched isoforms (dIF > 0.2) are highlighted. **(D)** Illustration of alternative splicing event types for the switched isoforms and distribution of isoform gain (increased dIF) or loss (decreased dIF) in each types. **(E)** Enrichment analysis for alternative splicing types in isoform fraction gain or loss. Intron retention (IR) and alternative splicing at termination site (ATSS) categories are enriched in loss switches, while A5 and A3 are significantly enriched in gain. **(F)** Distribution of dIF changes with or without IR and A3 events. Isoforms showed less usages in IR type and more usage in A3 type.

Therefore, we can write the RWMG model on a multi- and heterogeneous- graph $MG(V, E)$ as:

$$p^{t+1} = (1 - \alpha)Bp^t + \alpha p_s$$

where vectors p^{t+1} *and* p^t are N-dimensional column vectors where $p^t[i]$ denotes the probability of being at node i and t iteration, and α is the probability of restart (we set $\alpha = 0.5$ in this paper). p_s is an N-dimensional column vector with n lncRNA and m AS gene with p_s (seed) = 1 and others are 0. After a restart step, the particle can go back either to a seed lncRNA feature or to a seed AS gene. We implemented the RWR algorithm on the final multi-graphic network by R package dnet and igraph (Csardi and Nepusz, 2006; Fang and Gough, 2014). Network visualization was performed by R package visNetwork (Guerrieri, 2015). Those genes with known roles in regulating AS network will be set as the "seed" nodes in advance to predict the "new" lncRNAs, based on move probabilities from the current node to any of their randomly selected neighbors.

To evaluate our approach's sensitivity, we simulated different random walk strategies for optimization. We created a list of experimentally validated AS-associated genes as "gold-standard" true positive genes (TPG) curated from the careful literature

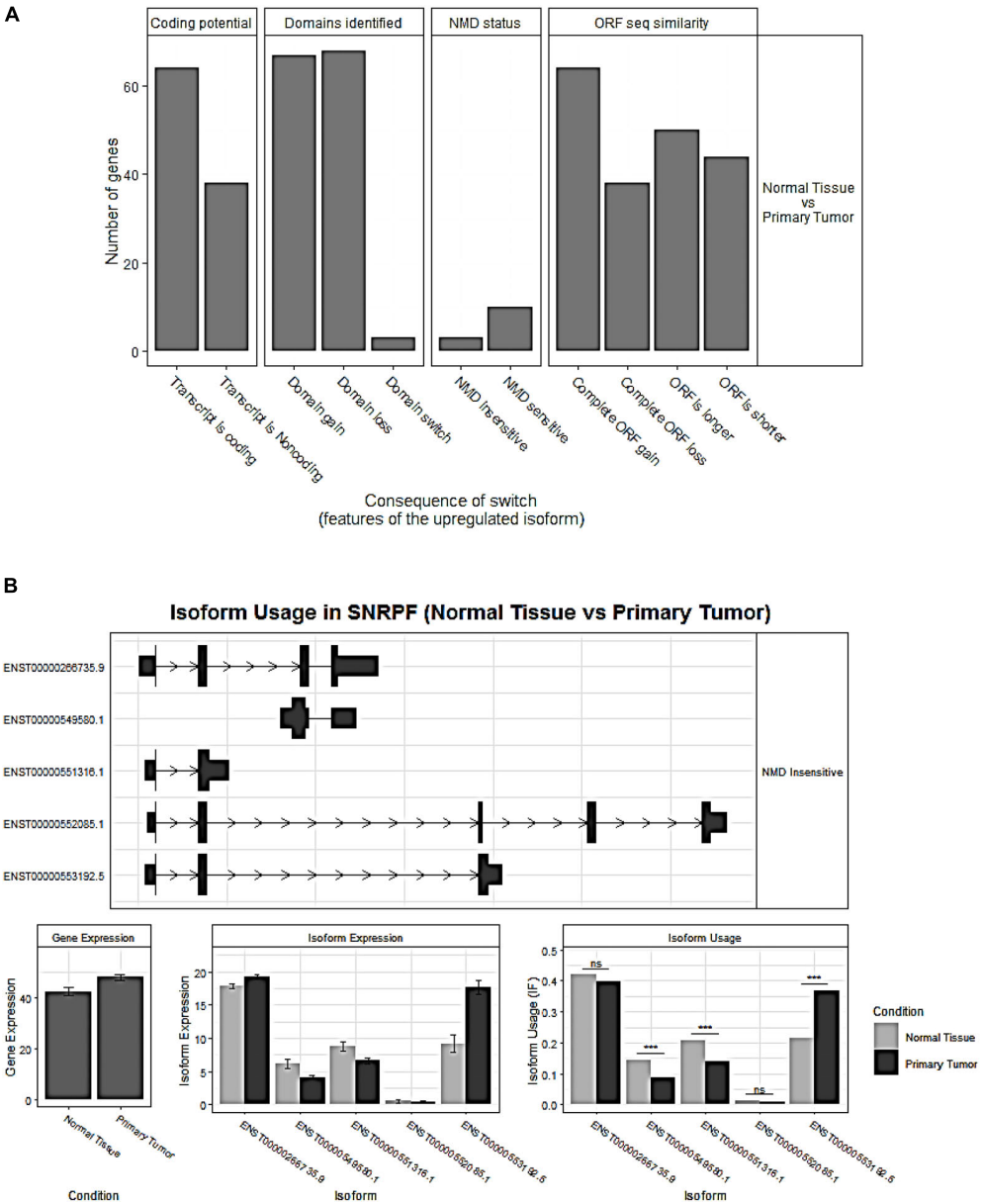


FIGURE 3 | (A) Overview of the number of switched isoforms predicted to have functional consequences. **(B)** Visualization of switched isoform structure. Taking a splicing factor gene, SNRPF, for example, its isoform ENST00000553192.5.1 showed opposite switching pattern compared to others. In addition, three out of five isoforms showed differential isoform expressions, although no difference for the overall gene expression.

review and randomly selected genes as the "gold-standard" true negative (TNG). We chose the "best" model that has the most candidates significantly enriched in the "gold-standard" gene list. In reality, the number of TPG is much smaller compared to TNG. To avoid bias from highly imbalanced data between these two sets, we performed a bootstrap resampling technique by selecting an equal number of data as TNG. This process was repeated 10 times, and the overall performances were calculated by the mean value of these performances.

Survival Analysis for Prognostic Confirmation of Identified Pathogenic lncRNAs and Pseudogenes

To confirm the pathogenic characteristics of identified lncRNAs and pseudogenes, univariate Cox proportional model was used to evaluate the association of selected genes with overall survival outcomes. Kaplan–Meier plots and log-rank test statistics were used to visualize the high- and low-risk groups. The cutoff of the high- and low-risk group was determined by the median value of the normalized count of selected genes.

RESULTS

Differentially Expressed lncRNAs and Pseudogenes in HCC

We identified 369 DE lncRNA genes and 171 DE pseudogenes from T/N comparison (**Supplementary Table S1**). The visualizations of DE lncRNAs and pseudogenes were shown in volcano plots (**Figure 1**). According to literature survey, many DE lncRNAs, such as MALAT1, CDKN2B-AS1, and HOTTIP, have been reported to be associated with liver cancers (Kunej et al., 2014; Quagliata et al., 2014; Guerrieri, 2015). In addition, we highlighted several important pseudogenes, such as HNRNPA1P4 and HNRNPA1P21, which are heterogeneous nuclear ribonucleoproteins A1 (hnRNPs) that play key roles in the regulation of alternative splicing. Furthermore, we performed DE analysis as an initial screen step to narrow the focus of the HCC-specific non-coding genes associated with AS for the downstream network analysis.

Identification of Significant Switched Isoforms and Prediction of Alternative Splicing Patterns

From the expression levels of isoform when comparing tumor and normal samples, we identified 1,375 isoforms that had switching properties and that mapped to 1,078 unique genes. Among these switched isoforms, 1,251 were protein-coding isoforms, and 124 were non-coding isoforms that included antisense, lncRNA and pseudogenes (**Supplementary Table S2**). We found that the proportion of switching rate for coding genes was much higher than that for non-coding genes (Fisher's exact test, $p = 8.4e\text{-}08$) (**Figures 2A,B**). In order to visualize the splicing composition of these switched isoforms, we broke down the dIF distribution according to isoform types such as lncRNA, antisense, and pseudogenes with the most significant switched isoforms (dIF > 0.2 or dIF < -0.2) highlighted in **Figure 2C**.

Figure 2D shows the eight splicing patterns for switched isoforms stratified by isoform usage gain or loss in the tumor. Some of the switched isoforms are predicted to have multiple AS events in HCC (**Supplementary Table S3**). Interestingly, we observed a global phenomenon that the AS events are not equally used—most prominently illustrated by the use of ATSS in HCC, where there was more losses than the gain of amino acid coding exons. It should be taken into consideration that IR and ATSS were enriched in significant low isoform usage in tumor, but A5 and A3 were significantly enriched in the gain isoform (**Figure 2E**). Here, IR events were of particular functional interest since they represented the largest changes in isoforms. As we show in the violin plots, the enriched IR and A3 splicing groups reported significant opposite directions of isoform usages between T/N samples (**Figure 2F**).

Analysis of Functional Consequences for Switched Isoforms

The overview of switched isoforms impacting the biological function alterations in HCC is shown in **Figure 3A**. The number of protein domain gains was comparable to domain loss, but is significantly more than domain "switch." Here, the "switch" term indicates both a gain and a loss occurrence. Also, switching resulting in ORF gain was significantly more than ORF loss. For the Gene Ontology analysis, both gain and loss switched isoforms were associated with different types of metabolic processes.

TABLE 1 | Statistic summary of splicing factor genes with alternative switched isoforms.

Isoform_id	Gene_id	Gene_name	dIF	q_value
ENST00000555295.1	ENSG00000100836.10	PABPN1	0.182	1.10E–32
ENST00000459687.5	ENSG00000100410.7	PHF5A	0.172	6.07E–18
ENST00000411938.1	ENSG00000128534.7	LSM8	0.169	2.49E–19
ENST00000553192.5	ENSG00000139343.10	SNRPF	0.152	6.22E–21
ENST00000297157.7	ENSG00000164610.8	RP9	0.145	2.68E–19
ENST00000491106.1	ENSG00000060688.12	SNRNP40	0.128	4.28E–19
ENST00000560313.2	ENSG00000090470.14	PDCD7	0.124	2.17E–06
ENST00000301785.5	ENSG00000214753.2	HNRNPUL2	0.116	1.19E–28
ENST00000402849.5	ENSG00000100028.11	SNRPD3	0.113	1.67E–11
ENST00000535326.1	ENSG00000110107.8	PRPF19	0.103	1.79E–07
ENST00000597776.1	ENSG00000130520.10	LSM4	0.102	2.31E–34
ENST00000472237.5	ENSG00000132792.18	CTNNBL1	0.102	5.80E–13
ENST00000548994.1	ENSG00000075188.8	NUP37	0.101	1.68E–11
ENST00000564651.5	ENSG00000102978.12	POLR2C	0.1	3.01E–14
ENST00000505885.1	ENSG00000096063.14	SRPK1	−0.108	1.94E–11
ENST00000404603.5	ENSG00000100028.11	SNRPD3	−0.109	2.30E–15
ENST00000540127.1	ENSG00000214753.2	HNRNPUL2	−0.116	2.99E–48
ENST00000367208.1	ENSG00000182004.12	SNRPE	−0.13	1.79E–31
ENST00000527554.2	ENSG00000100697.14	DICER1	−0.139	2.98E–21
ENST00000595761.1	ENSG00000213024.10	NUP62	−0.157	3.62E–31
ENST00000488937.1	ENSG00000136875.12	PRPF4	−0.159	6.94E–12
ENST00000559051.1	ENSG00000090470.14	PDCD7	−0.163	9.60E–15
ENST00000216252.3	ENSG00000100410.7	PHF5A	−0.216	4.30E–21

KEGG analysis showed that the isoform loss in tumor tissue was associated with virus infection, hepatitis C, etc, while isoform gain in the tumor is associated with base excision repair, apoptosis, etc. (**Supplementary Table S2**).

Importantly, we confirmed 20 genes with switched isoforms that were involved in AS regulatory functions (**Table 1**). **Figure 3B** shows one example of AS factor, SNRPF and its isoform structures, gene expression, and isoform usage when

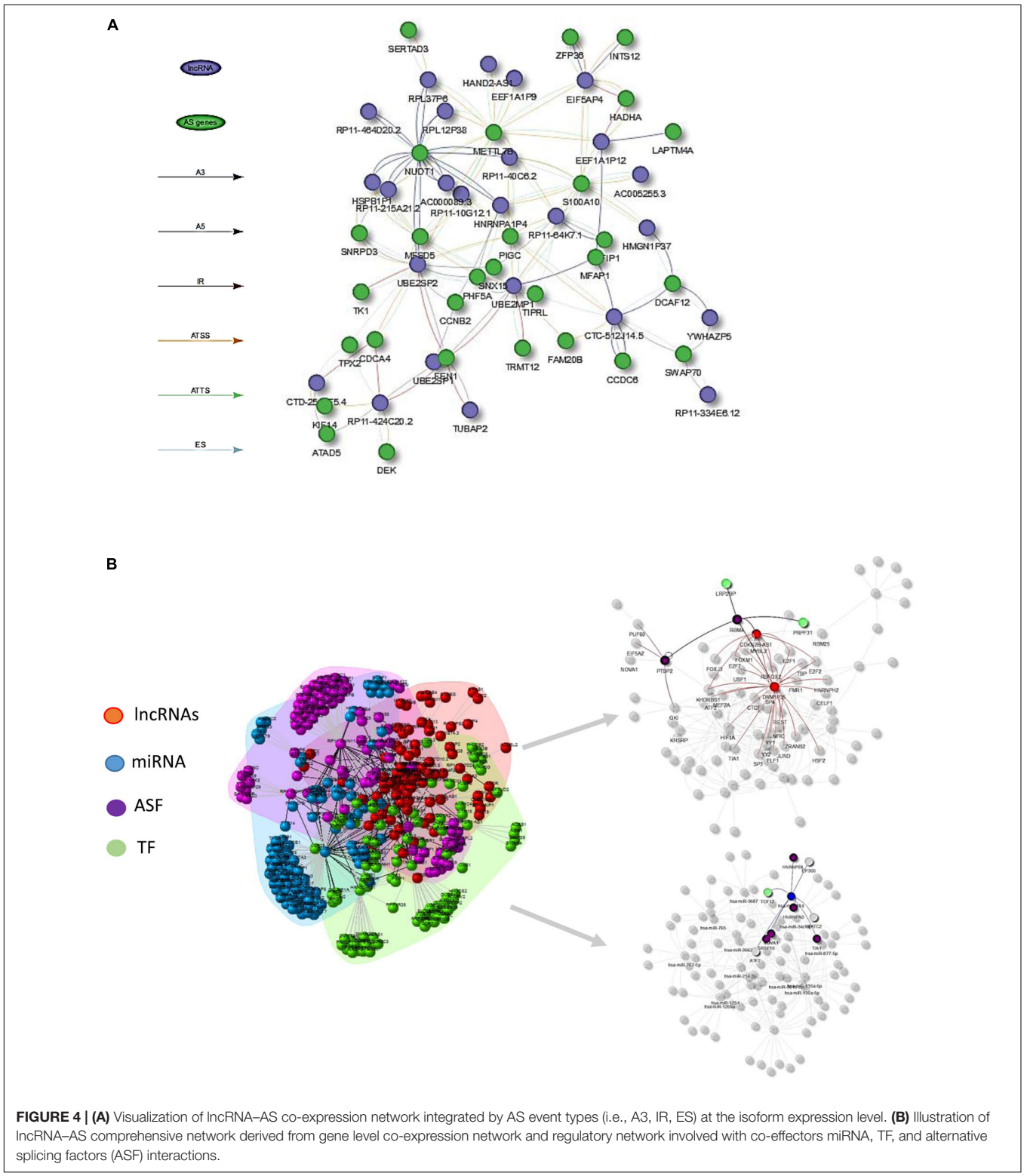


FIGURE 4 | (A) Visualization of lncRNA–AS co-expression network integrated by AS event types (i.e., A3, IR, ES) at the isoform expression level. **(B)** Illustration of lncRNA–AS comprehensive network derived from gene level co-expression network and regulatory network involved with co-effectors miRNA, TF, and alternative splicing factors (ASF) interactions.

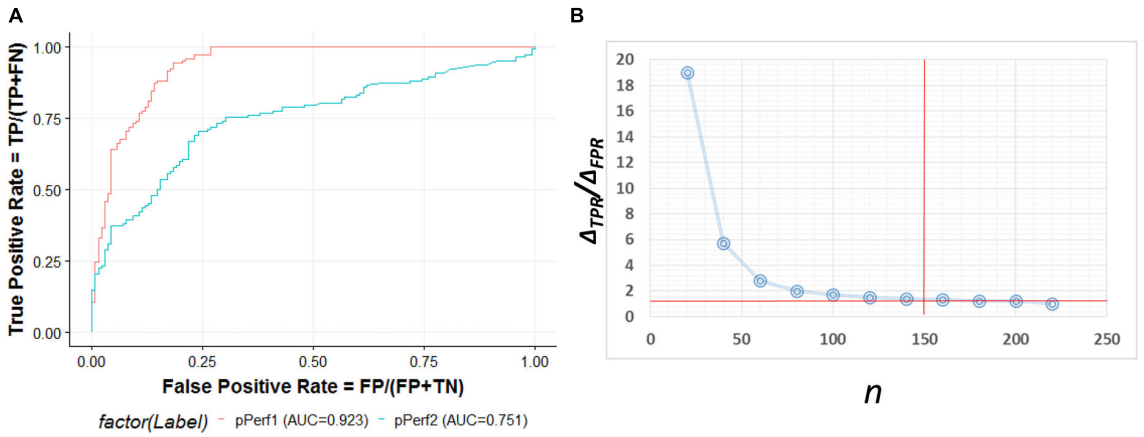


FIGURE 5 | (A) ROC curve for the predictive model evaluation. pPerf1 [area under curve (AUC) = 0.923] with the "seed" genes showed a better performance than pPerf2 (AUC = 0.751) without the "seed" genes. **(B)** Trade-off between the sensitivity and specificity with the number of top *n* genes. We can see that the best cutoff is *n* = 150, as the ΔTPR/ΔFPR value decreases very fast in the beginning and approaches smaller changes for n around 150.

comparing tumor vs normal. SNRPF is a core component of U small nuclear ribonucleoproteins that are key components of the pre-mRNA processing spliceosome. We found no significant difference for SNRPF gene expression; on the contrary, it had opposite directions in expression pattern for transcript ENST00000553192. The above evidences showed that genes with switched isoforms were often functionally important in tumorigenesis and had been ignored from previous reports.

Prediction of AS-Correlated Non-coding RNAs at Both Transcript and Gene Level

In order to identify which lncRNAs were associated to switched isoforms at the transcript level, we constructed a co-expression network that comprised lncRNA and genes with switched isoforms. Different from traditional gene level co-expression network, the connections between lncRNA and genes with multiple splicing isoforms could be singular or multiple when interacting between molecules. The lncRNAs-switched isoform connections are summarized in **Supplementary Table S3**. The relationships between lncRNAs and genes with enriched AS patterns is illustrated in **Figure 4A**.

However, since the lncRNA regulation mechanism involved in AS events was comprehensive, AS regulation may not directly be reflected from expression abundance, but through physical interaction or DNA/RNA binding sites. LncRNAs could influence gene-splicing patterns by inhibiting and activating the expression of ASFs, or through transcription factors that indirectly interact with splicing factors and ultimately cause changes in AS factor-targeted gene expression. Hence, constructing a comprehensive gene regulatory network that includes TF, AS regulators, and lncRNAs could allow better understanding of the mechanism of AS in cancers.

Figure 4B illustrates the HCC lncRNAs–AS network with interactors such as TFs, ASFs, and miRNAs based on evidence from publicly available resource and gene-level co-expression analysis. Only lncRNAs that directly altered AS gene expression or indirectly altered AS genes through TF, ASF, or miRNAs were included for downstream RWMG analysis. **Supplementary Table S4** provides the prediction of all AS-related genes ranked by RWMG-predicted score. **Supplementary Table S6** provides the total number of nodes and edges for the three types of networks.

Computational and Clinical Validation for Predicted Pathogenic lncRNAs Involved in AS Regulation

The ROC curve shown in **Figure 5A** contains an optimized averaged area under curve (AUC) value from 0.751 to 0.923 based on bootstrapping algorithm. In order to select the best number of top *n* ranked genes that corresponded to a fair tradeoff between sensitivity and specificity, we selected a cutoff based on the trend of the changes at $\Delta TPR/\Delta FPR$ that exhibited a sudden drop (**Figure 5B**). We also see from the figure that $n = 150$ is the best number for selecting genes. The top ranked lncRNAs associated with AS functions are described in **Table 2**.

Among the top predicted lncRNAs that were involved in AS, we further confirmed their clinical significance. As a result of univariate survival analysis screening, a total of 51 lncRNAs and 24 pseudogenes were found to be associated with HCC overall 5-year survival, respectively (**Supplementary Table S5**). **Figures 6A,B** show the top 10 significant genes based on the Cox proportional regression model. **Figures 6C,D** show the survival curve and distribution of CDKN2B-AS1 and UBE2SP1.

DISCUSSION

In the last decade, studies have investigated the association of splicing isoforms and lncRNA profiles from deep sequencing technologies. For instance, it has been known that some small nuclear uridine (U)-rich RNAs (snRNPs) are core components of the pre-mRNA processing spliceosome and can collaborate with some splicing factors that are encoded by heterogeneous

TABLE 2 | Statistic summary of predicted top-ranked non-coding RNAs associated with alternative splicing (AS) ranking by random walk-based multi-graphic (RWMG) score.

Gene. symbol	Ranking	Score	Types
LINC00675	1	0.00368174	LincRNA
CTD-2171N6.1	2	0.002824633	LincRNA
HOTTIP	3	0.002677841	Antisense
DNM1P35	4	0.002483954	Antisense
LEF1-AS1	5	0.002397948	Antisense
AP006285.7	6	0.002123275	LincRNA
WARS2-IT1	7	0.002081091	Antisense
LINC00355	8	0.002063983	LincRNA
RP11-81H3.2	9	0.002032482	LincRNA
HOXA11-AS	10	0.002028198	Antisense
RP11-261N11.8	11	0.002005615	Antisense
RP3-355L5.4	12	0.001938399	Antisense
RP11-138J23.1	13	0.001929433	LincRNA
RP11-525K10.3	14	0.001923733	Antisense
RP11-495P10.7	15	0.001917542	LincRNA
DLX6-AS1	16	0.00188837	Antisense
RP11-356C4.5	17	0.001861006	LincRNA
CDKN2B-AS1	18	0.001856714	Antisense
RP11-495P10.5	19	0.001834267	LincRNA
SFTA1P	20	0.001751498	LincRNA
PRSS51	21	0.001750058	Antisense
MALAT1	22	0.001672339	LincRNA
FEZF1-AS1	23	0.001669135	Antisense
RP4-530I15.9	24	0.001619806	Antisense
RP11-158M2.5	25	0.001618054	Antisense
CTD-2374C24.1	26	0.001617345	LincRNA
PWRN1	27	0.001605646	LincRNA
CTC-573N18.1	28	0.001534221	LincRNA
RP11-284F21.9	29	0.001527715	LincRNA
RP11-3J1.1	30	0.001523171	LincRNA
FENDRR	31	0.001509286	LincRNA

nuclear ribonucleoprotein complex subunits (hnRNPs) in order to fine tune complex splicing regulations (Romero-Barrios et al., 2018). Impressively, we found a number of core snRNP isoforms including SNRPE, SNRPD3, SNRPD3, SNRPF, and SNRNP40 that were switched even though their expression was not necessarily DE when comparing tumor vs. normal specific to HCC progression. SNRNP40 catalyzes the removal of introns from pre-messenger RNAs. Similarly, an hnRNP U like protein HNRNPUL2 that also has a scaffold attachment factor, plays an important role in the formation of a "transcriptional" complex binding through the scaffold attachment region and causes chromatin remodeling.

The primary mechanisms involving lncRNAs in AS modulation can be classified into three ways that include: (i) lncRNAs that directly influence isoform expression through activation or inhibition mechanism; (ii) lncRNAs that form RNA–RNA duplexes with pre-mRNA molecules, and (iii) lncRNAs that affect target AS genes through indirectly inhibiting or promoting the expression of splicing factors or through transcript factors. However, most previous studies only focus on individual genes and/or isoform switches regulated by lncRNAs. More comprehensive interactions can be detected at the isoform level besides the gene level. Our predictions identified several candidates that were either oncogenes or tumor suppressors and lncRNAs whose somatic alterations were associated with AS at both isoform and gene level in addition to showing clinical significance in HCC patients.

At the transcriptional level correlation network, we found that majority of lncRNA isoforms were correlated with more than one AS event, among which some were showing opposite roles in the AS regulations. In addition, we can see that many lncRNAs may partially compete with the same AS event. For example, the pseudogenes of UBE2S, which are UBE2SP1, UBE2SP2, and UBE2MP1, are significantly correlated with FEN1's intron retention and Alternative 5′ donor site mechanisms (**Figure 4A**). The FEN1 gene plays an important role in removing 5' overhanging flaps and the 5–3 exonuclease activities involved in DNA replication and repair (Wang et al., 2017), while the UBE2S is involved in ubiquitination and subsequent degradation of VHL, which results in an accumulation of HIF1A (Jung et al., 2006). However, the reason for pseudogenes being associated with FEN1 is not yet clear. Further research in regard to perform experimental validation for predicted mechanisms from our analysis is necessary. Taken together, these results confirmed that the identified lncRNAs need to be better investigated in experimental settings. Our results provided a better resolution of AS-correlated lncRNAs at the isoform level.

AS events are mainly regulated by splicing factors, which bind to pre-mRNAs and influence exon selection and splicing site choice. Moreover, TFs activate or suppress the expression of ASF. Importantly, we found ASF that may have switched isoforms. A switched ASF RP9, which can be bound by the proto-oncogene PIM1 product, a serine/threonine protein kinase, also can cause its target PIM1 to get switched. Although TFs were usually thought for a long time to encode a single protein that changes the expression of their target genes, more and more TFs are now found to be alternatively spliced (Marcel and Hainaut, 2009). Here, we also found a group of TFs in the ETS family (E26 transformation-specific), which are ETS1, ETS2, ETV3, ELF4, which were switched simultaneously. These ETS genes have been confirmed to be associated with cancer through gene fusion (Tomlins et al., 2005) and are involved in a wide variety of regulatory functions such as cell migration, proliferation, and cancer progression (Sharrocks, 2001; Lee et al., 2005). Interestingly, the ETS1 targets splicing factor QKI, and ETV3 targets splicing factor CELF1. Furthermore, lncRNA FAM99B is predicted to be associated within the ETS family genes, and their low expression is associated with HCC patients that had poor prognosis.

The association of CDKN2B-AS1, also known as ANRIL, with HCC has been reported in several studies (Hua et al., 2015; Chiu et al., 2018; Ma et al., 2018). CDKN2B-AS1 has both linear and circular isoforms, and their functions are different. For example, its linear isoform can regulate the c-myc-enhancer-binding factor RBMS1 (Hubberten et al., 2018), while its circular isoform is confirmed to be an important AS

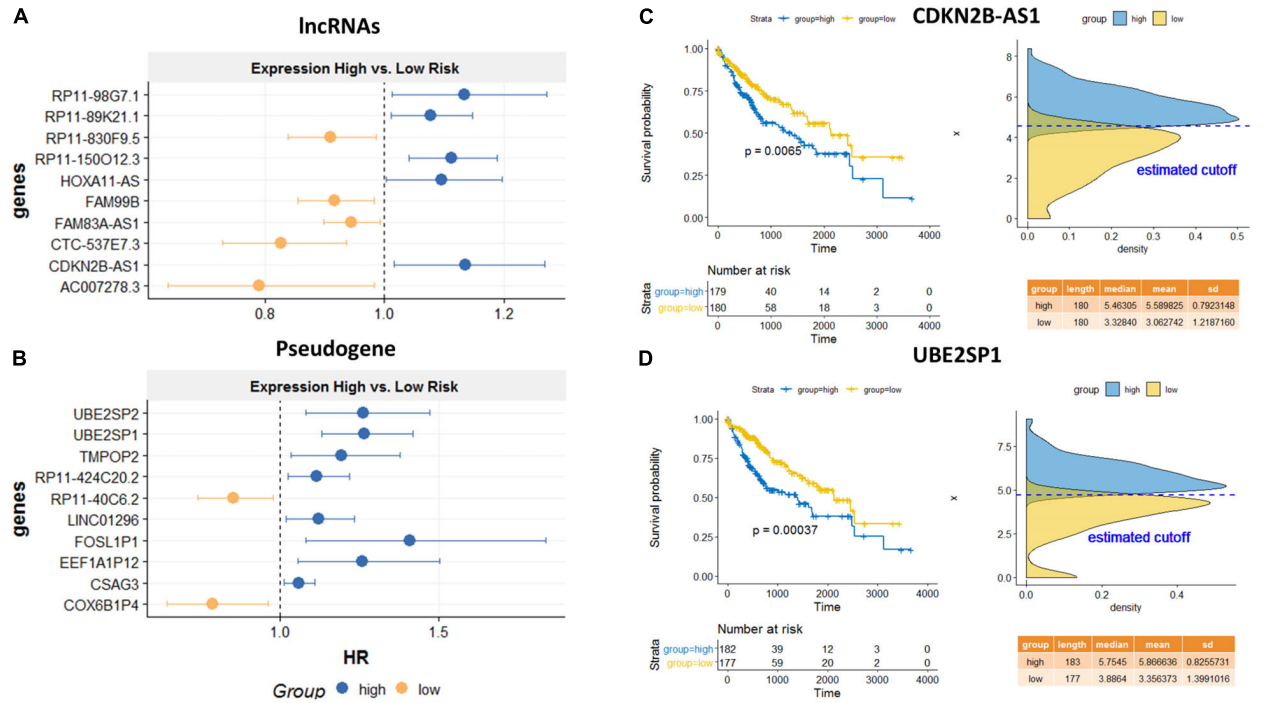


FIGURE 6 | Survival analysis for the identified lncRNA and pseudogenes involved in AS mechanisms. Hazard ratio plots from Cox regression analysis for top 10 lncRNAs **(A)** and top 10 pseudogenes **(B)** associated with overall survival. K–M curves for a lncRNA, CDKN2B-AS1 **(C)**, and a pseudogene, UBE2SP1 **(D)** using the median value as the cutoff. High expressions of both genes are significantly associated with poor prognosis.

regulator that causes skipping of exons (Holdt et al., 2016) and are mainly found in cardiovascular disease (Burd et al., 2010; Sarkar et al., 2017). However, this is the first time we found that ANRIL can activate alternative splicing genes in liver cancer. A potential explanation could be because of being functionally related to lipid metabolism and a majority of liver cancer subtypes. In addition, the prognostic value of CDKN2B-AS1 was revealed in our project. However, how exactly CDKN2B-AS1 controls this gene splicing is not yet clear. Further experimental validation is warranted. We identified HAND2-AS1 gene to show consistent alternative splicing pattern at the start sites and termination site for METTL7B especially at the isoform level. METTL7B is a membrane-associated protein that resides in hepatic lipid droplets. An explanation for this is that HAND2-AS1 activates the METTL7B spliced isoform lipid disordered and is associated with HCC, which was not reported before. Gene-level RWMG network analysis further revealed that both CDKN2B-AS1 and HAND2-AS1 can influence AS either through TFs and ASFs some of which include HAND2-AS1 TFs (i.e., ETS1, SP1, E2F7) or ASFs (i.e., SRSF7, SFRP1, HNRNPK); and CDKN2B-AS1-associated TFs (SP4, E2F7) and ASF (SRSF1, SRSF2).

In this project, we extended a previous existing algorithm into multiplex and heterogeneous networks. The research community can explore different layers of the epigenetic regulatory network, expression correlation network, and protein interaction network. A recent Nature Review paper by Cowen et al. (2017) also suggested that the "network-propagation" method was a "powerful" and "accurate" refined approach in the network biology, since it is capable of dealing with "noise" and "incomplete" observations by simultaneously considering all possible paths among vertices. Analyzing these heterogeneous data together will significantly improve the prediction accuracy of our method. By using this gene-ranking strategy, potentially spurious predictions (false positives) that are supported by a single (shortest) path are down-weighted, and true high-ranked genes that are potentially missed, even though they are well connected to the prior list (false negatives), are promoted.

To our best knowledge, this is the first attempt to predict lncRNA regulations on AS using a rigorous, multi-graphic approach by the integration of large-scale and complex networks. Of interest for potentially limiting the accuracy of random walk and network propagation methods are an incomplete collection of known lncRNAs, especially pseudogenes, used to supervise prediction of new candidates. As such, we addressed several unique challenges associated with these dataset complexities in each step. For example, in the data preprocessing steps, we carefully address the challenges by collecting as many as experimentally verified and predicted lncRNAs that were taking account of AS. In our statistical modeling steps, we specifically addressed the robustness of complex data integration, especially for non-informative or noisy datasets. Also, we investigated several random walk strategies by trying different groups of vertices such as lncRNAs, ASFs, and TFs as a starting point to optimize our models.

However, the lncRNA regulatory mechanism is complicated, as its mechanism differs with different stages, such as the pre-mRNA or post-mRNA stage. Therefore, the major limitation of this article is we were not able to consider other comprehensive mechanisms at different stages, such as recognition of the splicing site can be modulated by *cis*-regulatory sequences, known as splicing enhancers or silencers, which contribute to the generation of two or more alternatively spliced mRNAs from the same pre-mRNA. Also, lncRNA determines AS patterns through chromatin remodeling mechanism and shapes the three-dimensional genome organization. We will focus on interpreting these molecular mechanisms of lncRNA and associated AS at different stages of HCC in the near future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. TCGA and GTEx data can be found in data hubs in https://xenabrowser.net/.

AUTHOR CONTRIBUTIONS

JW contributed to the analysis and interpretation of data. XW contributed to mathematics model interpretation. AB reviewed and edited the manuscript. AB and SY contributed to data analysis, provided feedback, and edited the manuscript. YC and KX contributed to the machine learning predictive model design. YM contributed to the project design and interpretation of biological meanings. YZ led the project, provided the guidance, and prepared the manuscript. All the authors read and approved the manuscript.

FUNDING

This study was supported by grants from the University of Mississippi Medical Center Intramural Research Support Program for Clinical Population Science fund (No. 51002630519 to YZ), the National Natural Science Foundation of China (No. 81602544 to JW), and the Shanghai Pujiang Talent Project (No. 18PJD029 to JW).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at SSRN: https://ssrn.com/abstract=3335849 by Wang et al. (2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00659/full#supplementary-material

FIGURE S1 | (A) Illustrations of overall project design, and **(B)** explanation of biological mechanisms.

TABLE S1 | Statistic summaries for significantly differential expression genes between tumor and normal comparison for lncRNAs and pseudogenes.

TABLE S2 | Statistic summaries and functional analysis for significantly switched isoforms between tumor and normal comparison for lncRNA and protein-coding genes, as well as prediction of splicing event patterns for switched genes. Gene ontology and KEGG pathway enrichment analysis are performed for upregulated isoforms (gain) and downregulated isoforms (loss), respectively.

TABLE S3 | Predicted lncRNA interaction pairs at the transcriptional co-expression level.

TABLE S4 | Statistic summary of AS-associated lncRNAs and PCGs ranking by RWMG predictive score.

TABLE S5 | Statistic summary of significant lncRNAs and pseudogenes associated with overall survival.

TABLE S6 | Total number of node and edge distribution for the three networks.

REFERENCES

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2018). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* 47, D1034–D1037.

Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., et al. (2015). PathCards: multi-source consolidation of human biological pathways. *Database* 2015:bav006.

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* 13:405. doi: 10.1186/1471-2164-13-405

Burd, C. E., Jeck, W. R., Liu, Y., Sanoff, H. K., Wang, Z., and Sharpless, N. E. (2010). Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet.* 6:e1001233. doi: 10.1371/journal.pgen.1001233

Huang, M. D., Chen, W.-M., Qi, F.-Z., Xia, R., Xia, R., Sun, M., et al. (2015). Long non-coding RNA ANRIL is upregulated in hepatocellular carcinoma and regulates cell proliferation by epigenetic silencing of KLF2. *J. Hematol. Oncol.* 8:50.

Chiu, H.-S., Somvanshi, S., Patel, E., Chen, T.-W., Singh, V. P., Zorman, B., et al. (2018). Pan-Cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context. *Cell Rep.* 23, 297.e12–312.e12.

Climente-Gonzalez, H., Porta-Pardo, E., Godzik, A., and Eyras, E. (2017). The functional impact of alternative splicing in cancer. *Cell Rep.* 20, 2215–2226.

Consortium, E. P. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science* 306, 636–640.

Cowen, L., Ideker, T., Raphael, B., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* 18, 551–562. doi: 10.1038/nrg.2017.38

Croft, D., O'kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., et al. (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 39(Suppl._1), D691–D697.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Inter. J. Complex Syst.* 1695, 1–9.

Cuccurese, M., Russo, G., Russo, A., and Pietropaolo, C. (2005). Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucleic Acids Res.* 33, 5965–5977.

David, C. J., Chen, M., Assanah, M., Canoll, P., and Manley, J. L. (2010). HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature* 463, 364–368.

Dweep, H., and Gretz, N. (2015). miRWalk2. 0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods* 12:697.

Eskens, F. A., Ramos, F. J., Burger, H., O'Brien, J. P., Piera, A., de Jonge, M. J., et al. (2013). Phase I, pharmacokinetic and pharmacodynamic study of the first-in-class spliceosome inhibitor E7107 in patients with advanced solid tumors. *Clin. Cancer Res.* 19, 6296–6304.

Fang, H., and Gough, J. (2014). Thednet'approach promotes emerging research on cancer patient survival. *Genome Med.* 6:64.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Coggill, P., Heger, A., Pollington, J. E., et al. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2012). STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815.

Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., et al. (2009). The NCBI biosystems database. *Nucleic Acids Res.* 38(Suppl._1), D492–D496.

Giulietti, M., Piva, F., D'Antonio, M., D'Onorio De Meo, P., Paoletti, D., Castrignanò, T., et al. (2012). SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res.* 41, D125–D131.

Guerrieri, F. (2015). Long non-coding RNAs era in liver cancer. *World J. Hepatol.* 7, 1971–1973.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* 22, 1760–1774.

Holdt, L. M., Stahringer, A., Sass, K., Pichler, G., Kulak, N. A., Wilfert, W., et al. (2016). Circular non-coding RNA ANRIL modulates ribosomal RNA maturation and atherosclerosis in humans. *Nat. Commun.* 7:12429.

Hua, L., Wang, C.-Y., Yao, K.-H., Chen, J.-T., Zhang, J.-J., and Ma, W.-L. (2015). High expression of long non-coding RNA ANRIL is associated with poor prognosis in hepatocellular carcinoma. *Int. J. Clin. Exp. Pathol.* 8, 3076–3082.

Hubberten, M., Bochenek, G., Chen, H., Häsler, R., Wiehe, R., Rosenstiel, P., et al. (2018). Linear isoforms of the long noncoding RNA CDKN2B-AS1 regulate the c-myc-enhancer binding factor RBMS1. *Eur. J. Hum. Genet.* 27, 80–89.

Ji, Q., Zhang, L., Liu, X., Zhou, L., Wang, W., Han, Z., et al. (2014). Long non-coding RNA MALAT1 promotes tumour growth and metastasis in colorectal cancer through binding to SFPQ and releasing oncogene PTBP2 from SFPQ/PTBP2 complex. *Br. J. Cancer* 111, 736–748.

Jung, C.-R., Hwang, K.-S., Yoo, J., Cho, W. K., Kim, J. M., Kim, W. H., et al. (2006). E2-EPF UCP targets pVHL for degradation and associates with tumor growth and metastasis. *Nat. Med.* 12, 809–816.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Koh, C. M., Bezzi, M., Low, D. H., Ang, W. X., Teo, S. X., Gay, F. P., et al. (2015). MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. *Nature* 523, 96–100.

Kong, J., Sun, W., Li, C., Wan, L., Wang, S., Wu, Y., et al. (2016). Long non-coding RNA LINC01133 inhibits epithelial–mesenchymal transition and metastasis in colorectal cancer by interacting with SRSF6. *Cancer Lett.* 380, 476–484.

Kunej, T., Obsteter, J., Pogacar, Z., Horvat, S., and Calin, G. A. (2014). The decalog of long non-coding RNA involvement in cancer diagnosis and monitoring. *Crit. Rev. Clin. Lab. Sci.* 51, 344–357.

Lee, G. M., Donaldson, L. W., Pufall, M. A., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., et al. (2005). The structural and dynamic basis of Ets-1 DNA binding autoinhibition. *J. Biol. Chem.* 280, 7088–7099.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2013). starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97.

Ma, J., Li, T., Han, X., and Yuan, H. (2018). Knockdown of LncRNA ANRIL suppresses cell proliferation, metastasis, and invasion via regulating miR-122-5p expression in hepatocellular carcinoma. *J. Cancer Res. Clin. Oncol.* 144, 205–214.

Marcel, V., and Hainaut, P. (2009). p53 isoforms-a conspiracy to kidnap p53 tumor suppressor activity? *Cell. Mol. Life Sci.* 66, 391–406.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., et al. (2006). TRANSFAC® and its module TRANSCompel® : transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34(Suppl._1), D108–D110.

Paz, I., Akerman, M., Dror, I., Kosti, I., and Mandel-Gutfreund, Y. (2010). SFmap: a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Res.* 38(Suppl._2), W281–W285.

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204.

Quagliata, L., Matter, M. S., Piscuoglio, S., Arabi, L., Ruiz, C., Procino, A., et al. (2014). Long noncoding RNA HOTTIP/HOXA13 expression is associated with disease progression and predicts outcome in hepatocellular carcinoma patients. *Hepatology* 59, 911–923.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25.

Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F., and Crespi, M. (2018). Splicing regulation by long noncoding RNAs. *Nucleic Acids Res.* 46, 2169–2184.

Sarkar, D., Oghabian, A., Bodiyabadu, P. K., Joseph, W. R., Leung, E. Y., Finlay, G. J., et al. (2017). Multiple isoforms of ANRIL in melanoma cells: structural complexity suggests variations in processing. *Int. J. Mol. Sci.* 18:1378.

Sharrocks, A. D. (2001). The ETS-domain transcription factor family. *Nat. Rev. Mol. Cell Biol.* 2, 827–837.

Smyth, G. K. (2005). *Limma: Linear Models for Microarray data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Berlin: Springer, 397–420.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, A. M., Mehra, R., Sun, X.-W., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310, 644–648.

Vitting-Seerup, K., Porse, B. T., Sandelin, A., and Waage, J. (2014). spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics* 15:81. doi: 10.1186/1471-2105-15-81

Vitting-Seerup, K., and Sandelin, A. (2017). The landscape of isoform switches in human cancers. *Mol. Cancer Res.* 15, 1206–1220.

Vitting-Seerup, K., and Sandelin, A. (2018). IsoformSwitchAnalyzeR: Analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *bioRxiv* [Preprint]. doi: 10.1101/399642

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C. J., Rankin, T. L., Star, R. A., et al. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* 35, 314–316.

Wang, J., Cao, P., Qi, Y.-Y., Chen, X. P., Ma, L., Deng, R. R., et al. (2017). The relationship between cell apoptosis dysfunction and FEN1 E160D mutation in lupus nephritis patients. *Autoimmunity* 50, 476–480.

Wang, J., Chen, Y., Xu, K., Zhou, Y. (2019). *Comprehensive Network Analysis Reveals Alternative Splicing-Related lncRNAs in Hepatocellular Carcinoma*. Available online at: https://ssrn.com/abstract=3335849

Wang, J., Su, L., Chen, X., Li, P., Cai, Q., Yu, B., et al. (2014). MALAT1 promotes cell proliferation in gastric cancer by recruiting SF2/ASF. *Biomed. Pharmacother.* 68, 557–564.

Wang, J., Zhou, Y., Fei, X., Chen, X., and Chen, Y. (2018). Biostatistics mining associated method identifies AKR1B10 enhancing hepatocellular carcinoma cell growth and degenerated by miR-383-5p. *Sci. Rep.* 8:11094.

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., and Li, W. (2013). CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41:e74.

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J. S., et al. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. *Genome Biol.* 13:R35.

West, J. A., Davis, C. P., Sunwoo, H., Simon, M. D., Sadreyev, R. I., Wang, P. I., et al. (2014). The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell* 55, 791–802.

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., et al. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* 13:R50.

Zang, C., Nie, F.-Q., Wang, Q., Sun, M., Li, W., He, J., et al. (2016). Long non-coding RNA LINC01133 represses KLF2, P21 and E-cadherin transcription through binding with EZH2, LSD1 in non small cell lung cancer. *Oncotarget* 7, 11696–11707.

Zhang, L., Liu, X., Zhang, X., and Chen, R. (2016). Identification of important long non-coding RNAs and highly recurrent aberrant alternative splicing events in hepatocellular carcinoma through integrative analysis of multiple RNA-Seq datasets. *Mol. Genet. Genomics* 291, 1035–1051.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 17 July 2020
doi: 10.3389/fbioe.2020.00730

Investigation and Prediction of Human Interactome Based on Quantitative Features

Xiaoyong Pan[1,2], Tao Zeng[3], Yu-Hang Zhang[4], Lei Chen[5], Kaiyan Feng[6], Tao Huang[4] and Yu-Dong Cai[1]**

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Key Laboratory of System Control and Information Processing, Ministry of Education of China, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, [3] Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China, [4] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [5] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [6] Department of Computer Science, Guangdong AIB Polytechnic, Guangzhou, China

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Nagarajan Raju,
Vanderbilt University Medical Center, United States
Fuyi Li,
Monash University, Australia
Yun Li,
University of Pennsylvania, United States

***Correspondence:**
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 19 April 2020
Accepted: 09 June 2020
Published: 17 July 2020

Citation:
Pan X, Zeng T, Zhang Y-H, Chen L, Feng K, Huang T and Cai Y-D (2020) Investigation and Prediction of Human Interactome Based on Quantitative Features. Front. Bioeng. Biotechnol. 8:730. doi: 10.3389/fbioe.2020.00730

Protein is one of the most significant components of all living creatures. All significant and essential biological structures and functions relies on proteins and their respective biological functions. However, proteins cannot perform their unique biological significance independently. They have to interact with each other to realize the complicated biological processes in all living creatures including human beings. In other words, proteins depend on interactions (protein-protein interactions) to realize their significant effects. Thus, the significance comparison and quantitative contribution of candidate PPI features must be determined urgently. According to previous studies, 258 physical and chemical characteristics of proteins have been reported and confirmed to definitively affect the interaction efficiency of the related proteins. Among such features, essential physiochemical features of proteins like stoichiometric balance, protein abundance, molecular weight and charge distribution have been validated to be quite significant and irreplaceable for protein-protein interactions (PPIs). Therefore, in this study, we, on one hand, presented a novel computational framework to identify the key factors affecting PPIs with Boruta feature selection (BFS), Monte Carlo feature selection (MCFS), incremental feature selection (IFS), and on the other hand, built a quantitative decision-rule system to evaluate the potential PPIs under real conditions with random forest (RF) and RIPPER algorithms, thereby supplying several new insights into the detailed biological mechanisms of complicated PPIs. The main datasets and codes can be downloaded at https://github.com/xypan1232/Mass-PPI.

Keywords: decision tree, human interactome, prediction, protein–protein interaction, quantitative feature

INTRODUCTION

Protein–protein interactions (PPI) are core biochemical events that directly execute biological functions in all living creatures (Qian et al., 2014; Wang et al., 2014). As the major executor of various biological processes, proteins rarely act alone, and protein interactions guarantee the continuity and controllability of ordinary biological processes (De Las Rivas and Fontanillo, 2010).

On one hand, PPIs based on functional classification have multiple types, including signal transduction (Vinayagam et al., 2011), trans-membrane transport (Fairweather et al., 2015), cell metabolism (Gonzalez, 2012), and muscle contraction (Beqollari et al., 2015); these PPIs cover every detailed functional aspect in living cells. On the other hand, on the basis of chemical structure and stability, PPIs can be described as homo/hetero-oligomers, stable/transient interactions, and covalent/non-covalent interactions, thereby revealing the complicated chemical nature of common biochemical reactions that support protein interactions in all living cells (De Las Rivas and Fontanillo, 2010).

The complicated organization of PPIs can be clustered in multiple ways. Given the complexity and core regulatory role of protein interactions underlying biochemical processes in living cells, for a long time, many scientists have aimed to analyze and extract the key regulatory factors in the PPIs and describe their functional relationships and biological significance. According to previous studies, biochemical features of PPIs (e.g., protein concentration, protein binding ligands, presence of adaptors, and covalent modifications) have been recognized as candidate factors that may affect PPIs (Pan et al., 2010; Raj et al., 2013; Modell et al., 2016). However, most of such extracted features are ambiguous qualitative characteristics. These features may be directly or indirectly related to PPIs, but whether PPIs with optimal biological features may be determined in certain cell types is difficult. These features are not detailed differentiating indicators for the occurrence possibility of PPIs, rather than existence. Therefore, accurate and quantitative/semi-quantitative characteristics of PPIs must be identified through continuous studies and exploration.

In recent years, with the development of mass spectrometry and related analysis techniques, various omics features have been presented to describe the characteristics of PPIs and have been applied to evaluate the possibility and certain biological functions of cell-specific PPIs. In 2015, using high-throughput affinity-purification mass spectrometry, Huttlin et al. (2015) built a PPI network (BioPlex) and extracted various functional characteristics describing PPIs, thus providing us with a blueprint of quantitative human interactome in all living cells. In the same year, another study presented by Wan et al. focused on the macromolecular complexes' contribution to PPIs; these authors extracted the co-complex interactions using an integrative approach (Wan et al., 2015), thereby revealing the fundamental mechanistic significance of reconstructed interactomes. This study also extracted a group of parameters/features that can be used for a detailed quantitative description of PPI. In 2015, another study by Hein et al. (2015) further proposed nine features, such as NWD, Z, and Plate Z scores, which may quantitatively describe PPIs. Combining the datasets of the three studies, a systemic analysis of all reported human protein complexes based on mass spectrometry techniques has been recently presented (Drew et al., 2017). Such study summarized the identified features associated with PPIs (i.e., PPI features) and built a global map of all reported human protein complexes. It provided us with a database, namely hu.MAP (http://proteincomplexes.org/), as a new resource of a follow-up study on the core physical and pathological functions of human PPIs in normal and disease cells. Such features captured the specificity of real PPIs and were screened out by three independent studies (Hein et al., 2015; Huttlin et al., 2015; Wan et al., 2015). According to such studies (Hein et al., 2015; Huttlin et al., 2015; Wan et al., 2015), all candidate features are validated by large scale mass spectrometry and have been identified to contribute to the regulation and description of certain PPIs.

However, the original and combination studies of three datasets have not identified the key factors that may contribute to and appropriately describe the occurrence possibility of PPIs. Previous studies have merely identified and summarized potential PPI features, but the significance comparison and quantitative contribution of candidate PPI features remain to be identified. Thus, in this study, the PPI data obtained from multiple mass spectrometry experiments (Drew et al., 2017) is summarized by our newly presented decision tree-centered computational framework. Such PPI data contained one training dataset and one testing dataset, each of which consisted of proteins that can interact with each other, namely positive PPIs, and proteins that cannot interact with each other, namely negative PPIs. The core parameters of PPI features that may describe and judge the possibility of potential PPIs are accurately identified. The decision tree-based model with extracted core PPI features yielded better performance than the models with other classification algorithms, including nearest neighbor algorithm (NNA) (Cover and Hart, 1967) and recurrent neural network (RNN). Furthermore, a quantitative decision-rule system based on PPI features is built to supply several new insights into the detailed biological mechanisms of complicated PPIs. These quantified outcomes not only reveal the core regulatory factors in PPIs but also provide a new computational tool for investigating and predicting the potential of PPIs under different physical and pathological conditions.

MATERIALS AND METHODS

Datasets

The training and testing human PPI datasets were obtained from Drew et al. (2017) (http://proteincomplexes.org/download). The training dataset has 68,651 PPIs, in which 9,318 are actual positive PPIs (i.e., proteins that can interact with each other), and 59,333 are negative PPIs (i.e., proteins that cannot interact with each other). These PPIs cover 1,253 proteins. The testing dataset has 77,884 PPIs, in which 4,579 are actual positive PPIs, and 73,305 are negative PPIs. One thousand one hundred thirty-two proteins occur in the testing dataset, where 606 are also used in the training dataset. Each PPI was encoded with 258 features, which were downloaded from Drew et al. (2017) too. They were defined in three previous studies (Hein et al., 2015; Huttlin et al., 2015; Wan et al., 2015) and represented various biological characteristics of PPI. Only human proteins were included and the PPIs were literature-curated.

To describe the PPIs, we summarized the features described in three publications: Wan et al. (2015), BioPlex (Huttlin

et al., 2015), and Hein et al. (2015). There were 241 features from Wan et al. (2015), 11 features from BioPlex (Huttlin et al., 2015) and 6 features from Hein et al. (2015). These co-fractionation and physiochemical features described all the properties that may affect the potential interactions between the target protein either partially or as an entity. These features had been refined with mass spectrum results (Hein et al., 2015). The redundant and unimportant features had been removed to establish an effective framework for PPIs description using co-fractionation and physiochemical features. For instance, there is a specific feature named as spatiotemporal overlap (Hein et al., 2015), describing the temporal spatial interactions between two participators of PPIs. Interactions with either too high spatiotemporal overlap or too low overlap may indicate the interaction will not actually happen (Hein et al., 2015). All the features used in this study are summarized from existed datasets and derived from experimental results.

Feature Selection

In this study, a three-stage feature selection scheme was designed to identify important features for characterizing PPIs. In the first stage, all features were analyzed by the Boruta feature selection (BFS) (Kursa and Rudnicki, 2010) method, excluding irrelated features; then, the rest features were analyzed by the Monte Carlo feature selection (MCFS) (Draminski et al., 2008) method, producing a feature list; finally, the feature list was adopted in the incremental feature selection (IFS) (Liu and Setiono, 1998) method, incorporating a supervised classifier, to extract optimal features and build an optimal classifier.

Boruta Feature Selection Method

BFS method (Kursa and Rudnicki, 2010) is a wrapper method for selecting relevant features, which is based on random forest (RF) (Breiman, 2001). It evaluates feature importance by comparing with randomized features. Such method is different from most of the other wrapper feature selection methods that achieve a minimal error for a supervised classifier on a small subset of features, BFS selects all features either strongly or weakly relevant to the outcome variable.

The core idea of BFS is that it creates a shuffled version of original features, then uses a RF classifier to measure the importance score of the combined shuffled and original features. Only those features with importance score higher than that of the randomized features are selected. These selected features are considered significantly relevant to target variables. The difference between RF importance score and BFS importance score is that the statistical significance of the variable importance is introduced. Random permutation procedure is repeated to get statistically robust important features. BFS proceeds as follows by repeating multiple iterations:

1. Add randomness to the given dataset by shuffling original features.
2. Combine the shuffled dataset and original dataset.
3. Train a RF classifier on the combined dataset and evaluate the importance of each feature.
4. Calculate Z-scores of both original and shuffled features. The Z-scores of individual features are calculated as mean of importance scores divided by the standard error. For each real feature, evaluate whether it has a higher Z-score than the maximum of its shuffled feature. If yes, this feature is tagged as important, otherwise unimportant.
5. Finally, the algorithm stops until one of the two following condition is satisfied: (I) All features are either tagged "unimportant" or "important"; (II) Reach a predefined number of iterations.

In this study, we used the python implementation of BFS from https://github.com/scikit-learn-contrib/boruta_py, and the defaulted parameters are used.

Monte Carlo Feature Selection Method

As mentioned in section Boruta Feature Selection Method, features selected by BFS method are highly related to target variables. These features are further analyzed by the MCFS method (Draminski et al., 2008). MCFS is a powerful and widely used feature selection method (Chen L. et al., 2018a, 2019b; Pan et al., 2018, 2019; Wang et al., 2018), which consists of multiple decision trees, and constructs multiple bootstrap sets and randomly selects feature subsets. For each feature subset, new training samples are re-represented by using the features in this subset, and M decision trees are grown by using the bootstrap sets sampled from the new training samples. This process is repeated T times, thereby resulting in $M \times T$ trees. A relative importance (RI) score is calculated in accordance with the involvement of a feature in constructing $M \times T$ trees. Its equation is as follow:

$$RI_g = \sum_{\tau=1}^{MT} (wAcc)^u IG(n_g(\tau)) \left(\frac{no.in\ n_g(\tau)}{no.in\ \tau}\right)^v, \quad (1)$$

where g stands for a feature, wAcc denotes the weighted accuracy of the decision tree τ, $n_g(\tau)$ represents the node involving g in τ, $IG(n_g(\tau))$ represents the information gain of $n_g(\tau)$, $no.in\ \tau$ and $no.in\ n_g(\tau)$ denotes the number of samples in decision tree τ and node $n_g(\tau)$, respectively. u and v are weighting factors. Evidently, a high RI score indicates that one feature will be more frequently involved in learning these decision trees. Thus, this feature will have ranked relevance in characterizing PPIs. Based on the RI scores of features, a feature list, denoted as $F = [f_1, f_2, \ldots, f_N]$, can be built by the decreasing order of features' RI scores.

The MCFS program was downloaded from http://www.ipipan.eu/staff/m.draminski/files/dmLab_2.1.1.zip. We used the default parameters to execute such program, where u and v were set to 1, M and T were 2,000 and 5, respectively.

Incremental Feature Selection Method

A feature list can be generated according to the results of MCFS method, based on which incremental feature selection (IFS) (Liu and Setiono, 1998; Li et al., 2015, 2016, 2019; Chen et al., 2017b; Chen L. et al., 2018b, 2019a; Wang and Huang, 2018; Zhang et al., 2018), combining with a supervised classifier (i.e., RF), is adopted to further detect discriminative features for indicating PPIs. A series of feature subsets is generated from the ranked features

F from the MCFS. The first feature subset has feature f_1, the second feature subset has features $[f_1, f_2]$, and so on. RF is run to test these feature subsets with 10-fold cross validation. Finally, an RF classifier with the optimal classification performance is generated, such classifier was termed as the optimal classifier. And the features in the corresponding feature subset are called optimal features (i.e., PPI features).

SMOTE

It is easy to see that the negative PPIs were much more than positive PPIs in both training and testing datasets. In detail, in the training dataset, negative PPIs were about 6.37 times as many as positive PPIs, while such proportion was about 16 for the testing dataset. Thus, the investigated datasets were greatly imbalanced. For such type of dataset, it is not easy to build a perfect classifier. In this study, we employed Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to tackle such datasets.

SMOTE is a classic and widely used oversampling method. It generates predefined numbers of samples and pours them into the minority class. In detail, it first randomly selects a sample in one minority class, say x. Then, find k samples in such class, which have smallest distances to x. Randomly select a sample from these k samples, say y, and generate a new sample z, which is the linear combination of x and y. The generated new sample z is put into the minority class. Above procedures execute multiple times until predefined number of new samples have been produced.

In this study, we directly adopted the tool "SMOTE" in Weka (Version 3.6) (Witten and Frank, 2005), which implement above-mentioned SMOTE. For the training dataset, we used "SMOTE" generated lots of new samples and termed them as positive PPIs. Finally, the numbers of positive and negative PPIs were almost equal. We used the default value of parameter k, which was 3. As suggested in Blagus and Lusa (2013), feature selection should be performed before using SMOTE. Thus, in this study, the SMOTE was only adopted in IFS method. Samples yielded by SMOTE were not used in the BFS and MCFS methods.

Classifier

In IFS method, supervised classifiers are indispensable. Here, two classic classifiers were adopted. They were RF (Breiman, 2001) and RIPPER algorithm (Cohen, 1995). The first one was to build an efficient classifier. However, it cannot bring lots of information to uncover the essential differences between positive and negative PPIs. Thus, we further employed the second classifier, RIPPER algorithm, which is a rule learning algorithm. It can provide several rules to clearly display the classification procedures and differences between positive and negative PPIs.

Random Forest

As a supervised classifier, RF consists of multiple decision trees, and each decision tree is grown from a bootstrap set and a randomly selected feature subset. We assume a training set with N samples and M features. For each decision tree, the same number of samples is first randomly selected from the original training set with replacement and a feature subset with m features ($m << M$) is also randomly constructed. Each tree is grown from these selected samples with the selected feature subset. This process is repeated T times, and T decision trees comprising the RF are yielded. RF has much fewer parameters to tune; thus, this technique is extensively used in many biological problems with favorable performance (Pan et al., 2010, 2014; Zhao et al., 2018, 2019; Zhang et al., 2019). The RF classifier implemented by a tool "RandomForest" in Weka (Witten and Frank, 2005) software is used. Clearly, the number of decision trees is an important parameter of RF. Here, we tried four values: 10, 20, 50, and 100.

Repeated Incremental Pruning to Produce Error Reduction Algorithm

RIPPER algorithm (Cohen, 1995) is a classic rough set based rule learning algorithm. In fact, it is a generalized version of the Incremental Reduced Error Pruning (IREP) algorithm (Johannes and Widmer, 1994). The procedures of rule learning with RIPPER can be found in our previous study (**Figure 1**; Wang et al., 2018). Rules generated by RIPPER algorithm are represented by IF-THEN clauses. For example, IF (Feature 1 $\geq$2.333 and Feature 2 $\leq$1.234) THEN Positive PPI. Likewise, RIPPER algorithm is also implemented by a tool "JRip" in Weka (Witten and Frank, 2005). We directly used it and executed it with its default parameters.

Performance Measurement

The performance of the classifiers is evaluated using 10-fold cross validation. Several evaluation metrics, such as sensitivity (SN), specificity (SP), two types of accuracy (ACC1 and ACC2), Matthew correlation coefficient (MCC) (Matthews, 1975; Chen et al., 2017a; Chen Z. et al., 2018, 2019; Li et al., 2018; Song et al., 2018; Cui and Chen, 2019), recall, precision, and F-measure are calculated and formulated as follows:

$$SN = \frac{TP}{TP + FN}, \tag{2}$$

$$SP = \frac{TN}{TN + FP}, \tag{3}$$

$$ACC1 = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

$$ACC2 = (SN + SP)/2 \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{6}$$

$$Recall = \frac{TP}{TP + FN}, \tag{7}$$

$$Precision = \frac{TP}{TP + FP}, \tag{8}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{9}$$

where TP/TN are the numbers of true positives/negatives, and FP/FN are the numbers of false positives/negatives. Clearly, ACC1, ACC2, MCC, and F-measure can fully evaluate the performance of a classifier. This study selected F-measure as the key measurement.

In addition to above-mentioned measurements, we also employed ROC and PR curves to fully evaluate the performance

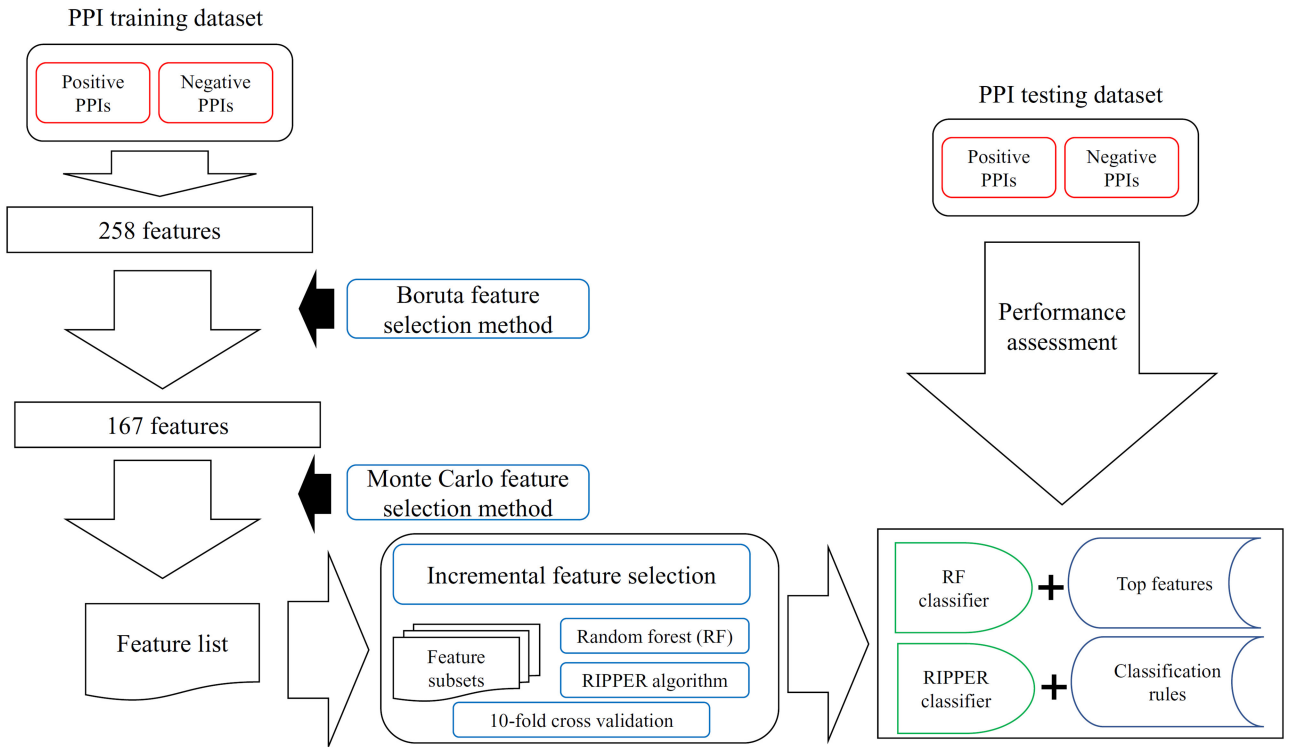


FIGURE 1 | The entire procedures to analyze PPI features with a three-stage feature selection scheme. The PPI samples comprise one training dataset and one testing dataset. For the training samples, they were represented by 258 features, which are processed by Boruta feature selection method. One hundred sixty-seven relevant features remain, which are further analyzed by the Monte Carlo feature selection method. A feature list is produced. The incremental feature selection uses such feature list to construct several feature subsets. On each subset, one random forest (RF) classifier and one RIPPER classifier are constructed, which are assessed by 10-fold cross-validation. With RF, the best RF classifier and top features are obtained; whereas with RIPPER, the best RIPPER classifier together with its rules is generated. The best RF and RIPPER classifiers are further evaluated on the testing dataset.

of different classifiers. The areas under these two curves are also important measurements to assess classifiers. They were called AUROC and AUPR, respectively, in this study.

RESULTS

In this study, the prior extracted 258 features were analyzed by a three-stage feature selection scheme. The entire procedures are illustrated in **Figure 1**.

Analysis of the Identity Between PPIs in the Training and Testing Datasets

Before performing the feature selection scheme, it is necessary to count the identity between PPIs in the training and testing datasets because PPIs with high identities will make the classification easily. Here, the identity between two PPIs was defined as the direction cosine of their 258-D feature vectors. We used 0.1 as the step to count the distribution of the obtained identities on the training and testing datasets, which is shown in **Figure 2**. It can be observed that the training and testing datasets gave the similar distribution on identities. The interval $[-0.1,0]$ contained the most identities and between -1 and 0.6, the distribution was quite similar to the normal distribution. It is also surprised that several identities were with high values (interval [0.9, 1]). However, more than 80% identities were <0.5, indicating that most PPIs were with low identities. The investigation on such datasets was quite reliable.

Results of Boruta Feature Selection (BFS) Method

In the training dataset, all PPIs were represented by 258 features. These features were analyzed by BFS method. As a result, 167 features were selected, as listed in **Table S1**.

Results of Monte Carlo Feature Selection (MCFS) Method

According to the three-stage feature selection scheme, remaining 167 features were analyzed by the powerful MCFS method. Each feature was assigned a RI score, which is also provided in **Table S1**. Accordingly, a feature list F was built, in which features were sorted by the decreasing order of their RI scores. This list is available in **Table S1**.

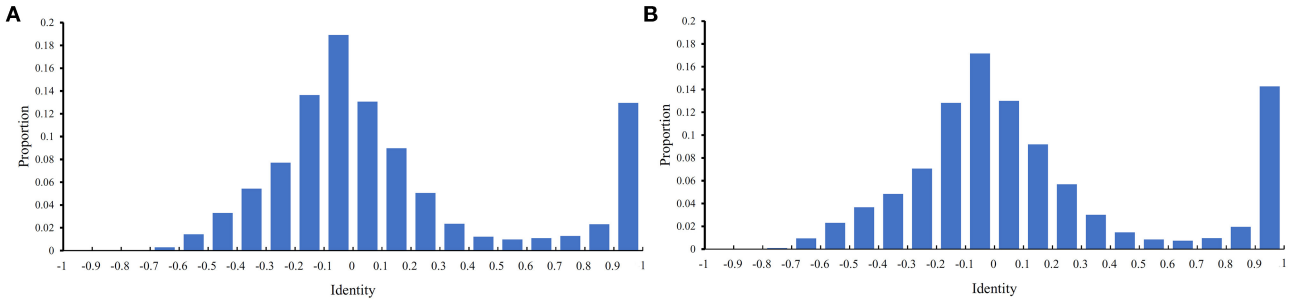


FIGURE 2 | Distribution of identities between PPIs in two datasets. The identify of two PPIs is defined as the direction cosine of their feature vectors. **(A)** Distribution on the training dataset; **(B)** Distribution on the testing dataset.

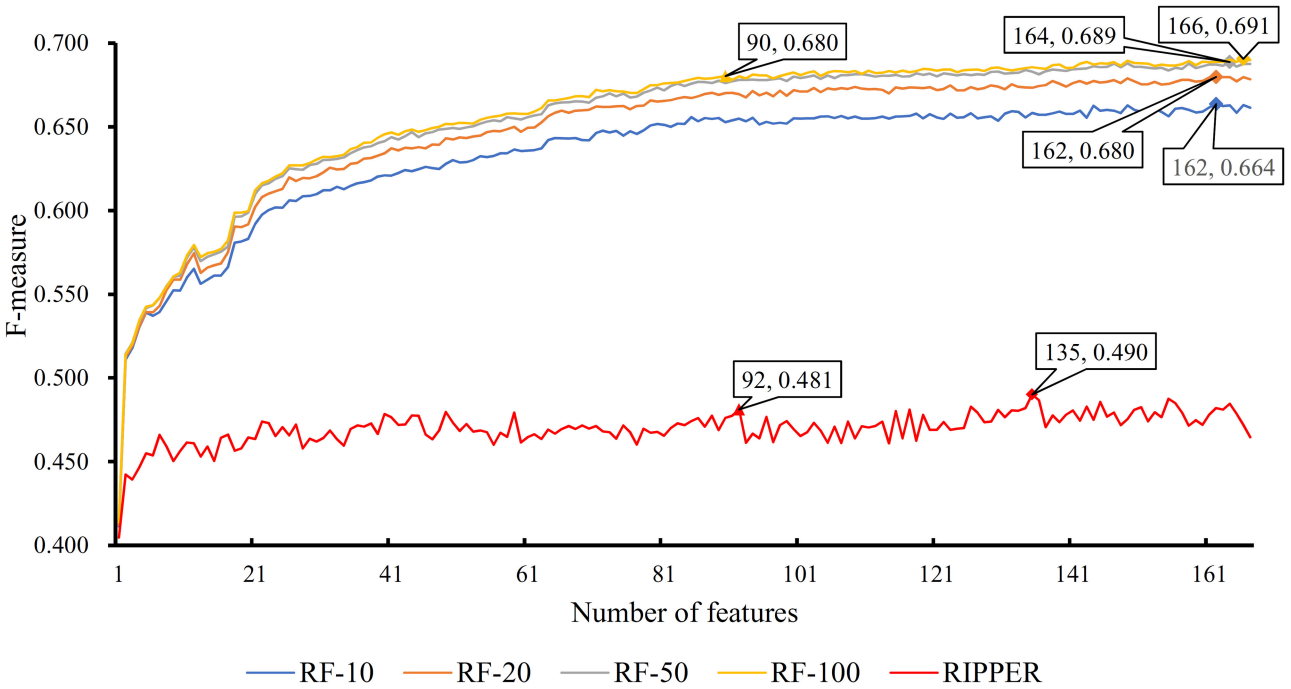


FIGURE 3 | IFS curves based on the IFS method with RF and RIPPER. The X-axis shows the number of features, and Y-axis shows the F-measure values. The numbers following RF indicate the number of decision trees of RF.

Results of Incremental Feature Selection (IFS) With Random Forest (RF)

The feature list only told us the importance of each feature. To extract optimal features for RF, IFS method was employed. For each feature subset constructed from *F*, RF classifiers with different number of decision trees (10, 20, 50, and 100) were built on the training dataset and evaluated through 10-fold cross validation. The results are provided in **Tables S2–S5**. To clearly display these RF classifiers on different feature subsets, four IFS-curves are plotted in **Figure 3**. It can be seen that the optimal F-measure value was 0.691 when the top 166 features in *F* were used and the number of decision trees was 100. Accordingly, the RF classifier containing 100 decision trees was built on the training dataset, in which PPIs were represented by top 166 features in *F*. Such classifier was called the optimal RF classifier. Other measurements yielded by such RF classifier are listed in **Table 1**. The SN, SP, ACC1, ACC2, MCC, and Precision were 0.794, 0.921, 0.903, 0.858, 0.642, and 0.611, respectively, suggesting the good performance of such classifier. Besides, we also used ROC curve and PR curve to evaluate the performance of such RF classifier, which are shown in **Figures 4A,B**. The AUROC and AUPR was 0.920 and 0.745, respectively.

To indicate the improvement of the RF with top 166 features, we conducted 10-fold cross-validation on this classifier 50 times. Also, the RF classifier with all 258 features were evaluated by 10-fold cross-validation 50 times. Obtained F-measures are shown in **Figure 5**, from which we can see that F-measures yielded by the RF classifier with top 166 features were evidently higher than

TABLE 1 | Performance of the RF and RIPPER classifiers on the training dataset evaluated by 10-fold cross-validation.

Classifier	Number of features	SN	SP	ACC1	ACC2	MCC	Precision	F-measure
RF	166	0.794	0.921	0.903	0.858	0.642	0.611	0.691
	90	0.786	0.918	0.900	0.852	0.630	0.600	0.680
RIPPER	135	0.701	0.818	0.802	0.760	0.409	0.377	0.490
	92	0.689	0.815	0.798	0.752	0.397	0.370	0.481
NNA	101	0.851	0.881	0.877	0.866	0.607	0.529	0.652
RNN	133	0.824	0.890	0.881	0.857	0.605	0.542	0.654

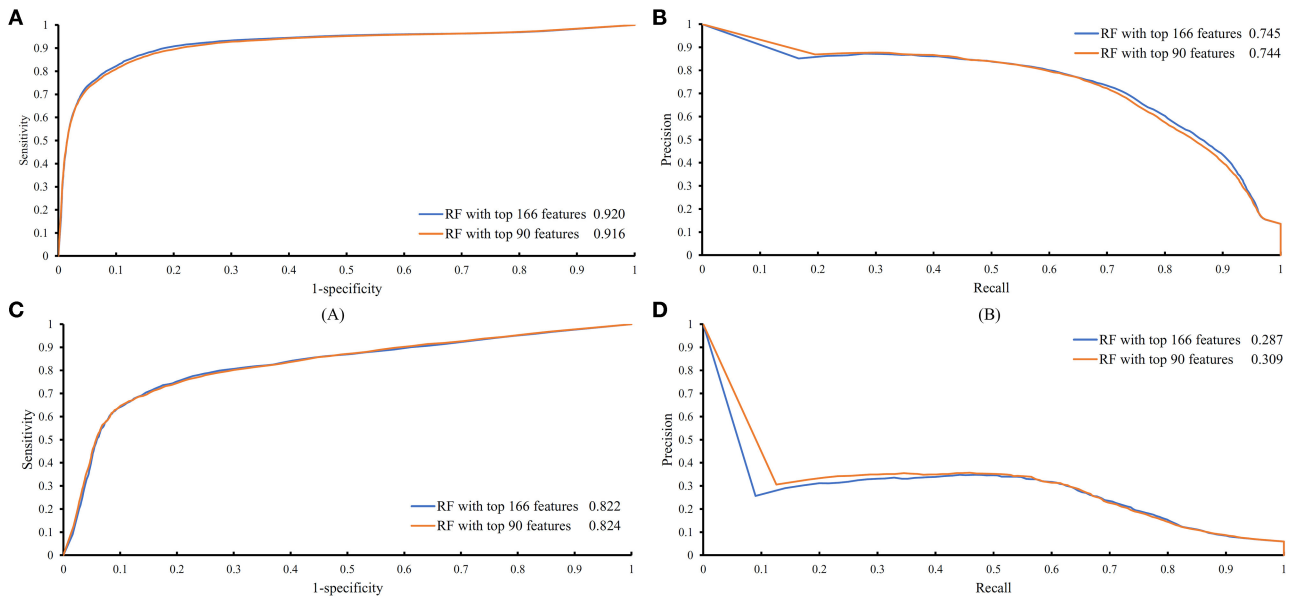


FIGURE 4 | ROC and PR curves of the RF classifiers with top 166 and 90 features on the training and testing datasets. **(A)** ROC curves of two RF classifiers on the training dataset; **(B)** PR curves of two RF classifiers on the training dataset; **(C)** ROC curves of two RF classifiers on the testing dataset; **(D)** PR curves of two RF classifiers on the testing dataset.

those produced by the RF classifier with all features. To confirm this result, a paired sample *t*-test was conducted, yielding the *p*-value of 1.309E-15, suggesting that the performance of the RF classifier was improved with statistical significance.

Above-constructed RF classifier was also applied to the testing dataset. The predicted results are listed in **Table 2**, from which we can see that the F-measure was 0.371. Its SN, SP, ACC1, ACC2, MCC and Precision were 0.674, 0.877, 0.865, 0.776, 0.358, and 0.256, respectively. The ROC and PR curves of the constructed RF classifier on the testing dataset are shown in **Figures 4C,D**. The AUROC and AUPR was 0.822 and 0.287, respectively. Although they were lower than those on training dataset, the ACC1 was still over 0.850.

As mentioned above, for RF with 100 decision trees, when top 166 features in *F* was used, it provided the best F-measure. However, after carefully checking the IFS results (**Table S2**), when top 90 features were used, RF can yield the F-measure of 0.680, which was a little lower than that yielded by the optimal RF classifier. Considering the efficiency of classifiers, we suggested the RF constructed on top 90 features as the proposed classifier. The detailed performance of this classifier, evaluated by 10-fold cross-validation, is provided in **Table 1** and the ROC and PR curves are shown in **Figures 4A,B**. Clearly, the performance of this classifier was almost equal to that of the optimal RF classifier. Besides, the proposed classifier was also performed on the testing dataset, obtained measurements are listed in **Table 2** and ROC and PR curves are shown in **Figures 4C,D**. Clearly, they all approximated to those of the optimal RF classifier. All of these indicated that the proposed RF classifier can provide similar results, however, it had high efficiency because much less features were involved.

Comparison of IFS With NNA and RNN

As mentioned above, the optimal RF classifier gave good performance. However, is the RF a proper choice? In fact, we also tried other two classification algorithms: NNA and RNN. NNA is a classic and simple classification algorithm, which makes prediction for a given sample according to its nearest

neighbor, while RNN is a kind of neural network with loop inside for sequential data. For each of these two algorithms, an IFS procedure was performed on the training dataset. Two IFS curves were obtained, as shown in **Figure 6**. The highest F-measure for NNA was 0.652 when top 101 features in F were used. For RNN, the highest F-measure was 0.654 when top 133 features were adopted. These F-measure values were all lower than that of the optimal RF classifier. The detailed performance of the best NNA and RNN classifiers is listed in **Table 1**. It can be observed that the optimal RF classifier produced higher values on most measurements, suggesting that RF is a more proper choice than NNA and RNN.

Results of IFS With RIPPER

In section Results of incremental feature selection (IFS) with random forest (RF), a RF classifier was built to identify PPIs. However, it is a black box. It is difficult to capture the classification principle. Thus, it provided limited biology insights for understanding PPIs. In view of this, we further employed a rule learning method, RIPPER algorithm, trying to partly uncover the differences between positive and negative PPIs.

Like RF, the RIPPER algorithm was also employed in the IFS method. The performance of the RIPPER algorithm on different feature subsets is available in **Table S6**. Also, an IFS-curve was plotted, as shown in **Figure 3**. The highest F-measure was 0.490 when top 135 features were used. Thus, the RIPPER classifier based on top 135 features was called the optimal RIPPER classifier. The detailed performance of such classifier, evaluated by 10-fold cross-validation, was provided in **Table 1**. Clearly, it was much inferior to the optimal RF classifier. In addition, the optimal RIPPER classifier was also executed on the testing dataset. The predicted results were listed in **Table 2**. The F-measure was 0.348, which was also much lower than that on the training dataset. Compared with the performance of the optimal RF classifier on the testing dataset, the performance of the optimal RIPPER classifier was only a little lower.

Likewise, the RIPPER classifier can yield the F-measure 0.481 on the training dataset when top 92 features were used after checking the predicted results listed in **Table S6**. It is a little lower than that generated by the optimal RIPPER classifier. Considering the efficiency of classifiers, we termed the RIPPER classifier with top 92 features as the proposed RIPPER classifier. The detailed performance of such classifier on the training dataset is listed in **Table 1**. All measurements were almost equal to those yielded by the optimal RIPPER classifier. Furthermore, the proposed RIPPER classifier was executed on the testing dataset. Predicted results are listed in **Table 2**. Obviously, the performances of the optimal and proposed classifiers were at the same level.

As mentioned above, the proposed RIPPER classifier adopted top 92 features to represent PPIs. Six rules were produced by the RIPPER algorithm when such algorithm was applied on all PPIs in the training dataset, which are listed in **Table 3**. These rules would be discussed in section Analysis of Optimal PPI Rules.

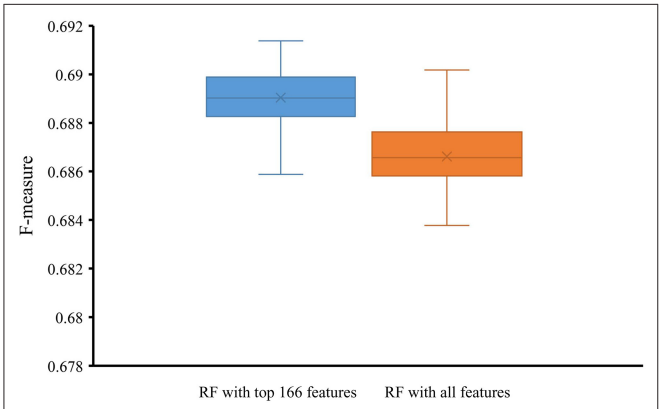


FIGURE 5 | Box plot to show F-measures yielded by RF classifiers with top 166 features and all features using 50 10-fold cross-validation. The F-measures obtained by RF classifier with top 166 features are evidently higher than those of the RF classifier with all features.

DISCUSSION

All PPI-associated features have been summarized in the three previously described datasets (Hein et al., 2015; Huttlin et al., 2015; Wan et al., 2015). In this study, we deeply analyzed these features. Based on some key features, a RF classifier was constructed and some classification rules were built. This section gave detailed analysis on some top features and classification rules. Several top features and all rules were supported by recent publications (Mitterhuber, 2008; Swiatkowska et al., 2008; Levin et al., 2013; Pinton et al., 2015).

Analysis of Optimal PPI Features

In the proposed RF classifier, top 90 features were used to represent PPIs. However, it is impossible to analyze them one by one due to our limited human resources. In fact, among these 90 features, some were more important than others. We did the

TABLE 2 | Performance of the RF and RIPPER classifiers on the testing dataset.

Classifier	Number of features	SN	SP	ACC1	ACC2	MCC	Precision	F-measure
RF	166	0.674	0.877	0.865	0.776	0.358	0.256	0.371
	90	0.677	0.874	0.863	0.776	0.356	0.252	0.367
RIPPER	135	0.797	0.826	0.825	0.812	0.360	0.223	0.348
	92	0.800	0.822	0.821	0.811	0.357	0.219	0.344

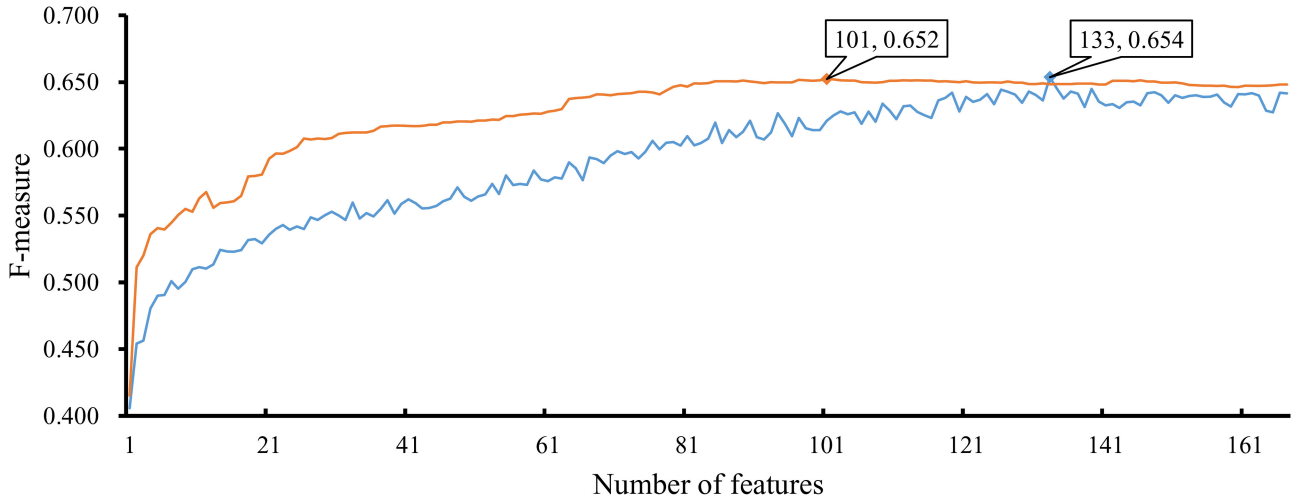


FIGURE 6 | IFS curves based on the IFS method with NNA and RNN. The X-axis shows the number of features, and Y-axis shows the F-measure values. The orange curve and blue curve were NNA and RNN, respectively.

TABLE 3 | Classification rules for predicting protein-protein interactions.

Rules	Criteria	Positive/Negative
Rule1	(neg_ln_pval ≤3.622) and (hein_neg_ln_pval ≤3.328)	Negative (non-interaction) PPI
Rule2	(hein_neg_ln_pval ≤6.955) and (Hs_G166_1104_pq_euc ≤0) and (neg_ln_pval ≤3.994)	Negative (non-interaction) PPI
Rule3	(hein_neg_ln_pval ≤6.960) and (neg_ln_pval ≤5.780) and (Hs_G166_1104_pq_euc ≤0) and (pair_count ≥2)	Negative (non-interaction) PPI
Rule4	(hein_neg_ln_pval ≤3.033) and (Hs_G166_1104_pq_euc ≤0) and (pair_count ≤3) and (neg_ln_pval ≤7.272)	Negative (non-interaction) PPI
Rule5	(hein_neg_ln_pval ≤0) and (Hs_G166_1104_pq_euc ≤0) and (pair_count ≤3) and (neg_ln_pval ≤8.611)	Negative (non-interaction) PPI
Rule6	Other conditions	Positive (interaction) PPI

following test to extract most important features. Firstly, 100 feature lists were randomly built, in which 167 features were randomly sorted. According to each feature list, we did the IFS method with RF (consisting of 100 decision trees) procedures. As a result, 100 IFS-curves were plotted, as shown in **Figure 7A**, in which the IFS-curve produced on the actual feature list *F* is also listed. It can be observed that when the number of used features was small, the F-measure on the actual feature list *F* was much higher than those on the randomly generated feature lists, indicating that some top features in *F* were related to identify PPIs with high statistical significance. Thus, given a feature number, we counted the mean values of 100 F-measures that were produced on 100 randomly generated feature lists. Then, an IFS-curve was plotted, as shown in **Figure 7B**. Furthermore, we also counted the critical values on 95% confidence interval for each feature number and plotted two IFS-curves on them, as shown in **Figure 7B**. It can be observed that top 14 features in *F* can produce the F-measure that was higher than the upper critical value on 95% confidence interval, indicating that these 14 features were highly related to identify PPIs. Furthermore, top 11 features in *F* can yield the F-measure that was higher than the upper critical value on 99% confidence interval. In the following text, we extensively analyzed top 14 features in *F*.

The first four features are "hein_neg_In_pval," "neg_In_pval," "hein_pair_count," and "pair count," reflecting the regulatory contribution of protein stoichiometric and abundant features. In accordance with a reference dataset presented by Hein et al. (2015), these features were confirmed to participate in and may affect the content of interactome. According to the stoichiometric and abundant levels, a stable protein complex denotes a probable involvement of such protein complex in functional PPIs. Two detailed features, namely, stoichiometric balance and protein abundance, might generally evaluate the stability of a protein complex and participate in describing PPIs. The stable PPIs formed by stoichiometric balance might be further shaped by the abundance of each protein that participates in such interactions.

To clearly describe what are stoichiometric and abundant features, here, we took two typical PPIs as effective examples to confirm the potential contribution of such two features on the PPIs.

Firstly, we took the effective PPIs during cell adhesion regulation and functioning as an example. The adhesive properties of endothelial cells have been confirmed to be regulated by various proteins and their potential interactions (Swiatkowska et al., 2008). According to recent publications (Swiatkowska et al., 2008; Levin et al., 2013), actually among such interactions, the abundance and stoichiometric balance

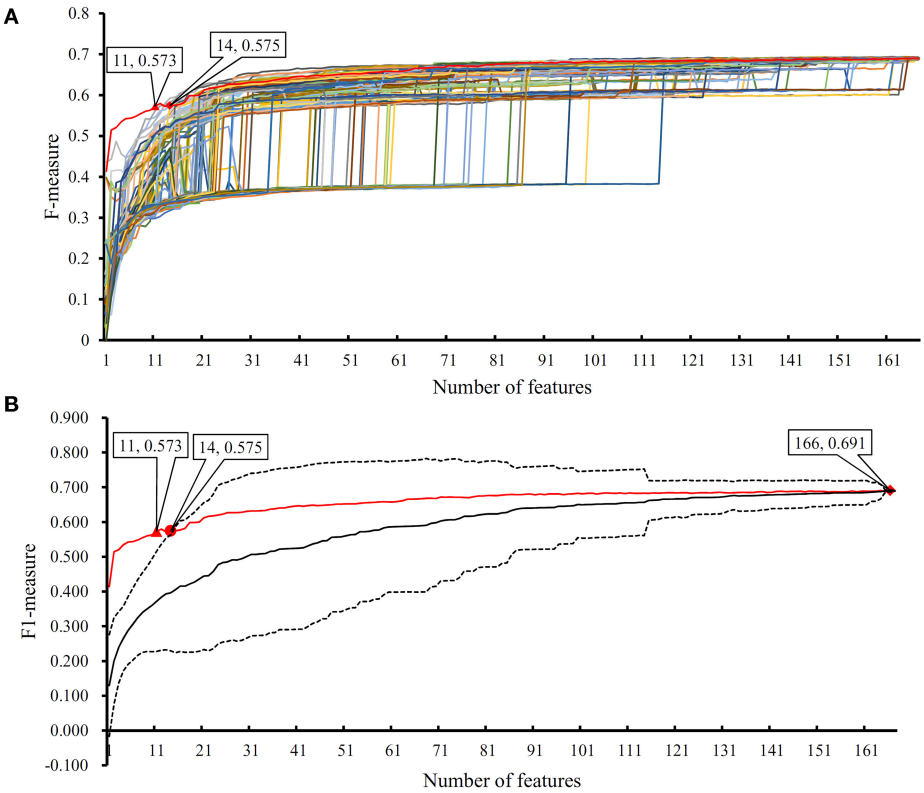


FIGURE 7 | The results of the IFS method with RF based on 100 randomly produced feature lists. **(A)** IFS curves on the actual feature list and 100 randomly produced feature lists; **(B)** the statistical analysis based on the results of randomly produced feature lists. The black curve indicates the average performance of RF on randomly produced feature lists. The red curve is the IFS curve of the actual feature list. Two dotted curves indicate the upper and low critical value on 95% confidence interval.

of disulfide isomerases and integrin may directly affect their PPIs and further interfere endothelial cell adhesion. Different abundance of disulfide isomerases caused different stoichiometric balance patterns between disulfide isomerases -integration interactions and therefore, induced different binding affinity, resulting in differential biological functions and regulatory effects (Swiatkowska et al., 2008). Therefore, stoichiometric balance is quite significant for PPIs.

Secondly, in addition to such PPI participants, the interactions between LamB and Odpq as another two effective proteins have also been influenced by the abundance of each protein and such abundance induced influences may further affect their potential biological functions, the antibiotic resistance in chlortetracycline-resistant Escherichia coli strain (Lin et al., 2014). Different abundance of such two participants may have totally opposite biological effects on such interactions: the interactions of lower concentration may improve the antibiotic sensitivity of E. coli, while the interactions at high concentration on the contrary directly induce the chlortetracycline-resistance. Therefore, the abundance of participants may be quite essential for PPIs. Similarly, another two features in the optimal feature list named as "Ce_CRF_wan_60_1209_poisson" and "Hs_helaC_mar_SGF_poisson" also contribute to the description of stoichiometric balance and protein abundance, validating their effective roles in the identification of actual PPIs.

Apart from such stoichiometric balance and protein abundance associated features, the following ten features can be further divided into two groups describing the molecular weight ("Ce_CRF_wan_60_1209_wcc," "Ce_BNF_wan_60_1209_wcc") and charge distribution ("Ce_CRF_wan_60_1209_pq_euc," "Ce_BNF_wan_60_1209_pq_euc," "Ce_beadsflow_1206 _pq_euc," "Ce_1111_pq_euc," "Ce_beadsL_1206_pq_euc," "Ce_6mg_1203_pq_euc") of related proteins, respectively. The features that possibly affect the PPIs might be the molecular weight and the charge distribution of each PPI participant. These features have been validated by recent publications.

For instance, a study on SG2NA protein variants confirmed that the molecular weight and structure of such protein may directly affect its binding affinity against its ligands (Mitterhuber, 2008; Soni et al., 2014; Pinton et al., 2015). Therefore, molecular weight induced by different amino acid substitution may affect PPIs. The associations among different proteins were reported to be possibly strongly affected by long-range electrostatic interactions, and similar proteins with different surface charges may have different interaction patterns (Twomey et al., 2013; Raut and Kalonia, 2015). Therefore, the charge distribution of PPI participants affected the interactions between proteins.

Analysis of Optimal PPI Rules

Based on the detailed parameter that corresponds to each optimal PPI feature extracted from the three datasets, the relatively quantitative rules to recognize potential PPIs were inferred (**Table 3**). The features that describe sensitivity gain factor confirmed that the PPI features and their parameters extracted from different datasets should be comparable, and the detailed analysis of each optimal PPI rule could be derived in the following discussions.

The literature confirmed rules with proper parameters may contribute to identifying potential PPIs and such predicted rules may act as reference for the prediction and screening of novel PPIs. In terms of the detailed quantitative features, two specific parameters, namely, "neg_ln_pval" and "hein_neg_ln_pval," were identified in Rule1-Rule5. High relative (parameter) value of such two features indicate the interaction may actually happen. Although the detailed parameter (threshold) cannot be validated through wet-experiments at present, proper stoichiometric balance and protein abundance indicated by the parameters were discussed previously and already confirmed to promote the PPIs according to recent publications (Vinayagam et al., 2011; Fairweather et al., 2015). These rules could also be grouped in accordance with their new insights into the detailed biological mechanisms:

Apart from such two features, another two features have also been screened out to contribute to the quantitative identification of actual PPIs: "Hs_G166_1104_pq_euc" (used in Rule2-Rule5) and "pair_count" (used in Rule3-Rule5). In all the top rules apart from the first one which only involves "neg_ln_pval" and "hein_neg_ln_pval" as we have mentioned above, the value of "Hs_G166_1104_pq_euc" turns out to be lower than zero according to our quantitative rules.

According to the analyses above, such parameter contributes to the description of the charge distribution of certain PPI participants. Although no accurate description of such parameter, it has been confirmed that the higher the value is, the lower surface charging the participants of potential PPIs carries. Considering that it has been reported that charge interactions play an irreplaceable role for actual PPIs, therefore, potential interactions with such parameter lower than zero may not be actual PPIs. As for another parameter named as "pair_count," in Rule3-Rule5, such parameter has a value >2, 3, and 3. It has been reported that the higher the value of such parameter may be, the less possible such interaction may actual happens (Hein et al., 2015). Therefore, interactions breaking such top five rules turns out to be actual PPIs, corresponding with our analyses above.

CONCLUSION

Protein is the basic molecule of life. Through protein-protein interactions, complex biological processes are carried out. Predict PPI is a fundamental problem in bioinformatics. In this study, we encoded protein with various physical and chemical features, such as stoichiometric balance, protein abundance, molecular weight, and charge distribution. Then with advanced feature selection methods, we identified the key factors affecting PPIs and built a quantitative decision-rule system to evaluate the potential of PPIs under real conditions. Our results provided novel insights of the molecular mechanisms of PPIs. The model can be extended to explore other molecular interaction questions. The main datasets and codes can be downloaded at https://github.com/xypan1232/Mass-PPI.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: the datasets for this study can be found in http://proteincomplexes.org/download.

AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. XP, KF, and LC performed the experiments. TZ and Y-HZ analyzed the results. XP and TZ wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This study was funded by the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Key R&D Program of China (2018YFC0910403), National Natural Science Foundation of China (31701151), Natural Science Foundation of Shanghai (17ZR1412500), Shanghai Sailing Program (16YF1413800), and the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00730/full#supplementary-material

REFERENCES

Beqollari, D., Romberg, C. F., Filipova, D., Meza, U., Papadopoulos, S., and Bannister, R. A. (2015). Rem uncouples excitation-contraction coupling in adult skeletal muscle fibers. *J. Gen. Physiol.* 146, 97–108. doi: 10.1085/jgp.201411314

Blagus, R., and Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* 14:106. doi: 10.1186/1471-2105-14-106

Breiman, L. (2001). Random forests. *Machine Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, L., Chu, C., Zhang, Y.-H., Zheng, M.-Y., Zhu, L., Kong, X., et al. (2017a). Identification of drug-drug interactions using chemical interactions. *Curr. Bioinformatics* 12, 526–534. doi: 10.2174/1574893611666160618094219

Chen, L., Li, J., Zhang, Y. H., Feng, K., Wang, S., Zhang, Y., et al. (2018a). Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J. Cell Biochem.* 119, 3394–3403. doi: 10.1002/jcb.26507

Chen, L., Pan, X., Hu, X., Zhang, Y.-H., Wang, S., Huang, T., et al. (2018b). Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* 143, 1731–1740. doi: 10.1002/ijc.31554

Chen, L., Pan, X., Zhang, Y.-H., Liu, M., Huang, T., and Cai, Y.-D. (2019a). Classification of widely and rarely expressed genes with recurrent neural network. *Comput. Struct. Biotechnol. J.* 17, 49–60. doi: 10.1016/j.csbj.2018.12.002

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., et al. (2017b). Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* 5, 26582–26590. doi: 10.1109/ACCESS.2017.2775703

Chen, L., Zhang, S., Pan, X., Hu, X., Zhang, Y. H., Yuan, F., et al. (2019b). HIV infection alters the human epigenetic landscape. *Gene Ther.* 26, 29–39. doi: 10.1038/s41434-018-0051-6

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., et al. (2019). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis *and modeling* of DNA, RNA and protein sequence data. *Brief. Bioinform.* 21, 1047–1057. doi: 10.1093/bib/bbz041

Chen, Z., Zhao, P., Li, F. Y., Leier, A., Marquez-Lago, T. T., Wang, Y. N., et al. (2018). iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

Cohen, W.W. (1995). "Fast effective rule induction," in *The Twelfth International Conference on Machine Learning* (Tahoe City, CA), 115–123.

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Cui, H., and Chen, L. (2019). A binary classifier for the prediction of EC numbers of enzymes. *Curr. Proteomics* 16, 381–389. doi: 10.2174/1570164616666190126103036

De Las Rivas, J., and Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.* 6:e1000807. doi: 10.1371/journal.pcbi.1000807

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2008). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Drew, K., Lee, C., Huizar, R. L., Tu, F., Borgeson, B., Mcwhite, C. D., et al. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.* 13:932. doi: 10.15252/msb.20167490

Fairweather, S. J., Broer, A., Subramanian, N., Tumer, E., Cheng, Q., Schmoll, D., et al. (2015). Molecular basis for the interaction of the mammalian amino acid transporters B0AT1 and B0AT3 with their ancillary protein collectrin. *J. Biol. Chem.* 290, 24308–24325. doi: 10.1074/jbc.M115.648519

Gonzalez, L. C. (2012). Protein microarrays, biosensors, and cell-based methods for secretome-wide extracellular protein-protein interaction mapping. *Methods* 57, 448–458. doi: 10.1016/j.ymeth.2012.06.004

Hein, M. Y., Hubner, N. C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* 163, 712–723. doi: 10.1016/j.cell.2015.09.053

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., et al. (2015). The bioPlex network: a systematic exploration of the human interactome. *Cell* 162, 425–440. doi: 10.1016/j.cell.2015.06.043

Johannes, F., and Widmer, G. (1994). *"Incremental Reduced Error Pruning," in Machine Learning: Proceedings of the Eleventh Annual Conference* (New Brunswick, NJ).

Kursa, M. B., and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J. Stat. Softw.* 36, 1–13. doi: 10.18637/jss.v036.i11

Levin, L., Zelzion, E., Nachliel, E., Gutman, M., Tsfadia, Y., and Einav, Y. (2013). A single disulfide bond disruption in the beta3 integrin subunit promotes thiol/disulfide exchange, a molecular dynamics study. *PLoS ONE* 8:e59175. doi: 10.1371/annotation/b4e96e4b-3106-4040-a63c-a3f018f0e5c0

Li, F., Li, C., Marquez-Lago, T. T., Leier, A., Akutsu, T., Purcell, A. W., et al. (2018). Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. *Bioinformatics* 34, 4223–4231. doi: 10.1093/bioinformatics/bty522

Li, F., Li, C., Revote, J., Zhang, Y., Webb, G. I., Li, J., et al. (2016). GlycoMine(struct): a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features. *Sci. Rep.* 6:34595. doi: 10.1038/srep34595

Li, F., Li, C., Wang, M., Webb, G. I., Zhang, Y., Whisstock, J. C., et al. (2015). GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics* 31, 1411–1419. doi: 10.1093/bioinformatics/btu852

Li, J., Lu, L., Zhang, Y. H., Liu, M., Chen, L., Huang, T., et al. (2019). Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell Biochem.* 120, 405–416. doi: 10.1002/jcb.27395

Lin, X. M., Yang, M. J., Li, H., Wang, C., and Peng, X. X. (2014). Decreased expression of LamB and Odp1 complex is crucial for antibiotic resistance in *Escherichia coli*. *J. Proteomics* 98, 244–253. doi: 10.1016/j.jprot.2013.12.024

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230. doi: 10.1023/A:1008363719778

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Mitterhuber, M. (2008). *The Role of PP2A Phosphatase Activator (PTPA) in the Biogenesis of PP2A in Mammalian Cells*. Vienna: University of Vienna.

Modell, A. E., Blosser, S. L., and Arora, P. S. (2016). Systematic targeting of protein-protein interactions. *Trends Pharmacol. Sci.* 37, 702–713. doi: 10.1016/j.tips.2016.05.008

Pan, X., Hu, X., Zhang, Y.-H., Chen, L., Zhu, L., Wan, S., et al. (2019). Identification of the copy number variant biomarkers for breast cancer subtypes. *Mol. Genet. Genomics* 294, 95–110. doi: 10.1007/s00438-018-1488-4

Pan, X., Hu, X., Zhang, Y. H., Feng, K., Wang, S. P., Chen, L., et al. (2018). Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes* 9:208. doi: 10.3390/genes9040208

Pan, X. Y., Zhang, Y. N., and Shen, H. B. (2010). Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *J. Proteome Res.* 9, 4992–5001. doi: 10.1021/pr100618t

Pan, X. Y., Zhu, L., Fan, Y. X., and Yan, J. C. (2014). Predicting protein-RNA interaction amino acids using random forest based on submodularity subset selection. *Comp. Biol. Chem.* 53, 324–330. doi: 10.1016/j.compbiolchem.2014.11.002

Pinton, L., Borroto-Escuela, D. O., Narváez, M., Oflijan, J., Agnati, L. F., and Fuxe, K. (2015). Evidence for the existence of dopamine D2R and Sigma 1 allosteric receptor-receptor interaction in the rat brain: role in brain plasticity and cocaine action. *SpringerPlus* 4:P37. doi: 10.1186/2193-1801-4-S1-P37

Qian, W., Zhou, H., and Tang, K. (2014). Recent coselection in human populations revealed by protein-protein interaction network. *Genome Biol. Evol.* 7, 136–153. doi: 10.1093/gbe/evu270

Raj, M., Bullock, B. N., and Arora, P. S. (2013). Plucking the high hanging fruit: a systematic approach for targeting protein-protein interactions. *Bioorg. Med. Chem.* 21, 4051–4057. doi: 10.1016/j.bmc.2012.11.023

Raut, A. S., and Kalonia, D. S. (2015). Liquid-liquid phase separation in a dual variable domain immunoglobulin protein solution: effect of formulation factors and protein-protein interactions. *Mol. Pharm.* 12, 3261–3271. doi: 10.1021/acs.molpharmaceut.5b00256

Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N. D., Webb, G. I., et al. (2018). iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinform.* 20, 638–658. doi: 10.1093/bib/bby028

Soni, S., Tyagi, C., Grover, A., and Goswami, S. K. (2014). Molecular modeling and molecular dynamics simulations based structural analysis of the SG2NA protein variants. *BMC Res. Notes* 7:446. doi: 10.1186/1756-0500-7-446

Swiatkowska, M., Szymanski, J., Padula, G., and Cierniewski, C. S. (2008). Interaction and functional association of protein disulfide isomerase with alphaVbeta3 integrin on endothelial cells. *FEBS J.* 275, 1813–1823. doi: 10.1111/j.1742-4658.2008.06339.x

Twomey, E. C., Cordasco, D. F., Kozuch, S. D., and Wei, Y. (2013). Substantial conformational change mediated by charge-triad residues of the death effector domain in protein-protein interactions. *PLoS ONE* 8:e83421. doi: 10.1371/journal.pone.0083421

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., et al. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* 4:rs8. doi: 10.1126/scisignal.2001699

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., et al. (2015). Panorama of ancient metazoan macromolecular complexes. *Nature* 525, 339–344. doi: 10.1038/nature14877

Wang, D., Li, J.-R., Zhang, Y.-H., Chen, L., Huang, T., and Cai, Y.-D. (2018). Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes* 9:155. doi: 10.3390/genes9030155

Wang, S. B., and Huang, T. (2018). The early detection of asthma based on blood gene expression. *Mol. Biol. Rep.* 46, 217–223. doi: 10.1007/s11033-018-4463-6

Wang, W., Li, X., Huang, J., Feng, L., Dolinta, K. G., and Chen, J. (2014). Defining the protein-protein interaction network of the human hippo pathway. *Mol. Cell Proteomics* 13, 119–131. doi: 10.1074/mcp.M113.030049

Witten, I.H., and Frank, E. (eds.). (2005). *Data Mining: Practical Machine Learn. Tools and Techniques.* San Francisco, CA: Morgan, Kaufmann.

Zhang, T. M., Huang, T., and Wang, R. F. (2018). Cross talk of chromosome instability, CpG island methylator phenotype and mismatch repair in colorectal cancer. *Oncol. Lett.* 16, 1736–1746. doi: 10.3892/ol.2018.8860

Zhang, X., Chen, L., Guo, Z.-H., and Liang, H. (2019). Identification of human membrane protein types by incorporating network embedding methods. *IEEE Access* 7, 140794–140805. doi: 10.1109/ACCESS.2019.2944177

Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinformatics* 14, 709–720. doi: 10.2174/1574893614666190220114644

Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* 306, 136–144. doi: 10.1016/j.mbs.2018.09.010

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 July 2020
doi: 10.3389/fgene.2020.00814

Integrated Bioinformatics Analysis Reveals Key Candidate Genes and Pathways Associated With Clinical Outcome in Hepatocellular Carcinoma

Yubin Li[1†], Runzhe Chen[2,3†], Jian Yang[1†], Shaowei Mo[4], Kelly Quek[2,3,5], Chung H. Kok[6,7], Xiang-Dong Cheng[8,9,10], Saisai Tian[1], Weidong Zhang[1*] and Jiang-Jiang Qin[4,8,9,10*]*

[1] School of Pharmacy, Naval Medical University, Shanghai, China, [2] Department of Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [3] Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [4] The First Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, China, [5] Accenture Applied Intelligence, ASEAN, Singapore, Singapore, [6] Precision Medicine Theme, South Australian Health and Medical Research Institute, Adelaide, SA, Australia, [7] Discipline of Medicine, Adelaide Medical School, The University of Adelaide, Adelaide, SA, Australia, [8] Institute of Cancer and Basic Medicine, Chinese Academy of Sciences, Hangzhou, China, [9] Cancer Hospital of the University of Chinese Academy of Sciences, Hangzhou, China, [10] Zhejiang Cancer Hospital, Hangzhou, China

OPEN ACCESS

Edited by:
Yucong Duan,
Hainan University, China

Reviewed by:
Manal Said Fawzy,
Suez Canal University, Egypt
Laura Lynn Montier,
Baylor College of Medicine,
United States

***Correspondence:**
Saisai Tian
372479584@qq.com
Weidong Zhang
wdzhangy@hotmail.com
Jiang-Jiang Qin
jqin@zcmu.edu.cn;
zylyzcmu@126.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 26 February 2020
Accepted: 06 July 2020
Published: 24 July 2020

Citation:
Li Y, Chen R, Yang J, Mo S, Quek K, Kok CH, Cheng X-D, Tian S, Zhang W and Qin J-J (2020) Integrated Bioinformatics Analysis Reveals Key Candidate Genes and Pathways Associated With Clinical Outcome in Hepatocellular Carcinoma. Front. Genet. 11:814. doi: 10.3389/fgene.2020.00814

Hepatocellular carcinoma (HCC) accounts for approximately 85–90% of all liver cancer cases and has poor relapse-free survival. There are many gene expression studies that have been performed to elucidate the genetic landscape and driver pathways leading to HCC. However, existing studies have been limited by the sample size and thus the pathogenesis of HCC is still unclear. In this study, we performed an integrated characterization using four independent datasets including 320 HCC samples and 270 normal liver tissues to identify the candidate genes and pathways in the progression of HCC. A total of 89 consistent differentially expression genes (DEGs) were identified. Gene-set enrichment analysis revealed that these genes were significantly enriched for cellular response to zinc ion in biological process group, collagen trimer in the cellular component group, extracellular matrix (ECM) structural constituent conferring tensile strength in the molecular function group, protein digestion and absorption, mineral absorption and ECM-receptor interaction. Network system biology based on the protein–protein interaction (PPI) network was also performed to identify the most connected and important genes based on our DEGs. The top five hub genes including osteopontin (*SPP1*), Collagen alpha-2(I) chain (*COL1A2*), Insulin-like growth factor I (*IGF1*), lipoprotein A (*LPA*), and Galectin-3 (*LGALS3*) were identified. Western blot and immunohistochemistry analysis were employed to verify the differential protein expression of hub genes in HCC patients. More importantly, we identified that these five hub genes were significantly associated with poor disease-free survival and overall survival. In summary, we have identified a potential clinical significance of these genes as prognostic biomarkers for HCC patients who would benefit from experimental approaches to obtain optimal outcome.

Keywords: hepatocellular carcinoma, differentially expression genes, enrichment analysis, survival analysis, prognosis

INTRODUCTION

Liver cancer is the fourth leading cause of cancer-related death worldwide and ranks sixth in terms of incidence (Bray et al., 2018; Villanueva, 2019). Among all types of primary malignant liver tumors, hepatocellular carcinoma (HCC) accounts for approximately 85–90% of all cases. The major risk factors including chronic infections by hepatitis B virus (HBV) and hepatitis C virus (HCV), aflatoxin exposure, smoking, type 2 diabetes, obesity, and so on (Marengo et al., 2016; Bray et al., 2018; Phukan et al., 2018). As a highly heterogeneous cancer disease, localized HCC patients often have poor prognosis with 5-year overall survival (OS) rate of 30%, and this rate drops below 5% for those with distant metastases (Oweira et al., 2017). For patients at early disease stages, liver resection is the most effective treatment option, however, only less than 30% of HCC patients are eligible surgery, and among those around 70% eventually relapse within 5 years after treatment (Waghray et al., 2015). Over the past few decades, despite advances in chemotherapy, targeted therapy, radiation therapy, and immunotherapy in the clinical arena, the survival of HCC patients has not significantly increased, and translational studies to understand the mechanisms and prognosis remain underwhelming to design novel therapeutic strategies (Visvader, 2011; Aravalli et al., 2013; Llovet et al., 2018).

Data, information, knowledge and wisdom (DIKW) model has been widely used in life in all aspects including medicine (Song et al., 2018, 2020; Duan, 2019a,b; Duan et al., 2019a,b). In recent years, genome-wide profiling has substantially advanced our understanding of the genetic landscape and driver pathways leading to HCC (Totoki et al., 2014; Schulze et al., 2015; Zucman-Rossi et al., 2015; Ally et al., 2017; Villanueva, 2019), revealing Cellular tumor antigen p53 (*TP53*), Catenin beta-1 (*CTNNB1*), Axin-1 (*AXIN1*), Telomerase reverse transcriptase (*TERT*) promoter and other key genes as driver mutations, and WNT/β-catenin, p53 cell cycle pathway, oxidative stress, PI3K/AKT/MTOR, and RAS/RAF/MAPK pathways as key signaling pathways involved in liver carcinogenesis. However, existing studies have been of limited sample size that failed to create molecular prognostic indices and also the inconsistent computational methods may have restricted the power to identify potential meaningful molecular biomarkers and new therapeutic targets. Therefore, an integrated bioinformatics study combining the most updated genomic data thus providing novel insight into the mechanisms underlying therapeutic resistance and disease progression is highly warranted.

Microarray technology has become an indispensable tool to monitor genome wide expression levels of genes in a given organism and has been successfully used to classify different types of cancer and predict clinical outcomes (Trevino et al., 2007). These microarray technologies have also been applied in many studies to define global gene expression patterns in primary human HCC in an attempt to gain insight into the mechanisms of hepatocarcinogenesis (Crawley and Furge, 2002; Woo et al., 2008; Hoshida et al., 2009; Villanueva et al., 2012; Jin et al., 2015). In the present study, we selected four independent datasets consisting a total of 320 HCC cases and 270 cases of normal liver tissues in the Gene Expression Omnibus (GEO) database to identify reliable markers and pathway alterations linked with the pathogenesis of HCC cases (Wurmbach et al., 2007; Mas et al., 2009; Roessler et al., 2010). We identified 89 differential expression genes (DEGs) including 31 up-regulated genes and 58 down-regulated genes. Gene ontology (GO) analysis revealed cellular response to zinc ion in biological process (BP) group, collagen trimer in the cellular component (CC) group, and extracellular matrix (ECM) structural constituent conferring tensile strength in the molecular function (MF) group. Further pathway enrichment analysis revealed that enrichment in protein digestion and absorption, mineral absorption, propanoate metabolism, and ECM-receptor interaction. Finally, the top five hub genes osteopontin (*SPP1*), Collagen alpha-2(I) chain (*COL1A2*), Insulin-like growth factor I (*IGF1*), lipoprotein A (*LPA*), and Galectin-3 (*LGALS3*) were identified from the protein–protein interaction (PPI) network and those highly altered genes were validated by western blot assay and Immunohistochemistry (IHC) analysis and found to be associated with clinical outcome of HCC patients.

MATERIALS AND METHODS

Data Source and Identification of DEGs

Microarrays data were obtained from the Oncomine 4.5 database[1] contains 715 datasets and 86,733 samples. Of which, we filtered four datasets comprising Mas liver (GSE14323, containing 19 liver tissues and 38 HCCs), Roessler liver (GSE14520 based on GPL571 platform, containing 21 liver tissues and 22 HCCs), Roessler liver 2 (GSE14520 based on GPL3921 platform, containing 220 liver tissues and 225 HCCs), and Wurmbach liver (GSE6764, containing 10 liver tissues and 35 HCCs) after using the following criteria: (a) Analysis type: cancer vs. normal analysis; (b) Cancer type: hepatocellular carcinoma; (c) Data type: mRNA; (d) Sample type: clinical specimen; (e) Microarray platform: Human Genome U133A, U133A 2.0, or U133 Plus 2.0. A total of 270 cases of normal liver tissues and 320 cases of HCCs were included in the integrated analysis. To analyze the DEGs between HCC and normal liver tissues, the data were then processed on GEO2R website[2]. The differentially expressed genes were identified using limma R package at a cutoff $|\log FC| > 1$ and adjusted p value < 0.05 (Benjamini & Hochberg).

GO and Pathways Enrichment Analysis

The annotation function of GO analysis is comprised of three categories: BP, CC, and MF. Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource for understanding high-level functions and utilities of the genes or proteins (Kanehisa and Goto, 2000; Kanehisa et al., 2012, 2016). GO analysis and KEGG pathway enrichment analysis of candidate DEGs were performed using the R package "clusterProfiler."

[1]https://www.oncomine.org/

[2]https://www.ncbi.nlm.nih.gov/geo/geo2r/

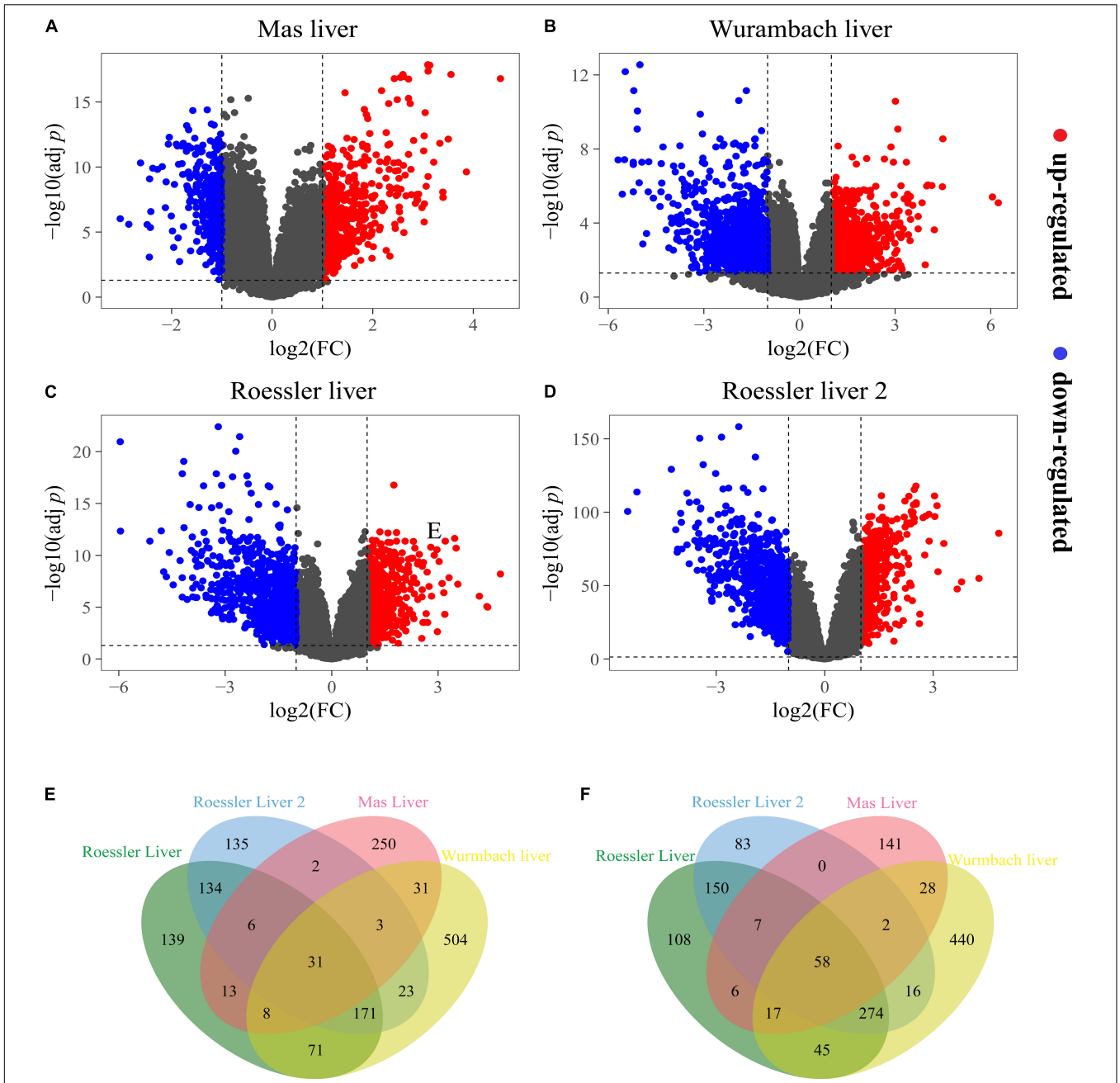


FIGURE 1 | The DEGs screened from four independent datasets. Up-regulated DEGs (red-colored dots) and down-regulated (blue-colored dots) DEGs are selected with [logFC] > 1 and adjust *p*-value < 0.05 from the mRNA expression profiling sets **(A)** Mas liver, **(B)** Wurmbach liver, **(C)** Roessler liver, and **(D)** Roessler liver 2. Venn diagram showed **(E)** 31 consistently up-regulated DEGs and **(F)** 58 consistently down-regulated DEGs in four datasets.

TABLE 1 | Details of the four HCC datasets.

Datasets	GSE	Tumor	Normal	References
Mas liver	GSE14323	38	19	Mas et al., 2009
Roessler liver	GSE14520(GPL571 platform)	22	21	Roessler et al., 2010
Roessler liver 2	GSE14520(GPL3921 platform)	225	220	Roessler et al., 2010
Wurmbach liver	GSE6764	35	10	Wurmbach et al., 2007

HCC, Hepatocellular carcinoma.

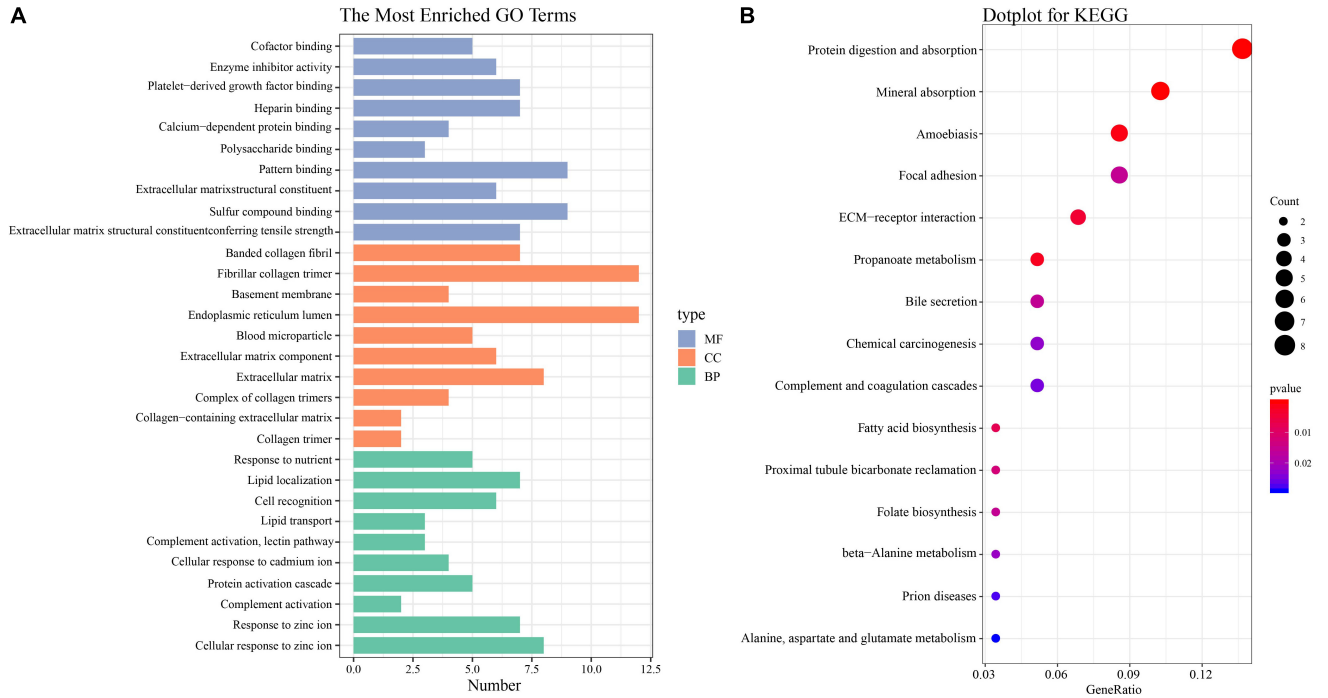


FIGURE 2 | GO and KEGG pathway enrichment analysis **(A)** GO analysis of all DEGs. The most enriched GO terms are listed in the diagram. **(B)** KEGG pathway analysis of all DEGs. The most enriched KEGG pathways are shown in the picture.

Reactome[3] was also used for pathway enrichment analysis (Fabregat et al., 2018). Adjusted *p*-value less than 0.05 was considered as the cut-off criterion for both GO analysis and pathway enrichment analysis.

PPI Network and Modular Analysis

Protein–protein interaction network was constructed to determine the importance of these DEGs by comparing the interactions between different DEGs. STRING database[4] and Cytoscape software (3.7.2 version) were applied to construct and visualize the PPI networks (Szklarczyk et al., 2017), followed by Molecular Complex Detection (MCODE) plug-in in Cytoscape for selecting significant modules of hub genes from the PPI network (Bader and Hogue, 2003), with the following criteria: degree cutoff (number of connections with other nodes) $\geq$ 2, node score cutoff (the most influential parameter for cluster size) $\geq$ 2, K-core (This parameter filters out clusters that do not contain a maximally interconnected sub-cluster of at least k degrees. For example, a triangle including three nodes and three edges is a two-core representing two connections per node. Two nodes with two edges between them meet the two-core rule as well) $\geq$ 2 and max depth (this parameter limits the distance from the seed node within which MCODE can search for cluster members) = 100. KEGG pathway enrichment analysis of the modules was carried out using the online DAVID database[5] (Huang da et al., 2009).

Hub Gene Selection and Prognostic Analysis

Hub genes were selected based on comparison of top 10 genes ranked by degree and betweenness centrality Network of hub genes. Their co-expressed genes were then analyzed using cBioPortal online platform[6] (Gao et al., 2013). Genetic alterations of these hub genes were explored and compared using the cBioPortal database. Biological process analysis of hub genes was then performed and visualized using plug-in Biological Networks Gene Oncology tool (BiNGO) app in Cytoscape software (Maere et al., 2005). Stage-related information analysis based on gene expression was performed in UALCAN[7], a comprehensive web resource for analyzing omics data (Chandrashekar et al., 2017). Disease-free survival (DFS) is a concept used to describe the period after a successful treatment of cancer. OS means the length of time from either the date of diagnosis or the start of treatment for HCC. DFS and OS are both measured to see how well a new treatment works. DFS and OS analysis associated with these hub genes were performed using the Kaplan–Meier Plotter online database[8].

[3]https://reactome.org/

[4]https://string-db.org/

[5]https://david.ncifcrf.gov/

[6]https://www.cbioportal.org/

[7]http://ualcan.path.uab.edu/analysis.html

[8]http://www.kmplot.com/

TABLE 2 | Significantly enriched GO terms of DEGs associated with HCC with adjust *p*-value < 0.01.

Expression	Category	Term	Count	*p*-value	Adj *p*-value
Up-regulated	CC	GO:0062023~collagen-containing extracellular matrix	9	<0.001	<0.001
	CC	GO:0044420~extracellular matrix component	5	<0.001	<0.001
	CC	GO:0098644~complex of collagen trimers	4	<0.001	<0.001
	CC	GO:0031012~extracellular matrix	9	<0.001	<0.001
	CC	GO:0005581~collagen trimer	5	<0.001	<0.001
	CC	GO:0005788~endoplasmic reticulum lumen	7	<0.001	<0.001
	CC	GO:0005604~basement membrane	4	<0.001	<0.001
	CC	GO:0042470~melanosome	4	<0.001	<0.001
	CC	GO:0048770~pigment granule	4	<0.001	<0.001
	CC	GO:0005583~fibrillar collagen trimer	2	<0.001	0.001
	CC	GO:0098643~banded collagen fibril	2	<0.001	0.001
	MF	GO:0030020~extracellular matrix structural constituent conferring tensile strength	5	<0.001	<0.001
	MF	GO:0005201~extracellular matrix structural constituent	5	<0.001	0.001
	MF	GO:0048407~platelet-derived growth factor binding	2	<0.001	0.008
Down-regulated	BP	GO:0071294~cellular response to zinc ion	5	<0.001	<0.001
	BP	GO:0010043~response to zinc ion	6	<0.001	<0.001
	BP	GO:0006956~complement activation	7	<0.001	<0.001
	BP	GO:0072376~protein activation cascade	7	<0.001	<0.001
	BP	GO:0071276~cellular response to cadmium ion	4	<0.001	0.001
	BP	GO:0001867~complement activation, lectin pathway	3	<0.001	0.001
	BP	GO:0006959~humoral immune response	7	<0.001	0.003
	BP	GO:0046686~response to cadmium ion	4	<0.001	0.004
	BP	GO:0010460~positive regulation of heart rate	3	<0.001	0.008
	BP	GO:0010038~response to metal ion	7	<0.001	0.008
	CC	GO:0072562~blood microparticle	5	<0.001	0.008
	MF	GO:0001871~pattern binding	3	<0.001	0.004
	MF	GO:0030247~polysaccharide binding	3	<0.001	0.004
	MF	GO:1901681~sulfur compound binding	6	<0.001	0.005
	MF	GO:0050662~coenzyme binding	6	<0.001	0.008

BP, biological process; CC, cellular component; MF, molecular function.

TABLE 3 | KEGG pathway enrichment analysis of DEGs in HCC.

Expression	KEGG Term	Count	*p*-value	Adj *p*-value
Up-regulated	hsa04974~Protein digestion and absorption	6	<0.001	<0.001
	hsa04512~ECM-receptor interaction	4	<0.001	0.003
	hsa04964~Proximal tubule bicarbonate reclamation	2	0.002	0.028
	hsa04510~Focal adhesion	4	0.002	0.028
	hsa04151~*PI3K/Akt* signaling pathway	5	0.002	0.028
	hsa04933~*AGE-RAGE* signaling pathway in diabetic complications	3	0.003	0.028
	hsa05146~Amoebiasis	3	0.003	0.028
	hsa04926~Relaxin signaling pathway	3	0.005	0.048
Down-regulated	hsa04978~Mineral absorption	5	<0.001	0.001
	hsa00640~Propanoate metabolism	3	0.001	0.025

Cell Lines and Cell Culture

Hepatocellular carcinoma cell lines Hep3B and HepG2 and human normal liver cell line L02 were obtained from Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences. All of these cell lines were cultured in Dulbecco's modified Eagle's medium (DMEM) (catalog number 10569010, Gibco) containing 10% (v/v) fetal bovine serum (FBS) (catalog number 10091148, Gibco) supplemented with 1% (v/v) penicillin streptomycin solution (catalog number SV30010, Hyclone) (containing 100 U/ml penicillin and 100 μg/ml streptomycin) in a humidified incubator at 37°C with 5% CO_2.

Protein Preparation and Western Blot Analysis

Briefly, HCC cells were lysed with cold M-PER lysate buffer (catalog number 78501, Roche) [containing 1 × protease

TABLE 4 | Pathways enriched in Reactome analysis of DEGs in HCC (Adj *p*-value < 0.01).

Expression	Pathway name	Count	*p*-value	Adj *p*-value
Up-regulated	R-HSA-8948216~Collagen chain trimerization	5	<0.001	<0.001
	R-HSA-2022090~Assembly of collagen fibrils and other multimeric structures	5	<0.001	<0.001
	R-HSA-1442490~Collagen degradation	5	<0.001	<0.001
	R-HSA-1650814~Collagen biosynthesis and modifying enzymes	5	<0.001	<0.001
	R-HSA-216083~Integrin cell surface interactions	6	<0.001	<0.001
	R-HSA-1474228~Degradation of the extracellular matrix	6	<0.001	<0.001
	R-HSA-1474290~Collagen formation	5	<0.001	<0.001
	R-HSA-3000171~Non-integrin membrane-ECM interactions	4	<0.001	<0.001
	R-HSA-1474244~Extracellular matrix organization	6	<0.001	<0.001
	R-HSA-2214320~Anchoring fibril formation	3	<0.001	0.001
	R-HSA-422475~Axon guidance	8	<0.001	0.002
	R-HSA-8985801~Regulation of cortical dendrite branching	1	<0.001	0.002
	R-HSA-2243919~Crosslinking of collagen fibrils	3	<0.001	0.002
	R-HSA-186797~Signaling by *PDGF*	4	<0.001	0.003
	R-HSA-8941333~*RUNX2* regulates genes involved in differentiation of myeloid cells	1	<0.001	0.004
	R-HSA-3000178~ECM proteoglycans	4	<0.001	0.004
	R-HSA-69205~G1/S-Specific Transcription	2	0.001	0.009
	R-HSA-3769402~Deactivation of the beta-catenin transactivating complex	3	0.001	0.009
	R-HSA-419037~*NCAM1* interactions	3	0.001	0.009
	R-HSA-8949275~*RUNX3* Regulates Immune Response and Cell Migration	1	0.001	0.009
	R-HSA-8939246~*RUNX1* regulates transcription of genes involved in differentiation of myeloid cells	1	0.001	0.010
	R-HSA-3000480~Scavenging by Class A Receptors	3	0.001	0.010
Down-regulated	R-HSA-5661231~Metallothioneins bind metals	5	<0.001	<0.001
	R-HSA-5660526~Response to metal ions	5	<0.001	<0.001
	R-HSA-2855086~Ficolins bind to repetitive carbohydrate structures on the target cell surface	3	<0.001	<0.001
	R-HSA-166662~Lectin pathway of complement activation	3	<0.001	<0.001
	R-HSA-166658~Complement cascade	6	<0.001	0.003

inhibitors (catalog number 11836153001, Roche) and phosphatase inhibitor cocktail (catalog number 78420, Roche)] and centrifuged at 4°C for 10 min. The protein concentrations of collected supernatants were determined by the BCA protein assay kit (catalog number P0011, Beyotime). Equal amounts of total proteins were separated in 10 or 12% SDS-PAGE and transblotted onto the 0.45 μm PVDF membranes (catalog number 1620177, BIO-RAD). The membranes were blocked in 5% fat-free milk in TBST (150 mM NaCl, 50 mM Tris, pH 7.2) for 1 h at room temperature and subsequently incubated with corresponding primary antibodies as following: anti-SPP1 (catalog number ab8448, Abcam, 1:1000), anti-COL1A2 (catalog number 66761-1-lg, Proteintech, 1:1000), anti-IGF1 (catalog number ab9572, Abcam, 1:1000), anti-LGALS3 (catalog number ab209344, Abcam, 1:1000), anti-GADPH (catalog number 10494-1-AP, Proteintech, 1:10000), and anti-β-Tubulin (catalog number T0023, Affinity, 1:20000) at 4°C overnight, followed by incubation with a donkey anti-mouse (catalog number C61116-02, LI-COR) or goat anti-rabbit (catalog number C80118-05, LI-COR) secondary antibody for 1 h at room temperature. Then membranes were scanned using the Odyssey infrared imaging system (LI-COR) and the images were captured. The gray levels of the bands were determined by Image J software. The expression of proteins was normalized using the GADPH or β-Tubulin values. The assay was performed three independent times.

Patient Samples

This study was approved by Cancer Hospital of the University of Chinese Academy of Sciences; Zhejiang Cancer Hospital. Twenty formalin-fixed, paraffin-embedded (FFPE) HCC tissues and corresponding adjacent non-cancerous tissues were collected from the Department of Abdominal Surgery, Zhejiang Cancer Hospital. All FFPE HCC tissues were screened by two pathologists independently to confirm the diagnosis of HCC. The most representative tumor and non-cancerous tissues were selected for immunohistochemistry analysis.

IHC Analysis

Neutral 10% buffered formalin-fixed tissue specimens were embedded in paraffin wax and then sliced to 4-micron thick sections by a microtome. In brief, the tissue slices were firstly deparaffinized, followed by rehydration and a 10-min boiling in 10 mmol/L citrate buffer (pH = 6.4) for antigen retrieval. Then, the sections were treated in methanol containing 3% H_2O_2 for 20 min to inhibit the endogenous tissue peroxidase activity. After being blocked with 1% bovine serum albumin (BSA) at 37°C for 30 min, IHC staining was carried out for the protein expression of SPP1, COL1A2, IGF1, and LGALS3 using specific primary antibodies at 4°C overnight, followed by staining with species-specific secondary antibodies labeled with horseradish peroxidase (HRP). The slides were developed in diaminobenzidine (DAB)

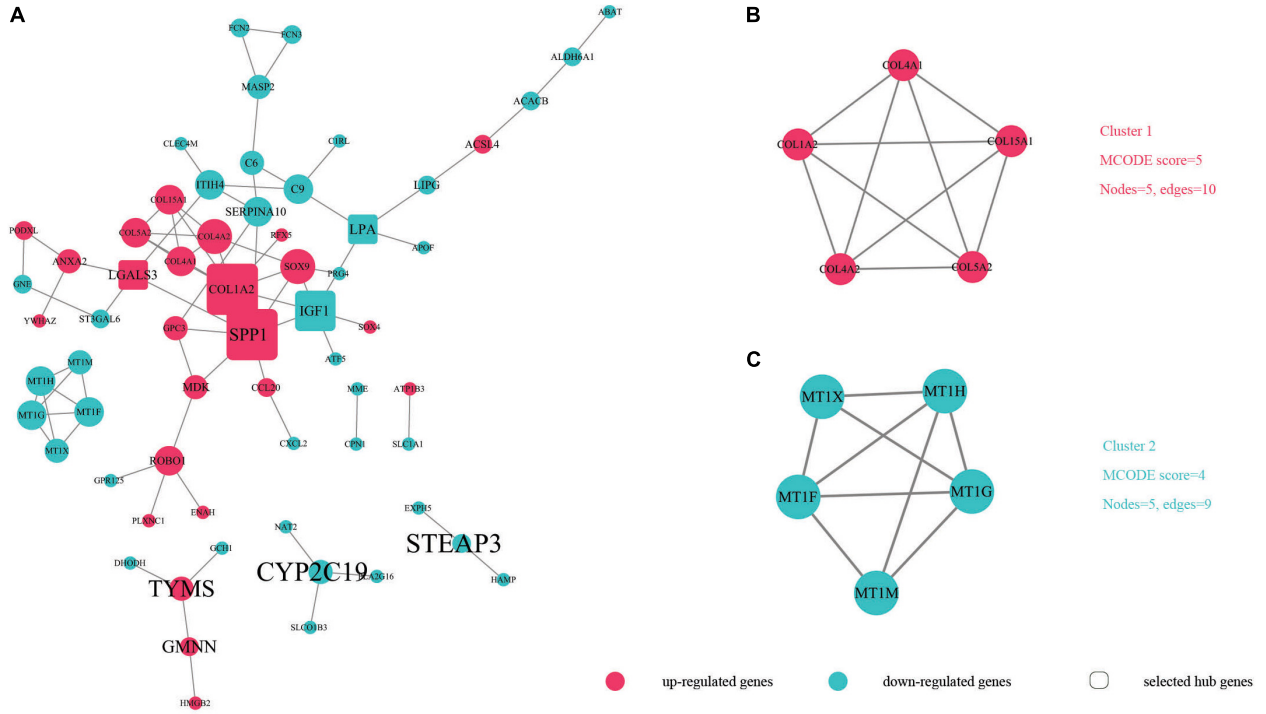


FIGURE 3 | The protein–protein interaction (PPI) of DEGs exported from STRING is visualized using Cytoscape software. Up-regulated DEGs are shown in red while down-regulated DEGs are shown in green dots. **(A)** 63 nodes and 78 edges are displayed. Larger node sizes correspond to higher degrees of DEGs. Label font sizes are shown in black from small to large according to the Betweenness centrality (from low to high). The hub genes we selected are emphasized and shown in the shape of round rectangles. **(B,C)** Module 1, module 2 and their related specifications determined by MCODE plug-in app in Cytoscape software. Circles represent genes and the lines between genes indicate the gene-encoded PPIs.

TABLE 5 | Top 10 most degree values and betweenness centrality hub genes between HCC and normal samples.

Genes	Expression	Betweenness centrality	Genes	Expression	Degree
CYP2C19	Down	1	***SPP1***	Up	8
STEAP3	Down	1	***COL1A2***	Up	8
TYMS	Up	0.83333333	***IGF1***	Down	6
SPP1	Up	0.50873984	*SOX9*	Up	5
GMNN	Up	0.5	*COL4A2*	Up	5
LPA	Down	0.30264228	***LPA***	Down	4
IGF1	Down	0.29329268	***LGALS3***	Up	4
LGALS3	Up	0.24573171	*C9*	Down	4
COL1A2	Up	0.19756098	*SERPINA10*	Down	4
MDK	Up	0.1804878	*ROBO1*	Up	4

Genes were ranked by betweenness centrality and degree, respectively. The genes in red are the shared genes with the two analysis methods.

and counter-stained with hematoxylin. Then images of the sections were photographed using an Olympus microscope (Olympus Life Science). The clinical specimen data of LPA were obtained from The Human Protein Atlas database[9].

Statistical Analysis

Statistical analysis was performed using GraphPad Prism software (version 8.0.1) and R software (version 3.4.2[10]). p-value < 0.05 was considered statistically significant. The column diagram was graphed with GraphPad Prism software (version 8.0.1).

RESULTS

Data Source and Analysis

A total of 603, 1,238, 1,095, and 1,722 DEGs have been extracted from the four independent expression datasets Mas liver, Roessler liver, Roessler liver 2, and Wurmbach liver,

[9]https://www.proteinatlas.org/
[10]https://www.r-project.org/

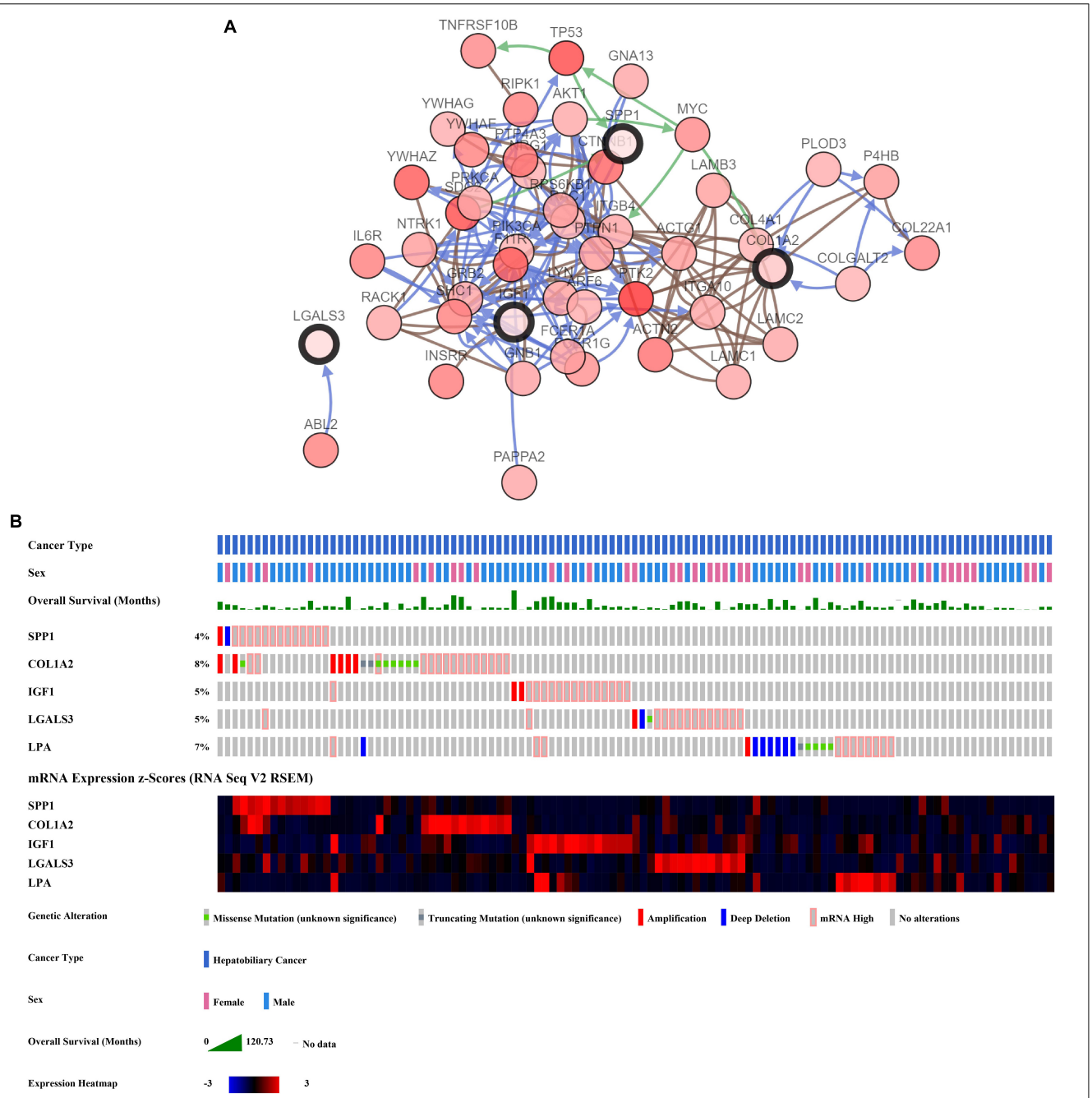


FIGURE 4 | Analysis of hub genes. **(A)** Hub genes and their co-expressed genes are analyzed using cBioPortal. Hub genes are represented with a thick border. Darker red indicates increased frequency of alteration in HCC. The blue connection indicates that the first protein controlled a reaction that changes the state of the second protein; the red connection suggests that the proteins belongs to members of the same complex. The green arrows represent "Controls expression of." The gray arrows represent "In complex with." **(B)** Genetic alteration analysis toward these five genes, overall survivals are also shown.

respectively (**Figures 1A–D**, **Table 1**, and **Supplementary Tables S1, S2**). The comparison of all genes from these datasets identified 89 consistently and significantly dysregulated genes, including 31 up-regulated genes and 58 down-regulated genes in HCC compared to normal liver tissues (**Figures 1E,F** and **Supplementary Table S3**). Notably, several genes such as Glypican-3 (*GPC3*) (Wurmbach et al., 2007) and *SPP1* (Roessler et al., 2010) have been reported previously, proving the feasibility of the method.

GO Analysis of DEGs in HCC

The functional characteristics of these 89 DEGs were explored using GO analysis and were grouped into BP, cell component and MF (**Figure 2A**). Overall, cellular response to zinc ion covering

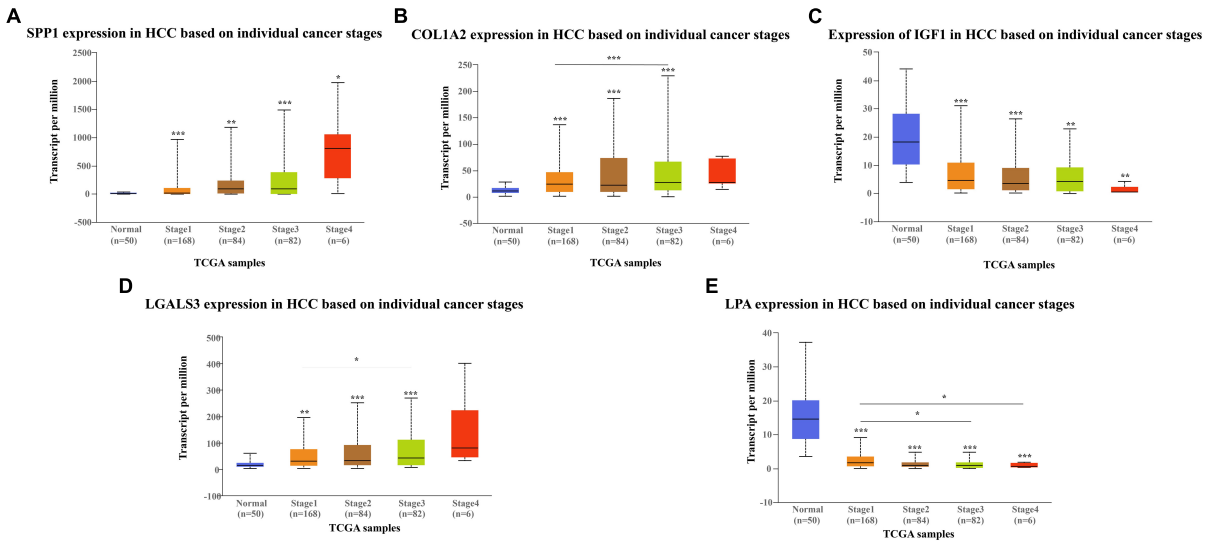


FIGURE 5 | Expression of hub genes in different HCC stages in TGCA database: **(A)** *SPP1*, **(B)** *COL1A2*, **(C)** *IGF1*, **(D)** *LGALS3*, and **(E)** *LPA*. *SPP1*, *COL1A2*, and *LGAIS3* are overexpressed in HCC tissues while *IGF1* and *LPA* are downregulated in HCC tissues compared to the control.

five genes was found to be the dominant BP. Collagen trimer covering seven genes was found to be the top CC. ECM structural constituent conferring tensile strength covering five genes was the top MF. As shown in **Table 2**, in the BP group, up-regulated genes were mainly enriched in ECM organization, epithelial tube morphogenesis, and positive regulation of leukocyte migration while down-regulated genes were mainly enriched in cellular response to zinc ion and humoral immune response. In the CC group, up-regulated genes were mainly enriched in collagen-containing ECM and ECM components while down-regulated genes mainly enriched in blood microparticle. In the MF group, up-regulated genes were mainly enriched in ECM structural constituents while down-regulated genes were mainly enriched in pattern binding. Taken together, these data suggest that those identified DEGs are mainly enriched in ECM-related items affecting the BP of negative regulation of growth, humoral immune response and so on.

Signaling Pathway Enrichment Analysis

To understand the biological changes during HCC pathogenesis, we performed pathway enrichment analysis using KEGG and Reactome. KEGG pathways enrichment analysis showed that those candidate DEGs were primarily enriched in protein digestion and absorption, mineral absorption, and ECM-receptor interaction (**Figure 2B**). Among them, up-regulated genes were mainly enriched in protein digestion and absorption and ECM-receptor interaction while down-regulated genes were mainly enriched in mineral absorption and metabolic pathways (**Table 3**). Furthermore, Reactome pathway enrichment analysis showed that the DEGs were enriched in collagen chain trimerization, collagen degradation, metallothioneins bind metals, and response to metal ions (**Supplementary Table S4**). Among them, up-regulated genes were primarily enriched in collagen chain trimerization, assembly of collagen fibrils and other multimeric structures and collagen degradation, while down-regulated genes were enriched in metallothioneins bind metals, response to metal ions and ficolins bind to repetitive carbohydrate structures on the target cell surface (**Table 4**).

Key Candidate Genes and Pathways Identified by DEGs PPI and Modular Analysis

In order to identify key candidate genes, 63 DEGs (24 up-regulated genes and 39 down-regulated genes) were filtered into the PPI network complex, including 63 nodes and 78 edges. Among the 63 nodes, only genes ranking in top 10 of both degrees (the number of interactions of each node) and betweenness centrality (degree of impact on interactions between other nodes in the network) parameters were recognized as hub genes. Finally, five genes *SPP1*, *COL1A2*, *IGF1*, *LGALS3*, and *LPA* were selected (**Figure 3A** and **Table 5**). Utilizing MCODE plug-in app in cytoscape, two modules were applied for further KEGG pathway enrichment analysis. Module 1 consisted of 5 nodes and 10 edges with genes enriched in protein digestion and absorption, ECM-receptor interaction, amoebiasis and focal adhesion. Module 2 consisted of five nodes and nine edges with the genes mainly enriched in mineral absorption (**Figures 3B,C** and **Supplementary Table S5**).

Hub Genes and Associations With Clinical Outcome

The network of hub genes constructed by cBioPortal contained 55 nodes, including five query genes (five hub genes) and the 50 most frequently altered neighbor genes (**Figure 4A**). After visualizing BP using BiNGO in Cytoscape software (**Supplementary Figure S1**), genetic alteration analysis of five hub genes in TCGA HCC patients was performed in the

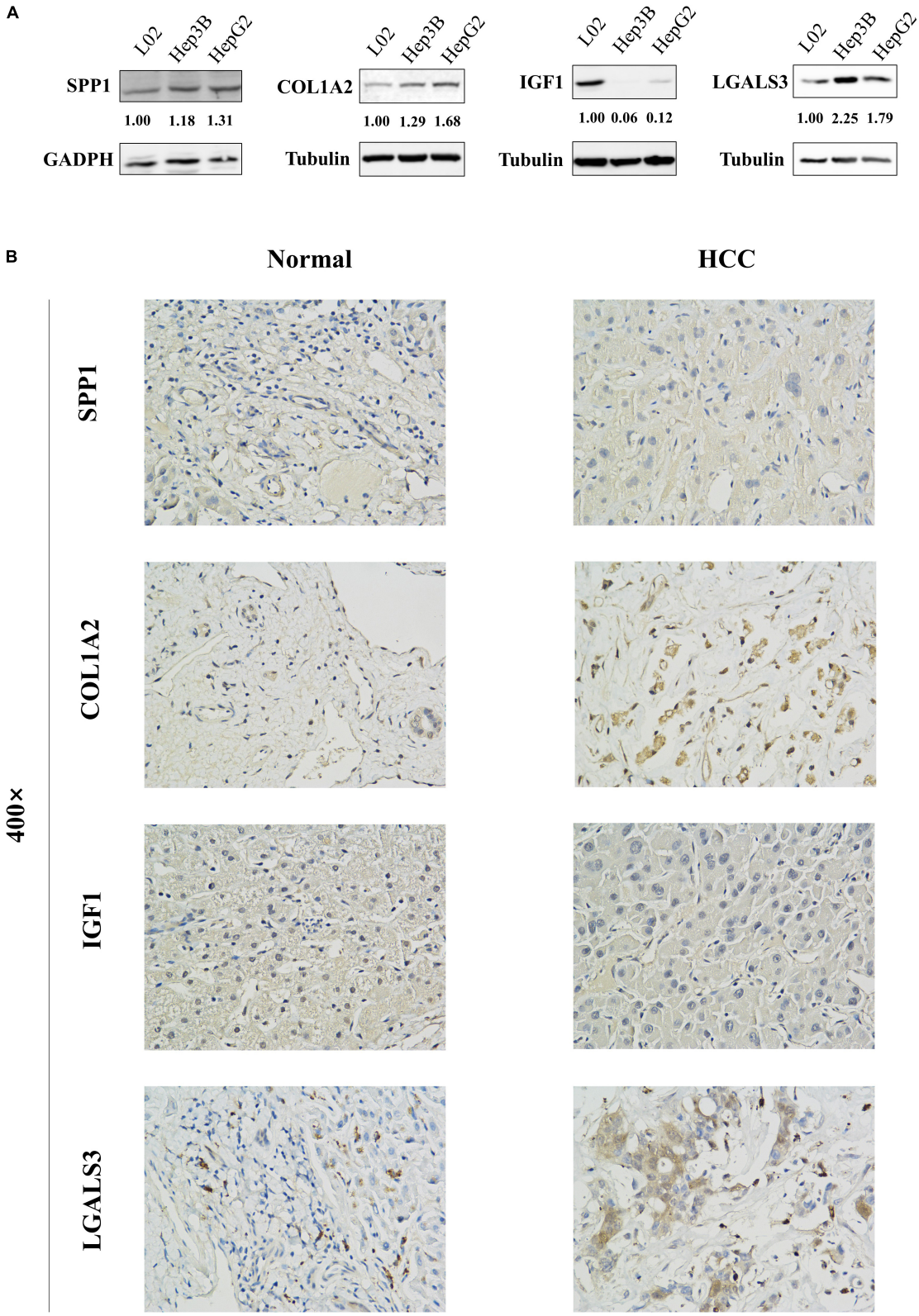


FIGURE 6 | Protein expression levels of these hub genes in normal liver cell line L02 and HCC cell lines Hep3B and HepG2 were examined by western blot analysis. The bands were analyzed and normalized using the GADPH or β-Tubulin values and processed by Image J software. **(A)** The representative cases of HCC patient tissues were detected by **(B)** Immunohistochemistry. The western blot and Immunohistochemistry results showed that SPP1, COL1A2, and LGALS3 are overexpressed in HCC samples while IGF1 is downregulated in HCC samples.

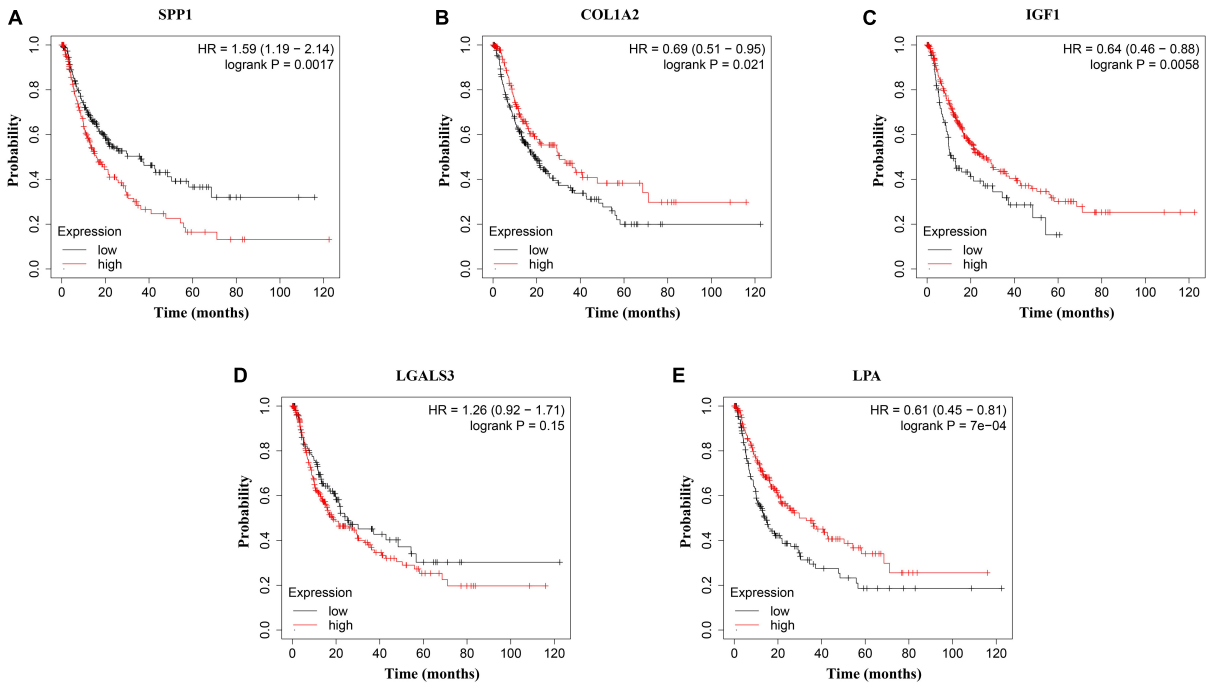


FIGURE 7 | Disease-free survival (DFS) analysis of **(A)** *SPP1*, **(B)** *COL1A2*, **(C)** *IGF1*, **(D)** *LGALS3*, and **(E)** *LPA* in HCC patients. HCC patients with high expressions of *COL1A2*, *IGF1*, and *LPA* as well as low expression of *SPP1* were found to be associated with the improved DFS (p = 0.0017, p = 0.0021, p = 0.0058, and $p = 7e^{-04}$, respectively).

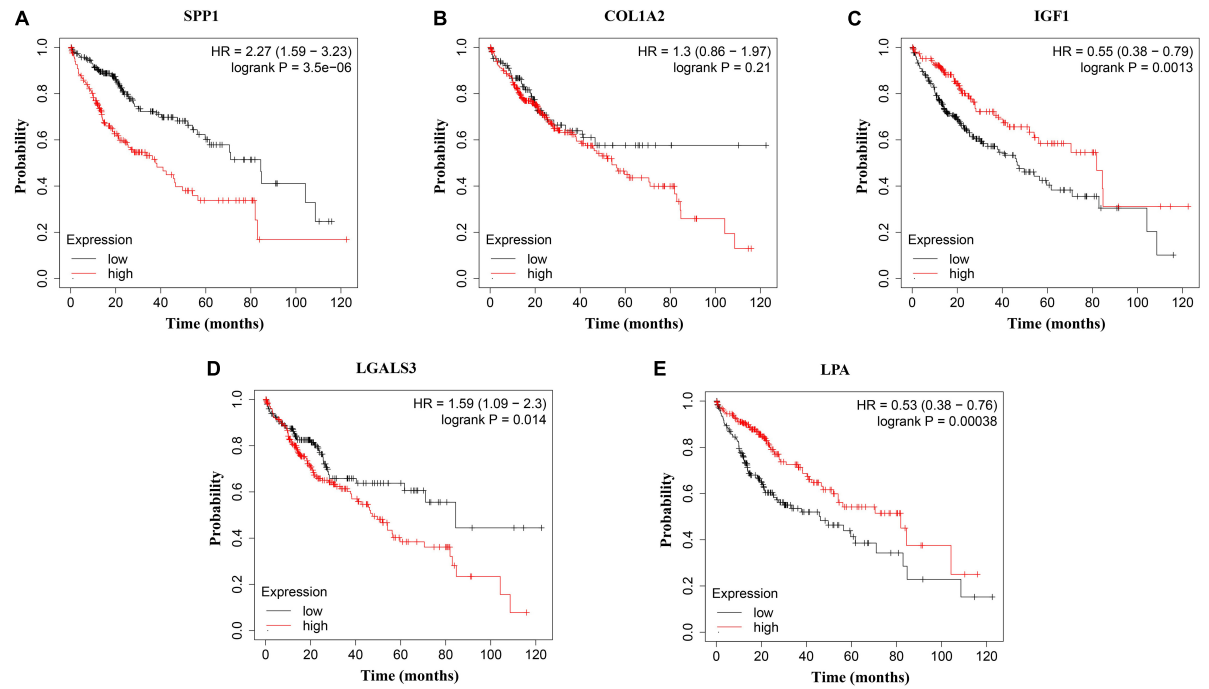


FIGURE 8 | Overall survival (OS) analysis of **(A)** *SPP1*, **(B)** *COL1A2*, **(C)** *IGF1*, **(D)** *LGALS3*, and **(E)** *LPA* in HCC patients. High expression of *SPP1* and *LGALS3* were linked with the disfavored OS ($p = 3.5e^{-06}$ and p = 0.014, respectively) **(A,D)**, while high expression of *IGF1* and *LPA* were associated with improved OS (p = 0.0013 and p = 0.00038, respectively).

cBioPortal database. The hub genes *SPP1*, *IGF1*, *LGALS3*, *LPA*, and *COL1A2* were altered in 4, 5, 5, 7, and 8% in a total population of HCC patients respectively, without significantly discrepancy in both sexes (**Figure 4B**). TCGA data analysis showed that *SPP1*, *COL1A2*, and LGALS3 are more highly expressed in HCC regardless of stages compared with normal tissues, while *IGF1* and *LPA* were low expressed (**Figure 5**). Further western blot analysis showed that the protein expression levels of *SPP1*, *COL1A2*, and *LGALS3* were highly expressed in HCC cell lines while IGF1 was down-regulated in HCC cell lines (**Figure 6A**). IHC analysis of HCC patient tissues showed similar results as western blot analysis (**Figure 6B**). The online human protein atlas showed the LPA protein expression was higher in normal liver tissues than in HCC tissues. For the identified five top hub genes, HCC patients with high expressions of *COL1A2*, *IGF1*, and *LPA* as well as low expression of *SPP1* were found to be associated with the improved DFS ($p = 0.0017$, $p = 0.0021$, $p = 0.0058$, and $p = 7e^{-04}$, respectively) (**Figure 7**). High expression of *SPP1* and *LGALS3* were linked with the disfavored OS ($p = 3.5e^{-06}$ and $p = 0.014$, respectively) (**Figures 8A,D**), while high expression of *IGF1* and *LPA* were associated with improved OS ($p = 0.0013$ and $p = 0.00038$, respectively) (**Figures 8C,E**). However, expression of *COL1A2* didn't show a significant correlation with clinical outcome ($p = 0.21$) (**Figure 8B**).

DISCUSSION

Hepatocellular carcinoma remains an aggressive form of cancer worldwide with high incidence and morbidity. Therefore, substantial efforts have been made to unveil mutational processes, pathogenesis and possible mechanisms underlying treatment resistance in order to expand the therapeutic landscape of this disease (Llovet et al., 2018; Cheng et al., 2019). However, most of these studies were based on single institutions with limited sample size, restricting the power to identify potential meaningful therapeutic targets (Zhang C. et al., 2017; Li et al., 2019; Zhang et al., 2019). Different HCC studies showed different results for different datasets chosen. In previous studies, some only chose one dataset and others chose datasets without performing explicit infiltration, leading to totally different outcome (Zhang C. et al., 2017; Li et al., 2019; Zhang et al., 2019). Here, we conducted an integrative analysis from four microarray datasets of HCC screened in Oncomine database and downloaded in GEO database to describe key candidate genes and pathways associated with clinical outcome in HCC patients.

In the present study, a total of 89 DEGs were identified between HCC and normal tissues, including 31 up-regulated genes and 58 down-regulated genes. Up-regulated DEGs of HCC were found to be enriched in GO categories such as epithelial tube morphogenesis, ECM organization, and positive regulation of leukocyte migration, and dysregulation of these processes have been found to contribute to several pathological conditions including cancer and may lead to disfavored clinical outcomes (Payne and Huang, 2013; Bonnans et al., 2014). While down-regulated genes were associated with GO categories such as cellular response to zinc ion where members of metallothionein family (*MT1M*, *MT1H*, *MT1X*, *MT1G*, and *MT1F*) play important roles in carcinogenesis of various cancer types (Si and Lang, 2018). KEGG pathway enrichment analysis demonstrated that up-regulated genes were significantly enriched in protein digestion and absorption, ECM-receptor interaction and PI3K/Akt signaling pathway while down-regulated genes were enriched in mineral absorption and metabolic pathways, and all those are significant pathways in various cancer types been reported previously (Boroughs and DeBerardinis, 2015; Dimitrova and Arcaro, 2015; Wang S.S. et al., 2017; Slattery et al., 2018). Intriguingly, a host of altered genes were found to be associated with ECM related pathways. The ECM, an extensive part of the microenvironment in all tissues, providing a physical scaffold for its surrounding cells, bind growth factors and regulate cell behavior, plays a vital part in tumor progression (Kalluri, 2016; Nissen et al., 2019).

We also constructed PPI network and identified five hub genes *SPP1*, *COL1A2*, *IGF1*, *LPA*, and *LGALS3* as key candidate genes potentially linked with pathogenesis of HCC. Their co-expressed genes were then analyzed using cBioPortal online platform. The results contained five query genes and the 50 most frequently altered neighbor genes. Among the five hub genes, *SPP1*, *COL1A2*, *IGF1*, and *LGALS3* and their co-expressed genes constructed a network. *LPA* and its co-expressed genes were isolated from the main network and didn't have directly interaction with it. That's why only four out of five hub genes contained in **Figure 4A**. Both SPP1 and COL1A2 are members belonging to PI3K/Akt signaling pathway and ECM-receptor interaction pathway regulating cell growth (Fang et al., 2017) and drug resistance (Zhang et al., 2016). SPP1, also known as osteopontin, has been reported to have the capability of regulating cell behaviors (Rowe et al., 2014). Previous data also showed that targeting SPP1 could inhibit gastric cancer cell epithelial–mesenchymal transition through inhibition of the PI3K/AKT signaling pathway (Song et al., 2019). In lung adenocarcinoma, SPP1 was found to up-regulate PD-L1 and subsequently facilitated the escape of immunity (Zhang Y. et al., 2017). These studies demonstrated that SPP1 was highly associated with the cancer invasion and progression, suggesting its potential to serve as a biomarker and target for the diagnosis and treatment of HCC. COL1A2, a member of group I collagen family, has once been reported as a target of Let-7g thus inhibiting cell migration in HCC (Ji et al., 2010) and gastric cancer cell proliferation (Ao et al., 2018). In 2018, a study found that the silencing of *COL1A2* could inhibit the proliferation, migration, and invasion of gastric cancer through regulating PI3K/AKT signaling pathway, revealing the potency of COL1A2 in HCC (Ao et al., 2018). IGF1, insulin-like growth factor 1, has the capability of maintaining the stemness in HCC, and its role of serving as an anticancer target has been confirmed by several studies (Kaseb et al., 2011, 2014; Chen and Sharon, 2013; Bu et al., 2014). IGF1 and IGF2 comprise of the IGF family, contributing largely to the activation of the PI3K/Akt signaling pathway, which was also found dysregulated by the KEGG analysis, thus enhancing the cancerogenesis of HCC (Kasprzak et al., 2017). LPA is lipoprotein A, a special kind of low-density lipoprotein, has shown the evidence of causing inflammation and regulating HCC cell proliferation (Pirro et al., 2017; Xu et al., 2017). Patients with HCC showed a statistically significant serum LPA level higher

than the healthy subjects, indicating its important role in HCC patients (Malaguarnera et al., 2017). *LGALS3*, which encodes the Galectin-3 protein, is regarded as a guardian of the tumor microenvironment (Ruvolo, 2016). Recent studies have shown that *LGALS3* is tightly associated with several malignancies such as Hodgkin's lymphoma (Koh et al., 2014), acute myeloid leukemia (Cheng et al., 2013), and HCC (Song et al., 2014). More importantly, *LGALS3* could increase the metastatic potential of breast cancer, might accounting for the metastatic potential of HCC (Pereira et al., 2019). Through validation in western blot and IHC assays, we found that the protein expression of these five hub genes was in accordance with their mRNA expression in HCC patient tissues. Strikingly, LPA has not been tested by western blot and IHC assays due to its large molecular weight of 501 kDa, exerting huge difficulty in performing these assays. Previous studies have shown that these genes are implicated in the tumorigenesis and transformation (Oates et al., 1997; Orsó and Schmitz, 2017; Wang Y.A. et al., 2017; Diao et al., 2018; Ma et al., 2019). In our study, correlations of *SPP1*, *COL1A2*, *IGF1*, *LPA*, and *LGALS3* with patient prognosis highlight the importance of these five genes as potential biomarkers to stratify HCC patients as well as potential therapeutic targets, but concrete roles of these genes need further investigation. In the future studies, we will develop knockdown and overexpression HCC cell lines and mouse models of these five hub genes to demonstrate their importance in the progression of HCC *in vitro* and *in vivo*.

Taken together, this study integrated four datasets to screen for reliable and accurate biomarkers of HCC and demonstrated that several pathways are altered. Several hub genes with the expression levels have significantly associated with clinical outcome in HCC patients. Further functional study on the mechanisms of those genes leading to HCC is under way.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.oncomine.org.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Cancer Hospital of the University of Chinese Academy of Sciences; Zhejiang Cancer Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ST, WZ, and J-JQ: conceptualization and supervision. YL, JY, SM, and X-DC: investigations. YL, JY, KQ, CK, and RC: methodology. YL and JY: data curation. YL and RC: original draft writing. YL, RC, and J-JQ: writing review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by Prof. Chang Jiang Scholars Program, National Natural Science Foundation of China (81520108030, 21472238, and 81903842), Shanghai Engineering Research Center for the Preparation of Bioactive Natural Products (16DZ2280200), Scientific Foundation of Shanghai China (13401900103 and 13401900101), National Key Research and Development Program of China (2019YFC1711000 and 2017YFC1700200), the Shanghai Sailing Program (No. 20YF1459000), Program of Zhejiang Provincial TCM Sci-tech Plan (2020ZZ005), and Zhejiang Chinese Medical University Startup Funding (111100E014).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00814/full#supplementary-material

FIGURE S1 | The biological process analysis of hub genes using BINGO.

TABLE S1 | The DEGs identified in each dataset.

TABLE S2 | Information of DEGs screened from each dataset.

TABLE S3 | 31 consistently up-regulated and 58 down-regulated DEGs.

TABLE S4 | Reactome pathway enrichment analysis.

TABLE S5 | KEGG pathway enrichment analysis of modules.

REFERENCES

Ally, A., Balasundaram, M., Carlsen, R., Chuah, E., Clarke, A., Dhalla, N., et al. (2017). Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* 169, 1327–1341.e1323.

Ao, R., Guan, L., Wang, Y., and Wang, J.-N. (2018). Silencing of COL1A2, COL6A3, and THBS2 inhibits gastric cancer cell proliferation, migration, and invasion while promoting apoptosis through the PI3k-Akt signaling pathway. *J. Cell. Biochem.* 119, 4420–4434. doi: 10.1002/jcb.26524

Aravalli, R. N., Cressman, E. N., and Steer, C. J. (2013). Cellular and molecular mechanisms of hepatocellular carcinoma: an update. *Arch. Toxicol.* 87, 227–247. doi: 10.1007/s00204-012-0931-2

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Bonnans, C., Chou, J., and Werb, Z. (2014). Remodelling the extracellular matrix in development and disease. *Nat. Rev. Mol. Cell Biol.* 15, 786–801. doi: 10.1038/nrm3904

Boroughs, L. K., and DeBerardinis, R. J. (2015). Metabolic pathways promoting cancer cell survival and growth. *Nat. Cell Biol.* 17, 351–359. doi: 10.1038/ncb3124

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Bu, Y., Jia, Q. A., Ren, Z. G., Zhang, J. B., Jiang, X. M., Liang, L., et al. (2014). Maintenance of stemness in oxaliplatin-resistant hepatocellular carcinoma is associated with increased autocrine of IGF1. *PLoS One* 9:e89686. doi: 10.1371/journal.pone.0089686

Chandrashekar, D. S., Bashel, B., Balasubramanya, S. A. H., Creighton, C. J., Ponce-Rodriguez, I., Chakravarthi, B., et al. (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 19, 649–658. doi: 10.1016/j.neo.2017.05.002

Chen, H. X., and Sharon, E. (2013). IGF-1R as an anti-cancer target–trials and tribulations. *Chin. J. Cancer* 32, 242–252. doi: 10.5732/cjc.012.10263

Cheng, C. L., Hou, H. A., Lee, M. C., Liu, C. Y., Jhuang, J. Y., Lai, Y. J., et al. (2013). Higher bone marrow LGALS3 expression is an independent unfavorable prognostic factor for overall survival in patients with acute myeloid leukemia. *Blood* 121, 3172–3180. doi: 10.1182/blood-2012-07-443762

Cheng, H., Sun, G., Chen, H., Li, Y., Han, Z., Li, Y., et al. (2019). Trends in the treatment of advanced hepatocellular carcinoma: immune checkpoint blockade immunotherapy and related combination therapies. *Am. J. Cancer Res.* 9, 1536–1545.

Crawley, J. J., and Furge, K. A. (2002). Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. *Genome Biol.* 3:research0075.0071.

Diao, B., Liu, Y., Xu, G.-Z., Zhang, Y., Xie, J., and Gong, J. (2018). The role of galectin-3 in the tumorigenesis and progression of pituitary tumors. *Oncol. Lett.* 15, 4919–4925. doi: 10.3892/ol.2018.7931

Dimitrova, V., and Arcaro, A. (2015). Targeting the PI3K/AKT/mTOR signaling pathway in medulloblastoma. *Curr. Mol. Med.* 15, 82–93. doi: 10.2174/1566524015666150114115427

Duan, Y. (2019a). *Existence Computation: Relationship Defined Everything Underlying Semantic Computation*. Toyama, 139–144.

Duan, Y. (2019b). "Towards a periodic table of conceptualization and formalization on concepts of state, style, structure, pattern, framework, architecture, service, etc., based on existence computation and relationship defined everything of semantic," in *Proceedings of the 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Toyama.

Duan, Y., Lu, Z., Zhou, Z., Sun, X., and Wu, J. (2019a). Data privacy protection for edge computing of smart city in a DIKW architecture. *Eng. Appl. Artif. Intell.* 81, 323–335. doi: 10.1016/j.engappai.2019.03.002

Duan, Y., Sun, X., Che, H., Cao, C., Li, Z., and Yang, X. (2019b). Modeling data, information and knowledge for security protection of hybrid IoT and edge resources. *IEEE Access* 7, 99161–99176. doi: 10.1109/access.2019.2931365

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46, D649–D655. doi: 10.1093/nar/gkx1132

Fang, X., Yang, D., Luo, H., Wu, S., Dong, W., Xiao, J., et al. (2017). SNORD126 promotes HCC and CRC cell growth by activating the PI3K-AKT pathway through FGFR2. *J. Mol. Cell Biol.* 9, 243–255. doi: 10.1093/jmcb/mjw048

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6:l1. doi: 10.1126/scisignal.2004088

Hoshida, Y., Nijman, S. M., Kobayashi, M., Chan, J. A., Brunet, J.-P., Chiang, D. Y., et al. (2009). Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res.* 69, 7385–7392. doi: 10.1158/0008-5472.can-09-1089

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Ji, J., Zhao, L., Budhu, A., Forgues, M., Jia, H. L., Qin, L. X., et al. (2010). Let-7g targets collagen type I alpha2 and inhibits cell migration in hepatocellular carcinoma. *J. Hepatol.* 52, 690–697. doi: 10.1016/j.jhep.2009.12.025

Jin, B., Wang, W., Du, G., Huang, G., Han, L., Tang, Z., et al. (2015). Identifying hub genes and dysregulated pathways in hepatocellular carcinoma. *Eur. Rev. Med. Pharmacol. Sci.* 19, 592–601.

Kalluri, R. (2016). The biology and function of fibroblasts in cancer. *Nat. Rev. Cancer* 16, 582–598. doi: 10.1038/nrc.2016.73

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi: 10.1093/nar/gkv1070

Kaseb, A. O., Morris, J. S., Hassan, M. M., Siddiqui, A. M., Lin, E., Xiao, L., et al. (2011). Clinical and prognostic implications of plasma insulin-like growth factor-1 and vascular endothelial growth factor in patients with hepatocellular carcinoma. *J. Clin. Oncol.* 29, 3892–3899. doi: 10.1200/jco.2011.36.0636

Kaseb, A. O., Xiao, L., Hassan, M. M., Chae, Y. K., Lee, J. S., Vauthey, J. N., et al. (2014). Development and validation of insulin-like growth factor-1 score to assess hepatic reserve in hepatocellular carcinoma. *J. Nat. Cancer Inst.* 106:dju088. doi: 10.1093/jnci/dju088

Kasprzak, A., Kwasniewski, W., Adamek, A., and Gozdzicka-Jozefiak, A. (2017). Insulin-like growth factor (IGF) axis in cancerogenesis. *Mutat. Res. Rev. Mutat. Res.* 772, 78–104. doi: 10.1016/j.mrrev.2016.08.007

Koh, Y. W., Jung, S. J., Park, C. S., Yoon, D. H., Suh, C., and Huh, J. (2014). LGALS3 as a prognostic factor for classical Hodgkin's lymphoma. *Mod. Pathol.* 27, 1338–1344. doi: 10.1038/modpathol.2014.38

Li, C., Zhou, D., Jiang, X., Liu, M., Tang, H., and Mei, Z. (2019). Identifying hepatocellular carcinoma-related hub genes by bioinformatics analysis and CYP2C8 is a potential prognostic biomarker. *Gene* 698, 9–18. doi: 10.1016/j.gene.2019.02.062

Llovet, J. M., Montal, R., Sia, D., and Finn, R. S. (2018). Molecular therapies and precision medicine for hepatocellular carcinoma. *Nat. Rev. Clin. Oncol.* 15, 599–616. doi: 10.1038/s41571-018-0073-4

Ma, H.-P., Chang, H.-L., Bamodu, O. A., Yadav, V. K., Huang, T.-Y., Wu, A. T. H., et al. (2019). Collagen 1A1 (COL1A1) is a reliable biomarker and putative therapeutic target for hepatocellular carcinogenesis and metastasis. *Cancers* 11:786. doi: 10.3390/cancers11060786

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448–3449. doi: 10.1093/bioinformatics/bti551

Malaguarnera, G., Catania, V. E., Francaviglia, A., Malaguarnera, M., Drago, F., Motta, M., et al. (2017). Lipoprotein(a) in patients with hepatocellular carcinoma and portal vein thrombosis. *Aging Clin. Exp. Res.* 29(Suppl. 1), 185–190. doi: 10.1007/s40520-016-0653-z

Marengo, A., Rosso, C., and Bugianesi, E. (2016). Liver cancer: connections with obesity, fatty liver, and cirrhosis. *Annu. Rev. Med.* 67, 103–117. doi: 10.1146/annurev-med-090514-013832

Mas, V. R., Maluf, D. G., Archer, K. J., Yanek, K., Kong, X., Kulik, L., et al. (2009). Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma. *Mol. Med.* 15, 85–94. doi: 10.2119/molmed.2008.00110

Nissen, N. I., Karsdal, M., and Willumsen, N. (2019). Collagens and Cancer associated fibroblasts in the reactive stroma and its relation to Cancer biology. *J. Exp. Clin. Cancer Res.* 38:115. doi: 10.1186/s13046-019-1110-6

Oates, A. J., Barraclough, R., and Rudland, P. S. (1997). The role of osteopontin in tumorigenesis and metastasis. *Invasion Metastasis* 17, 1–15.

Orsó, E., and Schmitz, G. (2017). Lipoprotein(a) and its role in inflammation, atherosclerosis and malignancies. *Clin. Res. Cardiol. Suppl.* 12(Suppl. 1), 31–37. doi: 10.1007/s11789-017-0084-1

Oweira, H., Petrausch, U., Helbling, D., Schmidt, J., Mehrabi, A., Schob, O., et al. (2017). Prognostic value of site-specific extra-hepatic disease in hepatocellular carcinoma: a SEER database analysis. *Exp. Rev. Gastroenterol. Hepatol.* 11, 695–701. doi: 10.1080/17474124.2017.1294485

Payne, L. S., and Huang, P. H. (2013). The pathobiology of collagens in glioma. *Mol. Cancer Res.* 11, 1129–1140. doi: 10.1158/1541-7786.MCR-13-0236

Pereira, J. X., Dos Santos, S. N., Pereira, T. C., Cabanel, M., Chammas, R., de Oliveira, F. L., et al. (2019). Galectin-3 regulates the expression of tumor glycosaminoglycans and increases the metastatic potential of breast cancer. *J. Oncol.* 2019:9827147. doi: 10.1155/2019/9827147

Phukan, R. K., Borkakoty, B. J., Phukan, S. K., Bhandari, K., Mahanta, J., Tawsik, S., et al. (2018). Association of processed food, synergistic effect of alcohol and

HBV with Hepatocellular Carcinoma in a high incidence region of India. *Cancer Epidemiol.* 53, 35–41. doi: 10.1016/j.canep.2018.01.005

Pirro, M., Bianconi, V., Paciullo, F., Mannarino, M. R., Bagaglia, F., and Sahebkar, A. (2017). Lipoprotein(a) and inflammation: a dangerous duet leading to endothelial loss of integrity. *Pharmacol. Res.* 119, 178–187. doi: 10.1016/j.phrs.2017.02.001

Roessler, S., Jia, H. L., Budhu, A., Forgues, M., Ye, Q. H., Lee, J. S., et al. (2010). A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res.* 70, 10202–10212. doi: 10.1158/0008-5472.Can-10-2607

Rowe, G. C., Raghuram, S., Jang, C., Nagy, J. A., Patten, I. S., Goyal, A., et al. (2014). PGC-1alpha induces SPP1 to activate macrophages and orchestrate functional angiogenesis in skeletal muscle. *Circ. Res.* 115, 504–517. doi: 10.1161/circresaha.115.303829

Ruvolo, P. P. (2016). Galectin 3 as a guardian of the tumor microenvironment. *Biochim. Biophys. Acta* 1863, 427–437. doi: 10.1016/j.bbamcr.2015.08.008

Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., et al. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* 47:505. doi: 10.1038/ng.3252

Si, M., and Lang, J. (2018). The roles of metallothioneins in carcinogenesis. *J. Hematol. Oncol.* 11:107. doi: 10.1186/s13045-018-0645-x

Slattery, M. L., Mullany, L. E., Sakoda, L. C., Wolff, R. K., Stevens, J. R., Samowitz, W. S., et al. (2018). The PI3K/AKT signaling pathway: associations of miRNAs with dysregulated gene expression in colorectal cancer. *Mol. Carcinog.* 57, 243–261. doi: 10.1002/mc.22752

Song, L., Mao, J., Zhang, J., Ibrahim, M. M., Li, L. H., and Tang, J. W. (2014). Annexin A7 and its binding protein galectin-3 influence mouse hepatocellular carcinoma cell line in vitro. *Biomed. Pharmacother.* 68, 377–384. doi: 10.1016/j.biopha.2013.10.011

Song, M., Duan, Y., Huang, T., and Zhan, L. (2020). Inter-Edge and cloud conversion accelerated user-generated content for virtual brand community. *EURASIP J. Wirel. Commun. Netw.* 2020:14. doi: 10.1186/s13638-019-1635-6

Song, S. Z., Lin, S., Liu, J. N., Zhang, M. B., Du, Y. T., Zhang, D. D., et al. (2019). Targeting of SPP1 by microRNA-340 inhibits gastric cancer cell epithelial-mesenchymal transition through inhibition of the PI3K/AKT signaling pathway. *J. Cell Physiol.* 234, 18587–18601. doi: 10.1002/jcp.28497

Song, Z., Duan, Y., Wan, S., Sun, X., Zou, Q., Gao, H., et al. (2018). Processing optimization of typed resources with synchronized storage and computation adaptation in Fog computing. *Wirel. Commun. Mob. Comput.* 2018, 1–13. doi: 10.1155/2018/3794175

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Totoki, Y., Tatsuno, K., Covington, K. R., Ueda, H., Creighton, C. J., Kato, M., et al. (2014). Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat. Genet.* 46:1267.

Trevino, V., Falciani, F., and Barrera-Saldaña, H. A. (2007). DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol. Med.* 13, 527–541. doi: 10.2119/2006-00107.Trevino

Villanueva, A. (2019). Hepatocellular carcinoma. *N. Engl. J. Med.* 380, 1450–1462. doi: 10.1056/NEJMra1713263

Villanueva, A., Alsinet, C., Yanger, K., Hoshida, Y., Zong, Y., Toffanin, S., et al. (2012). Notch signaling is activated in human hepatocellular carcinoma and induces tumor formation in mice. *Gastroenterology* 143, 1660–1669.e1667.

Visvader, J. E. (2011). Cells of origin in cancer. *Nature* 469, 314–322. doi: 10.1038/nature09781

Waghray, A., Murali, A. R., and Menon, K. N. (2015). Hepatocellular carcinoma: from diagnosis to treatment. *World J. Hepatol.* 7, 1020–1029. doi: 10.4254/wjh.v7.i8.1020

Wang, S. S., Chen, Y. H., Chen, N., Wang, L. J., Chen, D. X., Weng, H. L., et al. (2017). Hydrogen sulfide promotes autophagy of hepatocellular carcinoma cells through the PI3K/Akt/mTOR signaling pathway. *Cell Death Dis.* 8:e2688. doi: 10.1038/cddis.2017.18

Wang, Y. A., Sun, Y., Palmer, J., Solomides, C., Huang, L.-C., Shyr, Y., et al. (2017). IGFBP3 modulates lung tumorigenesis and cell growth through IGF1 signaling. *Mol. Cancer Res.* 15, 896–904. doi: 10.1158/1541-7786.MCR-16-0390

Woo, H. G., Park, E. S., Cheon, J. H., Kim, J. H., Lee, J.-S., Park, B. J., et al. (2008). Gene expression–based recurrence prediction of hepatitis B virus–related human hepatocellular carcinoma. *Clin. Cancer Res.* 14, 2056–2064.

Wurmbach, E., Chen, Y. B., Khitrov, G., Zhang, W., Roayaie, S., Schwartz, M., et al. (2007). Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. *Hepatology* 45, 938–947. doi: 10.1002/hep.21622

Xu, M., Liu, Z., Wang, C., Yao, B., and Zheng, X. (2017). EDG2 enhanced the progression of hepatocellular carcinoma by LPA/PI3K/AKT/ mTOR signaling. *Oncotarget* 8, 66154–66168. doi: 10.18632/oncotarget.19825

Zhang, C., Peng, L., Zhang, Y., Liu, Z., Li, W., Chen, S., et al. (2017). The identification of key genes and pathways in hepatocellular carcinoma by bioinformatics analysis of high-throughput data. *Med. Oncol.* 34:101. doi: 10.1007/s12032-017-0963-9

Zhang, P. F., Li, K. S., Shen, Y. H., Gao, P. T., Dong, Z. R., Cai, J. B., et al. (2016). Galectin-1 induces hepatocellular carcinoma EMT and sorafenib resistance by activating FAK/PI3K/AKT signaling. *Cell Death Dis.* 7:e2201. doi: 10.1038/cddis.2015.324

Zhang, Q., Sun, S., Zhu, C., Zheng, Y., Cai, Q., Liang, X., et al. (2019). Prediction and analysis of weighted genes in hepatocellular carcinoma using bioinformatics analysis. *Mol. Med. Rep.* 19, 2479–2488. doi: 10.3892/mmr.2019.9929

Zhang, Y., Du, W., Chen, Z., and Xiang, C. (2017). Upregulation of PD-L1 by SPP1 mediates macrophage polarization and facilitates immune escape in lung adenocarcinoma. *Exp. Cell Res.* 359, 449–457. doi: 10.1016/j.yexcr.2017.08.028

Zucman-Rossi, J., Villanueva, A., Nault, J. C., and Llovet, J. M. (2015). Genetic landscape and biomarkers of hepatocellular carcinoma. *Gastroenterology* 149, 1226–1239.e1224. doi: 10.1053/j.gastro.2015.05

Conflict of Interest: KQ was employed by the company Accenture Applied Intelligence, ASEAN.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 30 July 2020
doi: 10.3389/fgene.2020.00760

EnACP: An Ensemble Learning Model for Identification of Anticancer Peptides

Ruiquan Ge[1], Guanwen Feng[2], Xiaoyang Jing[3], Renfeng Zhang[4], Pu Wang[5] and Qing Wu[1]**

[1] Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, [2] Xi'an Key Laboratory of Big Data and Intelligent Vision, School of Computer Science and Technology, Xidian University, Xi'an, China, [3] Toyota Technological Institute at Chicago, Chicago, IL, United States, [4] Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China, [5] Computer School, Hubei University of Arts and Science, Xiangyang, China

As cancer remains one of the main threats of human life, developing efficient cancer treatments is urgent. Anticancer peptides, which could overcome the significant side effects and poor results of traditional cancer treatments, have become a new potential alternative these years. However, identifying anticancer peptides by experimental methods is time consuming and resource consuming, it is of great significance to develop effective computational tools to quickly and accurately identify potential anticancer peptides from amino acid sequences. For most current computational methods, feature representation plays a key role in their final successes. This study proposes a novel fast and accurate approach to identify anticancer peptides using diversified feature representations and ensemble learning method. For the feature representations, the information is encoded from multidimensional feature spaces, including sequence composition, sequence-order, physicochemical properties, etc. In order to better model the potential relationships of peptides, multiple ensemble classifiers, LightGBMs, are applied to detect the different feature sets at first. Then the obtained multiple outputs are used as inputs of the support vector machine classifier, which effectively identifies anticancer peptides. Experimental results on cross validation and independent test sets demonstrate that our method can achieve better or comparable performances compared with other state-of-the-art methods.

Keywords: anticancer peptides, feature representation, ensemble learning, pseudo amino acid composition, system biology

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Sotiris Kotsiantis,
University of Patras, Greece
Tao Huang,
Shanghai Institute for Biological Sciences (CAS), China

***Correspondence:**
Pu Wang
nywangpu@yeah.net
Qing Wu
wuqing@hdu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *20 April 2020*
Accepted: *26 June 2020*
Published: *30 July 2020*

Citation:
Ge R, Feng G, Jing X, Zhang R, Wang P and Wu Q (2020) EnACP: An Ensemble Learning Model for Identification of Anticancer Peptides. Front. Genet. 11:760. doi: 10.3389/fgene.2020.00760

INTRODUCTION

Cancer has become a common disease in humans, and it often leads to a higher mortality rate, especially in developing and developed countries (Ortega-Garcia et al., 2020). The complexity and heterogeneity of cancer are major obstacles for anticancer therapy development (Kasak and Laan, 2020; Umbreit et al., 2020). Traditional cancer treatments, such as radiation therapy, targeted therapy and chemotherapy, often fail to distinguish cancer cells from normal cells. Traditional surgery could not guarantee the precise removal of the diseased part, which is seriously harmful to the patient's body (An et al., 2019). At the same time, the risk of recurrence after surgery is

high. In addition, cancer cells have developed resistance to traditional anticancer drugs due to their overuse. Overall, traditional treatment methods have obvious side effects and poor results. In view of these problems, there is an urgent to discover and design novel cancer treatments and anticancer agents to fight against this deadly disease (Esfandiari Mazandaran et al., 2019; Sima et al., 2019; Bahuguna et al., 2020).

In recent years, peptide-based therapy has become a potential method of cancer treatments. This method can target and kill cancer cells while do not impair the normal cells (Harris, 2020). Anticancer peptides (ACPs) with short amino acid sequences can avoid the disadvantages of traditional cancer treatments. They generally have the characteristics of high specificity, high tissue penetration, low production cost, toxic under normal physiological functions, ease of synthesis and modification, etc. And natural ACPs are safer than synthetic drugs (Feng and Wang, 2019). The electrostatic interactions between ACPs and cancer cell membranes are considered to be one of the main factors for the selective killing of cancer cells (Lin et al., 2018; Naguib et al., 2018). They are believed to play a vital role in the selective toxicity of ACPs to cancer. Currently, many approved peptide-based drugs are being evaluated in various stages of clinical trials (Tesauro et al., 2019; Brunetti et al., 2020). As more and more ACPs are identified and verified by experiments, it is found that most ACPs are derived from protein sequences (Tyagi et al., 2013). However, the discovery of novel ACPs from wet-lab experimentation is laborious, time-consuming and expensive. So, it is essential to develop efficient computational methods to rapidly identify potential ACPs from the peptide sequences.

In the past decade, the accurate identification of ACPs from peptide sequences remains an open research topic in the field of bioinformatics and immunoinformatic. Machine learning methods have been widely used to identify ACPs in many researches. It mainly includes two key techniques which are feature representation and classifier. For feature representation, if the features of peptide sequences are well-extracted, it will be easier to precisely predict the ACPs (Jing et al., 2019). At present, some tools in the prediction of ACPs have been developed. The first computational tool is called Anti-CP (Tyagi et al., 2013), which encoded peptides with sequence-based features and binary profiles to predict ACPs based on Support Vector Machine (SVM). In another work, Hajisharifi et al. considered two kinds features from the local correlation and Chou's pseudo acid amino composition (PseAAC) to improve the prediction of ACPs (Hajisharifi et al., 2014). ACPP used an improved feature encoding method via three type of protein relatedness measure, integrating compositional information, centroidal and distributional information of amino acids (Vijayakumar and Lakshmi, 2015). iACP has referred that membrane interactions are related to their conformation or the order of amino acids. And, it can get better results through cross validation and optimizing the g-gap dipeptide components method compared to the previous predictors (Chen W. et al., 2016). Li et al. indicated that the different types of feature combinations can improve the prediction for ACPs (Chen W. et al., 2016). MLACP constructed features using amino acid composition, atomic composition, dipeptide composition, and physicochemical properties and developed SVM and random forest (RF) methods to predict ACPs (Manavalan et al., 2017). SAP employed 400D features with g-gap dipeptide information and feature selection to identify ACPs (Xu et al., 2018). ACPred-FL can orderly extract effective features from sequence-based feature and a group of SVM models (Wei et al., 2018). mACPpred explored seven feature encodings and a two-step feature selection method to exclude irrelevant features (Ge et al., 2016; Boopathi et al., 2019). Then, the obtained features are input into SVM classifier to gain the predicted result. In addition, a special repository named CancerPPD was collected and created with the manually verified ACPs from the published literature, patents and other databases (Tyagi et al., 2015). It provides a wealth of information related to the peptide for research and experimental personnel to use for reference such as its origin, the nature of the peptide, anticancer activity, terminal modification, conformation, etc. The information is helpful to understand the comprehensive properties of ACPs. And it also provides a reference for the design and identification of ACPs (Lin et al., 2015).

In this paper, we propose a novel two-step prediction model EnACP to accurately identify the ACPs. At first, feature representation is composed of four categories: amino acid composition, autocorrelation, pseudo amino acid composition and profile-based features (Chen et al., 2018). Each type includes a few modes. Finally, 19 kinds of feature patterns are generated. For each feature pattern, LightGBM (Light Gradient Boosting Machine) classifier is employed to generate the initial prediction (Ke et al., 2017). The former predicted results as the new features are input to SVM classifier to get the final prediction. Cross validation results showed that the proposed EnACP model performed better than the previous methods. Furthermore, EnACP achieved comparable performances compared with the existing methods on a new independent dataset. EnACP is available at https://github.com/greyspring/EnACP.

MATERIALS AND METHODS

Dataset

In this study, we use two groups of ACP datasets from the existed literatures to evaluate the performance of the proposed method. For them, one dataset is used to test the cross-validation performance compared with the existing models (Hajisharifi et al., 2014). The other with an independent test dataset can better measure the generalization capability of the model (Boopathi et al., 2019).

For the two datasets, one is called ZH dataset including 138 ACPs and 206 non-ACPs for the 5-fold cross-validation test. The other is from mACPpred for the independent test. In mACPpred dataset, the training dataset consists of 266 ACPs and 266 non-ACPs, and the independent dataset consists of 157 ACPs and 157 non-ACPs. The two group datasets have the low redundancy which were processed to prevent homology bias and high similarity in the related literatures. Amino acid frequency distribution of ACP and non-ACP in the two datasets are shown in **Figure 1**. The sequences containing not 20 natural amino acids are eliminated. From **Figure 2**, most of the peptide sequences are between 5 and 50 in length in the

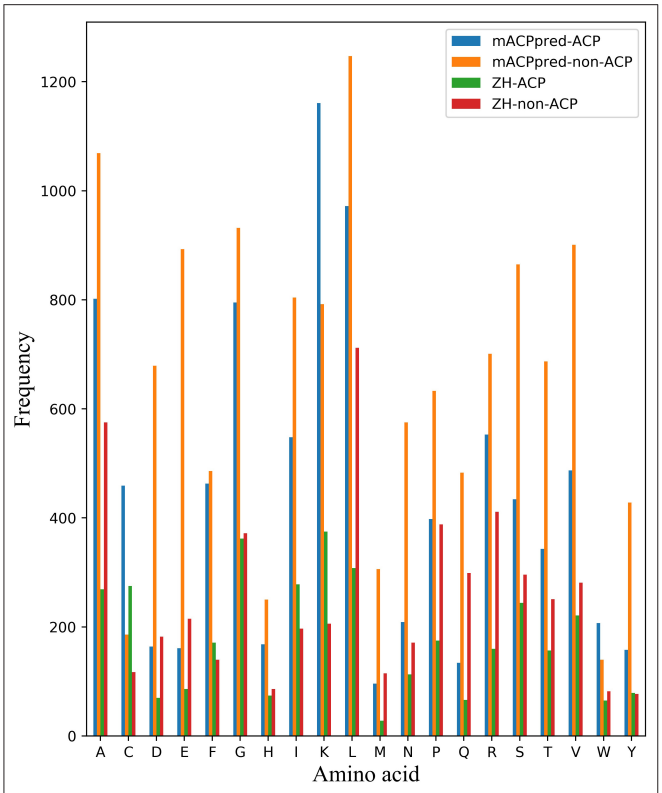


FIGURE 1 | Amino acid frequency distribution on cross-validation and independent datasets. The number of 20 amino acids are counted in mACPpred and ZH datasets. The horizontal axis represents the abbreviation of 20 amino acids. The ordinate represents the number of amino acids.

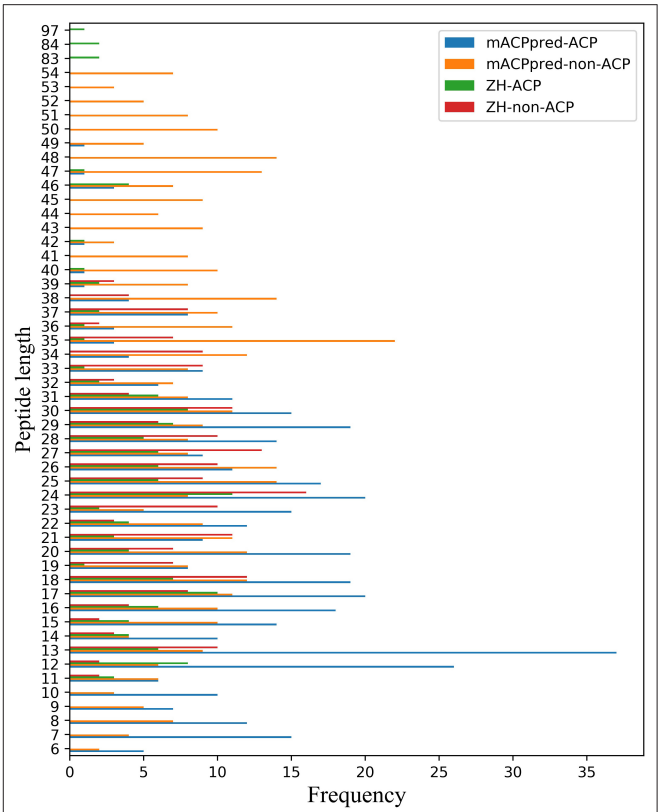


FIGURE 2 | Peptide length distribution of ACP and non-ACP on mACPpred and ZH datasets. The horizontal axis represents the number of statistics. The ordinate represents the length of the peptide sequence.

two datasets especially in mACPpred-ACP and ZH-non-ACP. For mACPpred-non-ACP and ZH-ACP, their ratio is 94.6 and 96.4%, respectively.

Features Representation

There are 19 kinds of features in total used in this study, three of which belong to amino acid composition, four of which belong to autocorrelation features, four of which belong to pseudo amino acid composition, and eight of which belong to profile-based features (Liu et al., 2015, 2017; Liu, 2019).

Amino Acid Composition

Basic kmer (Kmer) (Liu et al., 2008) is a very simple feature extraction method that represents any peptide sequence as a vector consisted of occurrence frequencies of k neighboring amino acids. Distance-based Residue (DR) (Liu et al., 2014b) extracts features from sequence by counting the occurrence frequencies of all possible residue pairs within a certain distance. Just like the DR method, the method of Distance-Pairs and reduced alphabet scheme (Distance Pair) (Liu et al., 2014a) also extracts features from sequence by counting the occurrence frequencies of residue pairs within a certain distance, except that the residue types are reduced by clustering.

Autocorrelation Features

A peptide sequence P is often formulated in the following format, with the N-terminus at the left, and the C-terminus at the right.

$$P = R_1R_2R_3\cdots R_L$$

where R_1 represents the 1st amino acid, R_2 represents the 2nd amino acid, and so forth.

Given a physicochemical index of amino acids, The Auto covariance (AC) (Cao et al., 2013) approach measures the correlation between two residues separated by distance d, which can be calculated as:

$$AC(u,d) = \sum_{i=1}^{L-d} \left(I_u(R_i) - \bar{I}_u\right)\left(I_u(R_{i+d}) - \bar{I}_u\right)/\left(L-d\right)$$

where u indicates the physicochemical index, $I_u(R_i)$ means the index value of R_i, and $\bar{I}_u$ is the average index value along the whole sequence:

$$\bar{I}_u = \sum_{i=1}^{L} I_u(R_i)/L$$

The Cross covariance (CC) (Cao et al., 2013) approach measures the correlation between two residues separated by distance d

based on two different physicochemical indices, which can be calculated by:

$$\mathrm{CC}(u,v,d)=\sum_{i=1}^{L-d}\left(I_u(R_i)-\bar{I}_u\right)\left(I_v(R_{i+d})-\bar{I}_v\right)/\left(L-d\right)$$

where u and v indicate two different indices, $I_u(R_i)(I_v(R_i))$ means the index value of R_i, and $\bar{I}_u(\bar{I}_v)$ is the average index value along the whole sequence.

Auto-cross covariance (ACC) (Cao et al., 2013) is the combination of AC and CC. Physicochemical distance transformation (PDT) (Liu et al., 2012) is a sequence-based method, in which any peptide sequence is firstly encoded as a series of numbers by amino acid index (AAindex) (Kawashima et al., 2008), and then a fixed length vector is extracted through distance transformation.

Pseudo Amino Acid Composition

Parallel correlation pseudo amino acid composition (PC-PseAAC) (Chou, 2001) is an approach that takes the sequence-order information into account and represents any peptide sequence as:

$$\mathrm{P}=[x_1\ x_2\ x_3\ \cdots\ x_{20}\ x_{20+1}\ \cdots\ x_{20+\lambda}]$$

where

$$x_u=\begin{cases}\dfrac{f_u}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{\lambda}\theta_j} & (1\le u\le 20)\\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{\lambda}\theta_j} & (20+1\le u\le 20+\lambda)\end{cases}$$

where $f_i(i=1,2,\ldots,20)$ is the occurrence frequency of the 20 native amino acids in the peptide; the integer λ represents the highest tier of correlation along the sequence; w is the weight factor ranging from 0 to 1; $\theta_j(j=1,2,\ldots,\lambda)$ is the j-tier correlation factor that is defined as:

$$\theta_j=\frac{1}{L-j}\sum_{i=1}^{L-j}\Theta\left(R_i,R_{i+j}\right)\qquad\left(1\le j\le\lambda\right)$$

Where the correlation function is given by

$$\Theta\left(R_i,R_j\right)=\frac{1}{3}\left\{\left[H_1(R_i)-H_1(R_j)\right]^2+\left[H_2(R_i)-H_2(R_j)\right]^2+\left[M(R_i)-M(R_j)\right]^2\right\}$$

where $H^1(R_i)$, $H^2(R_i)$, and $M(R_i)$ are the standardized hydrophobicity value, hydrophilicity value, and side-chain mass of R_i, respectively.

Series correlation pseudo amino acid composition (SC-PseAAC) (Chou, 2005) is a variant of PC-PseAAC that represents any peptide sequence as:

$$\mathrm{P}=[x_1\ \cdots\ x_{20}\ x_{20+1}\ \cdots\ x_{20+\lambda}\ x_{20+\lambda+1}\ \cdots\ x_{20+2\lambda}]$$

where

$$x_u=\begin{cases}\dfrac{f_u}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{2\lambda}\theta_j} & (1\le u\le 20)\\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{2\lambda}\theta_j} & (20+1\le u\le 20+2\lambda)\end{cases}$$

where $f_i(i=1,2,\ldots,20)$ is the occurrence frequency of the 20 native amino acids in the peptide; the integer λ represents the highest tier of correlation along the sequence; w is the weight factor ranging from 0 to 1; $\theta_j(j=1,2,\ldots,2\lambda)$ is the j-tier correlation factor that is defined as:

$$\begin{cases}\theta_1=\frac{1}{L-1}\sum_{i=1}^{L-1}H^1_{i,i+1}\\ \theta_2=\frac{1}{L-1}\sum_{i=1}^{L-1}H^2_{i,i+1}\\ \cdots\\ \theta_{2\lambda-1}=\frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}H^1_{i,i+\lambda}\\ \theta_{2\lambda}=\frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}H^2_{i,i+\lambda}\end{cases}$$

where the correlation functions are given by

$$\begin{cases}H^1_{i,j}=h^1\left(R_i\right)\cdot h^1\left(R_j\right)\\ H^2_{i,j}=h^2\left(R_i\right)\cdot h^2\left(R_j\right)\end{cases}$$

where $h^1(R_i)$ and $h^2(R_i)$ are the standardized hydrophobicity and hydrophilicity values of R_i, respectively.

General parallel correlation pseudo amino acid composition (PC-PseAAC-General) is an enhanced version of PC-PseAAC, in which both the built-in indices extracted from AAindex and the indices provided by users can be incorporated. General series correlation pseudo amino acid composition (SC-PseAAC-General) is an enhanced version of SC-PseAAC, in which both the built-in indices extracted from AAindex and the indices provided by users can be incorporated.

Profile-Based Features

The Top-n-gram (Liu et al., 2014b) approach extracts evolutionary information from the frequency profiles calculated from the multiple sequence alignments outputted by PSI-BLAST (Altschul et al., 1997), and any peptide sequence is represented as a fixed dimension feature vector by counting the occurrence times of each Top-n-gram. Profile-based physicochemical distance transformation (PDT-Profile) is similar with PDT except that the features are extracted from frequency profiles. Distance-based Top-n-gram (DT) extends the original Top-n-gram approach by considering the relative position information of Top-n-gram pairs in peptide sequences, and the feature vector of peptide sequence was calculated by counting the occurrences of all possible Top-n-gram pairs within a certain distance threshold.

Profile-based Auto covariance (AC-PSSM) (Dong et al., 2009) transforms the PSSM of a peptide into fixed-length vector,

in which the AC variable measures the correlation of the same property between two residues separated by a distance. Profile-based Cross covariance (CC-PSSM) (Dong et al., 2009) transforms the PSSM of a peptide into fixed-length vector, in which the CC variables measure the correlation of two different properties between two residues separated by a distance. Profile-based Auto-cross covariance (ACC-PSSM) (Dong et al., 2009) represents any peptide sequence as a feature vector consisting of ACC variables that are the combination of AC variables and CC variables. PSSM distance transformation (PSSM-DT) (Xu et al., 2015) extracts features from the PSSM of a peptide which measure the occurrence probabilities of any amino acid pairs separated by a distance. PSSM relation transformation (PSSM-RT) (Zhou et al., 2017) extracts features from the PSSM of a peptide by utilizing the relationships of evolutionary information between residues.

Support Vector Machine and LightGBM

In this study, the dataset has exactly two class labels: anticancer peptides (positive) and non-anticancer peptides (negative). Support vector machines (SVMs) are very suitable for binary classification, and because of the strong generalization ability for small datasets, they are used extensively in biomedical data mining (Chen et al., 2019; Jiang et al., 2020). SVM classifies data by finding the best hyperplane to separate all data points of one class from these of another class. The best hyperplane of SVM is the hyperplane with the largest margin between two classes. SVM is firstly proposed for linearly separable data, and when the data are non-separable, the kernel functions such as radial basis function can be used.

LightGBM (Light Gradient Boosting Machine) is a distributed gradient lifting framework based on decision tree algorithm proposed by Microsoft in 2017 (Ke et al., 2017). In order to shorten the computation time, LightGBM as a good ensemble learning algorithm was designed for two main reasons (Xia et al., 2017). For one thing, it can reduce the use of memory and the communication cost, improves the efficiency when multiple machines are parallel. For another thing, it designs and implements a good strategy for feature selection.

Methodology

To develop an accurate predictor of ACPs, we present a two-step ensemble learning method called EnACP. The framework of the model is shown in **Figure 3**. In the first step, 19 feature encodings of the peptide sequences are extracted in terms of amino acid composition, autocorrelation, pseudo amino acid composition and profile-based features as descripted in section Features Representation. For each group of feature encodings, the initial prediction is obtained separately using an ensemble learning classifier LightGBM. In this way, the complex higher-dimensional features are dispersed to lower dimensions. Then, the outputs of all LightGBMs as combinative nineteen-dimensional feature vector are input into an optimized SVM classifier to capture the hidden relationships. At last, the peptide sequence is identified whether it is ACP or non-ACP.

For a given binary classification problem about a set of sequences $Q(s)$, the class labels $C=\{C_1, C_2, \ldots, C_s\}$, $C_i \epsilon \{0, 1\}$, and

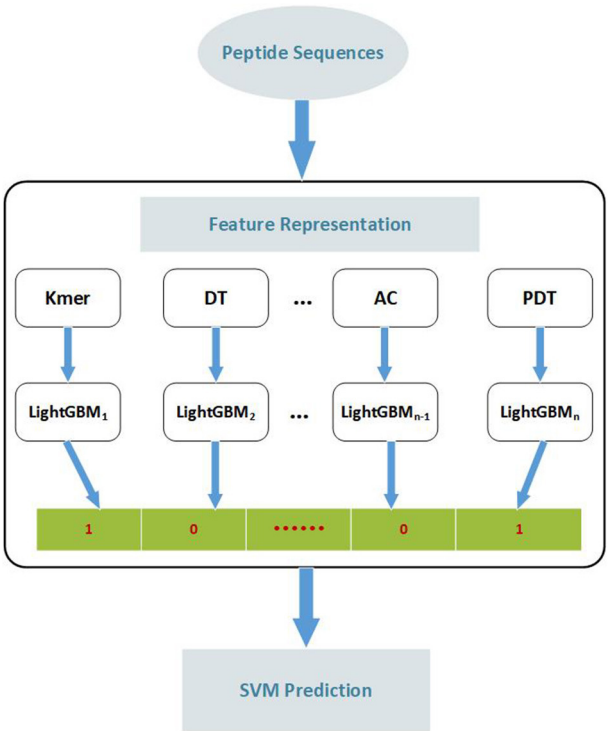


FIGURE 3 | The flow diagram of identifying anticancer peptides. Peptide sequences are represented by 19 feature extraction methods. According to the features obtained by each extraction method, LightGBM is employed to classify the peptide sequences. Then, the outputs of all LightGBMs are input into SVM classifier to predict the peptide sequence as ACP or non-ACP. Kmer: subsequences of length K contained within a peptide sequence; DT, Distance-based Top-n-gram; AC, Auto Covariance; PDT, Physicochemical Distance Transformation; SVM, Support Vector Machine.

each sample q_i has k group features $<F_1(q_i), F_2(q_i), \ldots F_k(q_i)>$, where F_j is the j^{th} group features. Each group has several related features. Firstly, all the features are generated by the 19 kinds of feature representation algorithm for all the sequences. For the train dataset, LightGBM is employed to classify each group features, respectively. The LightGBM classification results of k group features are input SVM to train the model. For the test dataset, the inputs are generated according to the first layer model of the train data set. Finally, ACPs or non-ACPs are identified for the test peptide sequences. The algorithm flow is described in the following pseudocode.

As shown from the pseudocode, there are three factors that affect the time complexity of the model EnACP, such as feature extraction, LightGBM and SVM algorithms. Let p and n be the numbers of the most features $F_i(q_i)$ and train samples Q_t, respectively. And the length of the longest sequence is l. Different feature extraction methods are relatively independent, and they can be generated in parallel. So, the most complex feature extraction method determines the time complexity of the feature extraction stage. For the 19 groups of feature extraction methods, the profile-based method with the highest complexity is $O(n^* l^3)$. LightGBM is implemented using three technologies to improve

Algorithm: *EnACP*

Input: a sequences set Q: (q_i, C_i), k groups of feature types, class label C_i={0, 1}, q_i is a peptide sequence, Q_t *is train dataset,* Q_v *is test dataset*.

Begin

1. **for** each sequence q_i in Q:
   // Initialize all features of q_i, *each* $F_i(q_i)$ *represent one group of features*
   2. $F(q_i)=<F_1(q_i), F_2(q_i),..., F_k(q_i)>=\{\}$
   // Initialize second level features
   3. $L2FK(q_i)[1..k] =\{\}$
   // Feature extract
   4. **for** $j = 1$ to k
      5. Generate features $F_j(q_i)$ according to feature representation algorithm F_j
   6. **endfor**
7. **endfor**
8. **for** train dataset Q_t: (q_t, C_t)
   9. **for** $m = 1$ to k
      // Classify the sequences Q_t *in the first level*
      10. $L1Model_m = LightGBM(F(q_t), C_t)$
      11. $L2FK(q_t)[m]= L1Model_m(F(q_t)$-)
   12. **endfor**
   // Train the model in the second level
   13. $L2Model=SVM\ (L2FK(q_t)\ [1..k], C_t)$
14. **endfor**
15. **for** test dataset Q_v: (q_v, C_v)
   16. **for** $n = 1$ to k
      // Classify the sequences Qv in the first level
      17. $L2FK(q_v)\ [n]= L1Modeln(F(q_v)$-)
   18. **endfor**
   // Predict the peptide sequence q_v: *ACP or non-ACP*
   19. $FinalPredict(q_v)= L2Model(L2FK(q_v)\ [1..k])$
20. **endfor**

End

the model efficiency: gradient-based one-side sampling, exclusive feature bundling, and histogram algorithm. These techniques have resulted in more or less a reduction in the number of samples and features. Moreover, it also supports feature parallel and data parallel processing. So, its worst time complexity will not exceed $O(p^* n)$. And the computational complexity of an SVM is $O(n^3)$ for the training dataset. So the worst-case time complexity of EnACP is $\max(O(n^* l^3), O(p^* n), O(n^3)$-). But most of the features will usually be excluded in the first layer. Then the SVM algorithm in the second layer will be significantly speeded up. So the actual calculation time will not reach the upper-bound in the train stage. For the test dataset, the time is mainly consumed in the feature extraction stage after the parameters of LightGBM and SVM are optimized.

Evaluation

The metrics for performance evaluation used in our experiments include Receiver Operating Characteristic curve (ROC), Area Under a ROC Curve (AUC), Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), and the Matthews correlation coefficient (MCC) (Plyusnin et al., 2019). Suppose TP, FP, TN and FN are the abbreviations for true positives, false positives, true negatives, and false negatives respectively, then the evaluation metrics can be calculated as:

$$Sp = \frac{TN}{TN + FP}$$

$$Sn = \frac{TP}{TP + FN}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

RESULTS

Performance on Different Feature Representations

In order to find the effective feature coding representation of the peptide sequence, four kinds of feature representation methods including 19 feature encodings were extracted in terms of amino acid composition, autocorrelation, pseudo amino acid composition and profile-based features. Referring to the first step of the model, the ACPs were identified by LightGBM classifier using various feature codes, respectively. From the overall results in **Figure 4**, they were ranked by pseudo amino acid composition, amino acid composition, profile-based features and autocorrelation. In terms of the various feature codes, pseudo amino acid composition worked best according to the value of the performance indexes Acc, AUC, Sp, Sn, and MCC. Its MCC was nearly 14 percentage points higher than the second place. And, its Acc, Sn, and Sp were about 7 percentage points higher than the second-place method from amino acid composition. Among them, autocorrelation encoding was the worst, and its performance indexes were all below 80%.

Performance Comparison on Cross-Validation Dataset

To verify the effect of our model, we compared the results of a few popular methods such as Li method (Li and Wang, 2016), ZH method (Hajisharifi et al., 2014) and iACP (Chen W. et al., 2016) on ZH dataset with 5-fold cross-validation. In order to compare the predictive capability, the predicted results of the four methods were showed in **Table 1**. Judging from the result, our predictor EnACP performed better than other three methods and reached the first place in the evaluation indexes on Sn, Acc, and MCC. In all the evaluation indexes, EnACP only lost to iACP in Sp index. Acc, Sn, and MCC of our method were about 0.6 to 5.7%, 2.2 to 7.6%, and 1.7 to 12.6% higher than the predictive results of other methods, respectively. In terms of Sp index, our method was only 0.9% lower than iACP method, but also much higher than other methods. From the discussion above, it can be seen that our method may automatically learn representative features from the numerous feature codes. The two step combined classifiers with LightGBM and SVM may improve the accuracy of prediction and achieve better identification efficiency between ACPs and non-ACPs.

Furthermore, for the stability of the model, 5-fold cross validation experiment was executed 30 times randomly.

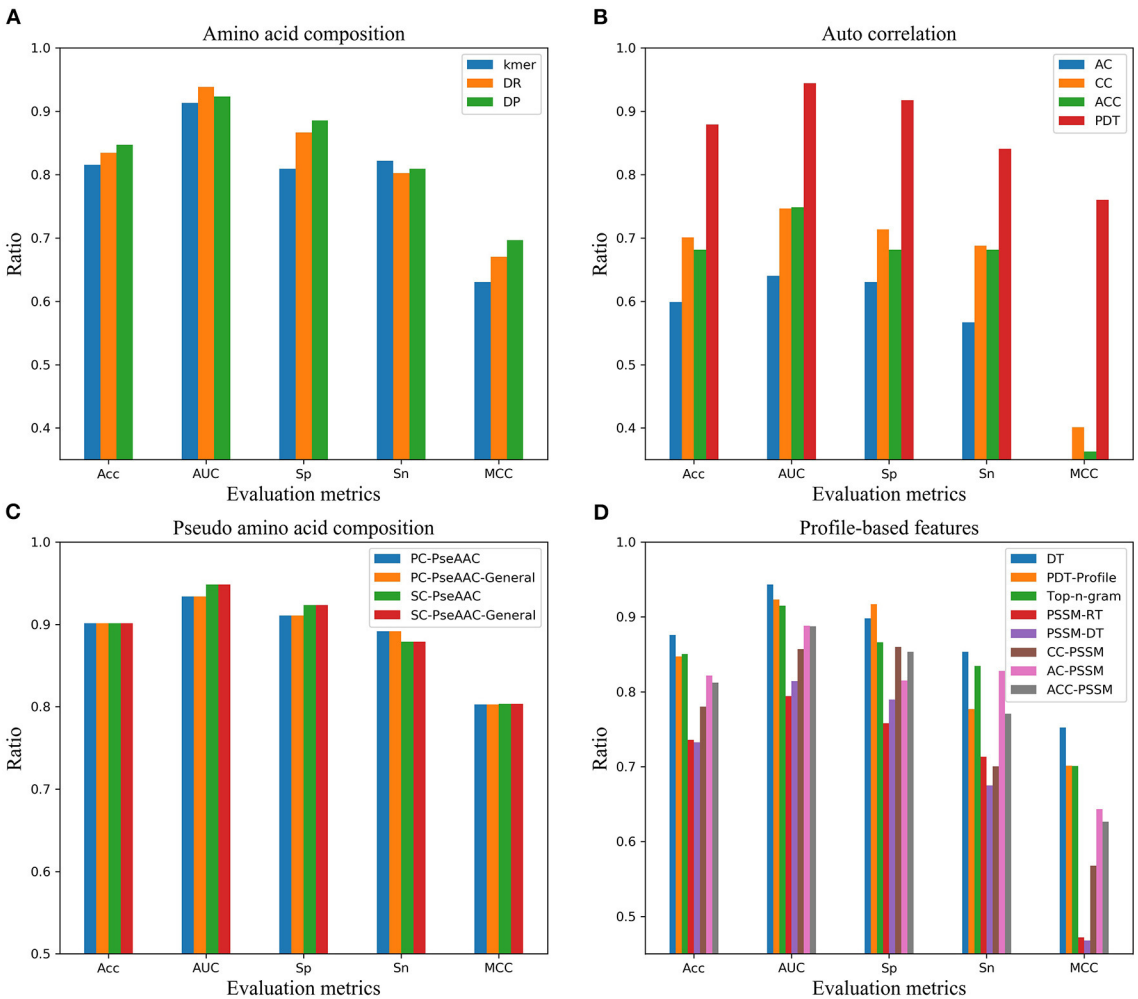


FIGURE 4 | Performance of four kinds of feature encodings on independent dataset in AUC, Sn, Sp, Acc, and MCC. **(A)** Performance of feature encodings based on amino acid composition**, (B)** Performance of feature encodings based on autocorrelation**, (C)** Performance of feature encodings based on pseudo amino acid composition, **(D)** Performance of feature encodings based on profile.

According to the statistical results of various evaluation metrics shown in **Figure 5**, several indicators fluctuate little. And the standard deviation of Acc, MCC, Sn, and Sp is 0.0005, 0.0012,0.0012, and 0.0011, respectively. Therefore, the cross-validation analysis showed the stability and robustness of our model EnACP.

Performance Comparison on Independent Test Datasets

To further verify the power of the current predictor, three independent datasets are analyzed from mACPpred (Boopathi et al., 2019), ACPP (Vijayakumar and Lakshmi, 2015), and Tyagi's paper (Tyagi et al., 2013) named mACP_Ind, ACPP_Ind, and Tyagi_Ind, respectively. For the independent test dataset mACPpred_Ind, SVMACP and RFACP belong to MLACP algorithm based on RF and SVM method, respectively. For this dataset, we refer to the experimental results from the literature mACPpred (**Table 2**). And for the independent test datasets ACPP_Ind and Tyagi_Ind, we compare our algorithms EnACP with mACPpred and iACP (**Table 3**). Experimental results on independent tests show that this proposed EnACP predictor is quite more effective and promising for identification of ACPs compared with the previous methods.

Compared with mACPpred method, our model EnACP had achieved excellent results, among which, MCC, Acc, Sn, and Sp were all about 2, 1, 0.7, and 1.2% higher, respectively, AUC was

TABLE 1 | Performance comparison of different methods on 5-fold cross-validation dataset.

Methods	Acc	Sn	Sp	MCC
EnACP	0.954	0.928	0.981	0.910
Li method	0.942	0.906	0.967	0.879
ZH method	0.897	0.852	0.927	0.784
iACP	0.948	0.884	0.990	0.893

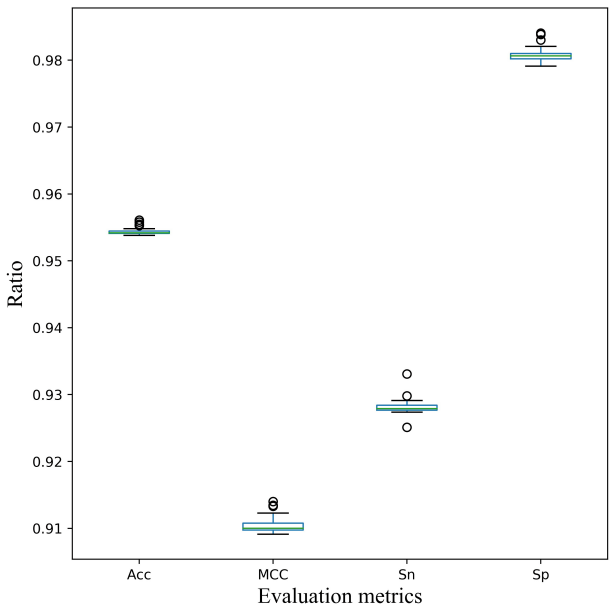


FIGURE 5 | Stability of the EnACP model on 5-fold cross-validation dataset. Five-fold cross validation experiment was executed 30 times randomly. And the metrics of Acc, MCC, Sn and Sp were plotted and analyzed.

TABLE 2 | Performance comparison of different methods on the independent test dataset mACPpred_Ind.

Methods	Acc	Sn	Sp	MCC	AUC
EnACP	0.924	0.892	0.955	0.849	0.968
mACPpred	0.914	0.885	0.943	0.829	0.967
SVMACP	0.768	0.554	0.981	0.592	0.896
RFACP	0.707	0.414	1.000	0.511	0.891
iACP	0.667	0.580	0.753	0.338	0.747

TABLE 3 | Performance comparison of different methods on the independent test datasets ACPP_Ind and Tyagi_Ind.

Datasets	Methods	Acc	Sn	Sp	MCC	AUC
	EnACP	0.948	1	0.9	0.901	0.992
ACPP_Ind	mACPpred	0.948	0.973	0.925	0.898	0.989
	iACP	0.74	0.919	0.575	0.558	0.875
	EnACP	0.853	1	0.708	0.739	0.996
Tyagi_Ind	mACPpred	0.884	0.957	0.813	0.777	0.948
	iACP	0.8	0.894	0.708	0.612	0.905

basically flat. MCC, Acc, Sn, and AUC obtained from our model EnACP were about 25.7 to 51.1%, 15.6 to 25.7%, 31.2 to 47.8%, 7.2 to 22.1% higher, respectively, compared with SVMACP, RFACP, and iACP. Additionally, it can also be seen from the results of **Figure 4** and **Table 2** that the EnACP method has an advantage over the pseudo amino acid composition method with one step prediction. Sn is only slightly lower less than a percentage point. And, MCC, Sp, Acc, and AUC obtained from EnACP model were about 4, 4, 2, 2% higher than the pseudo amino acid composition method with one step prediction. For ACPP_Ind and Tyagi_Ind datasets, EnACP achieves the similar performance advantages on AUC and Sn.

TABLE 4 | Pairwise comparison of ROC curves in three datasets.

Datasets	P(A, B)	EnACP	mACPpred	iACP
mACP_Ind	EnACP	—	0.9705	<0.0001
	mACPpred	—	—	<0.0001
	iACP	—	—	—
ACPP_Ind	EnACP	—	0.6612	0.0036
	mACPpred	—	—	0.0076
	iACP	—	—	—
Tyagi_Ind	EnACP	—	0.0384	0.0015
	mACPpred	—	—	0.2381
	iACP	—	—	—

The comparison P(A, B) is defined the statistical significance P-value of ROC curves between algorithm A and algorithm B.

TABLE 5 | The comparison triplets between algorithm pairs from EnACP, mACPpred and iACP.

T(A,B)	EnACP	mACPpred	iACP
EnACP	—	1/2/0	3/0/0
mACPpred	0/2/1	—	2/1/0
iACP	0/0/3	0/1/2	—

The comparison triplet T(A, B) is defined to be the numbers of the three datasets where algorithm A performs better, equally well and worse, compared with algorithm B in terms of P-value.

The statistical significance is evaluated using rank-based ROC curves comparison to determine whether EnACP performs better than, similarly to or worse than the other algorithms (DeLong et al., 1988; Hanley and Hajian-Tilaki, 1997). The results are shown in the following **Table 4**. For a confidence level of 0.95, EnACP perform statistically significantly better than iACP on all datasets. EnACP performs similarly or slightly better than mACPperd algorithms on mACP_Ind and ACPP_Ind. And mACPpred performs better than iACP on the previous two datasets. The algorithms EnACP and mACPperd perform better than iACP with statistical significance. The comparison triplets are also statistically tabulated between algorithm pairs from EnACP, mACPpred and iACP which show that one algorithm performs better, equally well and worse, compared with another algorithm in **Table 5**.

Comparison of Different Classification Methods

Based on many previous studies, using SVM classifier for task of peptide classification outperforms most of other classical classifiers such as AdaBoost, decision tree (DT), logistic regression (LR), Naïve Bayes (NB), random forest (RF) (Becker et al., 2011). We also conducted a comparative study on the two datasets and obtained the similar conclusion in the second step of

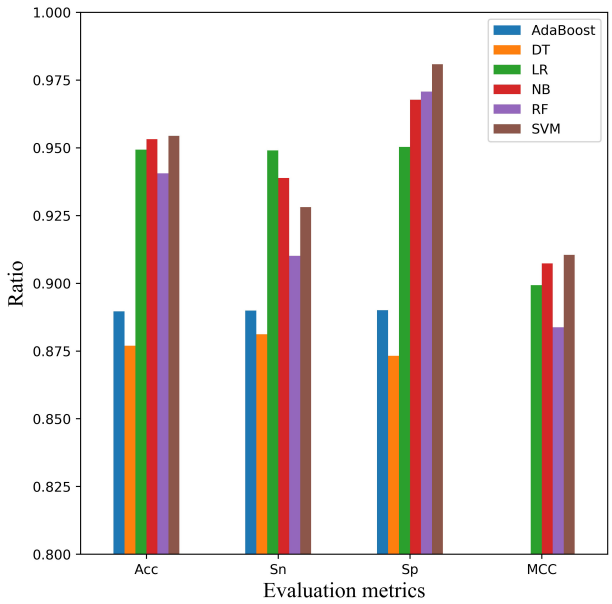


FIGURE 6 | Comparison of SVM with other classifiers on 5-fold cross-validation dataset. Four performance indicators which are Sn, Sp, Acc, and MCC are compared using six classifiers that are AdaBoost, decision tree (DT), logistic regression (LR), Naïve Bayes (NB), random forest (RF), and Support Vector Machine (SVM), respectively.

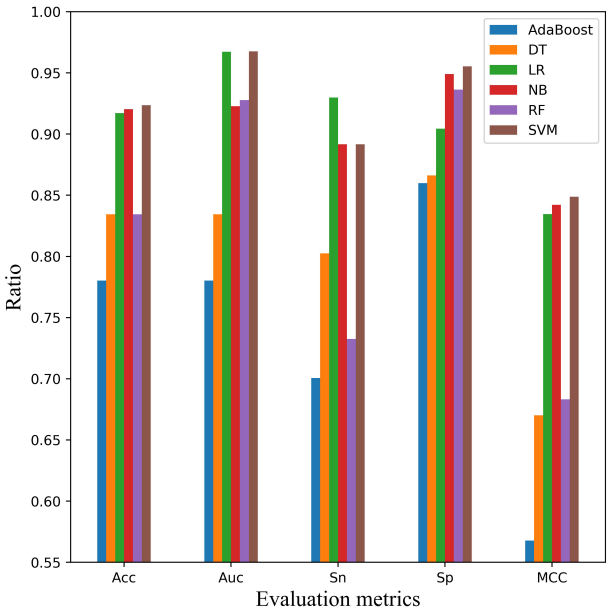


FIGURE 7 | Comparison of SVM with other classifiers on independent test dataset mACPpred_Ind. Five performance indicators which are AUC, Sn, Sp, Acc, and the Matthews correlation coefficient (MCC) are compared using six classifiers that are AdaBoost, decision tree (DT), logistic regression (LR), Naïve Bayes (NB), random forest (RF), and Support Vector Machine (SVM), respectively.

the model EnACP. Experimental results on both the 5-fold cross-validation and independent test showed that SVM, NB and LR were relatively stable, and SVM has the best overall effect.

In order to verify the performance of SVM classifier, we randomly selected scrambled data before 5-fold cross-validation. Finally, the average result of six classifiers were obtained after 30 times of 5-fold cross validation, as shown in **Figure 6**. Each classifier performed well, but in the comprehensive comparison, SVM, LR, and NB classifiers were better. On the whole, SVM classifier worked best. SVM achieved the first place in the three indexes of Acc, MCC, and Sp. For the Sn index, it was only about 1 and 2% lower than the classifier of NB and LR, respectively.

In addition, independent test dataset mACPpred_Ind was used to measure the performance and categorization capabilities of the optimal model in **Figure 7**. Compared with the cross-validation experiment, the AUC evaluation metric was added into this experiment except Acc, Sn, Sp, and MCC. Except for Sn, SVM classifier ranked the first place in Acc, AUC, Sp, and MCC, which was similar to the cross-validation result. But, SVM had better performance relative to cross validation tests. For example, for AUC index, SVM was more than 13 points higher than AdaBoost and DT. For Sp index, SVM is more than 5 points higher than AdaBoost, LR and DT. For MCC, SVM was 16% higher than RF and DT.

DISCUSSION

Even to this day, it is difficult to trace the cause of cancer because of its complex mechanisms. In spite of various treatment strategies, the effect was not ideal. Peptide-based therapy has become a research field of precision medicine. The rapid and accurate identification of ACPs from peptide sequences based on machine learning methods can be better applied to anticancer drug development and other biomedical experiments (Diller et al., 2018).

From the experimental results of the independent test datasets, our model EnACP performs well overall especially the high AUC and sensitivity. The higher the sensitivity is, the better the predicted model of ACPs is. The highly sensitive discovery of anticancer peptides plays an important role in the design of anticancer and anti-tumor synthetic drugs. The innovation of our model mainly includes the following points. The model EnACP is robust and easy to extend. Multi-group feature encodings contain abundant information. For each group of feature encoding, LightGBM as the first layer of EnACP can auto pre-learning and select the key features, respectively. Actually, for the higher-dimensional features, the computation is not very large. Meanwhile, the model implements the multi-layer feature learning strategy. Moreover, the second layer has fewer features and the model is more efficient to identify the ACPs and non-ACPs. The proposed EnACP performs better in identifying whether the peptide sequence is ACP compared with the existing methods. Its accuracy and stability may be attributed to the following reasons.

At first, how to effectively extract the valuable information of ACPs is a major challenge for all the predicted methods. It has been proved that the membrane interaction and insertion of

membrane-active peptides could be related to the order of amino acids. Systematic analysis revealed that some physiochemical properties of peptides are not clearly sufficient to predict their selectivity for example net positive charge, hydrophobicity, and hydrophobic moments (Chen W. et al., 2016). Some methods also are developed using amino acid composition and binary profiles as input features (Lin et al., 2015). Therefore, in order to find a suitable feature representation, EnACP extracts 19 kinds of features from four aspects, including amino acid composition, auto correlation, pseudo amino acid composition and profile-based features.

Then, in purpose to accurately identify the ACPs quickly, LightGBM classifier is applied to detect the peptide sequences with the 19 kinds of features. As an ensemble learning method, LightGBM can automatically optimize to achieve dimension reduction and effectively prevent overfitting. On the other hand, it can better discover the relationship of peptides and select the representative feature description from the integrating multiple groups features (Huang et al., 2010). In addition, the secondary structure and tertiary structure prediction characteristics of peptides can be added into this model as a part of basis feature description, which may further improve the performance of the model (Ma et al., 2015). Furthermore, neural network method can also be explored for the identification of ACPs with the increase of datasets (Hashemifar et al., 2018).

Finally, in terms of the used classifiers, many prediction tools have demonstrated the effectiveness of the SVM method. As a two-step prediction model, SVM finally outputs the identified results with grid search to optimize its parameters. Besides, in order to expedite the identification of ACPs, we called LightGBM with the default parameters in the scikit-learn package library. Better model parameters may be obtained by modern optimization methods to improve the prediction performance.

CONCLUSION

In order to effectively identify ACPs from amino acid sequences, a novel hybrid predicted model EnACP is proposed in this paper. EnACP involves two-step strategy based on ensemble learning method. Firstly, multi-type and multi-group feature descriptions were constructed based on amino acid composition, autocorrelation, pseudo amino acid composition and profile-based features. In purpose to find a suitable feature representation and accurately classify quickly, the ensemble classifier LightGBM was applied to detect the peptide sequences. Secondly, multiple groups of results from the output of LightGBMs were integrated as the input of SVM model to enhance the final prediction accuracy of ACP as well as non-ACP. To validate the performance of EnACP, two group experiments were performed on cross validate dataset and independent dataset. The experimental results indicated that the proposed EnACP model achieved competitive performance on some performance metrics. On the other hand, our model can be used to solve other protein sequence problems, such as homologous detection of proteins (Chen J. et al., 2016), prediction of various sites (Chou and Shen, 2008, 2010), prediction of protein-protein interaction (Wang et al., 2019), etc.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The datasets can be found in: https://github.com/greyspring/EnACP/tree/master/datasets.

AUTHOR CONTRIBUTIONS

RG and PW designed the method. RG and GF developed the prediction models. GF, XJ, RZ, and QW analyzed the data and results. All authors have read and approved the revised manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (Nos. 61702146, 61841104, 81602490), National key research and development program (No. 2019YFC0118404), Shandong province key research and development projects (No. 2017GSF218067), Zhejiang Postdoctoral Foundation (No. zj20180025) and China Scholarship Council (No. 201808330081).

ACKNOWLEDGMENTS

RG sincerely thanks Professor Jinbo Xu and his group from Toyota Technological Institute at Chicago (TTIC) for their creative discussions and valuable suggestions.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

An, Z., Flores-Borja, F., Irshad, S., Deng, J., and Ng, T. (2019). Pleiotropic role and bidirectional immunomodulation of innate lymphoid cells in cancer. *Front. Immunol.* 10:3111. doi: 10.3389/fimmu.2019.03111

Bahuguna, A., Singh, A., Kumar, P., Dhasmana, D., Krishnan, V., and Garg, N. (2020). Bisindolemethane derivatives as highly potent anticancer agents: synthesis, medicinal activity evaluation, cell-based compound discovery, and computational target predictions. *Comput. Biol. Med.* 116:103574. doi: 10.1016/j.compbiomed.2019.103574

Becker, N., Toedt, G., Lichter, P., and Benner, A. (2011). Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. *BMC Bioinform* 12:138. doi: 10.1186/1471-2105-12-138

Boopathi, V., Subramaniyam, S., Malik, A., Lee, G., Manavalan, B., and Yang, D. C. (2019). mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.* 20:1964. doi: 10.3390/ijms20081964

Brunetti, J., Piantini, S., Fragai, M., Scali, S., Cipriani, G., Depau, L., et al. (2020). A new NT4 peptide-based drug delivery system for cancer treatment. *Molecules* 25:1088. doi: 10.3390/molecules25051088

Cao, D. S., Xu, Q. S., and Liang, Y. Z. (2013). propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 29, 960–962. doi: 10.1093/bioinformatics/btt072

Chen, J., Guo, M., Wang, X., and Liu, B. (2018). A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief. Bioinform.* 19, 231–244. doi: 10.1093/bib/bbw108

Chen, J., Liu, B., and Huang, D. (2016). Protein Remote Homology Detection Based on an Ensemble Learning Approach. *Biomed Res. Int.* 2016, 5813645. doi: 10.1155/2016/5813645

Chen, T., Zhang, C., Liu, Y., Zhao, Y., Lin, D., Hu, Y., et al. (2019). A gastric cancer LncRNAs model for MSI and survival prediction based on support vector machine. *BMC Genomics* 20:846. doi: 10.1186/s12864-019-6135-x

Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K. C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 7, 16895–16909. doi: 10.18632/oncotarget.7815

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43, 246–255. doi: 10.1002/prot.1035

Chou, K. C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 21, 10–19. doi: 10.1093/bioinformatics/bth466

Chou, K. C., and Shen, H. B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162. doi: 10.1038/nprot.2007.494

Chou, K. C., and Shen, H. B. (2010). A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS ONE* 5:e9931. doi: 10.1371/journal.pone.0009931

DeLong, E. R., Delong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845. doi: 10.2307/2531595

Diller, K. I., Bayden, A. S., Audie, J., and Diller, D. J. (2018). PeptideNavigator: an interactive tool for exploring large and complex data sets generated during peptide-based drug design projects. *Comput. Biol. Med.* 92, 176–187. doi: 10.1016/j.compbiomed.2017.11.016

Dong, Q., Zhou, S., and Guan, J. (2009). A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics* 25, 2655–2662. doi: 10.1093/bioinformatics/btp500

Esfandiari Mazandaran, K., Mirshokraee, S. A., Didehban, K., and Houshdar Tehrani, M. H. (2019). Design, synthesis and biological evaluation of ciprofloxacin- peptide conjugates as anticancer agents. *Iran. J. Pharm. Res.* 18, 1823–1830. doi: 10.22037/ijpr.2019.111721.13319

Feng, P., and Wang, Z. (2019). Recent advances in computational methods for identifying anticancer peptides. *Curr. Drug Targets* 20, 481–487. doi: 10.2174/1389450119666180801121548

Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., et al. (2016). McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC Bioinform.* 17:142. doi: 10.1186/s12859-016-0990-0

Hajisharifi, Z., Piryaiee, M., Beigi, M. M., Behbahani, M., and Mohabatkar, H. (2014). Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* 341, 34–40. doi: 10.1016/j.jtbi.2013.08.037

Hanley, J. A., and Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Acad. Radiol.* 4, 49–58. doi: 10.1016/S1076-6332(97)80161-4

Harris, A. L. (2020). Development of cancer metabolism as a therapeutic target: new pathways, patient studies, stratification and combination therapy. *Br. J. Cancer* 122, 1–3. doi: 10.1038/s41416-019-0666-4

Hashemifar, S., Neyshabur, B., Khan, A. A., and Xu, J. (2018). Predicting protein-protein interactions through sequence-based deep learning. *Bioinformatics* 34, i802–10. doi: 10.1093/bioinformatics/bty573

Huang, T., Shi, X. H., Wang, P., He, Z., Feng, K. Y., Hu, L., et al. (2010). Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. *PLoS One* 5:e10972. doi: 10.1371/journal.pone.0010972

Jiang, H., Gu, J., Du, J., Qi, X., Qian, C., and Fei, B. (2020). A 21gene Support Vector Machine classifier and a 10gene risk score system constructed for patients with gastric cancer. *Mol. Med. Rep.* 21, 347–359. doi: 10.3892/mmr.2019.10841

Jing, X., Dong, Q., Hong, D. C., and Lu, R. (2019). Amino acid encoding methods for protein sequences: a comprehensive review and assessment. *IEEE/ACM Trans Comput Biol Bioinform.* doi: 10.1109/TCBB.2019.2911677. [Epub ahead of print].

Kasak, L., and Laan, M. (2020). Monogenic causes of non-obstructive azoospermia: challenges, established knowledge, limitations and perspectives. *Hum Genet.* doi: 10.1007/s00439-020-02112-y. [Epub ahead of print].

Kawashima, S., Pokarowski, P., Pokarowsika, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205. doi: 10.1093/nar/gkm998

Ke, G. L., Meng, Q., Finley, T., Wang, T. F., Chen, W., Ma, W. D., et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.

Li, F. M., and Wang, X. Q. (2016). Identifying anticancer peptides by using improved hybrid compositions. *Sci. Rep.* 6:33910. doi: 10.1038/srep33910

Lin, M. W., Tseng, Y. W., Shen, C. C., Hsu, M. N., Hwu, J. R., Chang, C. W., et al. (2018). Synthetic switch-based baculovirus for transgene expression control and selective killing of hepatocellular carcinoma cells. *Nucleic Acids Res.* 46:e93. doi: 10.1093/nar/gky447

Lin, Y. C., Lim, Y. F., Russo, E., Schneider, P., Bolliger, L., Edenharter, A., et al. (2015). Multidimensional design of anticancer peptides. *Angew. Chem. Int. Ed Engl.* 54, 10370–10374. doi: 10.1002/anie.201504018

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K. C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71. doi: 10.1093/nar/gkv458

Liu, B., Wang, X., Chen, Q., Dong, Q., and Lan, X. (2012). Using amino acid physicochemical distance transformation for fast protein remote homology detection. *PLoS ONE* 7:e46633. doi: 10.1371/journal.pone.0046633

Liu, B., Wang, X. L., Lin, L., Dong, Q. W., and Wang, X. (2008). A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. *BMC Bioinformatics* 9:510. doi: 10.1186/1471-2105-9-510

Liu, B., Wu, H., and Chou, K.-C. (2017). Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nat. Sci.* 9, 67–91. doi: 10.4236/ns.2017.94007

Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., et al. (2014a). iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS ONE* 9:e106691. doi: 10.1371/journal.pone.0106691

Liu, B., Xu, J., Zou, Q., Xu, R., Wang, X., and Chen, Q. (2014b). Using distances between Top-n-gram and residue pairs for protein remote homology detection. *BMC Bioinform.* 15(Suppl. 2):S3. doi: 10.1186/1471-2105-15-S2-S3

Ma, J., Wang, S., Wang, Z., and Xu, J. (2015). Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* 31, 3506–3513. doi: 10.1093/bioinformatics/btv472

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., and Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget* 8, 77121–77136. doi: 10.18632/oncotarget.20365

Naguib, A., Mathew, G., Reczek, C. R., Watrud, K., Ambrico, A., Herzka, T., et al. (2018). Mitochondrial complex I inhibitors expose a vulnerability for selective killing of Pten-null cells. *Cell Rep.* 23, 58–67. doi: 10.1016/j.celrep.2018.03.032

Ortega-Garcia, M. B., Mesa, A., Moya, E. L. J., Rueda, B., Lopez-Ordono, G., Garcia, J. A., et al. (2020). Uncovering tumour heterogeneity through PKR and nc886 analysis in metastatic colon cancer patients treated with 5-FU-based chemotherapy. *Cancers* 12:379. doi: 10.3390/cancers12020379

Plyusnin, I., Holm, L., and Toronen, P. (2019). Novel comparison of evaluation metrics for gene ontology classifiers reveals drastic performance differences. *PLoS Comput. Biol.* 15:e1007419. doi: 10.1371/journal.pcbi.1007419

Sima, P., Richter, J., and Vetvicka, V. (2019). Glucans as new anticancer agents. *Anticancer Res.* 39, 3373–3378. doi: 10.21873/anticanres.13480

Tesauro, D., Accardo, A., Diaferia, C., Milano, V., Guillon, J., Ronga, L., et al. (2019). Peptide-based drug-delivery systems in biotechnological applications: recent advances and perspectives. *Molecules* 24:351. doi: 10.3390/molecules24020351

Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., and Raghava, G. P. (2013). In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.* 3:2984. doi: 10.1038/srep02984

Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., et al. (2015). CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res.* 43, D837–D843. doi: 10.1093/nar/gku892

Umbreit, N. T., Zhang, C. Z., Lynch, L. D., Blaine, L. J., Cheng, A. M., Tourdot, R., et al. (2020). Mechanisms generating cancer genome complexity from a single cell division error. *Science* 368:aba0712. doi: 10.1126/science.aba0712

Vijayakumar, S., and Lakshmi, P. T. V. (2015). ACPP: a web server for prediction and design of anti-cancer peptides. *Int. J. Pept. Res. Ther.* 21, 99–106. doi: 10.1007/s10989-014-9435-7

Wang, X., Yu, B., Ma, A., Chen, C., Liu, B., and Ma, Q. (2019). Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* 35, 2395–2402. doi: 10.1093/bioinformatics/bty995

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Xia, J., Peng, Z., Qi, D., Mu, H., and Yang, J. (2017). An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier. *Bioinformatics* 33, 863–870. doi: 10.1093/bioinformatics/btw768

Xu, L., Liang, G., Wang, L., and Liao, C. (2018). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes (Basel)* 9:158. doi: 10.3390/genes9030158

Xu, R., Zhou, J., Wang, H., He, Y., Wang, X., and Liu, B. (2015). Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol.* 9(Suppl. 1):S10. doi: 10.1186/1752-0509-9-S1-S10

Zhou, J., Lu, Q., Xu, R., He, Y., and Wang, H. (2017). EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM relation transformation. *BMC Bioinformatics* 18:379. doi: 10.1186/s12859-017-1792-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 August 2020
doi: 10.3389/fgene.2020.00868

RIGD: A Database for Intronless Genes in the Rosaceae

Tianzhe Chen[1,2†], Dandan Meng[1,2†], Xin Liu[1,2], Xi Cheng[1,2], Han Wang[1,2], Qing Jin[1,2], Xiaoyu Xu[1,2], Yunpeng Cao[3] and Yongping Cai[1,2]**

[1] School of Life Sciences, Anhui Agricultural University, Hefei, China, [2] Anhui Provincial Engineering Technology Research Center for Development & Utilization of Regional Characteristic Plants, Anhui Agricultural University, Hefei, China, [3] Key Laboratory of Cultivation and Protection for Non-Wood Forest Trees, Ministry of Education, Central South University of Forestry and Technology, Changsha, China

OPEN ACCESS

Edited by:
Yungang Xu,
The University of Texas Health Science Center at Houston, United States

Reviewed by:
Wu Yuejin,
Hefei Institutes of Physical Science (CAS), China
Tao Huang,
Shanghai Institute for Biological Sciences (CAS), China

***Correspondence:**
Yunpeng Cao
xfcypeng@126.com
Yongping Cai
swkx12@ahau.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 23 January 2020
Accepted: 16 July 2020
Published: 07 August 2020

Citation:
Chen T, Meng D, Liu X, Cheng X, Wang H, Jin Q, Xu X, Cao Y and Cai Y (2020) RIGD: A Database for Intronless Genes in the Rosaceae. Front. Genet. 11:868. doi: 10.3389/fgene.2020.00868

Most eukaryotic genes are interrupted by one or more introns, and only prokaryotic genomes are composed of mainly single-exon genes without introns. Due to the absence of introns, intronless genes in eukaryotes have become important materials for comparative genomics and evolutionary biology. There is currently no cohesive database that collects intronless genes in plants into a single database, although many databases on exons and introns exist. In this study, we constructed the Rosaceae Intronless Genes Database (RIGD), a user-friendly web interface to explore and collect information on intronless genes from different plants. Six Rosaceae species, *Pyrus bretschneideri, Pyrus communis*, *Malus domestica*, *Prunus persica*, *Prunus mume*, and *Fragaria vesca*, are included in the current release of the RIGD. Sequence data and gene annotation were collected from different databases and integrated. The main purpose of this study is to provide gene sequence data. In addition, attribute analysis, functional annotations, subcellular localization prediction, and GO analysis are reported. The RIGD allows users to browse, search, and download data with ease. Blast and comparative analyses are also provided through this online database, which is available at http://www.rigdb.cn/.

Keywords: intronless genes, gene annotations, platform, database, Rosaceae

BACKGROUND

Genes in eukaryotes are generally composed of exons and introns, and according to the presence and absence of introns, they can be divided into intron-containing genes and intronless genes. It is generally believed that intron number is closely related to the complexity of the eukaryotic genome. If an organism is complex, it has more introns (Sakharkar et al., 2004). Most eukaryotic genes have two or more introns, while prokaryotes have a large number of intronless genes (Rogozin et al., 2005). Intronless genes are not interspaced by introns and can be sequentially encoded into proteins. Intronless genes can serve as focal point in analyses of gene function and evolution. For example, compared with intron-containing homologs, intronless genes can be used as a model to study the important role of introns, which are

only found in eukaryotes (Tine et al., 2011). Furthermore, studies on intronless genes help to solve some evolutionary issues, including (1) the main factors leading to the emergence of intronless genes (gene duplication, inheritance from ancient prokaryotes, retroposition or other mechanisms), (2) the evolutionary significance of retroposition (retrogenes are considered to be intronless), and (3) the biological origins of introns (is the introns-early hypothesis or introns-late hypothesis more correct) (Sakharkar and Kangueane, 2004).

In eukaryotes, the proportion of intronless genes varies from 2.7 to 97.7% of the genome (Louhichi et al., 2011). Currently, researchers have identified intronless genes in some species of mammals, hindmouths, bony fish, and plants (Agarwal and Gupta, 2005; Sakharkar et al., 2006; Jain et al., 2008; Zou et al., 2011). In Jain et al. (2006) studied the early auxin response SAUR (small auxin-up RNA) gene family in rice and found that all 58 members of the gene family were intronless genes. In the process of studying the functions of gene families in *Arabidopsis*, researchers also found a large number of intronless genes in the f-box protein family, DEAD box RNA helicase family, and PPR (pentatricopeptide repeat) gene family (Aubourg et al., 1999; Lecharny et al., 2003; Lurin et al., 2004). In addition, some of the largest families, such as the G-protein receptor family and the olfactory receptor family, are also composed of intronless genes (Gentles and Karlin, 1999; Takeda et al., 2002). Currently, the most studied intronless gene is the histone gene in the human genome. Researchers aim to explore the role of intronless genes in life processes by studying these gene families.

Since researching intronless genes in eukaryotes can help researchers better understand the evolutionary mechanism of related genes and genomes, the study of intronless genes has attracted more and more attention. In recent years, the construction of intronless gene databases has attracted great attention as the research on intronless genes. Relevant databases can provide important data resources for functional and evolutionary studies, facilitate researchers to carry out relevant research. So far, there are mainly databases on intronless genes: GENOME SEGE (Sakharkar and Kangueane, 2004), IGD (Louhichi et al., 2011), PIGD (Yan et al., 2014), and IGDD (Yan et al., 2016). GENOME SEGE contains NCBI data regarding the intronless genes of eukaryotes, however, the database website has stopped updating the data, and users are unable to access it. The IGD database, which includes 687 human intronless genes, was published in 2011. PIGD provides a platform for the collection, integration, and analysis of intronless genes in Poaceae. IGDD provides a comprehensive platform for researchers to explore intronless genes in dicot plants.

To build a centralized platform, we present the Rosaceae Intronless Genes Database (RIGD)[1]. This database, with a user-friendly web interface, covers a collection of intronless genes from six genome-sequenced Rosaceae species. The RIGD integrates functional and evolutionary annotations, making it easy for researchers to find content of interest and download detailed information. The RIGD provides a comparative analysis of genome data from six species in conjunction with the Blast program. Compared to the databases specifically for individual organisms, we expect the RIGD to be a useful resource for the research community, especially for studies on molecular function and the evolution of intronless genes.

TABLE 1 | The sources of six species in RIGD.

Species	Sources
Pyrus bretschneideri	GDR (ftp://ftp.bioinfo.wsu.edu/www.rosaceae.org/Pyrus_x_bretschneideri/Pbretschneideri-genome.v1.1)
Pyrus communis	GDR (ftp://ftp.bioinfo.wsu.edu/species/Pyrus_communis/Pcommunis_DH_genome.v2.0)
Malus domestica	NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/114/115/GCF_002114115.1_ASM211411v1)
Prunus persica	GDR (ftp://ftp.bioinfo.wsu.edu/species/Prunus_persica/Prunus_persica-genome.v2.0.a1)
Prunus mume	NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/346/735/GCF_000346735.1_P.mume_V1.0)
Fragaria vesca	GDR (ftp://ftp.bioinfo.wsu.edu/species/Fragaria_vesca/Fvesca-genome.v4.0.a1)

CONSTRUCTION AND CONTENT

Data Sources

Currently, the RIGD includes the following six Rosaceae species: *Pyrus bretschneideri*, *Pyrus communis*, *Malus domestica*, *Prunus persica*, *Prunus mume*, and *Fragaria vesca*. Genome data of *Malus domestica* and *Prunus mume* were downloaded via FTP from the NCBI genomes database[2]. Genome data of the other species were downloaded from the GDR database (Jung et al., 2019)[3] (**Table 1**).

Identification of Rosaceae Intronless Genes

A set of strict standards was used to identify Rosaceae intronless genes. First, we used a Perl script to extract genes containing only one line of "exon" from each genome information in the genome annotation files (GFF/GFF3 format files) and then used them as candidate intronless genes for further screening. The basis for the screening was: if there was only one row of "exons" in the genome information, indicating that the coding sequence is not disrupted by an intron, then the gene is an intronless gene. Since the mitochondrial genes and chloroplast DNA do not contain introns, the genes annotated as "Mt" (mitochondria) and "Pt" (chloroplast) were rejected. In addition, genes that are not mapped to the chromosome were removed. Genes defined as "pseudogene" or "transposable element" in the annotation files were deleted because a pseudogene cannot be transcribed or translated, it is usually not functional. Through the above steps, we obtained no redundant intronless genes in six Rosaceae species. Using the identified intronless gene number, we used a Perl script to extract the protein sequence and CDS sequence of the intronless genes and renumber them according to certain criteria.

[1]http://www.rigdb.cn/

[2]ftp://ftp.ncbi.nlm.nih.gov/genomes/all/

[3]https://www.rosaceae.org/

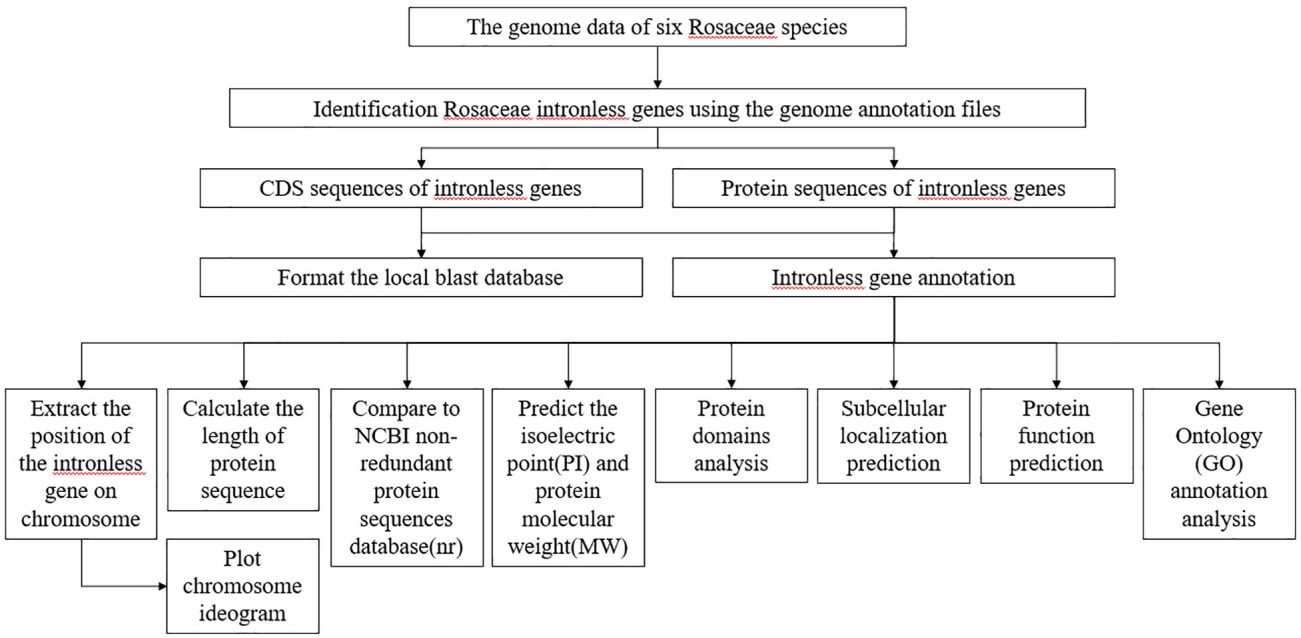


FIGURE 1 | The flowchart describing the analysis of RIGD. After the identification of intronless genes, attribute analysis, functional annotations, subcellular localization prediction, and GO analysis were accomplished.

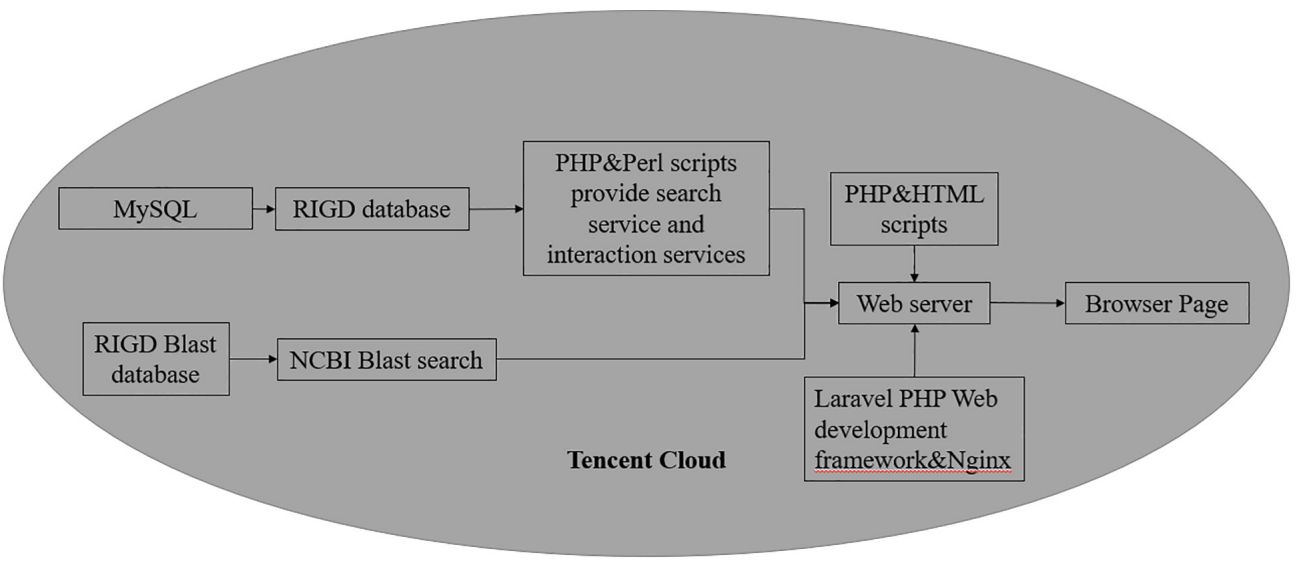


FIGURE 2 | The Flowchart of RIGD Sitemap. All the data is stored in MySQL database, the control of the platform is implemented by PHP and Perl scripts.

Intronless Gene Annotation

We established the following procedure to analyze each intronless gene stored in RIGD: (**Figure 1**). (1) A Perl script was used to extract the position information for intronless genes on corresponding chromosomes, and calculate the length of protein sequences. (2) Chromosome ideograms were plotted by using the chromosomeplot tool in MATLAB software[4] (Snijders et al., 2001). (3) The protein sequences were compared to the NCBI non-redundant protein sequences database (nr)[5] by using Diamond with default parameters[6] (Buchfink et al., 2015). The GI numbers were obtained, and then the bioperl module[7] was used to submit GI numbers to NCBI for the corresponding

[4]https://www.mathworks.com/help/bioinfo/ref/chromosomeplot.html

[5]ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz

[6]https://github.com/bbuchfink/diamond

[7]https://bioperl.org/

FIGURE 3 | An overview of the website page in the RIGD. **(A)** Home interface show photographs of six species, the RIGD's project description, author information and contact information. **(B)** In Species interface, the description of species, the download links of relevant analysis data and the analysis results are presented. **(C)** In Search interface, researchers can use different types of keywords to obtain the required data. **(D)** Blast interface can be used for Blast comparison of intronless genes in RIGD. **(E)** Statistics interface show the result of comparative analysis between Rosaceae species. **(F)** Tools interface support a tool that can batch submit sequences to ExPASy for pI/Mw prediction. **(G)** Upload&Download interface is designed to download and upload data. **(H)** Contact us interface. Researchers can contact us by email and link to some tools we used in our work.

FIGURE 4 | The website page of Species. In addition to the download links to the data, five statistical charts show the results of the number and proportion of intronless genes on each chromosome, the statistics of subcellular localization prediction, the chromosome ideogram, and the statistics of GO classifications.

annotation information. (4) A Python script was used to submit the protein sequences to ExPASy[8] (Artimo et al., 2012) to predict the isoelectric point (pI) and protein molecular weight (Mw). (5) Subcellular localization was predicted with MultiLoc2[9] (Blum et al., 2009). (The parameters are set as follows: -origin = plant, -predictor = HighRes, -output = advanced) (6) Protein domains were analyzed by using InterProScan (Jones et al., 2014) with default parameters and searched the Pfam database (El-Gebali et al., 2019). (7) Protein function was predicted based on sequences and structural templates from DeepGOPlus (Kulmanov and Hoehndorf, 2019), I-TASSER (Iterative Threading ASSEmbly Refinement) (Roy et al., 2010; Yang and Zhang, 2015; Yang et al., 2015) and InterProScan with default parameters. Finally, (8) eggNOG mapper (Huerta-Cepas et al., 2017, 2019) was used to implement Gene Ontology (GO) annotation analysis. (The parameters are set as follows: -m diamond –tax_scope auto –go_evidence non-electronic –target_orthologs all –seed_ortholog_evalue 0.001 –seed_ortholog_score 60 –query-cover 20 –subject-cover 0). Visualization of GO categories was performed by using the WEGO online tool (Ye et al., 2006, 2018).

Comparisons Between Rosaceae Species

In addition to analyzing the intronless genes of six Rosaceae species, we conducted the following comparative analysis of the intronless genes among species: (1) the number and percentage of intronless genes in each chromosome, (2) the distribution of protein length, (3) the distribution of pI, (4) the distribution of Mw, (5) the statistics of subcellular localization, and (6) the statistics of GO classifications.

The RIGD Implementation and Web Interface

As a web-based platform, the RIGD is constructed in a Tencent cloud server, and the operating system is Ubuntu Server 16.04.1 LTS 64-bit. The RIGD combines the MySQL (version 8.0.17) database management system with a dynamic web interface based on PHP (version 7.3.9-1), Laravel (version 5.8), Nginx (version 1.10.3), and Perl (version 5.22.1) scripts (**Figure 2**).

UTILITY AND DISCUSSION

Web Interface

The web interface of the RIGD is designed to comprise the following seven components: Home, Species, Search, Blast, Statistics, Upload&Download, and Contact Us. The RIGD provides a user-friendly interaction experience (**Figure 3**).

Home

The RIGD has seven navigation bars at the top. Scrolling through the home page reveals large photos of six species of Rosaceae.

[8]https://web.expasy.org/compute_pi/

[9]https://github.com/KohlbacherLab/MultiLoc2

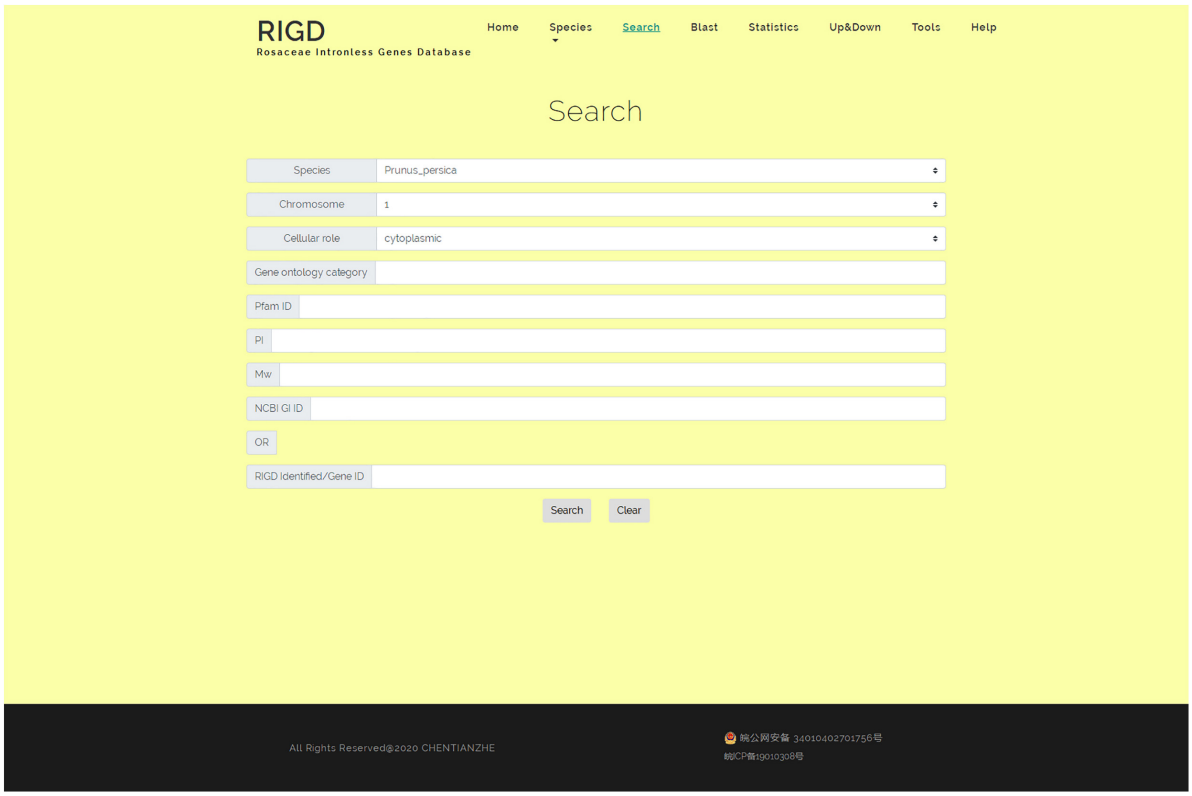


FIGURE 5 | The website page of Search. Researchers can search the RIGD database by species name, chromosome number, classification of subcellular location prediction, GO number, Pfam ID, the value of pI or Mw and NCBI GI ID.

There is a "detail" button on each photo that can be clicked to link to the "species" interface for each species. In addition, the RIGD's project description, author information and contact information are also available on the home page.

Species

The bar opens a drop-down menu with the names of the six Rosaceae species covered in the RIGD. Clicking to enter, you can then see a detailed description of the species and the picture on the page. There is also a table with download links, where much of data is available, including the CDS and protein sequence of intronless genes, the prediction of isoelectric point and protein molecular weight, the results of the sequence compared with the nr database, the results of protein domain analysis, the results of subcellular localization prediction, the results of protein function prediction, and the results of GO analysis. Some statistical charts are also shown on the page, such as the number and proportion of intronless genes on each chromosome, the statistics of subcellular localization prediction, the distribution of pI and Mw, and the statistics of GO classifications (**Figure 4**).

Search

In the search interface, users can search by species name, chromosome number, classification of subcellular location prediction, and even GO number. The program in the RIGD will search the database for eligible intronless genes and list them, and then users can click to view the detailed information. In addition, the RIGD will renumber the intronless genes after processing, and the rule is the abbreviation of the species name + "IG" + chromosome number + the order number of the gene (starting from 1). The gene number in the original data is still retained, and either the RIGD number or the gene number of the original data can be used for searching (**Figure 5**).

Blast

The RIGD has Blast software installed on the server, moreover, the intronless gene CDS and protein sequences of the six Rosaceae species stored in the RIGD were formatted into the Blast local database. In the Blast interface, users can paste a sequence or upload a fasta-format file to match with the RIGD's local Blast database and find the putative homologous sequences of these intronless genes in different species. The databases can be compared (CDS/protein), and Blast programs (Blastn/Blastp) and e-values can all be selected or entered into the interface (**Figure 6**).

Statistics

The results of comparative analysis among the six species are shown on the Statistics interface with statistical charts. Four pictures investigate the general trends in protein length

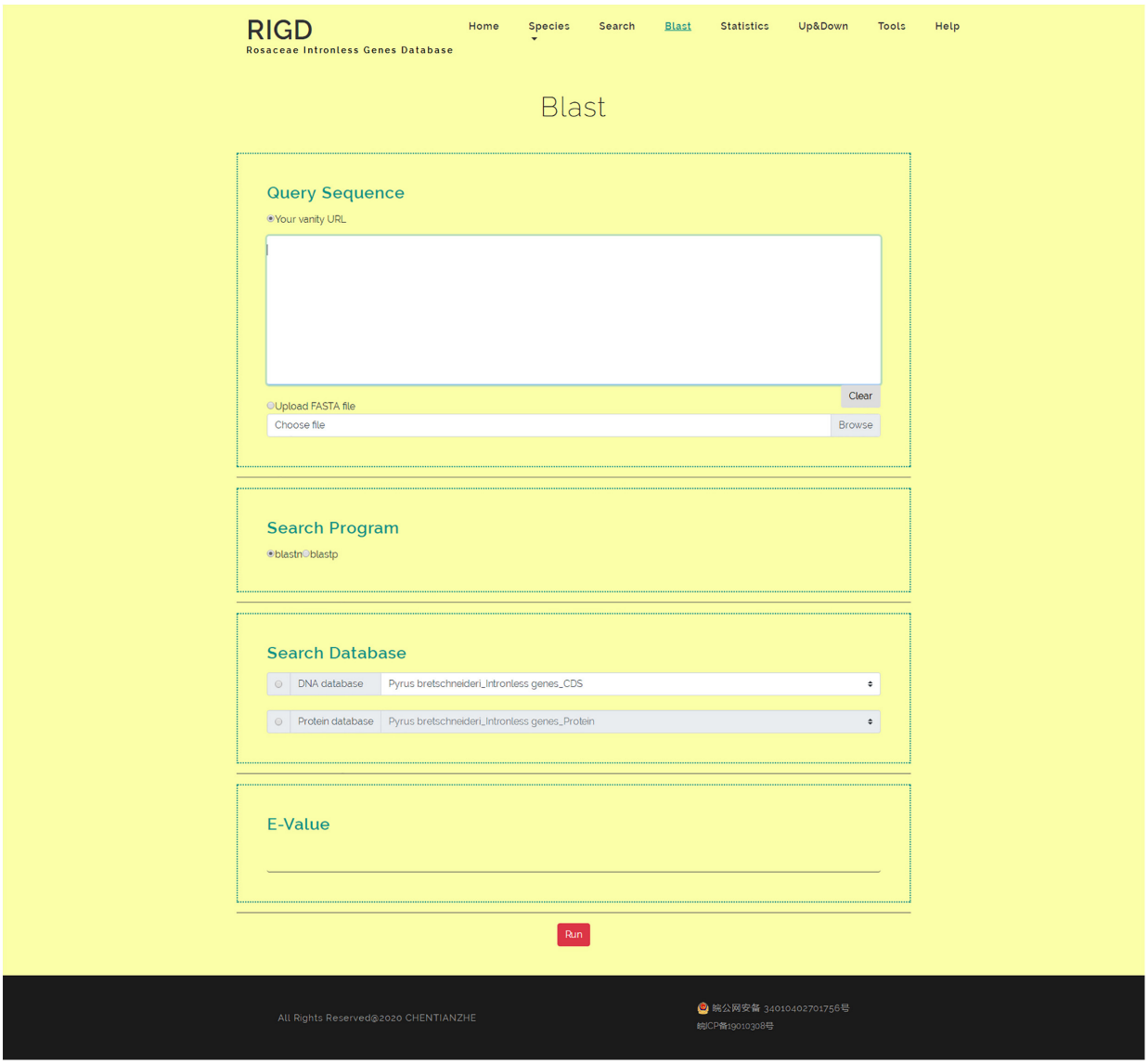


FIGURE 6 | The website page of Blast. The interface can be used for Blast comparison of intronless genes in RIGD, researchers can use blastp/blastn to analyze, and the E- value can be set.

distribution, show us the distribution of pI, the distribution of Mw and the statistics of subcellular localization.

Upload and Download

In the interface, according to the species name and the chromosome number in each species, users can download the following data in the "Download" section, namely, the CDS and protein sequence of intronless genes, the prediction of pI and Mw, the results of the sequence compared with the nr database, the results of protein domain analysis, the results of subcellular localization prediction, the results of protein function prediction, and the result of GO analysis, according to the species name and the chromosome number in each species. In the "Upload" section, users can upload intronless gene sequence files of other species or analysis result files to the RIGD server to expand the RIGD in the future.

Tools

We designed the Tools interface to collect some tools for intronless gene analysis or other practical bioinformatics analysis that will be developed in the future. The tool now available on this interface is a program that can batch submit sequences to ExPASy for pI/Mw prediction.

Contact Us

The Contact us interface is divided into "Contact us" and "Links." Users can email the RIGD's administrator in the "Contact us" interface to ask any questions or provide valuable suggestions.

TABLE 2 | The number of intronless genes reported for each species.

Species	Chromosome	Amount	Species	Chromosome	Amount	Species	Chromosome	Amount
Pyrus bretschneideri	Chr1	99	*Pyrus communis*	Chr1	298	*Malus domestica*	Chr1	191
	Chr2	107		Chr2	384		Chr2	246
	Chr3	147		Chr3	345		Chr3	260
	Chr4	118		Chr4	303		Chr4	195
	Chr5	138		Chr5	493		Chr5	292
	Chr6	104		Chr6	324		Chr6	185
	Chr7	151		Chr7	431		Chr7	247
	Chr8	76		Chr8	302		Chr8	185
	Chr9	105		Chr9	273		Chr9	243
	Chr10	150		Chr10	451		Chr10	300
	Chr11	107		Chr11	435		Chr11	270
	Chr12	91		Chr12	340		Chr12	210
	Chr13	96		Chr13	383		Chr13	254
	Chr14	99		Chr14	291		Chr14	224
	Chr15	179		Chr15	558		Chr15	345
	Chr16	75		Chr16	350		Chr16	245
	Chr17	124		Chr17	427		Chr17	251
	Total	1966		Total	6388		Total	4143
Prunus persica	Chr1	1272	*Prunus mume*	Chr1	399			
	Chr2	727		Chr2	570	*Fragaria vesca*	Chr1	523
	Chr3	709		Chr3	371		Chr2	708
	Chr4	643		Chr4	364		Chr3	811
	Chr5	517		Chr5	338		Chr4	640
	Chr6	855		Chr6	338		Chr5	672
	Chr7	577		Chr7	259		Chr6	997
	Chr8	651		Chr8	243		Chr7	579
	Total	5951		Total	2882		Total	4930

The "Links" interface contains links to external databases and analysis tools that the RIGD references.

CASE STUDY

The Results of Comparative Analysis Among the Six Species in Rosaceae

Twenty-six thousand two hundred sixty intronless genes were identified from six Rosaceae species. *Pyrus bretschneideri*, *Pyrus communis*, *Malus domestica*, *Prunus persica*, *Prunus mume*, and *Fragaria vesca* consist of 5.44% (1966), 17.79% (6388), 10.38% (4143), 22.20% (5951), 12.97% (2882), and 17.35% (4930) intronless genes, respectively (**Table 2**). The distribution of intronless genes on chromosomes was uneven in different species (**Supplementary Figures 1–6**). Although the number of intronless genes varied greatly from chromosome to chromosome, the proportion of intronless genes on each chromosome did not vary much among species (**Supplementary Figure 7**). The average protein length was ∼333.4 amino acids (aa) in *Pyrus bretschneideri*, 258.7 aa in *Pyrus communis*, 321.4 aa in *Malus domestica*, 277.5 aa in *Prunus persica*, 351.5 aa in *Prunus mume*, and 275.0 aa in *Fragaria vesca* (**Figure 7A**). The distribution of pI had three peaks (**Figure 7B**), and the distribution of Mw gathered at the front of the diagram, most predicted protein molecular weights were less than 100000 Da (**Figure 7C**). The largest number of intronless genes were categorized as cytoplasmic in their cellular role (**Figure 8**). The largest number of intronless genes in six species were predicted for pentatricopeptide repeat in their protein function. The second largest number of intronless genes were predicted for AP2/ERF domain in *Pyrus bretschneideri*, Leucine-rich repeat in *Pyrus communis*, Zinc finger (RING-type) in *Malus domestica*, *Prunus persica* and *Fragaria vesca*, and protein kinase domain in *Prunus mume*. Top 10 largest number of intronless genes in protein function were shown in **Figure 9**. The largest number of intronless genes were classified as biological process in GO categories. The largest proportion of intronless genes in six species were classified as cell and cell part (**Table 3** and **Supplementary Figures 8–13**).

Analysis of Intronless Pentatricopeptide Repeat Gene Family in *Pyrus bretschneideri*

In *Pyrus bretschneideri*, the largest intronless gene family is the Pentatricopeptide Repeat gene family. Meanwhile, PPR gene family is also one of the largest families found in most plants, which plays a wide and crucial role in plant growth and development. We searched RIGD database by using Pfam ID of Pentatricopeptide Repeat gene family (PF01535, PF13041,

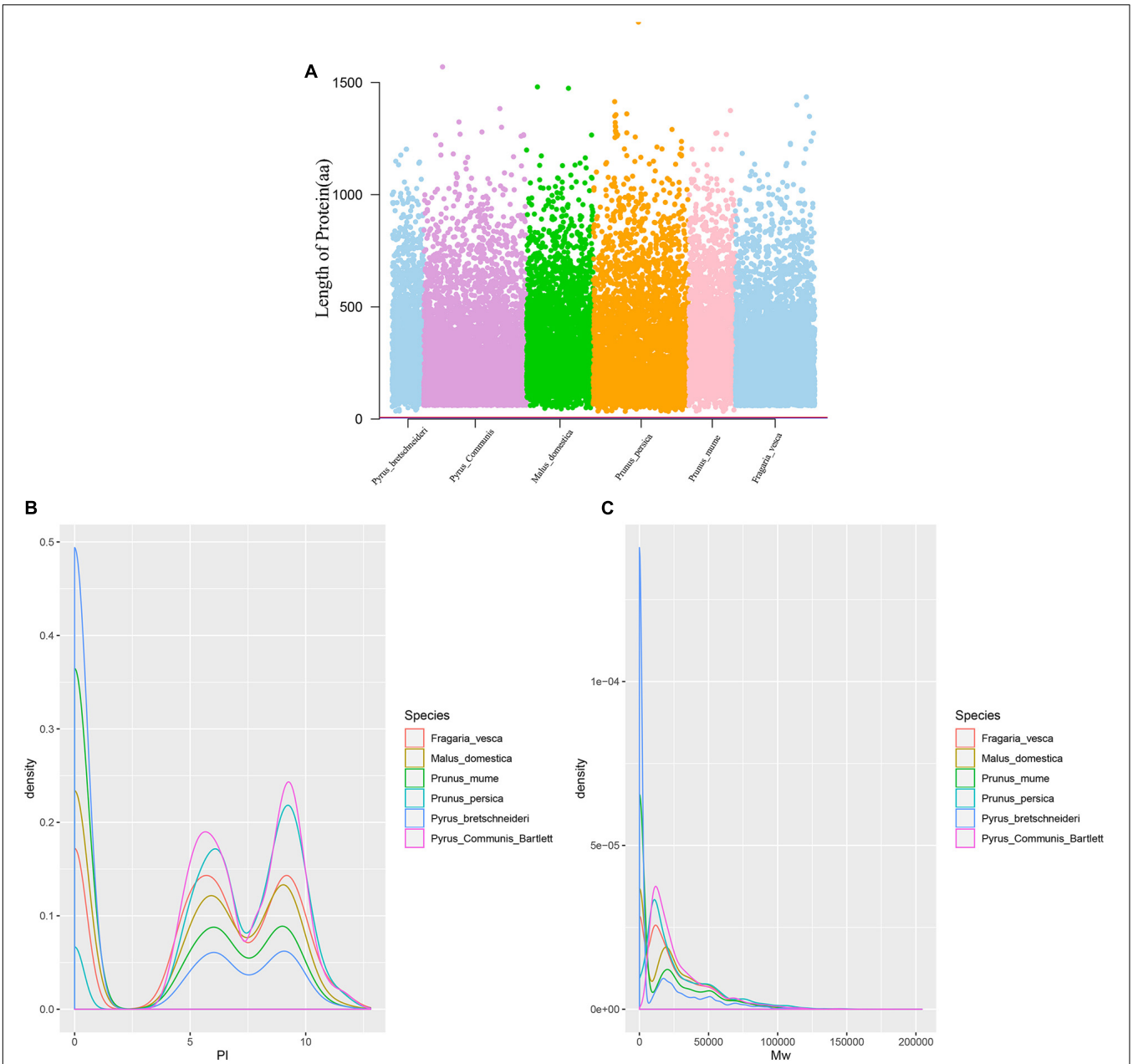


FIGURE 7 | (A) The length of protein (number of amino acids). Most proteins were less than 500 aa in length. **(B)** The distribution of pI. The graphic shows three peaks, their pI close to 0.6 and 8. **(C)** The distribution of Mw. Most predicted protein molecular weights were less than 100000 Da.

and PF13812), the predicted protein function was used to determine whether it belonged to PPR gene. The analysis results of isoelectric point, protein molecular weight and subcellular localization were obtained from RIGD by using the search interface. We downloaded the protein sequence, used the MEME SUITE (Bailey et al., 2009) and TBtools (Chen et al., 2020) to analysis the motif of intronless PPR gene in *Pyrus bretschneideri*.

We identified 120 intronless PPR genes in *Pyrus bretschneideri*. The relative molecular weight of each protein was between 11.5 and 113.7 kD. The molecular weight of gene named LOC103927494 was the smallest, while the molecular weight of LOC103947845 was far higher than that of other genes, 10 times the minimum molecular weight, and more than twice the average molecular weight of 120 amino acid sequences. In addition, the predicted results of theoretical isoelectric points were shown between 5.2 and 9.47. The isoelectric point of 46.3% members was less than 7 and belonged to acidic protein, while the other 53.7% were all basic proteins (**Supplementary Table 1**). The results of subcellular localization prediction showed that most genes were located in chloroplasts, some genes were in mitochondria and

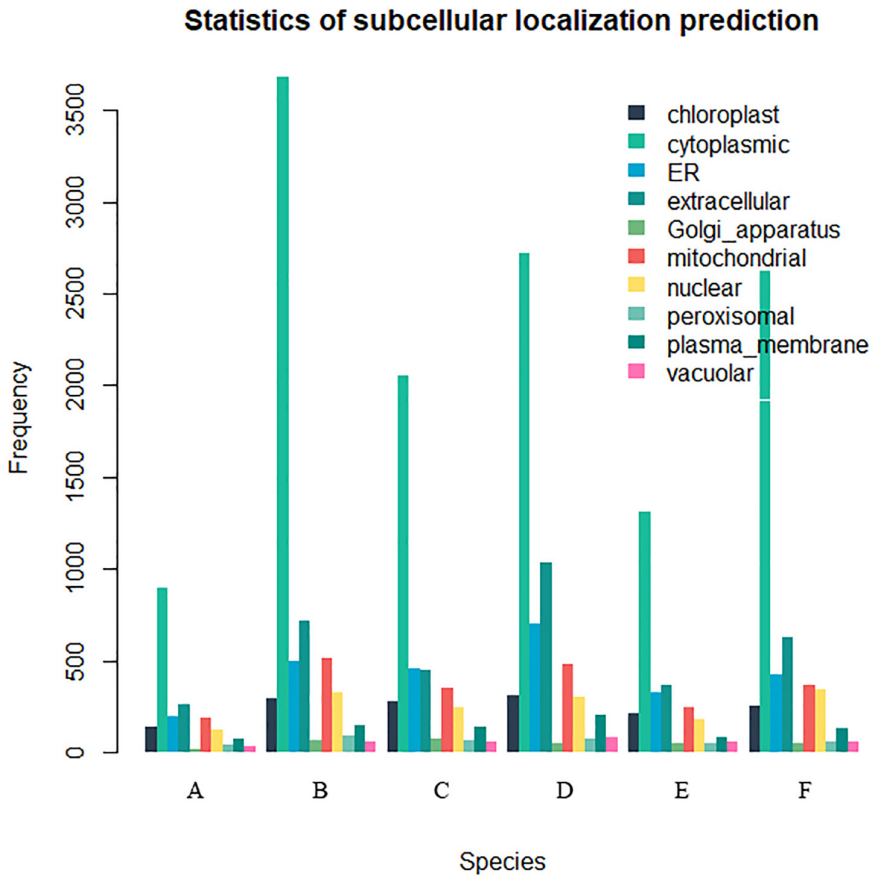


FIGURE 8 | The statistics of subcellular localization prediction. The largest number of intronless genes were categorized as cytoplasmic, extracellular also accounted for a large proportion. **(A)** *Pyrus bretschneideri*. **(B)** *Pyrus Communis*. **(C)** *Malus domestica*. **(D)** *Prunus persica*. **(E)** *Prunus mume*. **(F)** *Fragaria vesca*.

cytoplasm, a few genes were in nucleus, plastids, endoplasmic reticulum and extracellular regions. The above results showed that intronless PPR genes still has the characteristic of typical localization in semi-autonomous organelles, which was consistent with the localization of PPR protein in other plants (**Figure 10**). We identified three sequence motif: Motif1 (GIKPDVEHYGCMVDLLGRAGRLEEAEELIKEMPFK), Motif2 (IRVVKNLRVCGDCHSAIKLISKVVGREIIVRDANRFHHFKD GSCSCGDYW), and Motif3 (FVGNALIDMYAKCGSLEEARKV FDEMPERNVVSWNAMISGYAQ). Motif1 was covered in 120 intronless PPR genes, and was highly conserved. Thirty-three genes contained only Motif1 (27.5%), 58 genes contained Motif1 and Motif3 (48.3%) and 28 genes contained all three motif (23.3%). It is worth noting that Motif3 only existed at the end of amino acid sequence. In addition, LOC103956483 contained Motif1 and Motif2, which was the only one of the 120 intronless PPR genes contained only Motif1 and Motif2 (**Figure 11**).

DISCUSSION

In eukaryotes, there are intronless genes because there is no special structure of introns in genes, so studying the functions and evolutionary characteristics of these genes can help us to understand the evolution rules of related genes and genomes. Meanwhile, the exploration of intronless genes can help researchers to explore the effects of introns and selective splicing mechanisms on eukaryotes from the perspective of reverse thinking.

Because of the importance of intronless genes in comparative genomics and evolutionary biology, research on intronless genes in eukaryotes has been the focus of researchers for a long time. It is necessary to establish a centralized data platform for the integration, comparison, and analysis of the function and evolution of intronless genes on a larger scale. Little work has been done, as only a few databases exist, while Genome SEGE and IGDD have stopped providing services. IGD was limited to human intronless genes, which were annotated in different databases. PIGD focused on the intronless genes of Poaceae species and conducted a systematic comparative analysis from the perspective of comparative genomics, but the database has been damaged for providing retrieval services. As a result, users can only download the original data of the intronless gene sequences and the results of the analysis.

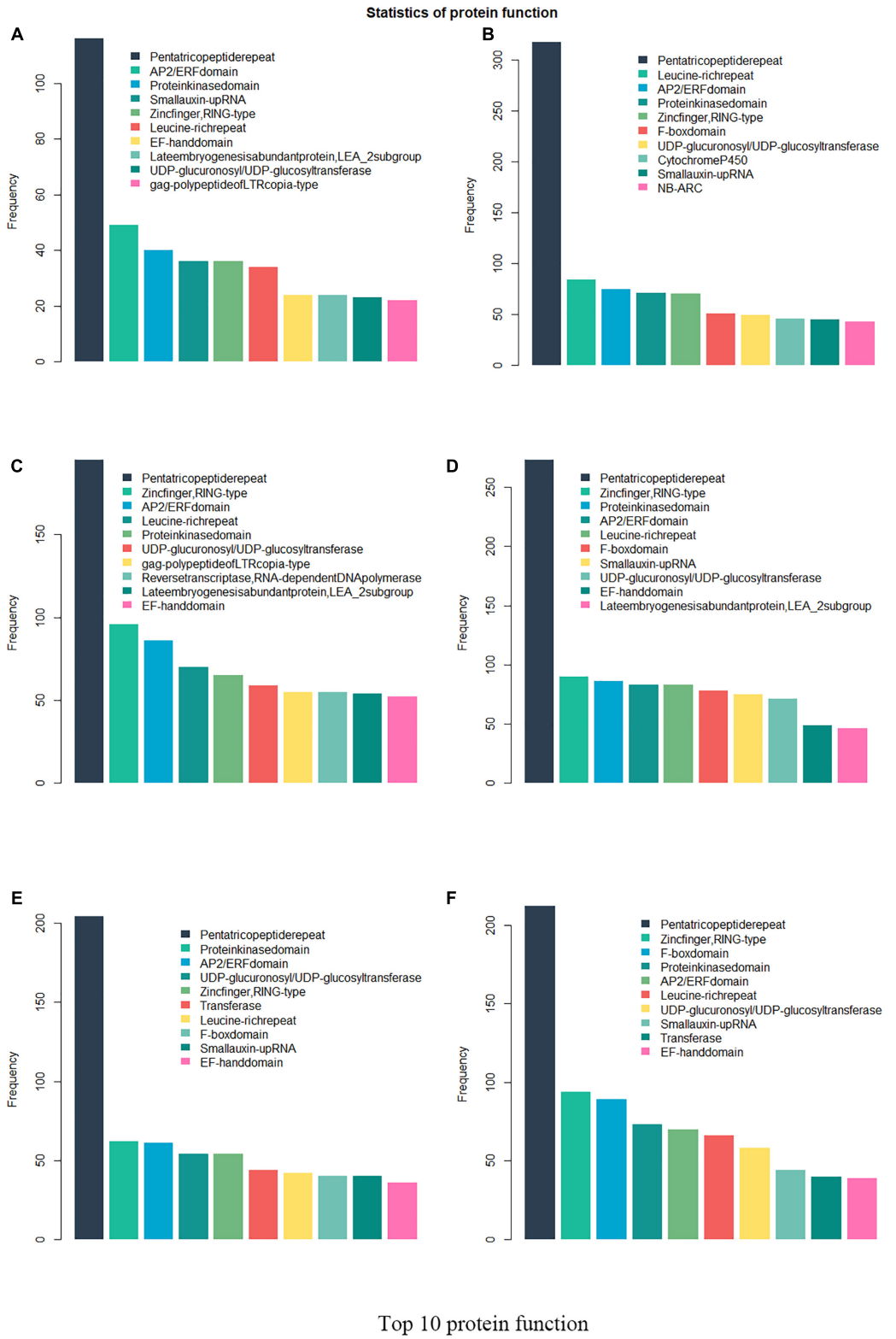


FIGURE 9 | Top 10 largest number of intronless genes in protein function. The largest number of intronless genes were predicted for pentatricopeptide repeat. AP2/ERF domain in *Pyrus bretschneideri*, Leucine-rich repeat in *Pyrus communis*, Zinc finger (RING-type) in *Malus domestica*, *Prunus persica* and *Fragaria vesca*, and protein kinase domain in *Prunus mume* accounted for a large proportion. **(A)** *Pyrus bretschneideri*. **(B)** *Pyrus communis*. **(C)** *Malus domestica*. **(D)** *Prunus persica*. **(E)** *Prunus mume*. **(F)** *Fragaria vesca*.

TABLE 3 | The number of intronless genes in GO categories.

Species	Annotated Genes	GO Terms			
		Biological	Cellular	Function	Total
Pyrus bretschneideri	780	666	630	539	1835
Pyrus communis	2550	2249	2140	1907	6296
Malus domestica	1720	1474	1385	1219	4078
Prunus persica	1805	1548	1479	1272	1299
Prunus mume	1151	983	941	840	2764
Fragaria vesca	1638	1429	1348	1203	3980

The RIGD, as the latest intronless gene database, integrates the intronless gene data of six species of Rosaceae and provides a systematic comparative analysis. The RIGD was designed as a simple, easy-to-use, and esthetically pleasing website interface that provides a feature-rich, user-friendly integrated data and analytics tool. The Species interface provides a download of the original data classified by chromosome number and analysis methods. The Statistics interface presents the results of systematic comparative genomics analysis of six species in the form of graphs. The Search interface allows users to search for data on intronless genes of interest. In addition, NCBI Blast, a common bioinformatics tool, is embedded in the RIGD to help researchers annotate new sequences and predict homology with genes in the RIGD. The RIGD also provides multiple interactive platforms, including Up&Down, Contact us and Links. Through these platforms, users can learn about the RIGD's analytical methods, download data of interest, and upload their important scientific findings to facilitate communication and data sharing among researchers in the same research field.

The RIGD is built on a Tencent cloud server with stable service and convenience for long-term maintenance and updating. In the future, we hope to update and expand the RIGD by communicating with researchers. The number of species collected is expected to increase, and more detailed annotation information on intronless genes, such as spatio-temporal expression data of intronless genes in different growth stages and tissues of plants, homologous genes in the genome, metabolic pathways of genes, and more, are expected to be added. This information will allow researchers to further explore the function and evolutionary mechanisms of intronless genes. Moreover, we are also committed to developing powerful comparative analysis tools to make the RIGD a centralized platform for intronless gene information and analysis, enabling researchers to use the database for data mining and analysis in various aspects.

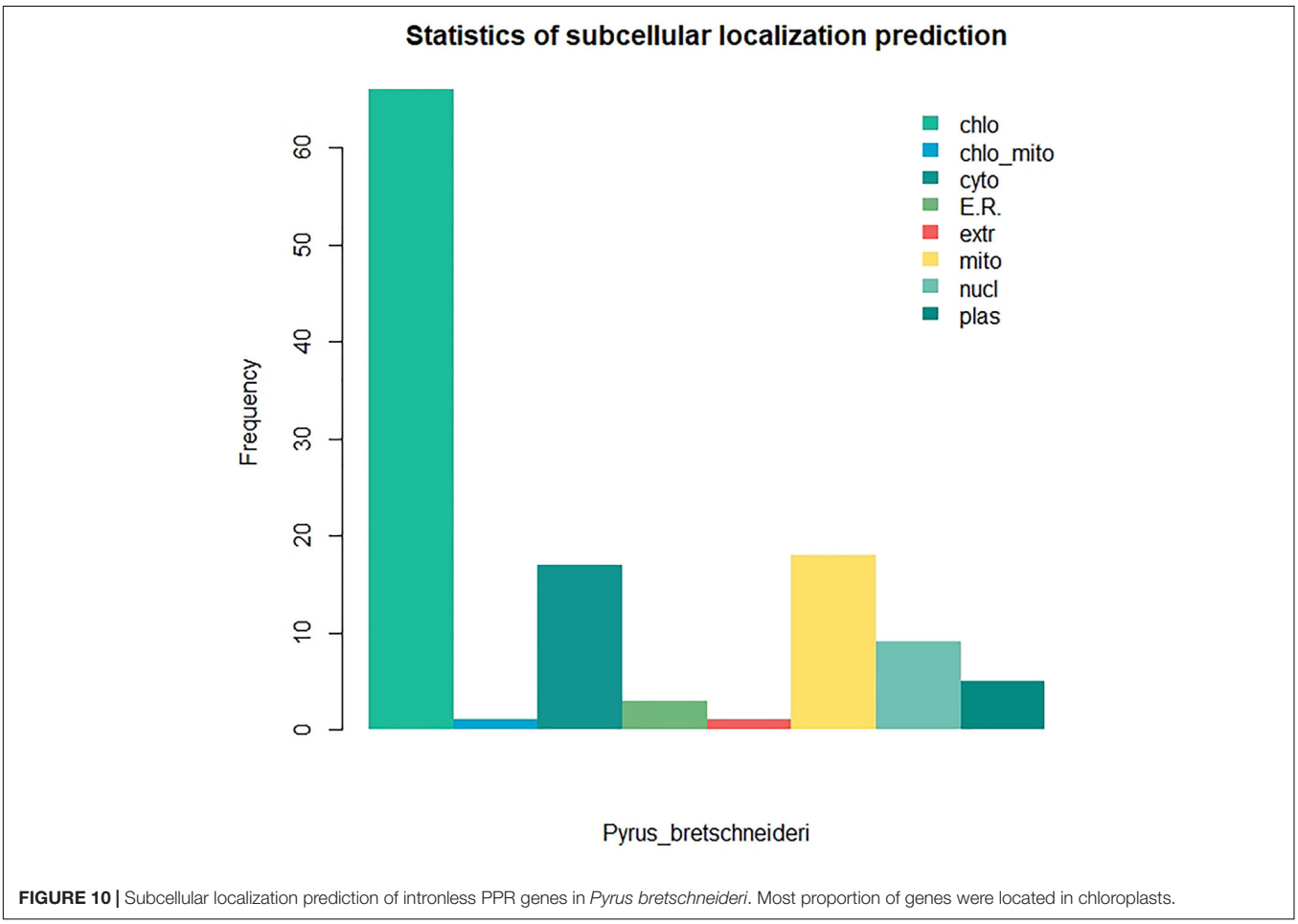


FIGURE 10 | Subcellular localization prediction of intronless PPR genes in *Pyrus bretschneideri*. Most proportion of genes were located in chloroplasts.

FIGURE 11 | Motif analysis and visualization. Motif1 is the most covered. Motif3 only existed at the end of amino acid sequence. LOC103956483 was the only one of the 120 intronless PPR genes contained only Motif1 and Motif2.

CONCLUSION

With the development of sequencing technology, an increasing number of plant genomes are sequenced and annotated, and there will be increasing data regarding intronless genes in the future. It is feasible to integrate, compare and analyze the function and evolution of intronless genes in a wide range. We developed the RIGD platform, collected and systematically analyzed the data from intronless genes in six species of Rosaceae, and provided a series of tools for users to search the data of intronless genes of interest and communicate with us. With the support of researchers, we eventually hope to develop a platform for integrating data from eukaryotic intronless genes with tools for comparative genomics analysis, which can greatly promote the research of intronless genes in plants, thus mining valuable genomic resources and helping researchers find more interesting discoveries.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: ftp://ftp.bioinfo.wsu.edu/www.rosaceae.org/Pyrus_x_bretschneideri/Pbretschneideri-genome.v1.1, ftp://ftp.bioinfo.wsu.edu/species/Pyrus_communis/Pcommunis_DH_genome.v2.0, ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/114/115/GCF_002114115.1_ASM211411v1, ftp://ftp.bioinfo.wsu.edu/species/Prunus_persica/Prunus_persica-genome.v2.0.a1, ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/346/735/GCF_000346735.1_P.mume_V1.0, and ftp://ftp.bioinfo.wsu.edu/species/Fragaria_vesca/Fvesca-genome.v4.0.a1.

AUTHOR CONTRIBUTIONS

TC projected the study, constructed the platform, wrote programs in the website background and analysis, involved in the bioinformatics analysis, and drew up the manuscript. DM put into effect the mainly bioinformatics analysis, handled figures and tables, participated in the design of platform, and update of the database. XL, XC, HW, QJ, and XX collected and collated the data, helped with the design and update of the database, provided suggestions, and criticisms for improving the manuscript and website. YCo and YCi participated in the design, helped in writing the manuscript,

and supervised the whole project. All authors read and accepted the final manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (Grant No. 31640068), China Postdoctoral Science Foundation (Grant No. 2019M662135), Anhui Provincial Natural Science Foundation (Grant No. 2008085QC100), Scientific Research Foundation of Anhui Agricultural University (Grant Nos. 2019zd01 and yj2019-17), Anhui Provincial Postdoctoral Science Foundation (Grant No. 2019B319), and Graduate Student Innovation Foundation of Anhui Agricultural University (Grant No. 2020ysj-61). The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

ACKNOWLEDGMENTS

We thank Yingxu Fan from the MOE Key Laboratory of Bioinformatics in Tsinghua University, Chao Chen from the School of Life Sciences and Technology in ShanghaiTech University, and Zhichao Yu and Bingliang Fan from the College of Informatics in Huazhong Agricultural University for valuable suggestions in bioinformatics analysis. We thank Mengtian Gu from the Institute of Information Engineering, CAS for the help of platform construction. We also thank Shuoyu Fang from the School of Foreign Languages in Shanghai Jiao Tong University for the advice of the manuscript writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00868/full#supplementary-material

REFERENCES

Agarwal, S. M., and Gupta, J. (2005). Comparative analysis of human intronless proteins. *Biochem. Biophys. Res. Commun.* 331, 512–519. doi: 10.1016/j.bbrc.2005.03.209

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 40, W597–W603. doi: 10.1093/nar/gks400

Aubourg, S., Kreis, M., and Lecharny, A. (1999). The DEAD box RNA helicase family in *Arabidopsis thaliana*. *Nucleic Acids Res.* 27, 628–636. doi: 10.1093/nar/27.2.628

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Blum, T., Briesemeister, S., and Kohlbacher, O. (2009). MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinform.* 10:274. doi: 10.1186/1471-2105-10-274

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools - an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 2052, 30187–30188. doi: 10.1016/j.molp.2020.06.009

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Gentles, A. J., and Karlin, S. (1999). Why are human G-protein-coupled receptors predominantly intronless? *Trends Genet.* 15, 47–49. doi: 10.1016/s0168-9525(98)01648-5

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., et al. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi: 10.1093/molbev/msx148

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Jain, M., Khurana, P., Tyagi, A. K., and Khurana, J. P. (2008). Genome-wide analysis of intronless genes in rice and *Arabidopsis*. *Funct. Integr. Genomics* 8, 69–78. doi: 10.1007/s10142-007-0052-9

Jain, M., Tyagi, A. K., and Khurana, J. P. (2006). Genome-wide analysis, evolutionary expansion, and expression of early auxin-responsive SAUR gene family in rice (*Oryza sativa*). *Genomics* 88, 360–371. doi: 10.1016/j.ygeno.2006.04.008

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Jung, S., Lee, T., Cheng, C. H., Buble, K., Zheng, P., Yu, J., et al. (2019). 15 years of GDR: new data and functionality in the genome database for rosaceae. *Nucleic Acids Res.* 47, D1137–D1145. doi: 10.1093/nar/gky1000

Kulmanov, M., and Hoehndorf, R. (2019). DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* 36, 422–429. doi: 10.1093/bioinformatics/btz595

Lecharny, A., Boudet, N., Gy, I., Aubourg, S., and Kreis, M. (2003). Introns in, introns out in plant gene families: a genomic approach of the dynamics of gene structure. *J. Struct. Funct. Genomics* 3, 111–116. doi: 10.1007/978-94-010-0263-9_11

Louhichi, A., Fourati, A., and Rebai, A. (2011). IGD: a resource for intronless genes in the human genome. *Gene* 488, 35–40. doi: 10.1016/j.gene.2011.08.013

Lurin, C., Andres, C., Aubourg, S., Bellaoui, M., Bitton, F., Bruyere, C., et al. (2004). Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* 16, 2089–2103. doi: 10.1105/tpc.104.022236

Rogozin, I. B., Sverdlov, A. V., Babenko, V. N., and Koonin, E. V. (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. *Brief Bioinform.* 6, 118–134. doi: 10.1093/bib/6.2.118

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5, 725–738. doi: 10.1038/nprot.2010.5

Sakharkar, K. R., Sakharkar, M. K., Culiat, C. T., Chow, V. T., and Pervaiz, S. (2006). Functional and evolutionary analyses on expressed intronless genes in the mouse genome. *FEBS Lett.* 580, 1472–1478. doi: 10.1016/j.febslet.2006.01.070

Sakharkar, M. K., Chow, V. T., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. *Silico. Biol.* 4, 387–393.

Sakharkar, M. K., and Kangueane, P. (2004). Genome SEGE: a database for 'intronless' genes in eukaryotic genomes. *BMC Bioinformatics* 5:67. doi: 10.1186/1471-2105-5-67

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., et al. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.* 29, 263–264. doi: 10.1038/ng754

Takeda, S., Kadowaki, S., Haga, T., Takaesu, H., and Mitaku, S. (2002). Identification of G protein-coupled receptor genes from the human genome sequence. *FEBS Lett.* 520, 97–101. doi: 10.1016/s0014-5793(02)02775-8

Tine, M., Kuhl, H., Beck, A., Bargelloni, L., and Reinhardt, R. (2011). Comparative analysis of intronless genes in teleost fish genomes: insights into their evolution and molecular function. *Mar. Genom.* 4, 109–119. doi: 10.1016/j.margen.2011.03.004

Yan, H., Dai, X., Feng, K., Ma, Q., and Yin, T. (2016). IGDD: a database of intronless genes in dicots. *BMC Bioinform.* 17:289. doi: 10.1186/s12859-016-1148-9

Yan, H., Jiang, C., Li, X., Sheng, L., Dong, Q., Peng, X., et al. (2014). PIGD: a database for intronless genes in the Poaceae. *BMC Genomics* 15:832. doi: 10.1186/1471-2164-15-832

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* 12, 7–8. doi: 10.1038/nmeth.3213

Yang, J., and Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* 43, W174–W181. doi: 10.1093/nar/gkv342

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34, W293–W297. doi: 10.1093/nar/gkl031

Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., et al. (2018). WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Res.* 46, W71–W75. doi: 10.1093/nar/gky400

Zou, M., Guo, B., and He, S. (2011). The roles and evolutionary patterns of intronless genes in deuterostomes. *Comp. Funct. Genom.* 2011:680673. doi: 10.1155/2011/680673

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 14 August 2020
doi: 10.3389/fbioe.2020.00867

Explainable Prediction of Medical Codes With Knowledge Graphs

Fei Teng[1]*, *Wei Yang*[1], *Li Chen*[2], *LuFei Huang*[1,2] and *Qiang Xu*[3]

[1] School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China, [2] The Third People's Hospital of Chengdu, Chengdu, China, [3] School of Information Engineering, Chengdu University of Traditional Chinese Medicine, Chengdu, China

International Classification of Diseases (ICD) is an authoritative health care classification system of different diseases. It is widely used for disease and health records, assisted medical reimbursement decisions, and collecting morbidity and mortality statistics. The most existing ICD coding models only translate the simple diagnosis descriptions into ICD codes. And it obscures the reasons and details behind specific diagnoses. Besides, the label (code) distribution is uneven. And there is a dependency between labels. Based on the above considerations, the knowledge graph and attention mechanism were expanded into medical code prediction to improve interpretability. In this study, a new method called G_Coder was presented, which mainly consists of Multi-CNN, graph presentation, attentional matching, and adversarial learning. The medical knowledge graph was constructed by extracting entities related to ICD-9 from freebase. Ontology contains 5 entity classes, which are disease, symptom, medicine, surgery, and examination. The result of G_Coder on the MIMIC-III dataset showed that the micro-F1 score is 69.2% surpassing the state of art. The following conclusions can be obtained through the experiment: G_Coder integrates information across medical records using Multi-CNN and embeds knowledge into ICD codes. Adversarial learning is used to generate the adversarial samples to reconcile the writing styles of doctor. With the knowledge graph and attention mechanism, most relevant segments of medical codes can be explained. This suggests that the knowledge graph significantly improves the precision of code prediction and reduces the working pressure of the human coders.

Keywords: automated ICD coding, knowledge graphs, explainable, medical records, natural language processing

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Sijia Liu,
Mayo Clinic, United States
Erick Antezana,
Norwegian University of Science and Technology, Norway

***Correspondence:**
Fei Teng
fteng@swjtu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *19 April 2020*
Accepted: *06 July 2020*
Published: *14 August 2020*

Citation:
Teng F, Yang W, Chen L, Huang L and Xu Q (2020) Explainable Prediction of Medical Codes With Knowledge Graphs. Front. Bioeng. Biotechnol. 8:867. doi: 10.3389/fbioe.2020.00867

INTRODUCTION

The International Classification of Diseases (ICD) is a standard classification system according to the characteristics of diseases and the rules maintained by the World Health Organization. Each code represents a specific disease, symptom, or surgery. And a set of codes in the medical record represents uniquely diagnostic and procedural information during patient visits. As a significant part of the hospital information system, it is widely used for medical insurance payments, health reports, and mortality calculations. Therefore, the ICD coding task is an essential job in the medical record information department. While ICD codes are important for making clinical and financial decisions, ICD coding is time-consuming, error-prone, and expensive. In most cases, the human coders assign ICD codes to medical records according to the clinical diagnosis record of physician. It is difficult because the code assignment should consider overall the health condition in the long text-free medical records, including symptoms, signs, surgery, medication, body, etc.

Automatic coding uses medical records as input to predict the final ICD codes based on text content. But the automatic ICD coding task usually has the following difficulties: (1) The clinical records of patients are not always structured in the same way. And the vital information in the text is distributed in various segments. For the above two reasons, it is very difficult to extract important and relevant knowledge from various kinds of medical records effectively. (2) Most importantly, the medical field has a lot of terminologies, which is difficult for non-professionals to understand the meaning of these terminologies. Even for the same disease, there are many ways to describe it differently from ICD description. (3) Datasets in the medical field are often small, and doctors have different writing styles. Each physician usually has his way to describe medical terminologies.

In this paper, we proposed a new end-to-end method called G_Coder (Graph-based Coder) for automatic ICD code assignment using clinical records. The contributions of this paper are summarized as follows: (1) We utilize Multi-CNN (multiple convolutional neural networks) to capture local correlation, which extracts key features from the irregular text. (2) We build a knowledge graph, which enriches the meaning of terminologies through integrated related knowledge points. It is combined with the attention mechanism to help understand the meaning of related terminologies, making the coding results interpretable. (3) The adversarial learning is used to generate adversarial samples to increase samples and reconcile the different writing styles.

Our model has outperformed other models in micro-AUC and micro-F1 on MIMIC-III (Multi-parameter Intelligent Monitoring in Intensive Care) datasets with 46 K distinct hospital admissions and top 50 common ICD-9 codes.

RELATED WORKS

Automatic ICD Coding

It was 20 years ago that many researchers have explored how to automatically assign ICD codes based on clinical records. There are two major categories of approaches for automatically assigning ICD-9 codes using medical records. One category is rule-based and the other category is learning-based. Rule-based systems are manually extracted statistical features by humans. Chen et al. (2017) and Ning et al. (2016) presented an improved approach based on the Longest Common Subsequence (LCS) and semantic similarity for performing ICD-10 code assignment to Chinese diagnoses. But such approaches only consider the simple matching of strings, which is not a medical problem. Beyond that, researchers applied automatic and semi-automatic (Medori and Fairon, 2010) machine learning methods to automatically assign ICD codes. Automatic ICD-9-CM encoding consisted of support vector machines (SVM) (Yan et al., 2010; Adler et al., 2011; Ferrão et al., 2013; Wang et al., 2017), k-nearest neighbors (Ruch et al., 2008; Erraguntla et al., 2012), Naive Bayes (Pakhomov et al., 2006; Medori and Fairon, 2010), and other methods such as topic model (Ping et al., 2010; Perotte et al., 2013). Semi-automatic methods generally require more manual participation and may require manual data processing, feature selection, data verification, etc. Automatic methods generally use a series of operations in an end-to-end manner. Nevertheless, the development of automatic coding technology is not yet mature, and manual verification is inevitable. All the above methods only utilize the statistical characteristics of words and ignore the contextual meaning.

In recent years, many new methods are emerging with the development of deep neural network. Li et al. (2018) combined the convolutional neural network (CNN) and the "Document to Vector" technique to extract textual features. It solves the characteristics of CNN's indistinguishable word order while taking all the words into account. Baumel et al. (2017) applied a hierarchical approach which is Hierarchical Attention bidirectional Gated Recurrent Unit (HA-GRU) to tag a discharge summary by identifying the relevant sentences. It utilizes the Gated Recurrent Unit to encode text, which experimental effect is similar to long short-term memory networks (LSTM), but it is easier to calculate. Yu Y. et al. (2019) explored character features and word features based on bidirectional LSTM with attention mechanism and Xie and Xing (2018) applied tree LSTM with ICD hierarchy information for automatic ICD coding. Compared with ordinary LSTM, bidirectional LSTMs tend to have higher accuracy, and tree LSTM is more suitable for data that is a tree-like hierarchical structure. Mullenbach et al. (2018) proposed to extract per-code textual features across the document using a convolutional neural network and used an attention mechanism to select the most relevant segments for each possible code. Based on that, Li and Yu (2019) combined multi-filter convolutional layers and residual convolutional layers to enlarge the receptive field.

Deep learning methods improved the ability to capture semantic information but ignored the importance of medical knowledge and experience. In practical work, the human coders fully utilize the basic medical knowledge to provide decision support for the work. However, all the methods just mentioned are data-driven approaches or simple mapping, which lack of the theoretical support and suffer from the complicated preprocessing of the noisy text. To build a more explainable ICD coding system, we utilize the knowledge graph as supplementary knowledge to add to the model, which is equivalent to combining a data-driven approach with medical knowledge. What is more, we successively perform text preprocessing and Multi-CNN algorithm to extract text features to reduce text noise. Adversarial learning generates adversarial samples for training to reconcile the different writing styles. The attention mechanism selects the most relevant segments for each possible code.

Graph Embedding

Graph embedding technology expresses nodes in the form of low-dimensional dense vectors, which require similar nodes in the original graph to be similar in the low-dimensional expression space. The representative work of Graph Embedding is DeepWalk (Perozzi et al., 2014), LINE (Tang et al., 2015), Node2Vec (Grover and Leskovec, 2016), SDNE (Wang et al., 2016), and Struc2Vec (Ribeiro et al., 2017). The obtained expression vectors can be used for downstream tasks, such as node classification (Ye et al., 2018; Gong and Ai, 2019), link

prediction (Li et al., 2019a), or visualization (Liu et al., 2020). In the field of biomedicine, graphs are often used to predict drug interactions and predict drug target proteins. The knowledge graph embedding is used to calculate several similarity measures between all drugs in the scalable and distributed framework to obtain the interaction of drugs (Ibrahim et al., 2017). Mohamed et al. (2020) used knowledge graph embeddings to learn the vector representation of all drugs and targets to discover protein drug targets.

Attention Mechanism

The attention mechanism was first used for machine translation (Dzmitry et al., 2014). It calculates the attention weight of each word in the encoder sequence to each word in the decoder sequence to focus more on the most relevant part of the current word. The attention mechanism improves the effect and also increases the interpretability of the neural network. After adding attention, the weight of the data can be visualized to confirm the correctness of the method. Besides, attention mechanism has the ability to capture global features in long texts. The attention mechanism mimics the internal process of biological observation behavior, which is a mechanism that aligns internal experience and external sensation to increase the observation precision of some areas. It has been successfully used in medical tasks. Such as medical imaging (Ozan et al., 2018), clinical text information extraction (Li et al., 2019b; Xu et al., 2019), and DNA-related tasks (Yu W. et al., 2019; Hong et al., 2020).

Adversarial Learning

Adversarial learning is to make the two networks compete against each other. The generator network continuously captures the probability distribution of the real data in the training set and transforms the input random perturbation into new samples. The discriminator network observes both real and fake data to determine the authenticity of this data. Through repeated confrontation, the capabilities of the generator and discriminator will continue to increase until a balance is reached. Goodfellow et al. (2015) developed a method named FSGM that can effectively calculate the perturbation. They set the perturbation to the maximum value of the loss function along the direction of the gradient. FSGM takes the same step in each direction, and Goodfellow's subsequent FGM (Miyato et al., 2017) is scaled according to specific gradients to obtain better adversarial samples. Adversarial learning improves the robustness of the model through the idea of games. It randomly adds perturbation factors to the input to simulate unknown data to ensure that the model can work stably in any situation. Adversarial learning has been used for privacy protection (Max et al., 2019) of medical records and named entity recognition (Zhao et al., 2019) in clinical texts.

MATERIALS AND METHODS

As can be seen from **Figure 1**, this section will detail all the processes by combining data materials with the proposed methods.

Dataset and Preprocessing

We utilize the transfer knowledge graph to improve the interpretability and performance of automatic ICD coding. In the study, we select Multi-parameter Intelligent Monitoring in Intensive Care-III (MIMIC-III) dataset (Johnson et al., 2016) as an experimental dataset and Freebase dataset as a source of the knowledge graph. A brief introduction to these two data sets and related preprocessing techniques are as follows.

MIMIC-III Dataset

MIMIC-III dataset is the only public database for learning automated ICD-9 coding, which allows fair comparisons with different methods. It contains reliable and comprehensive 58,976 hospital admissions collected between 2001 and 2012 in the Beth Israel Deaconess Medical Center. Each medical record usually includes discharge summaries, survival data, diagnostic codes, vital signs, laboratory measurements, etc. Besides, the discharge summary always contains multiple information, such as "discharge diagnosis," "past medical history," physical examination," and "chief complaint," etc. **Table 1** shows a sample of a medical record in the dataset. The "HADMID" uniquely identifies each medical record. Each hospital admission has a group of ICD-9 codes given by the medical coders. For each medical record, codes distribute unevenly in numbers which varies from one to 39. The number of codes is usually not equal to the number of diagnosis descriptions. It invalidates the one-to-one method of allocating codes. The entire dataset contains 6,984 distinct codes and 943 categories. Each code has a short phrase or a sentence, articulating a disease, symptom, or condition.

We adopt a series of standard text pre-processing techniques, which contain regular expression matching and tokenization to reduce the noise in raw note texts. Firstly, we extract relevant data from MIMIC-III as input text, which contains "physical examination," "chief complaint," "final diagnosis," "history," "medication," "course," and "procedure." Secondly, we remove stop words from the input text and transform each token into its lowercase. Simultaneously removing words <3 and replacing unknown words with "UNK." Thirdly, medical records with associated labels that do not contain the top 50 code are discarded.

Freebase

With the rapid development of the knowledge graph in recent years, research-based on knowledge graphs has attracted widespread attention in the medical field. Freebase mainly extracts structured data from wikis and publishes them as RDF. It is fully structured, but the data source is not limited to wikis. It also imports a large number of professional data sets and provides data query and entry mechanisms.

We fuse ICD-9 description information with medical knowledge extracted from freebase to build the final knowledge graph. Freebase Medicine originate from Wikipedia and other datasets such as U.S. National Medical Data. One study has reported that 70% of junior doctors used Wikipedia for health knowledge every week (Trevena, 2011). Because the freebase is reliable, the information provided in Freebase is generally considered to be reliable. The matching method is used for

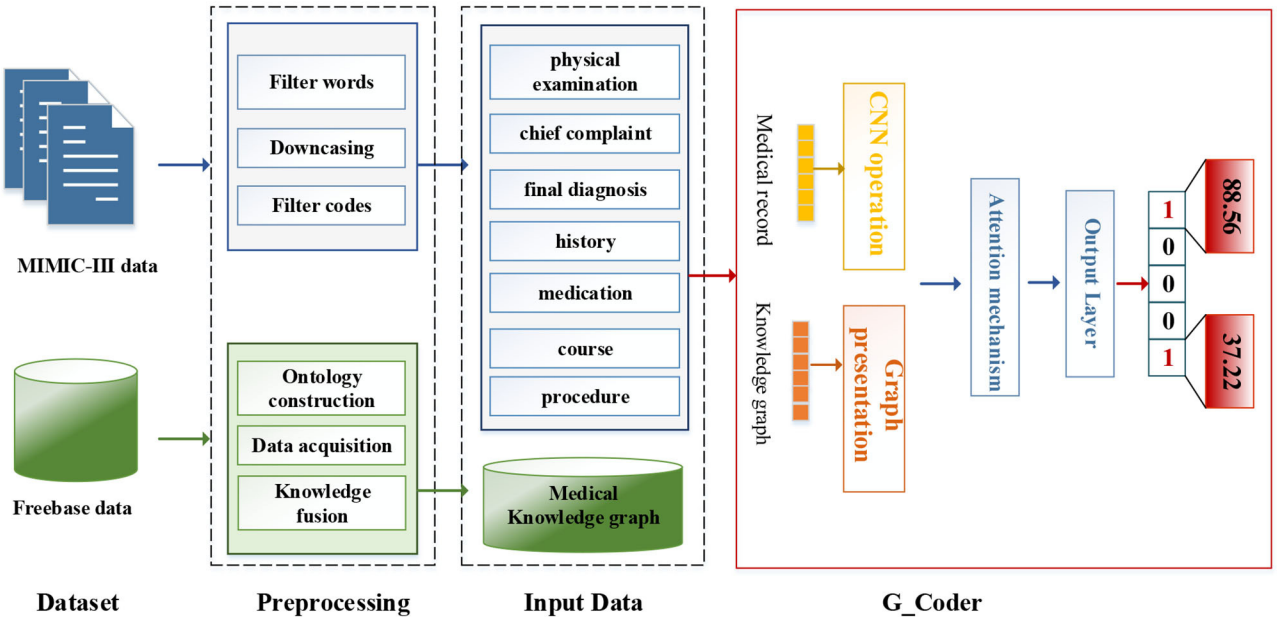


FIGURE 1 | The overall process of this method.

TABLE 1 | An example of a medical record.

Medical record (partially shown)
HADMID:105501
Admission Date: [**2172-7-6**] Discharge Date: [**2172-7-10**] Date of Birth: [**2096-4-25**] Sex: M Service: Cardiothoracic Surgery Service HISTORY OF PRESENT ILLNESS: The patient is a 75-year-old gentleman who is a patient of Dr. [**First Name4 (NamePattern1) **] [**Last Name (NamePattern1) 47696**] who was transferred in from [**Hospital3 3583**] status post a myocardial infarction for cardiac catheterization…… PAST MEDICAL HISTORY: 1. Hypertension. 2. Myocardial infarction. 3. Hypercholesterolemia. 4. Myocardial infarction in [**2158**]. ……
ICD-9 codes and description
88.56 Coronary arteriography using two catheters 39.61 Extracorporeal circulation auxiliary to open heart surgery 88.72 Diagnostic ultrasound of heart 36.15 Single internal mammary-coronary artery bypass 584.9 Acute renal failure, unspecified 37.22 Left heart cardiac catheterization 410.71 Acute myocardial infarction, subendocardial infarction, initial episode of care 414.01 Coronary atherosclerosis of native coronary artery 428.0 Congestive heart failure, unspecified 39.95 Hemodialysis

knowledge fusion. Since some diagnosis terms from ICD-9 description imperfectly match Freebase content, we use the ICD description text as the search terms to find the most relevant Freebase content by the Freebase API (http://freebase.gstore-pku.com/). The ontology that was constructed contains 5 entity classes, which are disease, symptom, medicine, surgery, and examination. The constructed ontology is shown in **Figure 2**, which contains the relationships (disease manifests as symptoms, medicine treats disease, surgery treats disease, and commonly used disease test data, etc.) and attribute types, such as id, name, ICD, etc. In the final knowledge graph, there are 1,560 nodes and more than 20,000 sets of relationships.

Methods

Overview

The modular method adopted in this study differs from the researchers used earlier. **Figure 3** shows an overview of our approach named G_Coder. The proposed approach mainly consists of four modules, which mainly contain Multi-CNN, Graph Presentation, Attentional Matching, and Adversarial Learning.

Input Layer

Considering that the pre-trained word vectors in the medical field are not yet perfect and the experimental data in this study are very long texts, the word embeddings were initialized randomly. Leveraging a token sequence $x = \{x_1, x_2, x_3 \ldots, x_n\}$ as input, where n denotes the sequence length. Assuming that the matrix W denotes the word embedding matrix, and $W = \{w_1, w_2, w_3, \ldots, w_v\} \in \mathbb{R}^{v \times d}$, where v represents the size of total vocabulary and d represents the token dimension. The vocabulary is obtained by pre-processing the MIMIC-III clinical text. A token x_i will correspond to a vector w_j by looking up W. The final input of the model is a matrix $X \in \mathbb{R}^{n \times d}$.

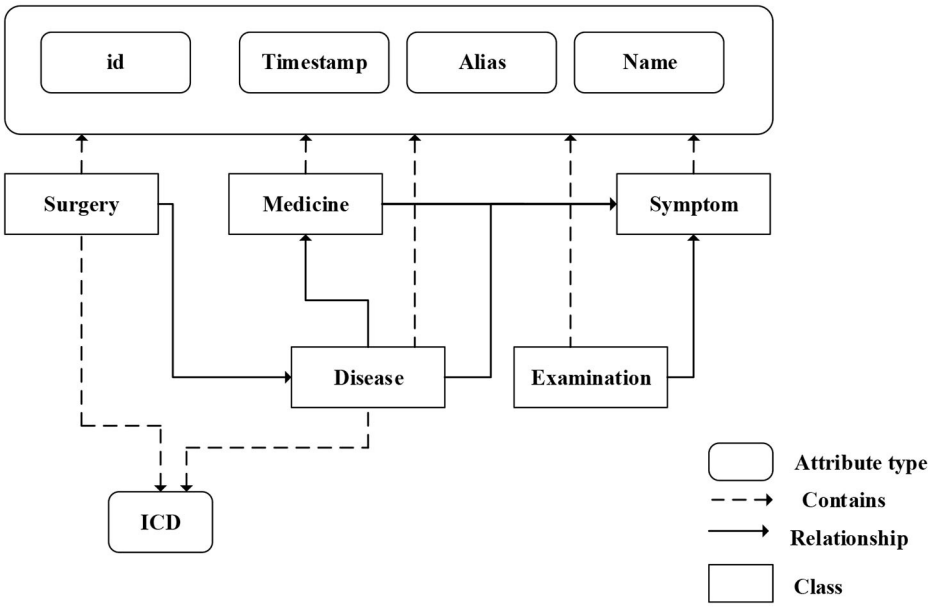


FIGURE 2 | The ontology of medical knowledge graph.

Multi-CNN

As can be seen from **Figure 3**, the structure of Multi-CNN is used to encode the input matrix X. Multi-CNN is a combination of multiple CNNs and MaxPooling. CNN is a kind of neural network algorithm that has successfully been applied to computer vision. MaxPooling reduces the dimension of the feature map, and effectively reduces the parameters required for subsequent layers. Besides, it magnifies the receptive field.

Multiple kernels of different sizes are used to extract key information in the sentence, which inspired by Kim (2014) who applied Text-CNN to the text classification task. Multi-CNN is used to better capture the local correlations. Assuming we have filters f_1, f_2, …, f_m where m denotes the filter number. Each kernel size of filters denotes as k_1, k_2, ..., k_m. The convolutional procedure can be formalized as formula (1),

$$\begin{aligned} H_1 &= g(W_{c1} * x_{i:i+k-1} + b_{c1}) \\ H_m &= g(W_{cm} * x_{i:i+k-1} + b_{cm}) \end{aligned} \tag{1}$$

where $*$ denotes the convolution operator, g is an element-wise non-linear transformation, W_{cm} is weight parameter and b_{cm} is the bias. Assuming that $H_m = \{h_1, h_2, h_3, \ldots, h_{n-k+1}\}$ is the output of m-th CNN and Hm' is the output of m-th MaxPooling. The result of Multi-CNN is $H' = [H1' \oplus H2' \oplus \ldots \oplus Hm'] \in \mathbb{R}^{\sum_1^m d_t}$, where $\oplus$ denotes concatenation operator and d_t denotes the dimension of Ht'.

Graph Presentation

In this study, we mainly adopt SDNE (Structural Deep Network Embedding) for medical knowledge graph node embedding. First-order proximity and second-order proximity are two crucial definitions in SDNE. The first-order proximity is used to describe the local similarity between paired nodes in the graph. If there are no directly connected edges, the first-order proximity is 0. The second-order proximity measures the similarity of their neighbor sets between two nodes. The optimization goal of SDNE is shown in formulas (2–4):

$$L_{1st} = \sum_{i,j=1}^{n_d} s_{i,j} ||r_i - r_j||_2^2 \tag{2}$$

Each s_i contains the neighbor structure information of the i-th node. The letter r denotes the vector representation of each node. Where n_d denotes the number of neighbors at nodes i.

$$L_{2st} = \sum_{i=1}^{n_d} ||\hat{s_i} - s_i||_2^2 \tag{3}$$

$$L = L_{1st} + \alpha L_{2st} + \beta L_{reg} \tag{4}$$

L_{1st} makes the embedding vectors corresponding to the two adjacent nodes in the graph close in the hidden space. L_{reg} is a regularization constraint, α is a parameter that controls the first- proximity loss, and β is a parameter that controls the regularization constraint. After SDNE, each node gets its own vector representation in the hidden space. Assuming that the matrix y_g is the result linked to ICD-9 of SDNE, which $\in \mathbb{R}^{l_g \times d_g}$. Where l_g denotes the number of ICD-9 and d_g denotes the dimensions of each node.

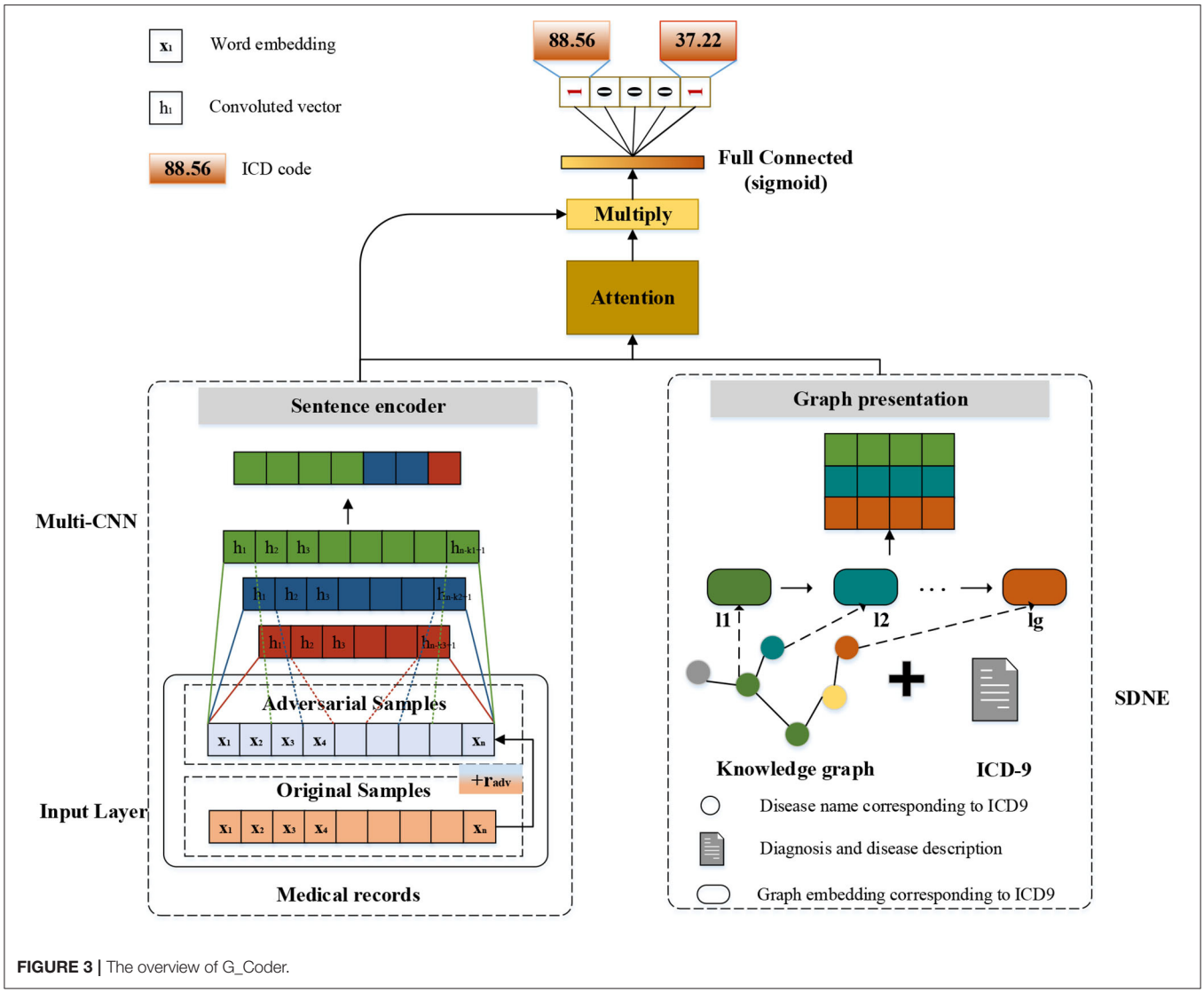


FIGURE 3 | The overview of G_Coder.

Attentional Matching

Human coders usually look for the most critical part of the medical record (Such as symptoms, complications, etc.) to determine the final coding result. In this task, we need to refine the text that most relevant to the ICD information and give higher weight. For the above reasons, we apply the attention mechanism. A benefit is that it selects the segments from the text that are most relevant to each predicted label. The specific algorithm details are shown in **Table 2**. It obtained the clinical text representation vector H' through preprocessing and Multi-CNN, and at the same time obtained the ICD coded representation y_g using the knowledge graph embedding results. A linear transformation was performed on the code representation to obtain the final code representation D, which has the same dimensions as the number of codes. The text representation H' and label representation D are used to calculate the weight a_i of the relationship between each label and each segment of the text. Finally, the text H' and weight a_i are used to weight the average of each part of the text to obtain the final clinical text representation v_i.

TABLE 2 | The algorithm details of attentional matching.

Algorithm1: Attentional matching

For each $\boldsymbol{H'}$ from Multi-CNN:

1. Calculate label representation vector $\boldsymbol{D}$;

$$\boldsymbol{D}=(\boldsymbol{W_g y_g}+\boldsymbol{b})$$

2.The $\boldsymbol{a_i}$ Measures how informative each n-gram is for the i-th label.;

$$\boldsymbol{a_i}=\boldsymbol{SoftMax}\left(\boldsymbol{H'^T D_i}\right),\boldsymbol{i}=\boldsymbol{1,2,3}\ldots\boldsymbol{l_g}$$

3. Calculate the weighted average $\boldsymbol{v_i}$ of the rows in $\boldsymbol{H'}$ forming a vector representation of the clinic text for the i-th label;

$$\boldsymbol{v_i}=\boldsymbol{a_i H'}$$

The results in **Table 2** can be summarized as follows:

$$A=SoftMax\left(H'^{T}W_g y_g\right), A=[a_1,a_2,\ldots,a_{\lg}] \quad (5)$$

$$V=AH', A=[v_1,v_2,\ldots,v_{\lg}] \quad (6)$$

TABLE 3 | The algorithm details of Adversarial Learning.

Algorithm 2: Adversarial learning
For each $\boldsymbol{X}$ in training samples: 1. Calculate the forward loss of $\boldsymbol{X}$ and get the gradient $\boldsymbol{g}$ by back propagation; $\boldsymbol{g}=\nabla_{\boldsymbol{X}}\boldsymbol{L}(\theta,\boldsymbol{X},\boldsymbol{Y})$ 2. Calculate $\boldsymbol{r_{adv}}$ according to the gradient of the embedding matrix $\boldsymbol{X}$ and add it to the current embedding, which is equivalent to $\boldsymbol{X}+\boldsymbol{r_{adv}}$; $\boldsymbol{r_{adv}}=\epsilon \bullet \boldsymbol{g}/\|\boldsymbol{g}\|\boldsymbol{2}$ $\boldsymbol{X_{adv}}=\boldsymbol{X}+\boldsymbol{r_{adv}}$ 3. Calculate the forward loss of $\boldsymbol{X_{adv}}$, backpropagate to obtain the gradient of the confrontation, and add to the gradient of step 1; 4. Restore embedding to the value at step 1; 5. Update the parameters according to the gradient of step 3.

Where $SoftMax\,(x) = \frac{\exp(x_i)}{\sum_i \exp(x_i)}$, and exp is an exponential function with natural constant e as a base. The matrix $W_g \in \mathbb{R}^{d_g \times l_g}$ is the weight parameter. And A denotes attention weights for each pair of an ICD code and the text. The letter $V \in \mathbb{R}^{l_g \times l_g}$ denotes the output of the attention. The concrete example can be found in **Table 7**.

Adversarial Learning

We apply FGM (fast gradient method) to reconcile the different writing styles of doctors and increase training samples (Miyato et al., 2017). The basic idea is: The writing of medical records follows the writing standards, but also contains different writing styles. Adversarial learning weakens the influence of writing style. The purpose of adversarial training is that the model will work steadily even if there are large differences in doctor writing styles. FGM uses a first-order Taylor expansion on the adversarial objective function to approximate to maximize the error output by the model, which is equivalent to using a single-step gradient descent method with a step size of ϵ to find the adversarial samples. The specific algorithm details are shown in **Table 3**. It calculates the gradient g of the clinic text embedding X after forward propagation and then back propagation. The gradient is used to calculate the perturbation r_{adv} added to X. After such a process, X_{adv} is an automatically generated adversarial sample. It uses the adversarial samples to calculate together with the original samples, increasing the number of samples, while mimicking the writing style of different doctors.

The goals of adversarial learning are as follows:

$$\min_{\theta} \mathbb{E}\ (X,Y) \sim D[\max_{\mathbf{r_{adv}} \varepsilon \mathbf{R}} (L(\theta, X_{adv}, Y))] \tag{7}$$

The formula (7) is divided into two parts, one is the maximization of the internal loss function, and the other is the minimization of the external risk. In the internal max, L is the defined loss function, D is the perturbation of input samples, and R is the space for a perturbation. The goal of adversarial learning is to find the amount of perturbation that makes the most judgment errors. For the above attacks, the most robust model parameters are found. After further optimizing the model parameters, the expected value of the entire data distribution is still minimal.

TABLE 4 | The hyperparameter settings of the experiment.

Hyperparameter	Value
d	100
d_g	128
d_f	50
lr	0.001
dp	0.4
λ	0.00001
Filters size	{4,5,6}

Output Layer

We compute a probability for label vector $\hat{Y} \in \mathbb{R}^{l_g}$ using full connection layer and a sigmoid transformation by the output of attention representation V:

$$\hat{Y} = \sigma\ (W_o V) \tag{8}$$

Where $W_o \in \mathbb{R}^{l_g \times l_g}$ is learnable weights of output layer and $\sigma\ (x) = \frac{1}{1+exp(-x)}$. The whole learning process minimizes the binary cross-entropy loss (9) of prediction probability $\hat{Y}_i$ and the target $Y_i \in (0, 1)$. The label i is selected when $\hat{Y}_i$ is >0.5.

$$L(\theta, X, Y) = -\sum_{i=1}^{l_g} Y_i \log\left(\hat{Y}_i\right) + (1-Y_i) \log\left(1-\hat{Y}_i\right) + \lambda||\gamma||_2^2 \tag{9}$$

Where X denotes the input word sequence, λ is the L2 regularization hyperparameter. And θ denotes all the parameters. We utilize the back-propagation algorithm and Adam optimizer (Kingma and Ba, 2014) to train the model.

EXPERIMENTS

Experimental Settings

A majority of codes are only assigned to too few medical records. Since the top 50 common ICD-9 codes covered 93.6% of the all dataset, we pick 50 most frequent codes to carry out the experiment while considering that our method can readily be extended to more codes as long as sufficient training data is available. The experimental dataset using top-50 codes has a total of 46,552 discharge summaries, which has 43,000 discharge summaries for training, 1,800 for validation, and 1,752 for the test. In this experiment, the settings are shown in **Table 4**. The token dimension d is 100; the knowledge graph embedding size d_g is 128; the out-channel size d_f of a filter in the Multi-CNN layer is 50; the learning rate lr is 0.001; the L2 regularization hyperparameter λ is 0.00001; the max length of each medical record is 1,800; the mini-batch size is 16 and the dropout rate dp is 0.4. We used three filters and the kernel size of filters is 4,5,6.

Evaluation Metrics

This task can be regarded as a multi-label classification problem. Therefore, we evaluate the method by $micro - F1$ and AUC

TABLE 5 | The experimental results of the top-50 codes.

Method	micro-F1	micro-AUC	P@5
CNN-Att	0.625	0.907	0.620
C-LSTM-Att Shi et al. (2017)	0.532	0.900	-
CAML Mullenbach et al. (2018)	0.614	0.909	0.609
DR-CAML Mullenbach et al. (2018)	0.633	0.916	0.618
MultiResCNN Li and Yu (2019)	0.673	0.928	0.641
No-knowledge-graph	0.670	0.923	0.637
No-adversarial-learning	0.681	0.929	0.647
G_Coder	**0.692**	**0.933**	**0.653**

Bold represent the current model result.

(Area under the curve). The $micro - F1$ is harmonic mean that calculated from *Precision* and *Recall*. All evaluation matrixes are calculated as follows:

$$Precision = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{10}$$

$$Recall = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{11}$$

$$micro{-}F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Precision} + \text{Recall}} \tag{12}$$

In these formulas, TP_i is the set of ground truth labels of each class, n is the number of samples, FN_i is the number of positive classes predicted as negative classes and FP_i is the number of negative classes predicted as positive classes. AUC is mainly used to evaluate the ranking ability of the current model. The higher the AUC, the better the ranking ability of the model. When the prediction probability values of all positive samples are higher than the negative samples, the AUC of the model is 1.

Results

Model Comparison

This section illustrates the performance of our approach. The experimental results of the top-50 codes show in **Table 5**, which show that our work has improved on previous work. CNN-Att is the baseline model for this experiment, which uses CNN to encode text. MultiResCNN has achieved the state-of-the-art results on the MIMIC-III datasets using unstructured text. Besides, their work is based on CAML and the model is improved. It mainly consists of a multi-filter convolutional layer and residual convolutional layer for multi-label classification. C-LSTM-Att applied LSTM-based language models to encode clinical notes and ICD codes and applied an attention method to solve the mismatch between clinical notes and codes. They focused on predicting the 50 codes that have the top frequencies for the medical records in the MIMIC-III dataset just like us.

Comparing our model with existing work for automatic ICD coding. As shown in **Table 5**, the conclusions are as follow:

TABLE 6 | The result of universality study.

Method	micro-F1	micro-AUC	P@5
CNN-Att	0.625	0.907	0.620
CNN-Att- graph	**0.651**	**0.920**	0.619

Bold represent the best results.

1) G_Coder obtains better results in the micro-AUC, micro-F1, and P@5. Compared with the state-of-the-art model MultiResCNN, G_Coder improves the micro-AUC by 0.005, the micro-F1 by 0.019, the P@5 by 0.012. P@5 measures the ability of the method to return the top 5 high-confidence subsets of codes. Our approach achieves relatively high precision of the five most confident predictions, on average 3.3 are correct.
2) CNN-based models are more suitable for this task. LSTM pay more attention to capture long sequence features, and cannot extract important local features from noise text. Simultaneously, the length of the medical record text makes the recurrent neural network have extremely high requirements for machine performance in this task. In contrast, it can be seen from the model construction that CNN can better extract long text features, and multilayer CNN with different convolution kernels can better capture local correlation.
3) The attention mechanism is essential. Each model utilizes the attention mechanism, which shows that the mechanism accurately highlights the information related to ICD in the text. The following content will prove the value of the knowledge graph and adversarial learning in this task.

Ablation Study

To gain more insight, the ablation study applied to verify the effectiveness of the adversarial learning and knowledge graph. To evaluate each module, we perform single variable experiments. The comparisons of the No-one module with the full model are given in **Table 6**. We remove one module from the full model without changing other modules and denote such a baseline by No-X. To evaluate them, we compared with the two configurations: (1) No-knowledge-graph, which removes the graph presentations and directly uses a randomly initialized vector as final representations of codes information; (2) No- adversarial-learning, which removes the adversarial learning form full model.

It can see from **Table 6** that our full model obtains better results in all evaluation matrix. Compared with the full model, No-knowledge-graph dropped the micro-AUC from 0.933 to 0.923, the micro-F1 from 0.692 to 0.670, the P@5 from 0.653 to 0.637. At the same time, No-adversarial-learning dropped the micro-AUC from 0.933 to 0.929, the micro-F1 from 0.692 to 0.681, the P@5 from 0.653 to 0.647. The above results show that the knowledge graph-based method can add clinical experience to make the results better. And adversarial learning generates adversarial samples through perturbation factors to enhance the generalization ability

TABLE 7 | Presentation of clinical text fragments and their corresponding ICD codes (The bold part indicates the highest weight).

ICD-9 codes and description	The highest weighted part
584.9 Acute renal failure, unspecified	…support **with acute renal failure secondary** to the prolong hypertension…
410.71 Acute myocardial infarction, subendocardial infarction, initial episode of care	…the patient experienced **right ventricular failure and went** back on bypass with drug manipulations…
414.01 Coronary atherosclerosis of native coronary artery	…with a right heart bypass cannulation in place. The patient **was profoundly hypoxic and acidotic**. …
428.0 Congestive heart failure, unspecified	…He also had lactic **acidosis and congestive heart failure**. The hypernatremia. …

TABLE 8 | The result of the evaluation of interpretability.

Type	Total	Correct	Accuracy
High weight (weight $\geq$ 0.8)	16	10	0.625
Others (weight $<$ 0.8)	84	60	0.714

of the model on the test set. From the results we have obtained, one can conclude that the combination of data-driven and medical knowledge can enhance the precision of ICD automatic coding.

Universality Study

To prove that the knowledge graph is universal in this task. We design the experiment, which is to add a knowledge graph to the basic baseline model and compare it with the baseline model.

According to the experimental results in **Table 6**, it can be seen that the knowledge graph not only performs well in G_Coder but also can be extended to other model structures. The knowledge graph improves the micro-F1 of the baseline model by 2.6%. This shows that the knowledge graph is universal and can be flexibly grafted into other model structures.

Evaluation of Interpretability

We use two methods to verify the interpretability. The first is an intuitive method that attention extracts keywords and displays the correlation between the code and the evidence. Examples can be found in **Table 7**. It can be seen from which words the basis of coding comes from. Taking 584.9 as an example, there is an information overlap between "acute renewal failure, unspecified" and "with acute renewal failure secondary" in clinical texts.

The second is a quantitative method where doctors judge the results of attention distribution. A clinical medical record was randomly selected, and segments were extracted based on the results of its attention. We select 5-words in this setting to emulate a span of attention over words likely to be given by a human reader. Since the segment may overlap, the most important 5-words were extracted according to attention weight. As can be seen from **Table 8**, the score is divided into two stages, one is high weight, that is $>$0.8, and the other is $<$0.8. In a total of 100 segments, there are 16 with a weight $>$0.8 and 84 with a weight $<$0.8. According to the evaluation results of human coders, 10 of the high weights are correct, and the remaining correct number is 60.

CONCLUSIONS AND DISCUSSIONS

Conclusions

Inspired by the structure of graphs that can model the relationships and knowledge between all things in the world, we think the graph structure can connect the parts of the data in this task and create a knowledge graph using medical-related data from the Freebase database. At the same time, the development of deep learning has also allowed further development of natural language processing such as automatic coding and text classification. In this paper, we propose a new explainable method for automatic ICD coding. The result of the micro-F1 score of 50 most frequent codes is 69.2%, which outperforms all the other models especially when raw clinical text data is used as input features to the prediction models.

The experimental evaluation of the MIMIC-III dataset shows the following points. First, we combined deep learning with knowledge graphs in ICD coding tasks. The medical knowledge graph supervises the coding process as a teacher. At the same time, we apply the SDNE algorithm to encode each entity of the knowledge graph and link it to the ICD-9 code. The Multi-CNN algorithm is utilized to encode long text information of MIMIC-III data. In the attention mechanism, we combine the two mentioned above to identify the segments of text that are most relevant to each ICD-9 code. Finally, we generate adversarial samples through adversarial training and send the samples to the training along with the original samples. It can weaken the influence of writing style and make model more stable. Moreover, in the ablation study and universality study, we use the single variable rule to verify the importance of adversarial learning and knowledge graph. The results prove that the knowledge graph can be flexibly grafted into the model structure to help understand the terminology. Two methods are used to verify the interpretability of the method. It is confirmed that this method is based on the important basis in the clinical text for ICD coding. G_Coder has a higher accuracy rate than the other method. And before the coder works, G_Coder can perform ICD pre-selection to save time for whole encoding work.

Discussions

The major limitation of this work is that it does not perform well on infrequent codes. To achieve fully automatic coding, infrequent coding has to be considered. And we hold that the method can readily be extended to more codes as long as sufficient training data is available. In addition, the new ICD version should also be considered, such as ICD10, ICD11, etc. ICD classification is a disease classification directory with hierarchical relationship. The structure of ICD is also a direction worth considering.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://freebase.gstore-pku.com/, https://mimic.physionet.org/, https://developers.google.com/freebase.

AUTHOR CONTRIBUTIONS

WY and FT provided total research ideas designed the experiments. WY performed the experiments and wrote the first draft of the manuscript. LH and LC guided the experiment as experts and analyzed the results. FT contributed to the High-performance experimental equipment. WY and QX contributed to manuscript revision, reading, and approving the submitted version. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Sichuan Science and Technology Program (No. 2017SZYZF0002), National Key R&D Program of China (No. 2019YFB2101802), and Sichuan Key R&D project (No. 2020YFG0035).

ACKNOWLEDGMENTS

We would like to thank the Beth Israel Deaconess Medical Center for providing data support. We would also like to thank the reviewers for their insightful comments.

REFERENCES

Adler, J. P., Frank, W., Noemie, E., and Nicholas, B. (2011). *Hierarchically Supervised Latent Dirichlet Allocation*. Advances in Neural Information Processing Systems, 2609–2617. Available online at: http://papers.nips.cc/paper/4313-hierarchically-supervised-latent-dirichlet-allocation

Baumel, T., Nassour-Kassis, J., Elhadad, M., and Elhadad, N. (2017). *Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment*. arXiv. Available online at: https://arxiv.org/abs/1709.09587 (accessed September 27, 2017).

Chen, Y., Lu, H., and Li, L. (2017). Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PLoS ONE*. 12:e0173410. doi: 10.1371/journal.pone.0173410

Dzmitry, B., Kyunghyun, C., and Yoshua, B. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv. Available online at: https://arxiv.org/abs/1409.0473 (accessed September 1, 2014).

Erraguntla, M., Gopal, B., Ramachandran, S., and Mayer, R. (2012). "Inference of missing ICD 9 codes using text mining and nearest neighbor techniques," in *2012 45th Hawaii International Conference on*. IEEE (HICSS), 1060–1069. doi: 10.1109/HICSS.2012.323

Ferrão, J., Janela, F., Oliveira, M., and Martins, H. (2013). "Using structured EHR data and SVM to support ICD-9-CM coding," in *Proceedings of the 2013 IEEE International Conference on Healthcare Informatics* (Philadelphia, PA), 511–516. doi: 10.1109/ICHI.2013.79

Gong, P., and Ai, L. (2019). *Neighborhood Adaptive Graph Convolutional Network for Node Classification*. New Jersey, NJ: IEEE Access, 170578–170588. doi: 10.1109/ACCESS.2019.2955487

Goodfellow, I., Shlens, J., and Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples*. Available onlibe at: https://arxiv.org/abs/1412.6572 (accessed March 20, 2015).

Grover, A., and Leskovec, J. (2016). "node2vec: scalable feature learning for networks," in *KDD: Proceedings. International Conference on Knowledge Discovery & Data Mining* (San Francisco, CA), 855–864. doi: 10.1145/2939672.2939754

Hong, Z., Zeng, X., Wei, L., and Liu, X. (2020). Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 36, 1037–1043. doi: 10.1093/bioinformatics/btz694

Ibrahim, A., Achille, F., Oktie, H., Ping, Z., and Mohammad, S. (2017). Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. *J. Web Sem.* 44, 104–117. doi: 10.1016/j.websem.2017.06.002

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scient. Data* 3:160035. doi: 10.1038/sdata.2016.35

Kim, Y. (2014). "Convolutional neural networks for sentence classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Doha), 1746–1751. doi: 10.3115/v1/D14-1181

Kingma, D., and Ba, J. (2014). "Adam: a method for stochastic optimization," in *International Conference on Learning Representations*. Available online at: https://arxiv.org/abs/1412.6980 (accessed December 22, 2014).

Li, F., and Yu, H. (2019). "ICD coding from clinical text using multi-filter residual convolutional neural network," in *AAAI Technical Track: Natural Language Processing* (New York, NY), 34. doi: 10.1609/aaai.v34i05.6331

Li, M., Fei, Z., Zeng, M., Wu, F., Li, Y., Pan, Y., et al. (2018). "Automated ICD-9 coding via a deep learning approach," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (New Jersey, NJ), 16, 1193–1202. doi: 10.1109/TCBB.2018.2817488

Li, Z., Liu, Z., Huang, J., Tang, G., Duan, Y., Zhang, Z., et al. (2019a). MV-GCN: multi-view graph convolutional networks for link prediction. *IEEE Access* 7, 176317–176328. doi: 10.1109/ACCESS.2019.2957306

Li, Z., Yanga, J., Goua, X., and Qi, X. (2019b). Recurrent neural networks with segment attention and entity description for relation extraction from clinical texts. *Artif. Intell. Med.* 97, 9–18. doi: 10.1016/j.artmed.2019.04.003

Liu, H., Li, Y., Hong, R., Li, Z., Li, M., Pan, W., et al. (2020). Knowledge graph analysis and visualization of research trends on driver behavior. *J. Intell. Fuzzy Syst.* 38, 495–511. doi: 10.3233/JIFS-179424

Max, F., Arne, K., Gregor, W., and Chris, B. (2019). "Adversarial learning of privacy-preserving text representations for de-identification of medical records," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence), 5829–5839.

Medori, J., and Fairon, C. (2010). "Machine learning and features selection for semi-automatic ICD-9-CM encoding," *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents* (Los Angeles, CA), 84–89.

Miyato, T., Dai, A., and Goodfellow, I. (2017). *Adversarial Training Methods for Semi-Supervised Text Classification*. Available online at: https://arxiv.org/abs/1605.07725 (accessed May 6, 2017).

Mohamed, S. K., Nováček, V., and Nounu, A. (2020). Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*. 36, 603–610. doi: 10.1093/bioinformatics/btz600

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. *NAACL* 1,1101–1111. doi: 10.18653/v1/N18-1100

Ning, W., Yu, M., and Zhang, R. (2016). A hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation. *BMC Med. Inform. Dec. Making.* 16, 1–12. doi: 10.1186/s12911-016-0269-4

Ozan, O., Jo, S., Loic, L. F., Matthew, L., Mattias, H., Kazunari, M., et al. (2018). *Attention U-Net: Learning Where to Look for the Pancreas*. Available online at: https://arxiv.org/abs/1804.03999 (accessed April 11, 2018).

Pakhomov, S., Buntrock, J., and Chute, C. (2006). automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *J. Am. Med. Inform. Assoc.* 13, 516–525. doi: 10.1197/jamia.M2077

Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., and Elhadad, N. (2013). Diagnosis code assignment: models and evaluation metrics. *JAMIA* 21, 231–237. doi: 10.1136/amiajnl-2013-002159

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "DeepWalk: online learning of social representations," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY). doi: 10.1145/2623330.2623732

Ping, C., Araly, B., and Chris, R. (2010). "Semantic analysis of free text and its application on automatically assigning ICD-9-CM codes to patient records," in *Proceedings of the 9th IEEE International Conference on Cognitive Informatics (ICCI)* (Beijing), 68–74.

Ribeiro, L., Saverese, P., and Figueiredo, D. (2017). "struc2vec: Learning node representations from structural identity," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS), 385–394. doi: 10.1145/3097983.3098061

Ruch, P., Gobeill, J., Tbahritia, I., and Geissbühlera, A. (2008). "From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding," in *AMIA. Annual Symposium Proceedings/AMIA Symposium* (Washington, DC: AMIA Symposium), 636–640.

Shi, H., Xie, P., Hu, Z., Zhang, M., and Xing, E. (2017). *Towards Automated ICD Coding Using Deep Learning*. Available at: https://arxiv.org/abs/1711.04075 (accessed November 11, 2017).

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). LINE: "Large-scale information network embedding," in *WWW* '15: *Proceedings of the 24th International Conference on World Wide Web* (Florence), 1067–1077. doi: 10.1145/2736277.2741093

Trevena, L. (2011). WikiProject Medicine. *BMJ* 342:d3387. doi: 10.1136/bmj.d3387

Wang, D., Cui, P., and Zhu, W. (2016). "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: KDD* (San Francisco, CA), 1225–1234. doi: 10.1145/2939672.2939753

Wang, S., Li, X., Chang, X., Yao, L., Sheng, Q., and Long, G. (2017). Learning multiple diagnosis codes for ICU patients with local disease correlation mining. *ACM Trans. Knowl. Disc. Data*. 11, 1–21. doi: 10.1145/3003729

Xie, P., and Xing, E. (2018). "A neural architecture for automated ICD coding," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Melbourne, VIC), 1066–1076. doi: 10.18653/v1/P18-1098

Xu, K., Yang, Z., Kang, P., Wang, Q., and Liu, W. (2019). Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. *Comp. Biol. Med*. 108, 122–132. doi: 10.1016/j.compbiomed.2019.04.002

Yan, Y., Fung, G., Dy, J., and Rosales, R. (2010). "Medical coding classification by leveraging inter-code relationships," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC), 193–202. doi: 10.1145/1835804.1835831

Ye, Q., Zhu, C., Li, G., Liu, Z., and Wang, F. (2018). Using node identifiers and community prior for graph-based classification. *Data Sci. Eng*. 3, 68–83. doi: 10.1007/s41019-018-0062-8

Yu, W., Yuan, C., Qin, X., Huang, Z. H., and Li, S. (2019). "Hierarchical attention network for predicting DNA-protein binding sites," in *Proceedings of the International Conference on Intelligent Computing (ICIC)* (Nanchang), 11644, 366–373. doi: 10.1007/978-3-030-26969-2_35

Yu, Y., Li, M., Liu, L., Fei, Z., Wu, F. X., and Wang, J. X. (2019). Automatic ICD code assignment of chinese clinical notes based on multilayer attention BiRNN. *J. Biomed. Inform*. 91:103114. doi: 10.1016/j.jbi.2019.103114

Zhao, S., Cai, Z., Chen, H., Wang, Y., Liu, F., and Liu, A. (2019). Adversarial training based lattice LSTM for Chinese clinical named entity recognition. *J. Biomed. Inform*. 99:103290. doi: 10.1016/j.jbi.2019.103290

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 18 September 2020
doi: 10.3389/fpls.2020.538311

Comparative Analysis of Soil Microbiome Profiles in the Companion Planting of White Clover and Orchard Grass Using 16S rRNA Gene Sequencing Data

Lijuan Chen[1], *Daojie Li*[1], *Ye Shao*[2], *Jannati Adni*[1], *Hui Wang*[1], *Yuqing Liu*[3] *and Yunhua Zhang*[3]*

[1] College of Animal Science and Technology, Anhui Agricultural University, Hefei, China, [2] School of Medicine, Huaqiao University, Quanzhou, China, [3] School of Resources and Environment, Anhui Agricultural University, Hefei, China

OPEN ACCESS

Edited by:
Yucong Duan,
Hainan University, China

Reviewed by:
Richard R. Rodrigues,
Oregon State University, United States
Yun Li,
University of Pennsylvania, United States
Qiang Wang,
Nanjing University, China

***Correspondence:**
Yunhua Zhang
yunhua9681@163.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Plant Science

Received: 28 February 2020
Accepted: 31 August 2020
Published: 18 September 2020

Citation:
Chen L, Li D, Shao Y, Adni J, Wang H, Liu Y and Zhang Y (2020) Comparative Analysis of Soil Microbiome Profiles in the Companion Planting of White Clover and Orchard Grass Using 16S rRNA Gene Sequencing Data. Front. Plant Sci. 11:538311. doi: 10.3389/fpls.2020.538311

Companion planting is one of the most common and effective planting methods in modern agriculture. White clover (*Trifolium repens* L.) and orchard grass (*Dactylis glomerata* L.) are two typical pastures planted together to promote each other's growth. However, the detailed biological foundations of companion planting remain unclear. In this study, we screened typical microbiome profiles under separate and combination planting conditions using 16s RNA gene sequencing techniques. We identified the typical distinctive microorganism subtypes based on the microbiome profiles and recognized the enriched functions of top abundant microorganisms in soil using different planting strategies with the help of Kyoto Encyclopedia of Genes and Genomes and Clusters of Orthologous Groups annotation. This analysis confirmed that the optimal microorganisms and screened functional annotations are correlated with nitrogen fixation; thus, companion planting may improve the yield and efficacy of plants by improving the efficiency of nitrogen fixation.

Keywords: companion planting, 16s RNA gene sequencing, microbiome, Clusters of Orthologous Groups (COGs), operational taxonomic unit classification, multiple variables analysis, machine learning models

INTRODUCTION

Companion planting is another typical agricultural pattern partially associated with organisms (Finch et al., 2003; Parker et al., 2013). Companion planting is a method of planting different kinds of plants at the same time in proximity (Finch et al., 2003; Szafirowska and Kolosowski, 2008). Companion planting can help in pest control (Parker et al., 2013), pollination (Hagiwara et al., 1995; Moeller, 2004), nutrition supply optimization (Mengel, 2001; George et al., 2013), and the maximization of the use of space (Bomford, 2004). For instance, soybeans can provide nitrogen with the help of certain microorganisms in soil (Oyekanmi et al., 2007; Chen et al., 2012). Soybeans can remodel soil microbiome and provide more nitrogen nutrition in proximity for plant growth (Chen et al., 2012). Therefore, the companion planting of soybean and *Medicago sativa* may help

improve the production of both plants through the modification of soil microorganisms (Plaza et al., 2003).

White clover (*Trifolium repens* L.) is a typical agricultural plant from the bean family Fabaceae (Davidson, 1969). The companion planting of white clover and grain crops or pasture grasses has been widely applied in poor soils to provide green cover (Sood et al., 2018). Orchard grass (*Dactylis glomerata* L.) also known as cat grass is a famous kind of pasture with high yield and great drought tolerance (Bybee-Finley and Ryan, 2018). White clover and orchard grass are two typical and traditional model plants for companion planting, which greatly improve their efficacy and yield rate. Early in 1962, a Canadian journal has reported the effects of companion planting on oats and confirmed that clover has equal to or greater yields when planted with the crop and orchard grass (Davis, 1962). However, the detailed mechanisms of the interactions between white clover and orchard grass still remain unclear. More progressions have been made in companion planting with the development of modern culture and sequencing technologies in the last 5 years. The improved yields can be attributed to the enhanced efficiency of nitrogen fixation induced by companion planting (Bybee-Finley and Ryan, 2018; Chalk, 2018; Payne, 2019). The improved nitrogen formulation efficiency is induced by the remodeling of microorganisms in proximity (Moore et al., 2019). However, the detailed mechanisms of the biological basis of companion planting are still unclear and require further studies.

In this study, we focused on the companion planting of white clover and orchard grass using 16S rRNA gene sequencing techniques (Pitombo et al., 2016; Smets et al., 2016). We monitored microbiome remodeling patterns in separate and companion planting conditions. The differential and altered microbiome distribution patterns confirmed the contribution of microbiome in the companion planting of the two plants and partially revealed the potential biological foundations for companion planting at least at the microbiome level. In this study, we revealed the biological foundation of the companion planting of white clover and orchard grass at the microbiome level, constructed a general workflow to study the contributions of microbiome on companion planting, and provided a new perspective on the biological foundations of companion planting.

MATERIALS AND METHODS

Experiment Site and Soil

Experiments were performed at the Dayangdian Experimental Station of Anhui Agricultural University (31°58′N, 117°24′E),

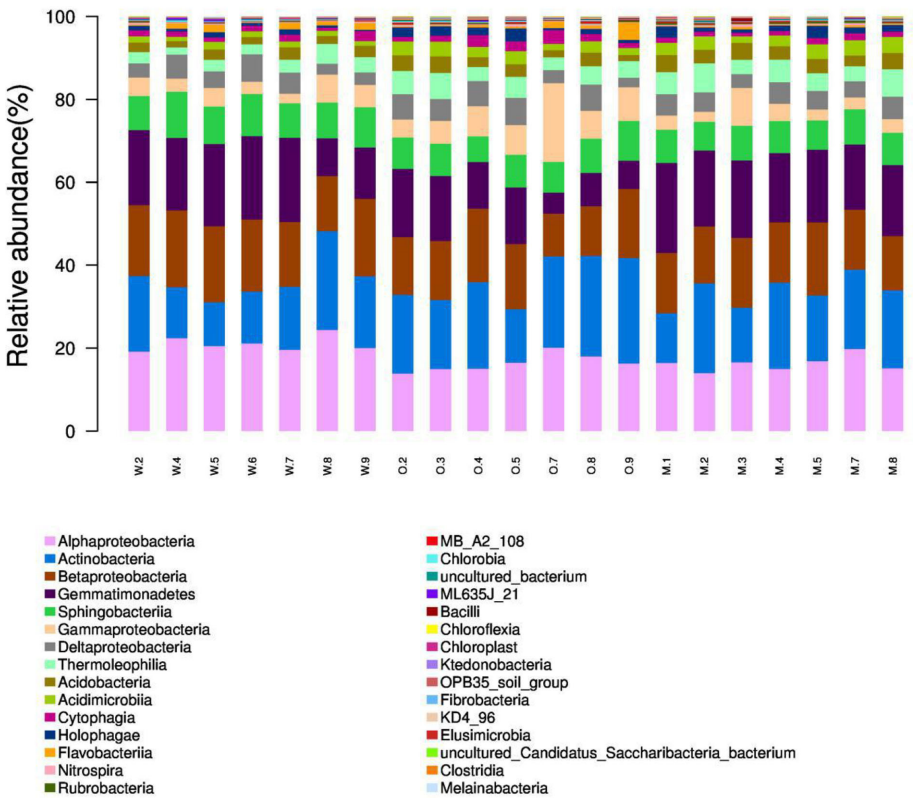


FIGURE 1 | Microbiome composition of the microorganisms with the top 30 abundances at the class level. We calculated the relative abundances (%) of the top 30 most abundant microorganisms at the class level. The proportion of top classes, such as Alphaproteobacteria, Actinobacteria, Betaproteobacteria, Gemmatimonadetes, was great in most samples regardless of groups and reflects the background microbiome pattern in the proximity. (Note: W2, W4–9 are WC samples; O2–5, O7–9 are OG samples; M1–5, M7–8 are Mixed samples).

Hefei City, Anhui Province, Southeast China. The study site is located between the Yangtze River and Huaihe River. The study area belongs to the transitional zone between the warm temperate zone and subtropical zone and has subtropical humid monsoon climate. Its annual temperature is cold in winter (8–17°C), hot in summer (21–29°C), and mild in spring and autumn, and its annual precipitation is 992 mm. The soil had the following physicochemical properties on a dry weight basis: 0.89% organic matter, 81.1 mg kg^{-1} available N, 16.3 mg kg^{-1} available P, and 100.5 mg kg^{-1} available K. In May 2019, 0.25 kg soil samples were collected in the rhizospheres of the white clover (WC) and orchard grass (OG) groups and soil samples of the companion planting of both plants (Mixed). All soil samples were preserved under the same conditions, and some fresh soil samples were further processed.

Treatments and Field Management

To explore the effects of white clover and orchard grass on the soil microorganisms, we established research sites in October 2018 where we applied three treatments: white clover, orchard grass, and the companion planting of both plants. The area of the land used in the experiment measured 4×4 m^2 in each plot. The amounts of white clover and orchard grass sowed were 10 and 20 kg ha^{-1}, respectively, and the Mixed group had 7.5 kg ha^{-1} white clove and 5 kg ha^{-1} orchard grass. The soil was watered when precipitation was insufficient.

DNA Extraction and Library Construction

Total genomic DNA was extracted using DNA Extraction Kit following the manufacturer's instructions. The quality and quantity of DNA were verified through spectrophotometry using NanoDrop spectrophotometer and *via* agarose gel electrophoresis. The extracted DNA was diluted to a concentration of 1 ng/μl and stored at −20°C until further processing. The diluted DNA was used as template for the polymerase chain reaction (PCR) amplification of bacterial 16S rRNA genes using barcoded primers and TaKaRa Ex Taq. The V3–V4 variable regions of 16S rRNA genes were amplified with universal primers 343F and 798R for bacterial diversity analysis.

Amplicon quality was visualized through gel electrophoresis, purified with AMPure XP beads (Agencourt), amplified for another round of PCR, and purified with AMPure XP beads again. The final amplicon was quantified using Qubit dsDNA assay kit. Equal amounts of purified amplicon were pooled for subsequent sequencing.

16S rRNA Gene Sequencing Result Analysis

Quality Control for Raw Sequencing Data

The raw image data obtained from high-throughput sequencing data was transformed into the original rRNA sequence in FASTQ file format by base calling analysis (Kao et al., 2009).

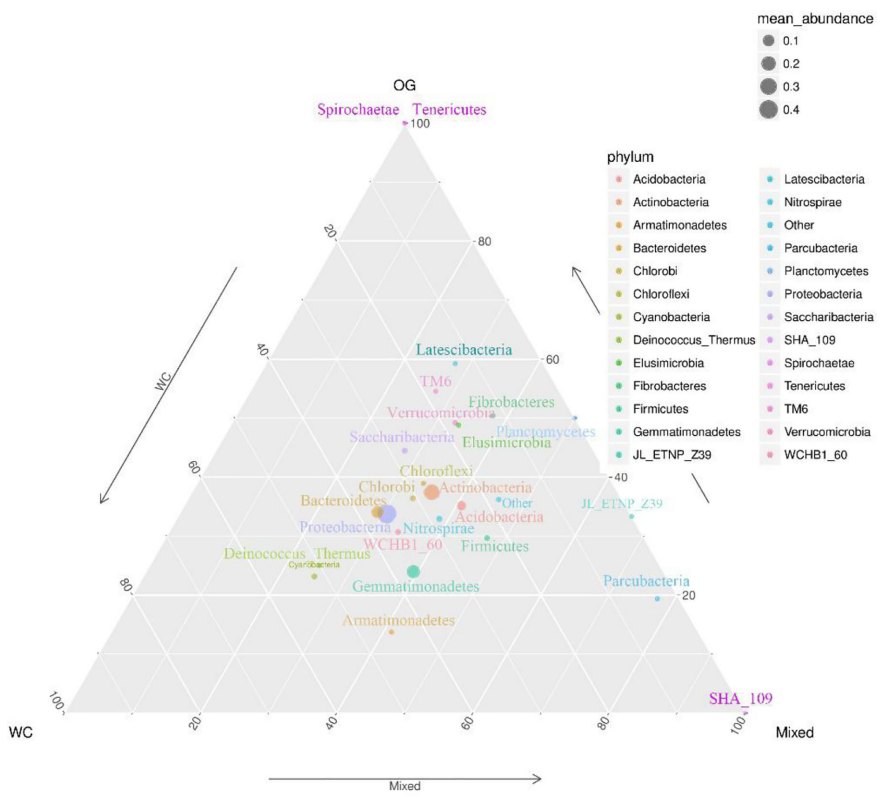


FIGURE 2 | Ternary plot of the contribution and relationship of different species in different groups. The ternary plot shows the unique species distribution patterns and abundances of the three groups.

The data in FASTQ format were further processed to remove the sequences with low quality and abnormal length using Trimmomatic software (Bolger et al., 2014). We also used UCHIME software to remove chimera in the raw FASTQ file to provide clean data for further analyses (Rognes et al., 2016). The distribution of sequence length after data cleaning is shown in the histogram and density map in **Figure S1**. Nearly all the reads distributed were within the length range of 400–450 bp with quite high quality; thus, our quality control procedure was efficient, and the clean data were eligible for further analysis.

Operational Taxonomic Unit Classification

We used Vsearch software to classify the high-quality sequence of the valid tags obtained by quality control according to 97% similarity. The most abundant sequence in each OTU was chosen as the representative sequence. We applied the Ribosomal Database Project classifier, the Naive Bayesian classification algorithm (Wang et al., 2007), to further align and annotate the representative sequences against the annotation database for the species information of each OTU. We further summarized the distribution of OTUs in different samples and the annotations of tags and OTUs based on the species results to show the general species distribution pattern of different samples. We used flower plot to show the numbers of shared and unique OTUs among different samples (**Figure S2**). Standardized the original data in OTU table file (the form of biom), and then the predicted functional (KEGG/COG) results were obtained by mapping the standardized data with the species functional genes from the online sequenced genome.

Analysis of Biome Structure From Soil in Proximity

Community structure or "biological community" refers to all the organisms that have a direct or indirect relationship with each other. Various groups in a microbial community interact with each other and can coexist in a regular manner but have their own distinct types of nutrition and metabolism. In this study, we summarized the composition of microbiome communities. We performed ternary plot analysis (Graffelman and Camarena, 2008) to compare and analyze the species composition of the three groups according to the classification results.

Alpha Diversity Analysis

Alpha diversity, which reflects the diversity of species in shared habitats, was calculated to present the species diversity in each sample (Huttenhower et al., 2012). Microhabitats have been tested for differences in estimated abundances with the Kruskal–Wallis significance test for all pairwise combinations. We measured the number of species and the uniformity of species abundance used the indexes of Shannon and Chao based on a rarefied (18,860 reads) dataset (**Figure S1**) to quantitatively evaluate species diversity.

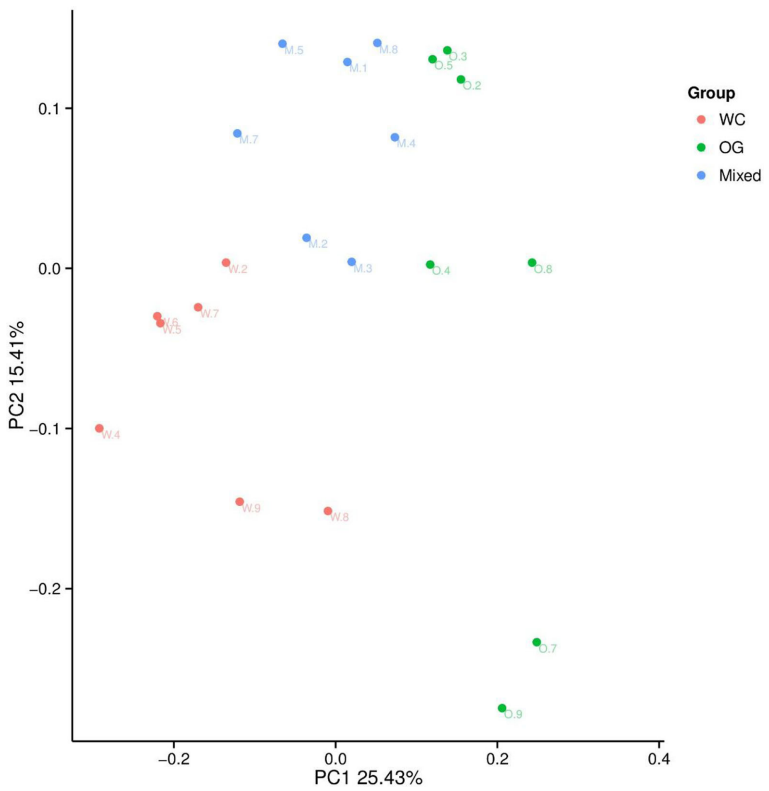


FIGURE 3 | PCoA analysis of the OTU composition differences of the three groups. The microbiome distribution and diversity of three groups were separated into different parts.

Beta Diversity Analysis

Beta diversity is the diversity of the relationships between organisms and environment in proximity (Kurilshikov et al., 2017). Similar with the alpha diversity analysis, we also used some quantitative parameters to evaluate the differential beta diversity patterns in different groups. In this study, we used principal co-ordinate analysis (PCoA) based on Bray Curtis to reveal the beta diversity among different groups.

Multiple Variable Analysis of Soil Microbiome

We used OTU and species data to identify the specific species that have statistically significant difference in abundances. We used ANOVA to identify the most substantial differentially existing species among the three groups (Rojewski et al., 2012).

Correlation Analysis and Prediction Using Machine Learning Models

We analyzed the correlations of different species and their contribution on the distinction of different groups using correlation analysis and machine learning methods. We also applied random forest apart from direct correlation analysis for further analysis. Random forest is the machine learning algorithm first proposed by Leo Breiman and Adele Cutler in 2001 (Breiman, 2001). Random forest is regarded as an integrated learning method with multiple decision trees. The output classification result is the result of "voting" by each decision tree. The classification results of random forests have high accuracy and do not need to "cut branches" to reduce overfitting because each tree uses random variables and random

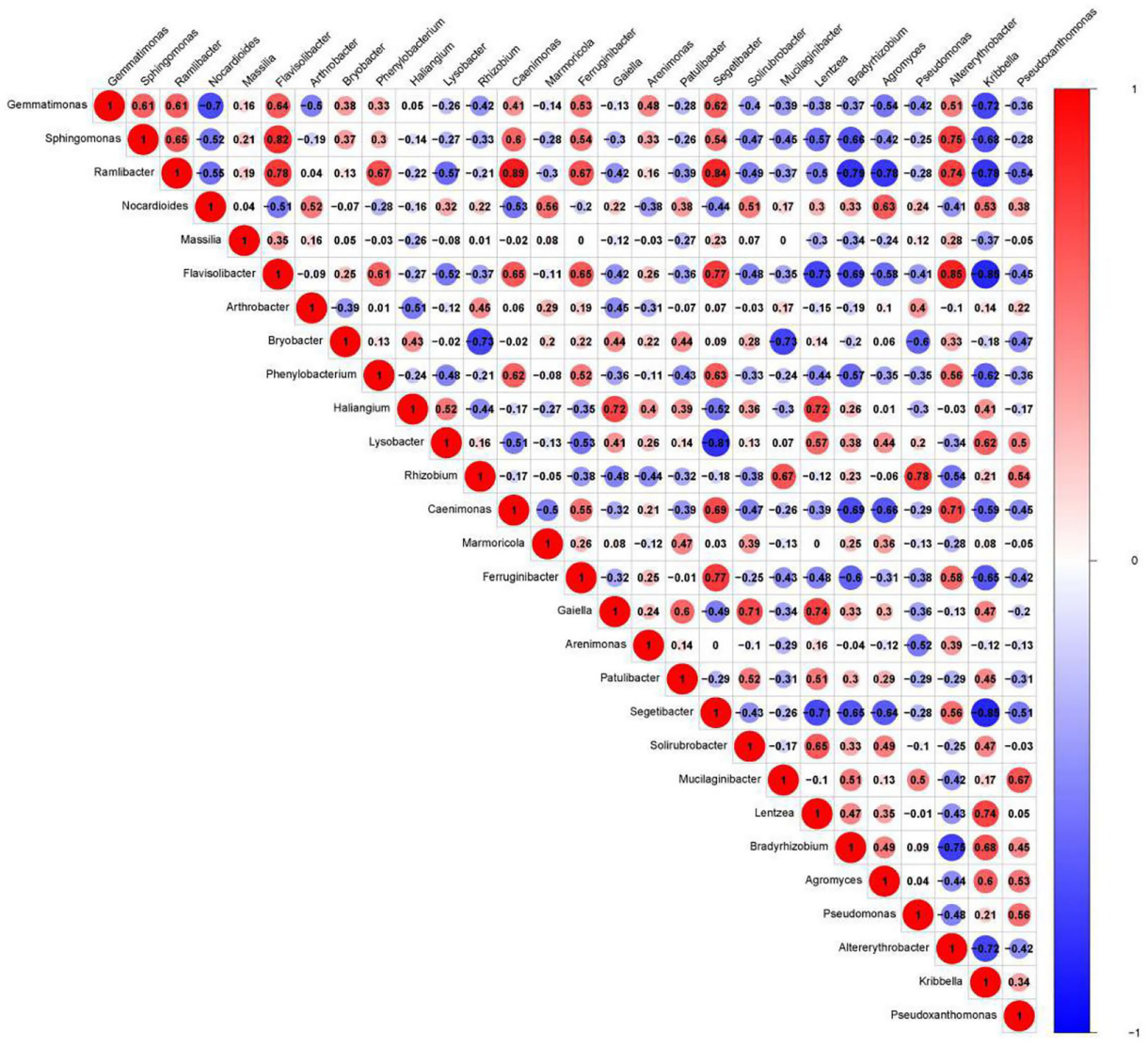


FIGURE 4 | Correlation plot of the top 30 genera with the highest abundance. The plot shows the inner correlation between different species (FDR < 0.1). Red indicates negative correlations, and blue indicates positive correlations at the abundance level. Size of the circle indicates absolute strength of correlation.

sampling methods in the construction process (Breiman, 2001; Segal, 2004). We used the proper R package (random forest) to perform the random forest algorithm (Segal, 2004).

Phylogenetic Investigation of Communities by Reconstruction of Unobserved States Analysis

PICRUSt functional predictive analysis is based on 16S rRNA gene sequencing data and annotated by Greengenes database (DeSantis et al., 2006). The PICRUSt software (Langille et al., 2013) is widely used to analyze the functional genetic composition of identified microorganism to reveal the functional diversity between different samples or groups. In this study, we applied PICRUSt analysis workflow (Langille et al., 2013) to reveal the functional distribution patterns of the different samples and groups.

RESULTS

Effect of Companion Planting on Microbiome Community Structure and Abundance

We summarized the composition of the microbiome community at the class level according to the microorganisms with the top 30 abundances (**Figure 1**, **Table S1**). According to the result, specific classes, such as Alphaproteobacteria, Actinobacteria, Betaproteobacteria, Gemmatimonadetes, had top abundances in nearly every sample and reflected the background microbiome distribution pattern in proximity. However, some specific classes, such as Gemmatimonadetes, had relatively higher abundance in the Mixed group compared with the OG group and indicated the potential microbiome remodeling effects of companion planting. However, the differential distribution patterns of the microbiomes of the three different groups were not clear. Therefore, we also used the ternary plot to reveal the contribution and relationship of different Phylum in different groups (**Figure 2**). According to **Figure 2**, Tenericutes and Spirochaetae were found in the specific distribution pattern of the OG group. This result indicated that these two microbiomes may be unique under the OG planting pattern and verified that companion planting affects the microbiome distribution pattern in proximity.

Effect of Companion Planting on Alpha Diversity

We used the boxplot to show the alpha diversity using Shannon and Chao1 parameters (**Figures S4**, **S5**). Results showed that the OG and Mixed groups had remarkably higher Chao1 (community richness) index values compared with the WC group ($P = 0.0103$; $P = 0.0029$). However, no significant difference was shown in Shannon index among there groups ($P = 0.0545$). Chao1 index describes and evaluates the number of species. A higher Chao1 index indicates a higher number of species in the sample. The Mixed group had the most diverse microbiome among the groups and had similar species abundance as the OG group ($P = 0.713$). The OG and Mixed groups had higher Chao1 index than the WC group. The differences of the

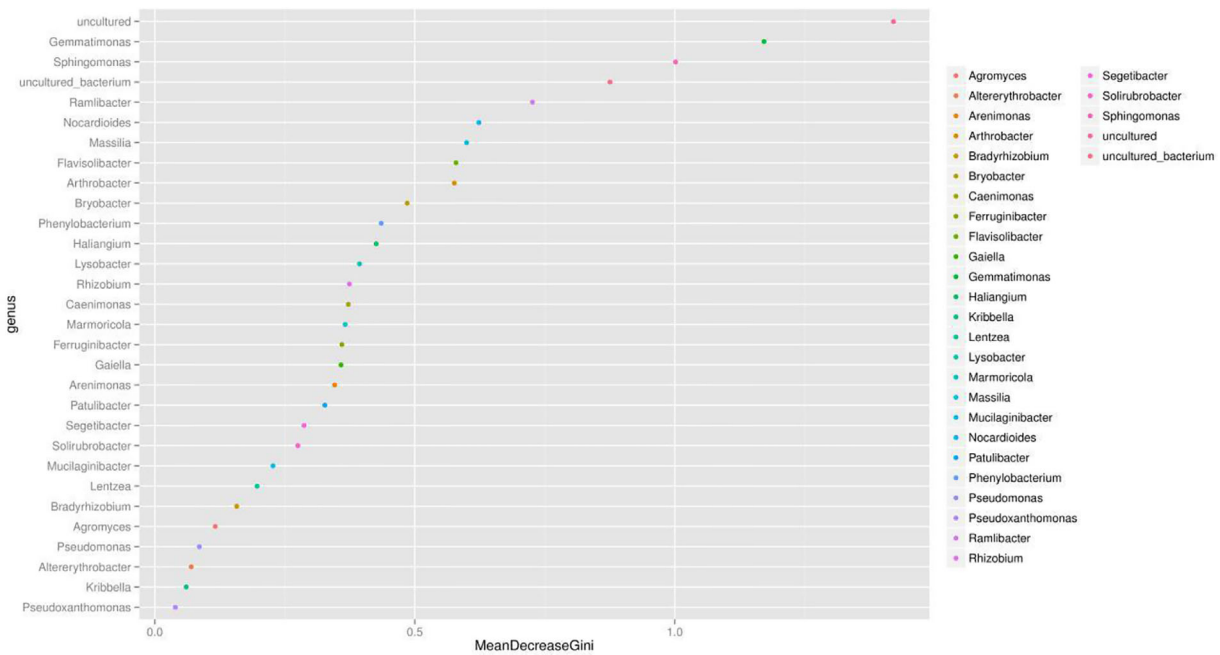


FIGURE 5 | Random forest evaluation of the top 30 genera for the distinction of different groups. We used random forest method and evaluated the importance of these features using the mean decrease grid. The point map of genus importance (variable) is shown. The abscissa is the measure of importance, and the ordinate is the name of the genus sorted by importance. Standardized importance values are used by default.

three groups in Chao1 index indicated that their microbial diversity and microbiome abundance are quite different from each other.

Effect of Companion Planting on Beta Diversity

We used Bray Curtis distance to evaluate the relationships between different samples and groups. The results were similar to those of the alpha diversity analysis. The three groups were divided into different parts. The mixed groups were distributed between WC and OG. However, the mixed groups were closer OG than WC, which indicated that the microbial community of OG played the main function in companion planting. Similar results were also shown by the PCoA results (**Figure 3**).

Effect of Correlation Analysis on Specific Contribution on Genus Level

We presented the correlation analysis results using a correlation plot (**Figure 4**) to show the inner relationship among the top 30 genera with the highest abundance. We also showed the specific contribution of each genus through a random forest in **Figure 5** based on the classification of the three subgroups. We ranked the contribution of each genus according to the mean decrease grid parameter and showed the top genera that contributed to the distinction of the three groups. Unique genera, such as *Gemmatimonas* and *Sphinomonas*, are quite important for the distinction of the three groups.

Effect of Companion Planting on Functional Differential Enrichment

We used Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2002; Aoki and Kanehisa, 2005) and Clusters of Orthologous Groups (COGs) (Galperin et al., 2017) functional annotation and prediction to show the distribution of functional clusters among different samples and groups. We could not distinguish the samples from the three groups using KEGG enrichment analysis (**Figure 6**). By contrast, we were able to distinguish the WC group from the OG and Mixed groups using COG annotation and enrichment analysis (**Figure 7**). This finding is similar with previous functional analysis, which indicated that the effect of planting only white clover on the

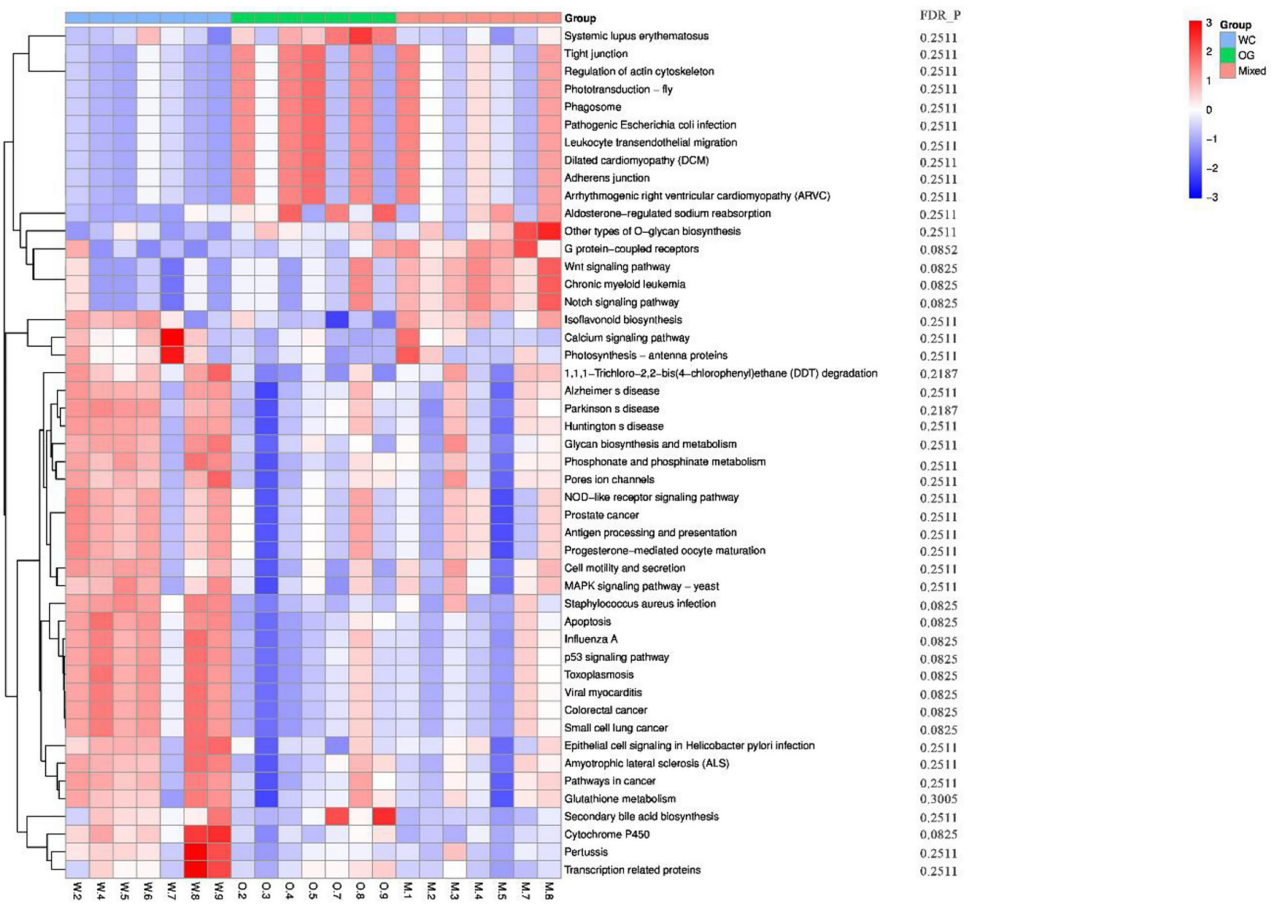


FIGURE 6 | KEGG functional annotation and differential enrichment analysis of the three groups. The functional annotation and clustering of the top KEGG terms were performed using the KEGG database. The samples from different groups are difficult to distinguish using KEGG terms.

microbiome is quite different from those of planting only orchard grass and companion planting.

DISCUSSION

Effect of Companion Planting on Soil Bacteria Community Structure and Diversity

The samples from different groups have different biome structures. **Figure 1** demonstrates that most of the samples from three groups have similar microbiome compositions, but relatively abundance differ among three groups at the OTU level (**Table S1**). Similar results have been reported in previous studies on soil microbiome (Gołębiewski et al., 2014; Samad et al., 2017). These results confirmed the complexity of soil at the microbiome level. We also identified some unique distribution patterns at the class level. Alphaproteobacteria, Actinobacteria, Betaproteobacteria, Gemmatimonadetes were detected in almost all the samples and reflect the general soil background in the proximity. Alphaproteobacteria, Actinobacteria, Betaproteobacteria, Gemmatimonadetes are widely detected in farmlands and pastures all over the world (Aguilar et al., 2004; Rosenblueth and Martínez-Romero, 2004; Reina-Bueno et al., 2012).

We identified two unique species from Tenericutes and Spirochaetae that contributed to distinguishing the OG group from the other two groups. Tenericutes has been identified in regions with various kinds of grass orchards worldwide (Liu et al., 2017; Deakin et al., 2018). Spirochaetae has also been identified in regions planted with grass orchards (Brown, 1943). Here we can't find Tenericutes and Spirochaetae in soil of WC and Mixed group. These findings may imply that some root exudates in WC inhibit the specific distribution of these microorganisms.

Effect of Companion Planting on Bacteria Groups

Genus *Gemmatimonas* contributed the most to the distinction of the groups (**Figure 4**). *Gemmatimonas* may participate in nitrogen fixation processes and inhibit plant pathogens in the soil (Abed et al., 2010; Peng et al., 2019). Therefore, the identification of this genus may indicate differential nitrogen fixation process efficacy among different groups and indicates that improving nitrogen fixation efficacy may be one of the biological foundations of companion planting. Other bacterial genera also participate in nitrogen fixation, such as *Sphingomonas* (Xie and Yokota, 2006), *Ramlibacter* (Thanh and Diep, 2014), and *Nocardioides* (Lim et al., 2014). Differential abundance analysis showed that the nitrogen fixation-associated bacteria of the different groups were different. Therefore, nitrogen fixation is of the biological bases and microbiome effects of improving the efficacy of planting by companion planting.

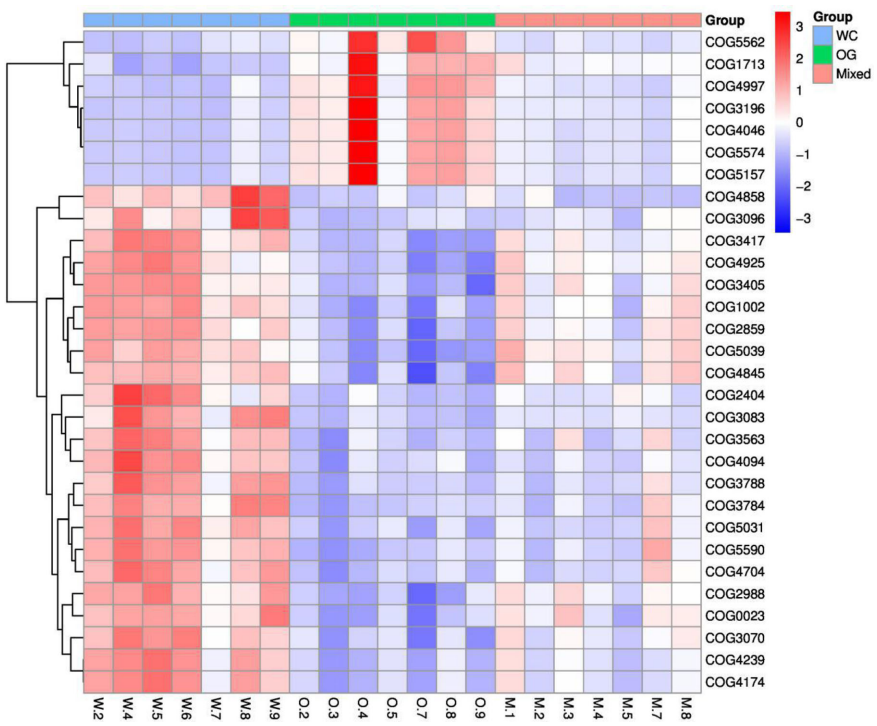


FIGURE 7 | COG functional annotation and differential enrichment analysis of the three groups. The samples were screened for the top enrichment functions using their annotation from the COG database (FDR = 0.0907). The samples from the WC group can be easily distinguished from the OG and Mixed groups. The different functional distributions of the different groups are shown.

Effect of Companion on Functional Differential Enrichment

According to the COG annotation and clustering results, some COG terms have different enrichment patterns in the different groups, especially in the Mixed group. For instance, COG1713 and COG5574 were enriched in the Mixed group, and COG1713 had a high enrichment pattern in the Mixed group. According to the EggNog database (Powell et al., 2014), COG1713 describes the co-enzyme transport and metabolism processes in bacteria, such as *Treponema azotonutricium* ZAS-9. According to an independent study on the symbiotic nitrogen fixation in New Zealand (Reid and Lloyd-Jones, 2009), bacteria plays an effective role in nitrogen fixation in pasture regions. Therefore, the activation of these biological processes may contribute to the improvement of nitrogen fixation.

The other COG term, COG5574, has been supported by Wani et al. (2007). COG5574 describes the post-translational modification, protein turnover, and chaperones involved in various ion binding processes. In 2007, a systematic analysis (Wani et al., 2007) on the molecular genetics of white clover confirmed that the binding of cadmium, chromium, and copper ion is functionally related to nitrogen fixation in this plant. Therefore, the identified biological process is also functionally related to nitrogen fixation processes.

CONCLUSION

We compared the microbiome distribution patterns of planting white clover and orchard grass under single planting and companion planting conditions using 16S rRNA gene sequencing techniques. The analysis results confirmed that the companion planting of white clover and orchard grass can remodel soil microbiome in proximity, especially when compared with the single planting of white clover. We identified a group of differentially distributed microorganisms, such as Gemmatimonadetes. We also identified a group of biological processes, namely, COG1713 and COG5574, using functional annotation and clustering. The screened microorganisms and functional enrichment patterns indicate the specific role of nitrogen fixation effects during companion planting. Therefore, we were able to screen the specific microbiome distribution patterns at the species and functional levels and confirm that nitrogen fixation is one of the most important biological mechanisms for companion planting.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in NCBI (https://www.ncbi.nlm.nih.gov/sra/PRJNA625872).

AUTHOR CONTRIBUTIONS

All authors contributed to the article and approved the submitted version. LC and YZ designed the study. LC and DL performed the experiments. YS, HW, and YL analyzed the results. LC wrote the manuscript.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (31872418), the Natural Science Foundation of Anhui Province (1808085MC60), the Science and Technology Research Projects of Anhui Province (201904b11020043, 201904e01020014), and the National Key Research and Development Program of China (2018YFD1100104).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.538311/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | The rarefaction plots for chao1.

SUPPLEMENTARY FIGURE 2 | Histogram and density map for clean tags length after data cleaning. It's easy to figure out that after data cleaning, all the high-quality reads locate in the range between 400 bp and 450 bp, indicating that the quality control procedure is effective and such clean data is eligible for the downstream analyses.

SUPPLEMENTARY FIGURE 3 | Flower plot for the OTU number distribution pattern among different groups. The numbers in core represent the common OTUs in all samples (*i.e.*, core OTUs), and the numbers on the petals represent the total OTUs of each sample minus the number of common OTUs.

SUPPLEMENTARY FIGURE 4 | Box plot for the Shannon index to evaluate the alpha diversity of different groups. Here, we compared the Shannon index of all the three groups (the mixed for white clover, the OG for orchard grass and the WC for companion planting). Planting only white clover may have lower species diversity comparing to only planting orchard grass and companion planting.

SUPPLEMENTARY FIGURE 5 | Box plot for the Chao (community richness) index to evaluate the alpha diversity of different groups. Here, we further compared the Chao index of all the three groups (the mixed for white clover, the OG for orchard grass and the WC for companion planting). Planting only white clover may have lower species abundancy comparing to only planting orchard grass and companion planting.

SUPPLEMENTARY TABLE 1 | The relative abundance of dominant community to each group on class level (%).Values with different lowercase superscript letters in the same row indicate the existence of a significant difference ($P < 0.05$); values with the same letters indicate no significant difference ($P > 0.05$).

REFERENCES

Abed, R. M., Al Kharusi, S., Schramm, A., and Robinson, M. D. (2010). Bacterial diversity, pigments and nitrogen fixation of biological desert crusts from the Sultanate of Oman. *FEMS Microbiol. Ecol.* 72 (3), 418–428. doi: 10.1111/j.1574-6941.2010.00854.x

Aguilar, O. M., Riva, O., and Peltzer, E. (2004). Analysis of Rhizobium etli and of its symbiosis with wild Phaseolus vulgaris supports coevolution in centers of host diversification. *Proc. Natl. Acad. Sci.* 101 (37), 13548–13553. doi: 10.1073/pnas.0405321101

Aoki, K. F., and Kanehisa, M. (2005). Using the KEGG database resource *Curr. Protoc. Bioinf.* 11 (1), 1.12. 1–1.12. doi: 10.1002/0471250953.bi0112s11

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Bomford, M. K. (2004). *Yield, pest density, and tomato flavor effects of companion planting in garden-scale studies incorporating tomato, basil, and brussels sprout* (Morgantown, WV: West Virginia University).

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi: 10.1023/A:1010933404324

Brown, E. M. (1943). Some effects of soil and air temperatures on the growth of certain grass species. *Sci. Monthly* 57 (3), 283–285. doi: 10.2307/17990

Bybee-Finley, K., and Ryan, M. R. (2018). Advancing intercropping research and practices in industrialized agricultural landscapes. *Agriculture* 8 (6), 80. doi: 10.3390/agriculture8060080

Chalk, P. M. (2018). The role of 15 N-depleted fertilizers as tracers in N cycling studies in agroecosystems. *Nutrient Cycling Agroecosyst.* 112 (1), 1–25. doi: 10.1007/s10705-018-9927-5

Chen, K.-I., Erh, M.-H., Su, N.-W., Liu, W.-H., Chou, C.-C., and Cheng, K.-C. (2012). Soyfoods and soybean products: from traditional use to modern applications. *Appl. Microbiol. Biotechnol.* 96 (1), 9–22. doi: 10.1007/s00253-012-4330-7

Davidson, R. (1969). Effects of soil nutrients and moisture on root/shoot ratios in Lolium perenne L. and Trifolium repens L. *Ann. Bot.* 33 (3), 571–577. doi: 10.1093/oxfordjournals.aob.a084309

Davis, W. (1962). Effects of using oats as a companion crop with orchardgrass, Dactylis glomerata L., and white clover, Trifolium repens L., sown for pasture. *Can. J. Plant Sci.* 42 (4), 582–588. doi: 10.4141/cjps62-100

Deakin, G., Tilston, E., Bennett, J., Passey, T., Harrion, N., Fernández-Fernández, F., et al. (2018). Soil microbiome data of two apple orchards in the UK. *Data Brief* 21, 2042–2050. doi: 10.1016/j.dib.2018.11.067

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, L., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72 (7), 5069–5072. doi: 10.1128/AEM.03006-05

Finch, S., Billiald, H., and Collier, R. (2003). Companion planting–do aromatic plants disrupt host-plant finding by the cabbage root fly and the onion fly more effectively than non-aromatic plants? *Entomol. Experimentalis Applicata* 109 (3), 183–195. doi: 10.1046/j.0013-8703.2003.00102.x

Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2017). Microbial genome analysis: the COG approach. *Briefings Bioinf.* 20 (4), 1063–1070. doi: 10.1093/bib/bbx117

George, D. R., Collier, R. H., and Whitehouse, D. M. (2013). Can imitation companion planting interfere with host selection by Brassica pest insects? *Agric. For. Entomol.* 15 (1), 106–109. doi: 10.1111/j.1461-9563.2012.00598.x

Gołębiewski, M., Deja-Sikora, E., Cichosz, M., Tretyn, A., and Wróbel, B. (2014). 16S rDNA pyrosequencing analysis of bacterial community in heavy metals polluted soils. *Microbial Ecol.* 67 (3), 635–647. doi: 10.1007/s00248-013-0344-7

Graffelman, J., and Camarena, J. M. (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Hum. Hered.* 65 (2), 77–84. doi: 10.1159/000108939

Hagiwara, M., Yoshida, T., and Matano, T. (1995). Effects of companion planting of two common buckwheat varieties on yield and yield concerning characters. *Curr. Adv. Buckwheat Res.* 1, 469–473.

Huttenhower, C., Geversm, D., Knight, R., Abubucker, S., and Badger, J. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486 (7402), 207. doi: 10.1038/nature11234

Kanehisa, M. (2002). The KEGG database. *Found. Symp.* 247, 91–101. doi: 10.1002/0470857897.ch8

Kao, W.-C., Stevens, K., and Song, Y. (2009). BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Res.* 19 (10), 1884–1895. doi: 10.1101/gr.095299.109

Kurilshikov, A., Wijmenga, C., Fu, J., and Zhernakova, A. (2017). Host genetics and gut microbiome: challenges and perspectives. *Trends Immunol.* 8 (9), 633–647. doi: 10.1016/j.it.2017.06.003

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31 (9), 814. doi: 10.1038/nbt.2676

Lim, J. M., Kim, S., Hanmada, M., Ahn, J., Weon, H., Suzuki, K., et al. (2014). Nocardioides daecheongensis sp. nov., isolated from soil. *Int. J. Syst. Evolution. Microbiol.* 64 (12), 4109–4114. doi: 10.1099/ijs.0.063610-0

Liu, K., Xu, Q., Wang, L., Wang, J., Guo, W., and Zhou, M. (2017). The impact of diet on the composition and relative abundance of rumen microbes in goat. *Asian-Australasian J. Anim. Sci.* 30 (4), 531. doi: 10.5713/ajas.16.0353

Mengel, K. (2001). Alternative or complementary role of foliar supply in mineral nutrition. International Symposium on Foliar Nutrition of Perennial Fruit Plants. *Acta. Sci. Pol-Hortoru.* 594, 33–47. doi: 10.17660/ActaHortic.2002.594.1

Moeller, D. A. (2004). Facilitative interactions among plants via shared pollinators. *Ecology* 85 (12), 3289–3301. doi: 10.1890/03-0810

Moore, K. J., Anex, R. P., Elobeid, A. E., Fei, S., Flora, C., Goggi, A., et al. (2019). Regenerating agricultural landscapes with perennial groundcover for intensive crop production. *Agronomy* 9 (8), 458. doi: 10.3390/agronomy9080458

Oyekanmia, E. O., Coyne, D. L., Fagadea, O. E., and Osonubia, O. (2007). Improving root-knot nematode management on two soybean genotypes through the application of Bradyrhizobium japonicum, Trichoderma pseudokoningii and Glomus mosseae in full factorial combinations. *Crop Prot.* 26 (7), 1006–1012. doi: 10.1016/j.cropro.2006.09.009

Parker, J. E., Snyder, W. E., Hamilton, G. C., and Rodriguez-Saona, C. (2013). *Companion planting and insect pest control. Weed and Pest Control-Conventional and New Challenges* (IntechOpen).

Payne, K. M. (2019). Enhanced Efficiency Nitrogen Formulation Effect on Grass-legume Pasture Productivity. [dissertation/master's thesis]. Lexington (Kentucky): University of Kentucky.

Peng, C., Gao, Y., Fan, X., Peng, P., Huang, H., and Zhang, X. (2019). Enhanced biofilm formation and denitrification in biofilters for advanced nitrogen removal by rhamnolipid addition. *Biores. Technol.* 287, 121387. doi: 10.1016/j.biortech.2019.121387

Pitombo, L. M., Carmo, J., Hollander, M., Rossetto, R., Maryeimy, V., and Cantarella, H. (2016). Exploring soil microbial 16S rRNA sequence data to increase carbon yield and nitrogen efficiency of a bioenergy crop. *Gcb Bioenergy* 8 (5), 867–879. doi: 10.1111/gcbb.12284

Plaza, L., Ancos, B., and Cano, P. (2003). Nutritional and health-related compounds in sprouts and seeds of soybean (Glycine max), wheat (Triticum aestivum. L) and alfalfa (Medicago sativa) treated by a new drying method. *Eur. Food Res. Technol.* 216 (2), 138–144. doi: 10.1007/s00217-002-0640-9

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., et al. (2014). eggNOG v4. 0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* 42 (D1), D231–D239. doi: 10.1093/nar/gkt1253

Reid, N. M., and Lloyd-Jones, G. (2009). Symbiotic nitrogen fixation in the New Zealand dampwood termite (Stolotermes ruficeps). *New Z. J. Ecol.* 33 (1), 90. doi: 10.1128/AEM.05609-11

Reina-Bueno, M., Argandoña, M., Nie, J., Hidalgo-García, A., Iglesias-Guerra, F., Delgado, M., et al. (2012). Role of trehalose in heat and desiccation tolerance in the soil bacterium Rhizobium etli. *BMC Microbiol.* 12 (1), 207. doi: 10.1186/1471-2180-12-207

Rognes, T., Flouri T., Nichols, B., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, e2584. doi: 10.7717/peerj.2584

Rojewski, J., Lee, I. H., and Gemici, S. (2012). Use of t-test and ANOVA in career-technical education research. *Career Tech. Educ. Res.* 37 (3), 263–275. doi: 10.5328/cter37.3.263

Rosenblueth, M., and Martínez-Romero, E. (2004). Rhizobium etli maize populations and their competitiveness for root colonization. *Arch. Microbiol.* 181 (5), 337–344. doi: 10.1007/s00203-004-0661-9

Samad, A., Trognitz, F., Compant, S., Antonielli, L., and Sessitsch, A. (2017). Shared and host-specific microbiome diversity and functioning of grapevine and accompanying weed plants. *Environ. Microbiol.* 19 (4), 1407–1424. doi: 10.1111/1462-2920.13618

Segal, M. R. (2004). "Machine learning benchmarks and random forest regression," in *Machine learning benchmarks and random forest regression, Technical report, eScholarship Repository* (University of California). Available at: http://escholarship.org/uc/item/35x3v9t4.

Smets, W., Leff, J., Bradford, M., McCulley, R., Lebeer, R., and Noth F. (2016). A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biol. Biochem.* 96, 145–151. doi: 10.1016/j.soilbio.2016.02.003

Sood, V., Chaudhary, H. K., Kumari, A., Singh, H. P., Devi, R., and Sharma, A. (2018). Genetic improvement of temperate grasses and legumes in indian himalayan region: A review. *Int. J. Curr. Microbiol. Appl. Sci.* 7 (6), 3454–3463. doi: 10.20546/ijcmas.2018.706.405

Szafirowska, A., and Kolosowski, S. (2008). The effect of companion plants on Lygus feeding damage to bean. *The effect of companion plants on Lygus feeding damage to bean. Poster at: Cultivating the Future Based on Science: 2nd Conference of the International Society of Organic Agriculture Research ISOFAR,* (Modena, Italy), 442–445.

Thanh, D. T. N., and Diep, C. N. (2014). Isolation and identification of rhizospheric bacteria in Acrisols of maize (Zea mays L.) in the eastern of South Vietnam. *Am. J. Life Sci.* 2 (2), 82–89. doi: 10.11648/j.ajls.20150302.18

Wang, Q., Garrity, G., Tiedje J., and Cole, J. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73 (16), 5261–5267. doi: 10.1128/AEM.00062-07

Wani, P. A., Khan, M. S., and Zaidi, A. (2007). Cadmium, chromium and copper in greengram plants. *Agron. Sustain. Dev.* 27 (2), 145–153. doi: 10.1051/agro:2007036

Xie, C. H., and Yokota, A. (2006). Sphingomonas azotifigens sp. nov., a nitrogen-fixing bacterium isolated from the roots of Oryza sativa. *Int. J. Syst. Evolution. Microbiol.* 56 (4), 889–893. doi: 10.1099/ijs.0.64056-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 29 September 2020
doi: 10.3389/fbioe.2020.553904

On the Logical Design of a Prototypical Data Lake System for Biological Resources

*Haoyang Che and Yucong Duan**

College of Information Science and Technology, Hainan University, Haikou, China

Biological resources are multifarious encompassing organisms, genetic materials, populations, or any other biotic components of ecosystems, and fine-grained data management and processing of these diverse types of resources proposes a tremendous challenge for both researchers and practitioners. Before the conceptualization of data lakes, former big data management platforms in the research fields of computational biology and biomedicine could not deal with many practical data management tasks very well. As an effective complement to those previous systems, data lakes were devised to store voluminous, varied, and diversely structured or unstructured data in their native formats, for the sake of various analyses like reporting, modeling, data exploration, knowledge discovery, data visualization, advanced analysis, and machine learning. Due to their intrinsic traits, data lakes are thought to be ideal technologies for processing of hybrid biological resources in the format of text, image, audio, video, and structured tabular data. This paper proposes a method for constructing a practical data lake system for processing multimodal biological data using a prototype system named ProtoDLS, especially from the explainability point of view, which is indispensable to the rigor, transparency, persuasiveness, and trustworthiness of the applications in the field. ProtoDLS adopts a horizontal pipeline to ensure the intra-component explainability factors from data acquisition to data presentation, and a vertical pipeline to ensure the inner-component explainability factors including mathematics, algorithm, execution time, memory consumption, network latency, security, and sampling size. The dual mechanism can ensure the explainability guarantees on the entirety of the data lake system. ProtoDLS proves that a single point of explainability cannot thoroughly expound the cause and effect of the matter from an overall perspective, and adopting a systematic, dynamic, and multisided way of thinking and a system-oriented analysis method is critical when designing a data processing system for biological resources.

Keywords: data lake, DIKW, biological resources, unstructured data, XAI, explainability, interpretability

OPEN ACCESS

Edited by:
Xiyin Wang,
North China University of Science and Technology, China

Reviewed by:
Marco Brandizi,
Rothamsted Research, United Kingdom
Simone Marini,
National Research Council (CNR), Italy

***Correspondence:**
Yucong Duan
duanyucong@hotmail.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *20 April 2020*
Accepted: *28 August 2020*
Published: *29 September 2020*

Citation:
Che H and Duan Y (2020) On the Logical Design of a Prototypical Data Lake System for Biological Resources. Front. Bioeng. Biotechnol. 8:553904. doi: 10.3389/fbioe.2020.553904

Abbreviations: AGI, Artificial general intelligence; AI, Artificial intelligence; AM, Algorithm metadata; AV, Algorithm visualization; BioBPX, Biological Pathway Exchange; DG, Data governance; DI, Data ingestion; DIKW, Data–information–knowledge–wisdom; DL, Data lake; DM, DIKW metadata; DP, Data pond; DPV, DIKW provenance visualization; DS, Dialogue system; EI, Explainable infrastructure; ELM, Extreme learning machine; EML, Explainable machine learning; HCLS, Health care and life sciences; IM, Infrastructure metadata; IV, Infrastructure visualization; JS, Job scheduler; KG, Knowledge graph; LIME, Local Interpretable Model-agnostic Explanations; MC, Metadata catalog; ML, Machine learning; MM, Mathematics metadata; NA, Narrator; PCA, Principal component analysis; ProtoDLS, Prototypical Data Lake System; SMV, Software Metrics Visualization; SP, Security and privacy; SRM, Software runtime metrics; ST, Sandbox training; SVD, Singular value decomposition; TA, Twin agent; TDW, Traditional data warehouse; VI, Visualization; XAI, Explainable artificial intelligence.

INTRODUCTION

Biological resources encompass a vast range of organisms (and parts thereof), the genetic materials they contain (known more specifically as genetic resources), and any other biological components of a population or ecosystem that has an actual or potential use or value for human beings. The digitization of biological resources has created a large volume of biological big data; however, those data are possibly coming from multiple sources and are heterogeneous, and in order to make an actionable decision, there needs to be a trustful data integration and an integrated analytical solution. Research scientists in bioinformatics around the world have pulled all out of their efforts to collaboratively solve the challenging problems (Afgan et al., 2016; da Veiga Leprevost et al., 2017; Bussery et al., 2018). A data lake is claimed to be capable of fulfilling descriptive analytics, exploratory analytics, and confirmatory factor analytics requirements in other application fields, yet it has not been introduced in bioinformatics or genetics to store large quantities of biological resource data or experimental data in a massive way.

As a newly emerging paradigm of modern data architectures, the data lake radically simplifies the enterprise-wide data infrastructure, and it is expected to accelerate technological innovation alongside with the deep penetration of artificial intelligence and machine learning capabilities into every industrial and social sector. In the past, almost all of the data involved in the operational products and the decision-making products come from structured data stored in the back-end databases or data warehouses, or semi- or unstructured data crawled from the Web, and nowadays, many innovative products are embedding AI in the unstructured data format of computer vision, speech recognition, and text mining. These new requirements differ a great deal from the requirements emerging in the era of data warehouses, which need a structured, subject-oriented, and relational database claiming to hold a single view of data without data silos (Stein and Morrison, 2014). For example, over the years, Irvine Medical Center, University of California, had accumulated a pile of patient health records owned by one million inpatients and outpatients. These different types of data included online spreadsheet data, semi-structured medical reports, unstructured prescriptions, and radiology images from its radio department. The medical center had to store, integrate, and access the big data, so they chose to use a Hadoop distribution as their initial data lake infrastructure, for the benefit of Hadoop open-source software stacks and low-price commodity hardware clusters (Stein and Morrison, 2014). When it comes to the processing of biological resources, as research institutions, labs, and pharmaceutical plants increasingly use mobile apps and cloud services, the application scenarios will be somewhat similar to what they have experienced at the UC Irvine Medical Center.

Due to their intrinsic traits, data lakes are thought to be ideal technologies for processing of hybrid biological resources in the format of text, image, audio, video, and structured tabular data. Unfortunately, facing with these voluminous and heterogeneous data, current data lake proposals cannot afford the system complexity and high tolerance for human errors, due mostly to their incipient design and low explainability. However, some research directions and application scenarios have received special attention on the explainability due to their specialties and critical states, especially those in medicine and pharmacy pertaining to human lives where decisions are literally a matter of life or death. Biology as a discipline also concerns much on the explainability of biological phenomena and effects. Thus, the data management of biological resources urgently needs to solve the following two problems: (i) efficient and effective management of heterogeneous data from multiple sources and (ii) reasonable explanation of applications running on the platform in terms of the overall system design. Usually, explainability cares more about the Explainable Artificial Intelligence and Machine Learning (XAI and ML) algorithm (Samek et al., 2017) and recommender systems (Schafer et al., 1999; Zhang et al., 2014). However, we consider that explainability is a very broad term that still includes engineering-related aspects like the data/information/knowledge/wisdom spectrum, or DIKW (Duan et al., 2019), network architecture, and development language, human-related aspects like human faults, and cognitive psychology, not just algorithm and mathematics-related aspects. A single point of explainability cannot thoroughly expound the cause and effect of the matter from an overall perspective; we must adopt a systematic view and system-oriented analysis method.

Also, the data lake approaches may learn lessons and experiences from other similar approaches, which are possibly coming from different application domains, for example, the virtual research environment approaches (Assante et al., 2019; Houze-Cerfon et al., 2019; Remy et al., 2019; Albani et al., 2020). As is known to all, the integration of domain knowledge from different application domains will bring different perspectives to the data lake solutions.

Consequently, we propose in this paper the following:

1. to construct a practical data lake system for processing multimodal biological data using a prototype system named ProtoDLS;
2. to adopt a horizontal pipeline to ensure the intra-component explainability factors from data acquisition to data presentation, and a vertical pipeline to ensure the inner-component explainability factors including mathematics, algorithm, execution time, memory consumption, network latency, and sampling size.

In order to better understand the meaning of explainability from the outset, in here we give a brief definition of explainability and interpretability (Adadi and Berrada, 2018; Arrieta et al., 2020).

Definition 1 Explainability denotes an account of the system, its workings, and the implicit and explicit knowledge it uses to arrive at conclusions in general and the specific decision at hand, which is sensitive to the end-user's understanding, context, and current needs.

Definition 2 Interpretability denotes the extent to which a cause and effect can be observed within a system. Or, to put it in another way, it is the extent to which you are able to

predict what is going to happen, given a change in input or algorithmic parameters.

In the context of this article, explainability and interpretability are used interchangeably.

The reminder of this paper is organized as follows. Section "Related Work" briefly surveys the current *status quo* of data lakes and XAI, as well as the research development of data management approaches in the field of bioinformatics, genetics, and phenomics. Section "The Prototype Architecture" presents the overall architecture of our prototype system, i.e., ProtoDLS and describes the specific features of each key component. Sections "Horizontal Pipeline" and "Vertical Pipeline" elucidate the detailed design especially from the point view of explainability in a horizontal and vertical pipeline, respectively. Section "Project Progress and Discussion" discusses the current project progress and makes a contrast between ProtoDLS and emerging data lake systems. In closing, Section "Conclusion and Future Work" comes to a conclusion of the paper and outlines future development directions about ProtoDLS.

RELATED WORK

In 2010, James Dixon, CTO of Pentaho (2020), firstly proposed the concept of data lake in one blog post, as a way trying to store voluminous and diversely structured data in their native formats, in an evolutionary storage place allowing later detailed analyses (Dixon, 2010). Although the concept was first coined in early 2010, academia adopted it a couple of years later. Until now, there has been no well-accepted definition of what a data lake is, and the corresponding underlying features vary differently according to the real-world contexts. Some early research advancements on data lakes for a time were ever bound up with on-demand data models, or widely called schema-on-read models (also known as late binding models) (Fang, 2015; Miloslavskaya and Tolstoy, 2016). The key reason for adopting the schema-on-read model in data lakes lies in the bulk workloads of manual schema extraction, which is inoperable in the face of machine learning tasks, especially deep learning tasks. At the same time, Suriarachchi and Plale (2016) found that with the continuing growth of data in top gear, a data swamp will soon appear from a meant-to-be data lake without the guidance of a clear-cut schema. Thus, to ensure data accessibility, exploration, and exploitation, an efficient and effective metadata system becomes an indispensible component in data lakes (Quix et al., 2016). Yet, most of the research work on data lakes still concentrate on structured data, or semi-structured data only (Farid et al., 2016; Farrugia et al., 2016; Madera and Laurent, 2016; Quix et al., 2016; Klettke et al., 2017). So far, unstructured data have not received enough consideration in the relevant research literature, while more often than not unstructured heterogeneous data occur frequently (Miloslavskaya and Tolstoy, 2016). Multimodality in data lake systems is estimated to come under the spotlight in the next research wave.

Almost at the same time, with the development of big data and deep learning, especially since the totemic year of 2012, AI algorithms have attained or surpassed the limits of human beings in many areas like chess games and drug discovery, which were computationally unimaginable in early years (Lecun et al., 2015). However, some black-box models like random forest (Breiman, 2001), GBDT (Friedman, 2001), and deep learning (Lecun et al., 2015) have extraordinarily complex inner working mechanisms and inexplicable outer input–output mappings. Even for a senior graduate student, to fully understand the rationale of a black-box model will cost him several days and make him go through a painful process of a conscientious manual formula derivation and a time-consuming experimental verification. The problem with these models is that they are devoid of transparency and explainability, although they will nearly gain superior performance after careful fine-tuning. In the healthcare and medical field, that would become a big problem since applications in these demanding fields require a full-fledged explanation of model rationales. Thus, research efforts in these fields have witnessed a burst of articles and papers in explainable artificial intelligence (XAI) (Došilović et al., 2018). Since XAI methods have extensive application scenarios, a full survey of XAI research and development is a difficult task to accomplish. On a large scale, the related research topics in XAI can be roughly divided into two major categories: integrated approaches and *post-hoc* approaches.

The integrated approaches usually keep an eye on the transparency factors, and transparency is a required means for the protection of human rights from unfairness and discrimination (Edwards and Veale, 2017). Similar to the idea, transparent models are expected to be both explainable and interpretable. As one of its subbranches, pure transparent approaches restrict the model choices to the model families that are considered transparent. For example, Himabindu et al. (2016) ever proposed a method to use separate if-then rules to effectively interpret decision-making sets. Based on region-specific predictive models, Wang et al. (2015) proposed an oblique treed sparse additive model, which exchanges a modest measure of interpretability for accuracy, but in SVM and some other non-linear models, it gains a satisfying degree of accuracy. As another subbranch, hybrid approaches combine pure transparent models and black-box models to get a balance between interpretability and performance. To develop internal rating models for banks, Gestel et al. (2005) used a progressive method balancing the requirements of predictability and interpretability.

Post-hoc approaches will not impact the model performance since it extracts information from the already learned model. Usually, *post-hoc* approaches are used in cases where model mechanisms are too complex to explain. For example, as for explainable recommendation, two diverse models generate recommendations and explanations, respectively. After the genuine recommendations have been performed, an explanation model independent of the recommendation algorithms will provide explanations for the recommendation model carried out just a while ago (so it is called as "*post-hoc*"). Likewise, to provide a *post-hoc* explainability for recommendations, Peake and Wang also presented a data mining method with several association rules (Peake and Wang, 2018). In addition to recommendation, *post-hoc* approaches were also used in

image recognition and text classification. To find out model defects in these fields, with the aid of several elastic nets, Guo et al. (2018) augmented a Bayesian regression mixture model and extracted explanations for a target model through global approximation.

XAI has received relatively little attention in the field of bioinformatics and biology, but ontology-based data management in this line has gleaned quite a few studies. In phenomics research, aiming to support adequate collaboration between teamworkers, Li Y.F. et al. (2010) presented PODD, a data reservoir based on ontology. Like its big brother, genomics, phenomics research uses imaging devices and measurement apparatuses to acquire vast amounts of generated data, which are subsequently used for analysis. Thus, in phenomics research, there are key challenges for data management of large amounts of raw data (image, video, raw text). Meanwhile, in genomics, Ashburner et al. (2000) constructed the famous Gene Ontology (Gene Ontology, 2020), a well-established and structured tool to represent gene ontology categories and terms, which has been successfully used for many years by researchers. The Gene Ontology includes three independent ontologies, molecular function ontology, cellular component ontology, and biological process, and can be used for all eukaryotes, even as we are gaining more knowledge of protein and gene functions in cells (Ashburner et al., 2000). The Bio2RDF (2020) project declares to transform silos of life science data into a globally distributed network of linked data for biomedical knowledge translation and discovery. Up until now, Bio2RDF has accumulated 378 datasets, including Bio2RDF:Drugbank and Bio2RDF:Pubmed. The EBI RDF platform (Ebi Rdf, 2020) claims to bring together a number of EMBL-EBI resources that provide access to their data using Semantic Web technologies. As a well-accepted exchange language, BioPAX (Biological Pathway Exchange) aims to enable integration, exchange, visualization, and analysis of biological pathway data (Demir et al., 2010). OntoLingua provides a distributed collaborative environment to browse, create, edit, modify, and use ontologies (OntoLingua Server, 2020). The BioSchemas project develops different types of schemas for the exchange of biological data and aims to reuse existing standards and reach consensus among a wide number of life science organizations (BioSchemas, 2020). Although closed, W3C's HCLS (Health Care and Life Sciences) group has done a great deal of work to use Semantic Web technologies across health care, life sciences, clinical research, and translational medicine. Amid questions about its feasibility and availability, IBM Watson brings to customers a cognitive computing platform which can understand, reason, and learn from a magnitude of unstructured medical literature, patents, genomics, and chemical and pharmacological data (Ying et al., 2016).

Apart from ontology research in biology, at different times, a collection of databases such as scientific publications (PubMed, 2020), genes (Ensembl, 2020), proteins (UniProt, 2020), and gene expression data (Ebi ArrayExpress, 2020; Gene Expression Atlas, 2020; GEO, 2020) have ever been created in order to store big quantities of bio-data for the purpose of refining and for systematic scientific research work. These data storage platforms have seldom based on data lakes since data lakes are not as mature as commercial databases and data warehouses; neither are open-source data management solutions like Hadoop software stack.

At the intersection of data lakes and explainability, research on explainable data lakes still remains unexplored. Also, for now, barely little literature in the field of data lakes has discussed explainability systematically. This paper tries to fill the gap between data lakes and explainability from a systematic view, not just a XAI view, and to borrow knowledge and experience from the research development on XAI and data lakes.

THE PROTOTYPE ARCHITECTURE

In this section, we will present the overall architecture of ProtoDLS (Prototypical Data Lake System for Biological Resources) we have designed. In the field of bioinformatics and genetics, ProtoDLS intends to answer the explainability problem in a systematic, dynamic, and multisided view instead of an isolated, static, and one-sided view. In order to explain certain questions about data, metrics, rules, and business objectives, ProtoDLS insists that only every component and module is self-explanatory itself, the unhindered explainability can be thoroughly implemented in the system level as a whole, and only after that, explainability can take real effect and solve real-world problems. ProtoDLS also disbelieves a single point of explainability such as XAI for that XAI also has input into, output out of, and interactions with other components, modules, or even machine learning algorithms in a data lake system as in **Figure 1**.

As in **Figure 2**, the overall architecture of ProtoDLS can be roughly divided into two major components: **Data Lake** (DL, A in **Figure 2**) and **Traditional Data Warehouse** (TDW, B in **Figure 2**). Initially debuted as a substitute for data marts in the topmost tier of data warehouses, data lakes have exhibited a relationship of complement to data warehouses rather than a competitive relationship with data warehouses. The complementary strengths and challenges between them in recent years also suggest the urgent needs to exchange ideas on opportunities, challenges, and cutting-edge techniques within them. In ProtoDLS, TDW is usually used to cleanse, integrate, store, and analyze the processed, trusted, and well-structured data or semi-structured data like website logs. Raw data is always discarded or stored in a NAS/SAN/Cloud storage area. TDW and DL transfers data back and forth; sometimes, DL can serve as a staging area for TDW, and vice versa. DL stores raw data in any format and outputs the deeply analyzed results in a schematic format to TDW for visualization, reporting, and *ad hoc* query. TDW also outputs some structured data to DL as its metadata and elementary elements. The detailed data flow between them is stored in **Metadata Catalog** (MC, A-6 in **Figure 2**) of DL for later explanation and traceability.

The **Data Ingestion** (DI, A-1 in **Figure 2**) component of ProtoDLS provides an appropriate data extraction, integration, transformation, and load mode for multiple heterogeneous data sources. DI has the following features:

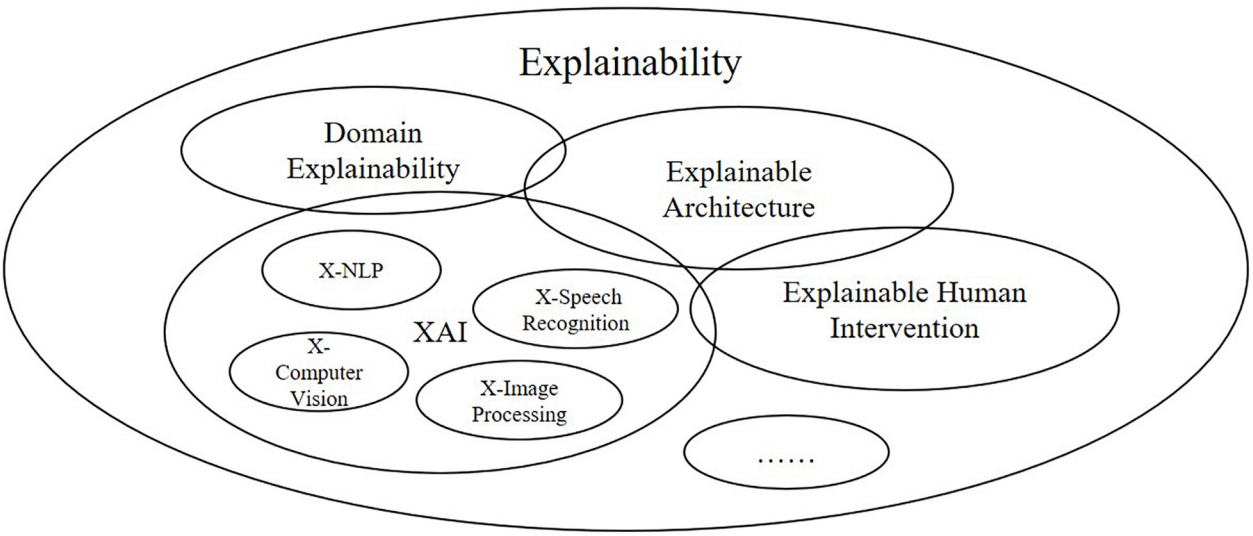


FIGURE 1 | Explainability is a comprehensive concept and XAI is just an aspect of explainability.

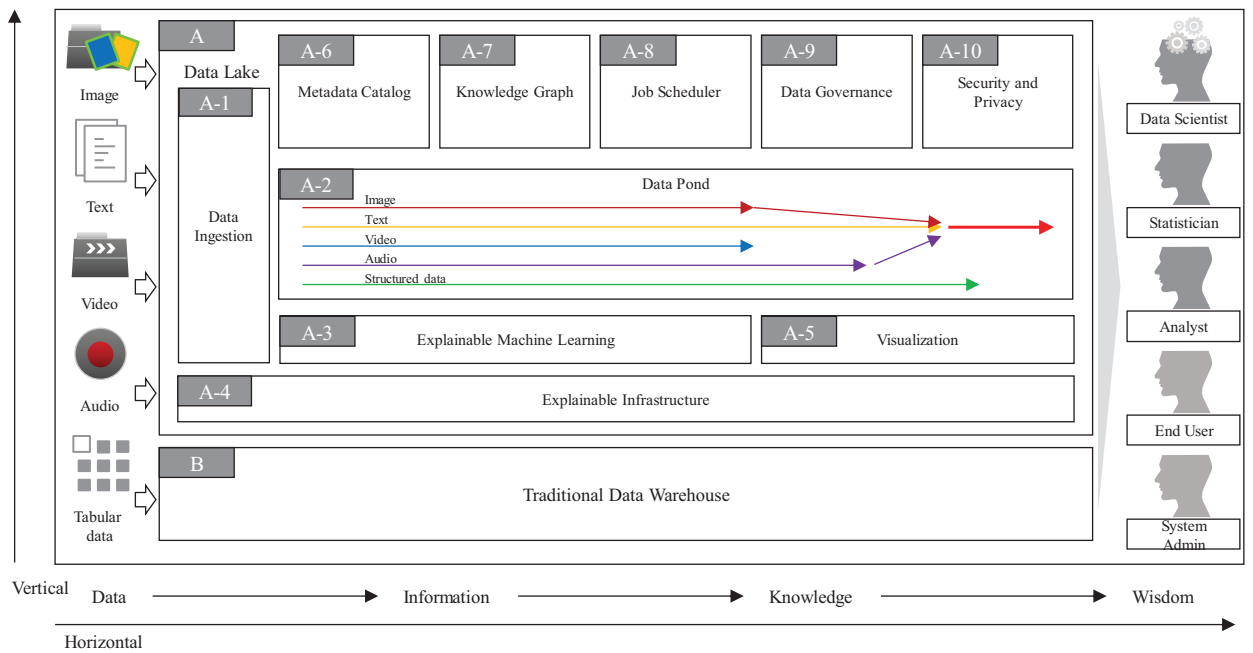


FIGURE 2 | The overall architecture of ProtoDLS.

1. **Data source configuration:** to support multiple data sources, including but not limited to TDW, databases, flat files, message queues, and protocol datagrams.
2. **Data collection:** to support the collection actions of the corresponding data source, and complete data structure analysis, data cleaning, data transforming, data normalization, data format standardization, etc.
3. **Data synchronization:** to support data synchronization to other data sources, including necessary cleaning, processing, and transforming.
4. **Data distribution:** to support data sharing and distribution, and publish data in various forms (object stores, APIs, etc.).
5. **Data preprocessing:** to support data encryption, desensitization, standardization, and other particular processing logic.

The **Explainable Infrastructure** (EI, A-4 in **Figure 2**) component of ProtoDLS is slightly different from the traditional infrastructure layer. EI is also composed of network unit, storage

unit, and computing unit. The generated data are all collected in these units, such as memory consumption by second, network latency by second, storage capacity by hour/day. The warning, alert, or other important system admin event will be triggered and displayed in an intuitive way, like using NLG (Natural Language Generation).

The **Visualization** (VI, A-5 in **Figure 2**) component of ProtoDLS empowers other components with visualization capabilities. With visualization, horizontal components and vertical modules can enhance explainability of DIKW flow (see below), statistical algorithms, and deep learning algorithms.

In one respect, the design of a data lake platform is fundamentally metadata driven, especially in terms of explainability. The **MC** component of ProtoDLS is very critical to explainability, since it globally stores all of the metadata generated locally in every component, of which type includes technical metadata, business metadata, and operational metadata. MC stores every and each data change and schema change. MC can represent metadata in tabular forms or human understandable sentences supported by the **Explainable Machine Learning** (EML, A-3 in **Figure 2**) component. MC provides a system-wide single point of truth for all kinds of users in ProtoDLS.

The **Knowledge Graph** (KG, A-7 in **Figure 2**) component of ProtoDLS visualizes knowledge entities and the relationships between the entities in a graph model. With the help of KG, EML, or any other component in ProtoDLS can extract named entity, relationship, and attributes from it. Knowledge representation, knowledge fusion, entity disambiguation, and knowledge/ontology reasoning in KG can enhance explainability in other components. The question and answer feature is critical to explainability, and KG can provide accurate and concise natural language abilities to aid it.

The **Job Scheduler** (JS, A-8 in **Figure 2**) component of ProtoDLS schedules jobs for execution (start, stop, terminate, invocate, replay, or sleep) at a specific time/date, or triggers jobs upon receiving some certain event, or records the execution orders of jobs and running status within jobs in DC. JS orchestrates the running jobs in ProtoDLS in a sequential or concurrent order, and its scheduling trigger scheme includes time-based, interval-based, and event-based.

In contrast to TDW, the biological data maintained in DL are more scattered, disordered, and schema-less, so it is more necessary to govern the data usability, availability, integrity, security, and flows in DL through the work of **Data Governance** (DG, A-9 in **Figure 2**), otherwise DL will gradually become corrupted and finally transforms into a data swamp. To efficiently and effectively drive data intelligence, DG is crucial and it is also one of the biggest challenges during the construction of DL. The core task of DG lies in improving the multimodal data quality by the aid of metadata management, data standard conformance, data lifecycle management, data security and privacy management, and data stewardship. Without the aid of DG, low-quality data will greatly lower the precision and recall of machine learning algorithms and thus will further restrict the interpretability and explainability of ProtoDLS as a whole.

The **Security and Privacy** (SP, A-10 in **Figure 2**) component of ProtoDLS deals with security and privacy issues since ProtoDLS will be flooded by the influx of numerous raw and unprocessed data, which will be very dangerous without some appropriate supervision, audit, and access control methods. Privacy preserving data mining can protect personal privacy data from leakage and damage, improve explainability, and reduce bias.

The **Data Pond** (DP, A-2 in **Figure 2**) component of ProtoDLS subdivides and processes the data exported by DI according to the incoming data format. ProtoDLS needs to provide a variety of data analysis engines to meet the needs of data computing. It needs to meet batch, real-time, streaming, and other specific computing scenarios. In addition, it also needs to provide access to massive data to meet the demand of high concurrency and improve the efficiency of real-time analysis. Heterogeneous data enters into DP according to the dispatch of JS. Initially, text data enter into text DP, and image data enter into image DP, and so on. When multimodality analysis is set, different types of data may enter into a hybrid DP, for example, text data and image data may enter into text-image DP for later coordinated processing. The partition of DP over data formats ensures the explainability and traceability in DP.

The EML component of ProtoDLS is responsible for executing NLP, image classification, video classification, audio recognition, and conventional machine learning and deep learning algorithms in an explainable way. The methods for explainability may include example illustration, analogy, visualization, model-agnostic, local approximation, or even human intervention.

Potentially, ProtoDLS has a wide range of platform users including system administrators, data scientists, statisticians, analysts, and ordinary end users, who have different explainability demands for ProtoDLS. For discovery and ideation, data scientists will currently focus more on the explainability of the black-box deep learning algorithms. Statisticians will pull all out to explore data patterns and identify data rules through tests, summaries, and higher-order statistics under some hypothesis. Data analysts may cost their efforts to explain the business intelligence metrics in their everyday life. System administrators will pay attention to the normal operation of ProtoDLS. When the system is down or a performance degradation occurs, EI in ProtoDLS will give system administrators an easy-to-understand explanation and system administrators will rephrase the explanation in less technical terms to other users of ProtoDLS, in order to mitigate the user anxieties and confusion. Ordinary end users usually are not technical experts in the abovementioned areas, and all they want is an easy-to-understand explanation. According to the explanation, they will make decisions and enact policies and rules. However, the requirement creates the most difficult part of explainability in ProtoDLS, since the generated explanation by the platform must be presented in an intuitive way prone to human understanding, without many technical terms or nomenclatures. ProtoDLS accumulates all the explanations in every component and module and ranks them in an important or critical order, and ProtoDLS will synthesize them into a paragraph that human can easily understand and accept. The training procedures will absorb insights and suggestions from experts in bioinformatics, genetics,

and phenomics, in algorithms, in computer architecture, or even in cognitive psychology.

ProtoDLS aims to help researchers and practitioners in bioinformatics, genetics, and phenomics finish the following tasks in an explainable way:

1. Multimodality data governance over datasets collected from biological resources, including data standard conformance, data security enforcement, metadata management, data quality improvement, data stewardship, and data lifecycle management, with an aim to reduce the difficulties of data analytics without data lakes.
2. Multimodality data exploration and exploitation using cutting-edge machine learning, deep learning, and artificial intelligence techniques, on the basis of ingesting, aggregating, cleaning, and managing datasets maintained in ProtoDLS.
3. To generate new data dimensions based on the analysis of previous usage histories.
4. To create a centralized multimodality data repository for data scientists and data analysts, etc., which is conducive to the realization of a data service optimized for data transmission.

HORIZONTAL PIPELINE

In ProtoDLS, the horizontal pipeline mainly concerns about when, where, how, by who, and to what degree the large amounts of instantly collected raw data are being transformed into meaningful information, useful knowledge, even insightful wisdom for all kinds of end users. In the horizontal direction, ProtoDLS is divided into several components according to specific functional requirements. Thus, ProtoDLS adopts a horizontal pipeline to ensure the intra-component (or subsystem-level) explainability across the horizontal landscape, from data acquisition, to data storage, all the way up to data processing, and finally to data presentation.

Data, Information, Knowledge, Wisdom

ProtoDLS observes and manages the data flow between horizontal units in light of the conceptual framework of DIKW (Duan et al., 2019). As seen in **Figure 3**, the DIKW model integrates data, information, knowledge, and wisdom in a set of related layers, each extending the ones underneath itself. The original observation and measurement activities obtain the raw data, in the format of image/text/video/audio, and the relationships between the raw data are analyzed accordingly to obtain the information, which in ProtoDLS is the integrated and formatted data from the raw data according to some ETL processing rules. The application of information in action produces knowledge, which in ProtoDLS is information applied to bioinformatics, genetics, and phenomics. Wisdom is concerned about the future, and it tries to understand things that have not been understood in the past, things that have not been done in the past.

Finding out how data, information, knowledge, and wisdom flow between components should act as the first step toward the complete explainability of a specific question posed toward ProtoDLS, since quick location of the need-to-explain points will be realized in terms of clear-and-cut DIKW flows.

ProtoDLS records any DIKW flow between every two components in MC in the following format: <Task_Name, Source_Component_Name | External_Source_Name, DIKW, Sink_Component_Name | External_Source_Name, Last Execution Time, Duration, Executed By >, i.e., DIKW flows from a source component or an external source into a sink component or flows out into an external source. MC can give back the DIKW flow path for any request from other components. System administrators can monitor and retrieve the DIKW flow in a single operational console. When incidents occur, DIKW flow monitoring capabilities can give admin teams a quick start and a good explanation for other platform users.

DIKW Provenance

Provenance was a concept originated from the database community decades ago (Buneman et al., 2001). The emergence of data provenance, or data lineage, is that database or data warehouse users need to find out the data origin and the data evolution process, where they are coming from, where they are going, and what is happening to them; also, they need to frequently execute impact analyses in order to make sure that certain actions to be performed will affect the system in a controlled way and within a controlled range, or trace back data quality issues and errors till to their root causes as fast as they can. Or, on the other hand, many senior technical users like data scientists and data analysts tend to use datasets in isolation or in a team, which may quickly create some explicit or implicit upstream and downstream dependencies and chaining of dependent data processing. In this regard, the system or the platform need to cover a broad spectrum of workload scenarios like batch jobs, streaming jobs, mini-batch queries, *ad hoc* queries, deep learning training tasks, and support programming languages like R, Python, and Scala, and even new programming languages like Julia. To perform provenance on the data lake, we need DIKW provenance as an upgraded version in place of data provenance. With DIKW provenance, the ProtoDLS users can track and understand how DIKW flows across the platform at every stage, where DIKW resources are sourced from, and how they are being consumed, thus allowing users to develop trust and confidence in the platform, algorithms, infrastructure, and other inner working mechanisms of ProtoDLS.

Basically, DIKW provenance has the following categories:

1. What—provenance answers the question: what does this do?
2. Who—provenance answers the question: who did this?
3. When—provenance answers the question: when did this happen?
4. Where—provenance answers the question: where did this happen?
5. How—provenance answers the question: how the knowledge is worked out?
6. Why—provenance answers the question: why the result is working?

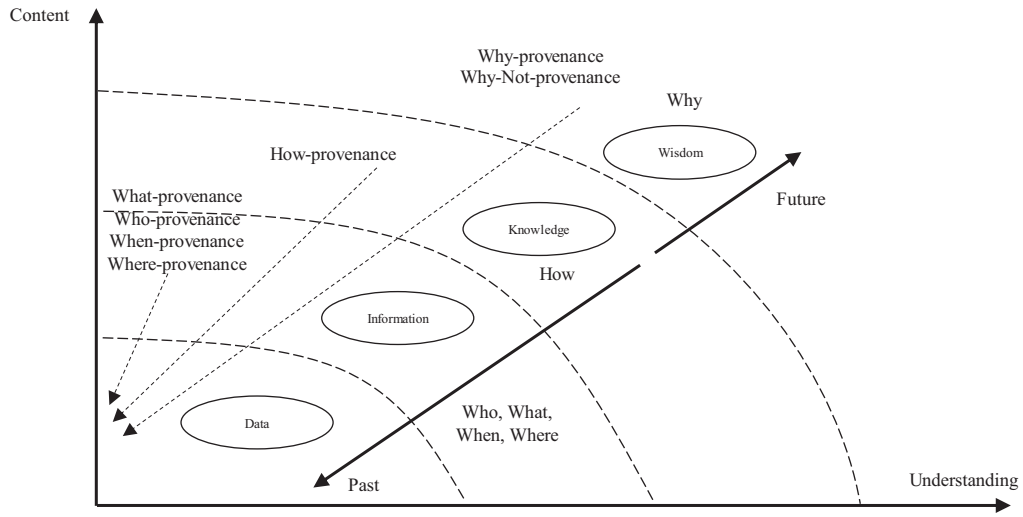


FIGURE 3 | The DIKW provenance.

7. Why Not—provenance answers the question: why the result is not achieved?

With DIKW provenance, many questions related to the horizontal pipeline can be answered and thus explainability on a horizontal level will be achieved to some extent in ProtoDLS. As seen in **Figure 4**, the DIKW provenance flows can be recorded in a module named **DIKW Metadata** (DM) in MC, and the **DIKW Provenance Visualization** (DPV) module in VI is responsible for replaying the provenance flows in a reverse direction using animation.

VERTICAL PIPELINE

In ProtoDLS, the vertical pipeline concerns more about the explainability in every component rather than the intra-component explainability. For example, the fusion algorithm implemented in DP may require an explanation mainly in the scope of DP, rather than an explanation of the DIKW flows between DI and DP which has been explained and can be queried in the horizontal pipeline.

Mathematics

For certain users, mathematics can be annoyingly inevitable when they conduct data analytics; however, mathematics is central to the area of computational biology and biomedical research. Some mathematical equations are relatively straightforward and easy to explain to the ordinary users, just like the famous SVD (Singular Value Decomposition) theorem:

$$M = U\Sigma V^T$$

Although the formula is simple in its form, the meaning behind it is quite wide and deep. That is, a seemingly simple formula also needs thorough explanation for ordinary end users. Furthermore, when stepping into the territory of machine learning, we will find that this area is glutted with so many cranky mathematical equations and formulas in all sorts of complex and profound algorithms, for example, convex optimization algorithms (Boyd et al., 2006) and the ELM (Extreme Learning Machine) algorithm (Huang et al., 2006), just like the following one excerpted from ELM:

$$min\ L_{RELM} = \frac{1}{2}||\beta||^2 + \frac{C}{2}||Y - H\beta||^2$$

Some mathematical formulas are very hard to comprehend even by seasoned machine learning experts. Therefore, it is necessary to explain different mathematical formulas, even those seemingly ones in the system. To understand mathematical formulas is fundamental to understanding how a complex algorithm works as a whole. ProtoDLS thinks about the problem in three aspects:

1. to explain the interpretation of each symbol in the mathematical formula.
2. to explain the denotation of the entire mathematical formula including its user, context, etc.
3. to explain the connotation lying behind the entire mathematical formula.

Thus, in ProtoDLS, MC also records all of the bits and pieces about every mathematical formula and the associated formula derivation process in a particular region of it, namely, the **Mathematics Metadata** (MM) region, as seen in **Figure 5**.

With MM in MC, all the components and users of ProtoDLS can easily query, retrieve, and check in mathematics-related problems. At the same time, with the aid of VI, ProtoDLS can offer its users with an intuitive visualization presentation for better explainability.

Algorithms

Likewise, algorithms, especially black-box deep learning algorithms are also hard to comprehend. There roughly exist two

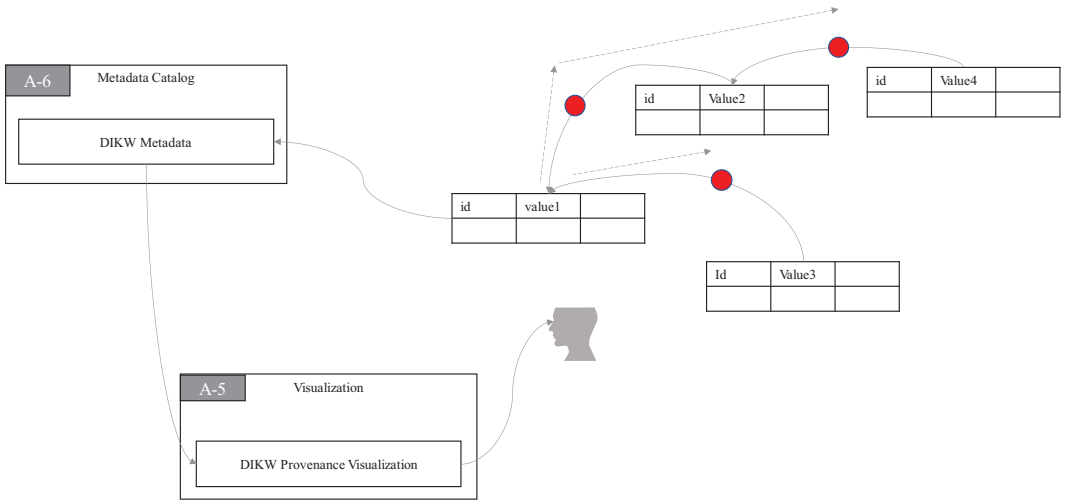


FIGURE 4 | The DIKW provenance visualization.

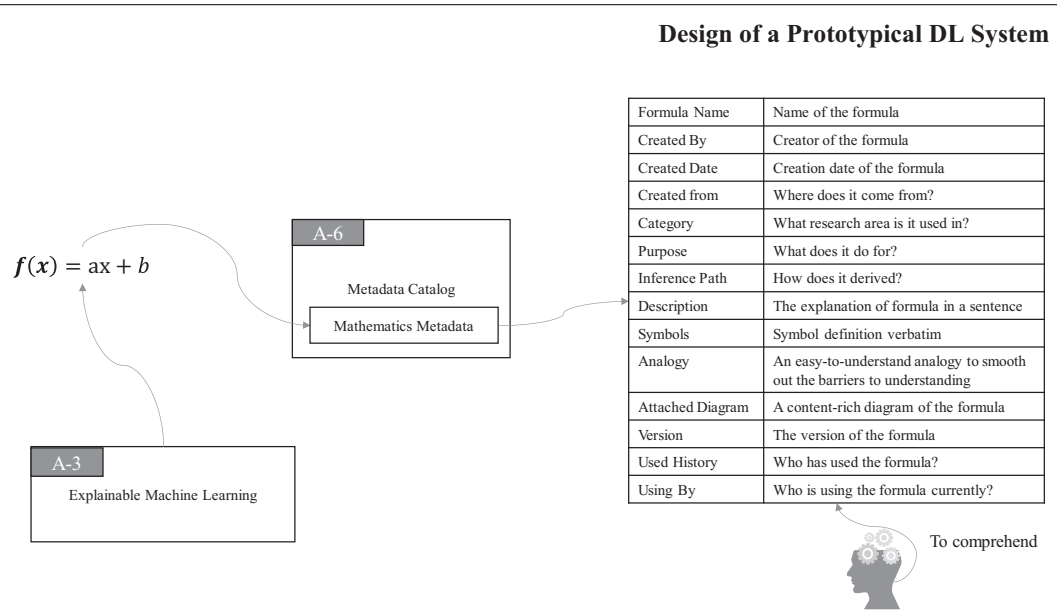


FIGURE 5 | The MC component records all about mathematical formulas and formula inference processes.

different cases when we are dealing with algorithms in practical applications:

1. Algorithms cannot be thoroughly understood by ordinary end users.
2. Algorithms cannot be clearly explained by current technical advancements, like black-box series of deep learning algorithms.

The first case can be somewhat eased by technical measures, like visualization and narrative storytelling. When it comes to the second case, explainability will soon become the bottlenecks of the whole platform. The methods we could adopt are reproducing algorithms and methods in the latest literature in ProtoDLS.

As you can see in **Figure 6**, there is a module named **Algorithm Metadata** (AM) in MC. AM preserves all the information related to algorithms. The **Algorithm Visualization** (AV) module in VI plays a very special role in the algorithmic framework of ProtoDLS, since visualization is very critical to the explainability of algorithms, especially the deep learning algorithms. In the algorithmic framework, four modules exist in EML, i.e., **Sandbox Training** (ST), **Dialogue System** (DS), **Narrator** (NA), and **Twin Agent** (TA). At the same time, EML opens up a sandbox region in DP, specifically for model training, which has the following benefits: (i) algorithms can

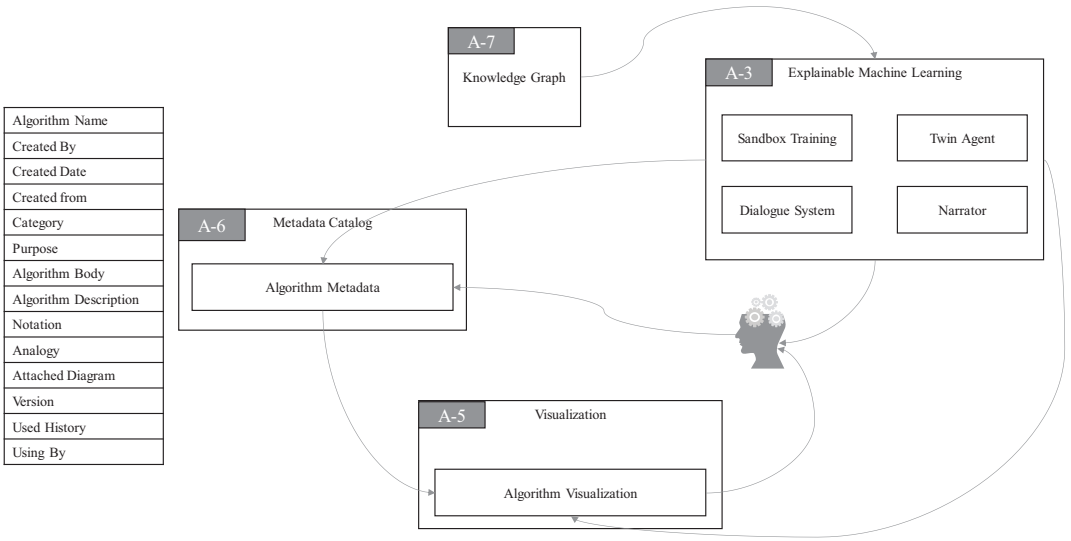


FIGURE 6 | Algorithmic design for explainability in ProtoDLS.

be trained and tested with real data sets in the production environment, and training the algorithm with the same data distribution will enhance the explainability; (ii) deployment from the sandbox region to the production environment is relatively straightforward; and (iii) the sandbox is isolated from the production environment and thus faults or halts of sandbox will not affect the production environment. The events that occurred in ST will go into the AM and be stored for later explanation. The dialogue bots behind DS interact with algorithm users with multiround natural language dialogues, with the help of KG. NA tells users how the algorithms work in a narrative storytelling mode. For example, it is well-known that beginners to NLP usually find it very hard to understand the concept of embedding. Embedding is a technique of mapping an object onto a vector. Without any explanation, this definition cannot be thoroughly understood by beginners. However, we may use a narrative of the algorithm to exchange for understandability, as seen in **Figure 7**. The idea of TA in EML is imitated by a reinforcement framework proposed by Wang et al. (2018). Sometimes, the machine learning algorithm is too complex to be thoroughly grasped. In such cases, we usually choose an explanation model to give *post-hoc* explanations about the real algorithm model. Based on this basic idea, the twin agents go further and select the best explanations about the algorithm model according to some reinforcement learning and adversarial learning rules. Drawing inspiration from this idea, TA simulates a reinforcement learning framework to enhance the explainability of some hard black-box algorithms.

Engineering Explainability

In ProtoDLS, engineering factors of explainability consist of two major categories: the infrastructure (storage, memory, network, and computation) explainability and the software development explainability (programming language, process, thread, software methodology). The infrastructure explainability concentrates on the explanation of the running status of the infrastructure, like for example the question: how much memory do the underlying cluster nodes consume now? Meanwhile, the software development explainability concentrates on the explanation of program-related problems and questions.

EI uses a simple audit log to record all file access requests of the file system, intended to be easily written and non-intrusive. The log details include operation status (success, halted, failed, etc.), user name, client address, operation command, and operation directory. Through the audit log, system admins can view all kinds of operation status of EI in real time, track all kinds of warnings, errors, alerts, and incorrect operations, and execute some metric monitoring.

At the same time, EI daemons will generate a series of monitoring logs. The monitoring log monitors and collects the measurable information of EI according to some predefined rules. For example, the following metrics will be collected by EI: the number of bytes written, the number of file blocks copied, and the number of requests from the client. EI daemons also monitor the network latency, memory consumption, and storage consumption. The X-Storage, X-Memory, X-Computation, and X-Network modules of EI continuously monitor their metrics and output the monitored metrics to the **Infrastructure Metadata** (IM) module of MC, respectively, as their names indicate. IM in MC transfers metrics to the **Infrastructure Visualization** (IV) module of VI to monitor the running status of EI on a visual interface for system administrators in real time on one side. On the other side, the DS module in EMI can asynchronously request metrics from IM to finish multiround natural language dialogue with users, as seen in **Figure 8**.

Programming languages used for the computations also would affect our correct understanding of runtime contexts, algorithms, and running results. The runtime context of a process is composed of its program code, data structure, and hardware environment needed for program running. To collect the runtime

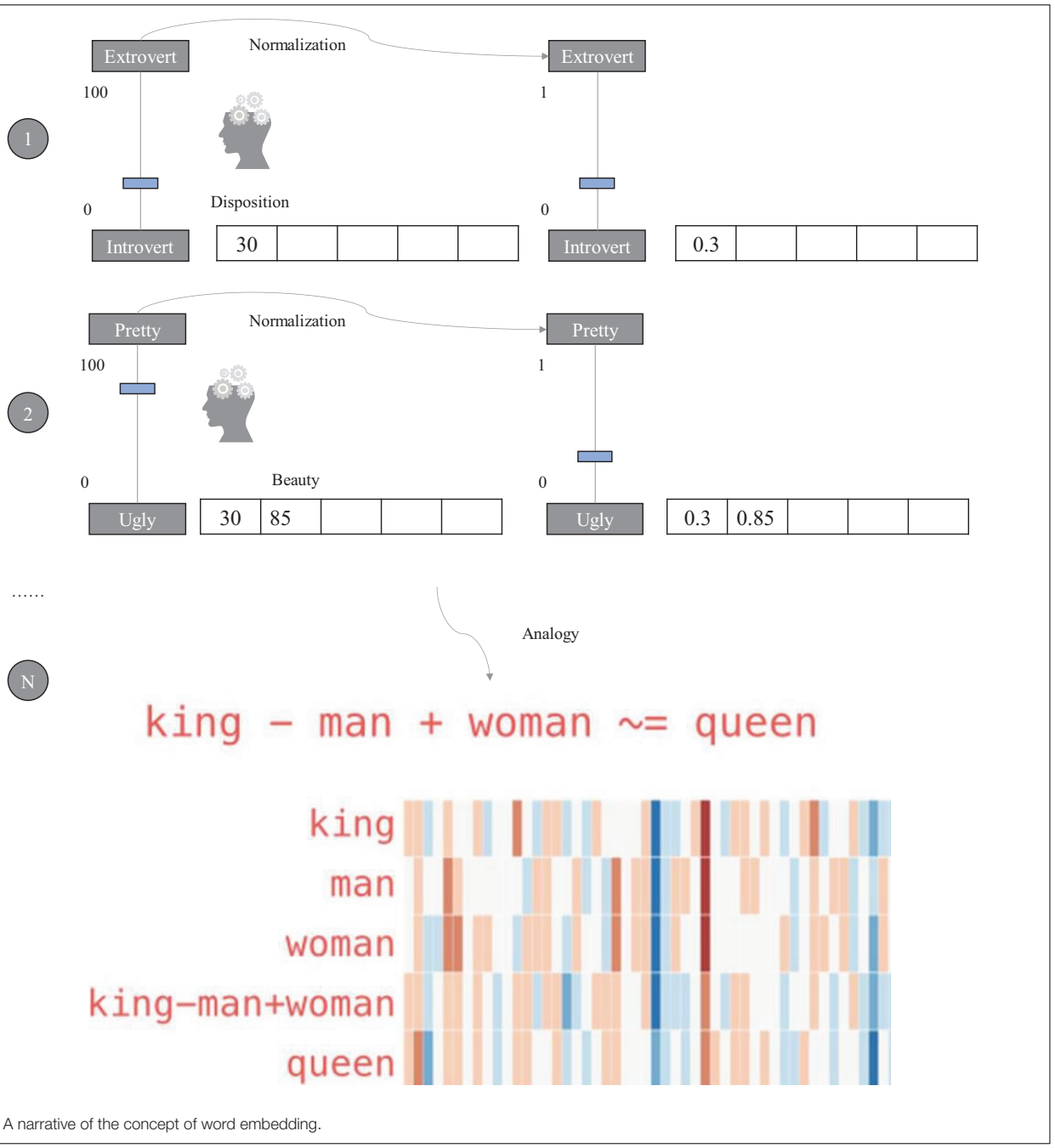


FIGURE 7 | A narrative of the concept of word embedding.

context of programs, the JS component in ProtoDLS will start a job to monitor in real time, and it will store the monitored metrics in MC, as seen in **Figure 9**. Similarly, the collected metrics will store in a module named **Software Runtime Metrics** (SRM) in MC. SRM transfers the processed metrics to the **Software Metrics Visualization** (SMV) module of VI to monitor the running status of software contexts on a visual interface for software users or developers in real time on one side. On the other side, the DS module in EMI can request metrics from SRM to finish multiround natural language dialogue with software users or developers, as seen in **Figure 9**.

Modality Explainability

ProtoDLS is created to support multimodality in the first place since it was chartered and initially designed for multimodal bioinformation processing for research purposes. Multimodality itself is a hard problem both in techniques and in applications, which causes several aspects of explainability needs. The explainability of multimodality lies in multimodal fusion. Multimodal fusion refers to the synthesis of information from two or more modalities for preprocessing. Multimodal fusion can be roughly divided into three types of fusion: early fusion, late fusion, and intermediate fusion.

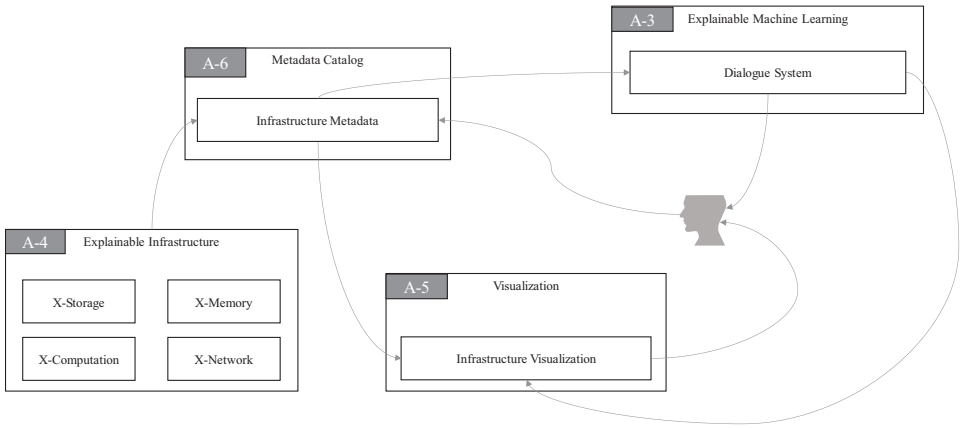


FIGURE 8 | Infrastructure design for explainability in ProtoDLS.

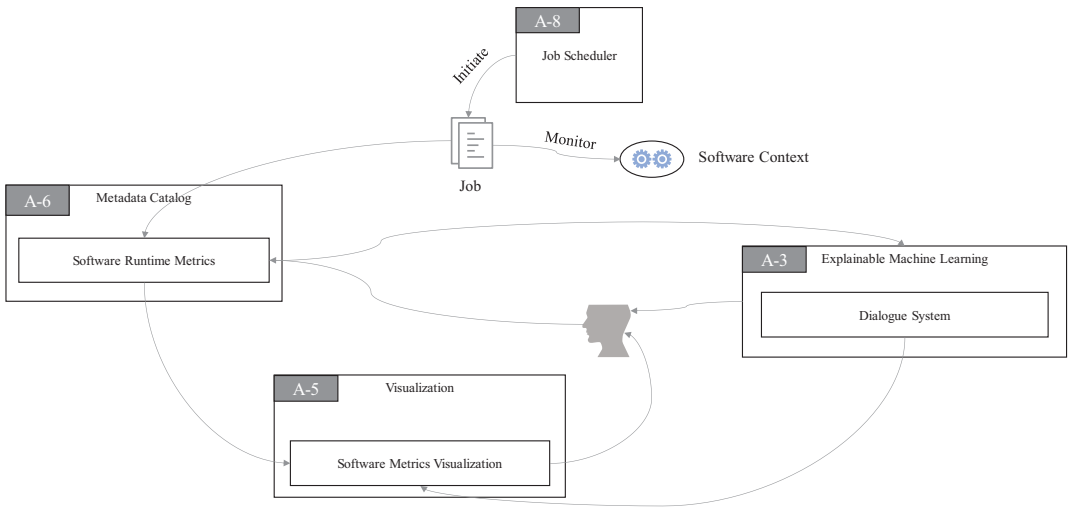


FIGURE 9 | Software runtime status design for explainability in ProtoDLS.

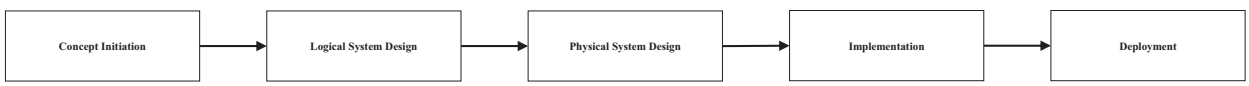


FIGURE 10 | Software development process of ProtoDLS.

The early fusion or data-level fusion combines multiple independent data sets into a single feature vector and then inputs it into a machine learning classifier. Because the early fusion of multimodal data often cannot fully utilize the complementarity of multimodal data, the original data of early fusion usually contain a great deal of redundant information. Therefore, early fusion methods are often combined with feature extraction methods to eliminate redundant information, such as principal component analysis (PCA) (Jolliffe, 2005), mRMR (Peng et al., 2005), and autoencoders (Vincent et al., 2008). In this regard, feature-level explainability largely determines the overall explainability of the fusion model.

The late fusion or decision-level fusion fuses the output scores of classifiers for decision-making trained by different modal data. The advantage of this method lies in that the errors of the fusion model come from different classifiers, while the errors from different classifiers are often separate and independent, which will not cause further accumulation of errors. Common late fusion methods include max fusion, average fusion, Bayes-based fusion, and ensemble learning fusion. As a typical representative of late fusion, ensemble learning is widely used in communication, computer recognition, speech recognition, and many other research fields. As a classical model-agnostic method, LIME (Local Interpretable Model-agnostic

Explanations) will help to explain the late fusion of multimodality data (Ribeiro et al., 2018).

Intermediate fusion refers to the transformation of different modal data into high-dimensional feature expression, and then fusion in the middle layer of the model. Taking the neural networks as an example, the intermediate fusion first uses the neural network to transform the original data into high-dimensional feature representation and then obtains the commonalities of different modal data in high-dimensional space. One of the advantages of the intermediate fusion method is that it can flexibly choose the location where fusion happens.

The location where multimodal data fusion happens directly relates to the explainability. The late fusion method is a little bit easier to explain than the intermediate fusion and the early fusion. Moreover, the explainability also relates to the classifiers and is constrained by the explainability of the participated classifiers, which engenders new difficulties. In ProtoDLS, we treat modality explainability in a quite straightforward way, so we view the modality explainability independently and explain the participated algorithms or classifiers irrelevantly at first. In terms of its complexities and the current technical limitations, we leave the modality explainability as an open problem to be tackled in the future.

PROJECT PROGRESS AND DISCUSSION

ProtoDLS started from an initial intent to build a platform supporting multimodal bioinformation processing for research and experiment purposes, with explainability in the heart of its design goals. In addition, some other minor design goals might include performance, robustness, security, privacy-preserving, and extensibility. However, meeting these design goals at the same time will greatly increase the complexity of system construction. Thus, we focus completely on the explainability of ProtoDLS in the first stage.

As seen in **Figure 10**, the whole software development process roughly includes the following phases: concept initiation, logical design, physical design, implementation, and deployment. Currently, we have just finished the logical system design and entered into the physical system design phase. During the phase, we will determine the physical structure of TDW and DL and subsequently evaluate the performance of the physical design. In order to guarantee the explainability design goal, we should observe the following steps during the implementation phase:

1. Finish the implementation tasks in the horizontal pipeline in the first place, then switch to finish the tasks in the vertical pipeline.
2. In the horizontal pipeline, DIKW provenance is among the top priorities to implement since DIKW provenance acts as a backbone for the explainability of ProtoDLS, and meanwhile the technologies behind it are relatively mature and engineering oriented.
3. In the vertical pipeline, mathematical formulas and algorithms are the first two implementation factors before we start implementing engineering-related factors.
4. In the last step, finish the implementation tasks for multimodality explainability.

TABLE 1 | Feature comparison with some popular data lake systems on the market.

	Delta Lake	Apache Iceberg	Apache Hudi	Apache Kudu	AWS Data Lake (Dremio)	ProtoDLS
Hadoop support	√	√	√	–	–	√
Metadata management	√	√	√	√	√	√
Workload management	–	–	–	–	√	√
Data governance	–	–	–	–	√	√
Streaming	√	√	√	√	√	√
Versioning	√	–	√	–	√	–
Spark SQL	√	–	√	√	–	–
Index	–	√	√	√	√	√
Row-level update	√	√	√	√	√	–
ACID transactions	√	√	√	√	√	–
Standard compliance	–	–	–	–	√	–
Security	–	–	–	√	√	√
S3 support	√	√	–	–	√	–
Explainability	–	–	–	–	–	√

ProtoDLS is an ambitious and challenging project with uncertain risks, which requires a continuous investment of capital and human resources. Only after a process of thoughtful and considerable design and implementation, it is estimated that ProtoDLS will reach a preliminary stage in 10 months and implement a primary overall explainability. At that stage, compared with some popular data lake systems on the market, such as Apache Hudi (2020), Apache Iceberg (2020), Apache Kudu (2020), Aws Data Lake (2020), and Delta Lake (2020) ProtoDLS will gain some competitive advantages, as illustrated in **Table 1**.

CONCLUSION AND FUTURE WORK

The large amounts of data continuously generated from heterogeneous types of biological resources cause great challenges for advancing biological research and development; accordingly, these challenges will further incur great difficulties for biological data processing subsequently. To attack these challenges, this paper presents a design scheme for constructing a practical data lake platform for processing multimodal biological data using a prototype system named ProtoDLS. Explainability is a major concern when we deploy and use such a platform oriented for processing of biological resources, ProtoDLS adopts a dual mechanism to ensure explainability across the platform. On the horizontal landscape, ProtoDLS ensures the intra-component

explainability from data acquisition to data presentation. On the other hand, on the vertical axis, ProtoDLS ensures the inner-component explainability including mathematics, algorithm, execution time, memory consumption, network latency, security, and sampling size.

The explainability is a rather broad concept, with multiple meanings in diverse scenarios, in a degree, to realize a full spectrum of explainability is somewhat close to the realization of artificial general intelligence (AGI), which will cost substantial human resources and capital investment. Also, the design of ProtoDLS is only a little step toward this. So many aspects need to be considered for ProtoDLS. For example, to design a typed DIKW resource framework will stand on a more abstract level to explain DIKW provenance, which will enhance the degrees of explainability on the horizontal axis in ProtoDLS. Every vertical module of each component leaves a huge gap for further fine-tuning that will require considerable research efforts and sometimes need several times of practical experiments. Finally, we should start from the logical prototype design given by this paper and begin implementing some subsets of ProtoDLS. For example, with the help of NLP techniques, an extensible and highly concurrent metadata management component can be designed and implemented, with a dialogue module supporting human understandable sentences. Upon the submission of this paper, the physical design of ProtoDLS has already started off, and implementation also has initiated simultaneously to prepare some initial verification.

To the best of our knowledge, this may be the first time that a logical design of a prototypical data lake is proposed in terms of the explainability around the data processing in a data lake. Although this paper is relatively elementary, we also hope to provide a starting point and a stepping stone for any academic researchers and industrial practitioners in bioinformatics, genetics, and phenomics, or people interested in data lake research and deployment in any other fields. For people who are doing research on the data lake explainability, this paper also may be beneficial and helpful.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HC proposed the conceptual framework, designed the initial version of the prototype system, revised the manuscript, and provided the final submission version. YD corrected the conceptual framework, provided some insightful contributions to the design details of ProtoDLS, and proposed an upgraded version of it. HC and YD edited and modified the manuscript, figures, and table. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Shanghai Innovation Action Plan Project under Grant Nos. 18510732000, 18510760200, and 19510710500 and the National Natural Science Foundation of China under Grant Nos. 61572137, 61728202, and 61873309.

ACKNOWLEDGMENTS

Thanks are due to the reviewers for their hard work at reading and proofreading the early version of the manuscript. We thank Sheng Zhang for helping submit the final manuscript.

REFERENCES

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainableartificial Iitelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/access.2018.2870052

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Èech, M., et al. (2016). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 2016, W3–W10.

Albani, M., Leone, R., Foglini, F., De Leo, F., Marelli, F., and Maggio, I. (2020). Everest: The platform allowing scientists to cross-fertilize and cross-validate data. *Data Sci. J.* 19:21.

Apache Hudi (2020). http://hudi.apache.org/. Accessed 10 Jul. 2020.

Apache Iceberg (2020). http://iceberg.apache.org/. Accessed 10 Jul. 2020.

Apache Kudu (2020). http://kudu.apache.org/. Accessed 10 Jul. 2020.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fus.* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Assante, M., Candela, L., Castelli, D., Cirillo, R., Coro, G., Frosini, L., et al. (2019). Enacting open science by D4Science. *Future Generat. Comp. Syst.* 101, 555–563. doi: 10.1016/j.future.2019.05.063

Aws Data Lake (2020). http://dremio.com/aws/. (accessed July 10, 2020).

Bio2RDF (2020). https://old.datahub.io/organization/bio2rdf. (accessed July 13, 2020).

BioSchemas (2020). https:// bioschemas.org/. (accessed July 13, 2020).

Boyd, S., Vandenberghe, L., and Faybusovich, L. (2006). Convext Optimization. *IEEE Transact. Automat. Contr.* 51, 1859–1859.

Breiman, L. (2001). Random forest. *Mach. Learn.* 45, 5–32.

Buneman, P., Khanna, S., and Tan, W.-C. (2001). Why and where: a characterization of data provenance. *Proc. ICDT* 2001, 316–330. doi: 10.1007/3-540-44503-x_20

Bussery, J., Denis, L.-A., Guillon, B., Liu, P., Marchetti, G., and Rahal, G. (2018). eTRIKS platform: conception and operation of a highly scalable cloud-based platform for translational research and applications development. *Comput. Biol. Med* 2018, 99–106. doi: 10.1016/j.compbiomed.2018.02.006

da Veiga, Leprevost, F., Grüning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., et al. (2017). BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics.* 33, 2580–2582. doi: 10.1093/bioinformatics/btx192

Delta Lake (2020). http://delta.io/. (accessed July 10, 2020).

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., et al. (2010). The BioPAX community standard for pathway data sharing. *Nat. Biotechnol* 28, 935–942.

Dixon, J. (2010). *Pentaho, Hadoop and Data Lakes.* Available online at: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (accessed January 27, 2019).

Došilović, F. K., Brcic, M., and Hlupic, N. (2018). "Explainable Artificial Intelligence: A Survey," in *proceedings of the MIPRO* 2018 - 41st International Convention Proceedings, (Opatija: MIPRO), 210–215.

Duan, Y., Sun, X., Che, H., Cao, C., Li, Z., and Yang, X. (2019). Modeling data. information and knowledge for security protection of hybrid iot and edge resources. *IEEE Access*. 7, 99161–99176. doi: 10.1109/access.2019.2931365

Ebi ArrayExpress (2020). https:// www.ebi.ac.uk/arrayexpress/. (accessed July 17, 2020).

Ebi Rdf (2020). https://www.ebi.ac.uk/rdf/. (accessed July 13, 2020).

Edwards, L., and Veale, M. (2017). Slave to the algorithm? why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law Technol. Rev*. 16, 1–65.

Ensembl (2020). http://www.ensembl.org/. (accessed July 19, 2020).

Fang, H. (2015). "Managing data lakes in big data era: what's a data lake and why has it became popular in data management ecosystem," in *Proceedings of the 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems (CYBER 2015)*, (Shenyang: IEEE), 820–824.

Farid, M., Roatis, A., Ilyas, I. F., Hoffmann, H.-F., and Chu, X. (2016). "CLAMS: bringing quality to data lakes," in *Proceedings of the 2016 International Conference on Management of Data (SIGMOD 2016)*, (San Francisco, CA: ACM), 2089–2092.

Farrugia, A., Claxton, R., and Thompson, S. (2016). "Towards social network analytics for understanding and managing enterprise data lakes," in *Advances in Social Networks Analysis and Mining (ASONAM* 2016), (San Francisco, CA: IEEE), 1213–1220.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29, 1189–1232.

Gene Expression Atlas (2020). http://www.ebi.ac.uk/gxa. (accessed July 30, 2020).

Gene Ontology (2020). http://www.geneontology.org/. (accessed July 30, 2020).

GEO (2020). http://www.ncbi.nih.gov/geo/. (accessed July 19, 2020).

Gestel, T. V., Baesens, B., Dijcke, P. V., Suykens, J. A. K., and Garcia, J. (2005). Linear and non-linear credit scoring by combining logistic regression and support vector machines. *J. Credit Risk* 1, 31–60. doi: 10.21314/jcr.2005.025

Guo, W., Huang, S., Tao, Y., Xing, X., and Lin, L. (2018). Explaining deep learning models - a bayesian non-parametric approach. *NeurIPS* 2018, 4519–4529.

Himabindu, L., Bach, S. H., and Leskovec, J. (2016). "Interpretable decision sets: a joint framework for description and prediction," in *Proceedings of ACM SigKDD International Conference*, (New York, NY: ACM).

Houze-Cerfon, C.-H., Vaissié, C., Gout, L., Bastiani, B., Charpentier, S., and Lauque, D. (2019). Development and evaluation of a virtual research environment to improve quality of care in overcrowded emergency departments: observational study. *J. Med. Internet Res*. 21:e13993. doi: 10.2196/13993

Huang, G. B., Zhu, Q. Y., and Siew, C. K. (2006). Extreme learning machine: theory and applications. *Neurocomputing* 70, 489–501. doi: 10.1016/j.neucom.2005.12.126

Jolliffe, I. T. (2005). *Principal Component Analysis*. Berlin: Springer-Verlag.

Klettke, M., Awolin, H., Sturl, U., Müller, D., and Scherzinger, S. (2017). "Uncovering the evolution history of data lakes," in *Proceedings of the 2017 IEEE International Conference on Big Data (BIGDATA 2017)*, (Boston, MA: IEEE), 2462–2471.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.

Li, Y. F., Kennedy, G., Davies, F., and Hunter, J. (2010). "podd: an ontology-driven data repository for collaborative phenomics research," in *The Role of Digital Libraries in a Time of Global Change. Proceedings of ICADL 2010. Lecture Notes in Computer Science*, Vol. 6102, eds G. Chowdhury, C. Koo, and J. Hunter (Berlin: Springer).

Madera, C., and Laurent, A. (2016). "The next information architecture evolution: the data lake wave," in *Proceedings of the 8th International Conference on Management of Digital EcoSystems (MEDES 2016)*, (Biarritz: ACM), 174–180.

Miloslavskaya, N., and Tolstoy, A. (2016). Big data. fast data and data lake concepts. *Procedia Comp. Sci*. 88, 300–305. doi: 10.1016/j.procs.2016.07.439

OntoLingua Server (2020). http://www.ksl.standford.edu/software/ontolingua/. (accessed July 13, 2020).

Peake, G., and Wang, J. (2018). "Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (New York, NY: ACM), 2060–2069.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transact. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/tpami.2005.159

Pentaho (2020). https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform.html. (accessed July 19, 2020).

PubMed (2020). http://www.ncbi.nih.gov/pubmed/. (accessed July 19, 2020).

Quix, C., Hai, R., and Vatov, I. (2016). Metadata extraction and management in data lakes with GEMMS. *Compl. Sys. Inform. Model. Q.* 9, 67–83. doi: 10.7250/csimq.2016-9.04

Remy, L., Ivanoviæ, D., Theodoridou, M., Kritsotaki, A., Martin, P., Bailo, D., et al. (2019). Building an integrated enhanced virtual research environment metadata catalogue. *Electronic Library* 37, 929–951. doi: 10.1108/el-09-2018-0183

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: high-precision model-agnostic explanations. *AAAI* 2018, 1527–1535.

Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. ITU J. ICT Discov. -Spec. Issue 1 – Impact Artif. Intell. AI Commun. Netw. Serv. 1, 1–10. doi: 10.21037/jmai.2018.07.01

Schafer, J. B., Konstan, J., and Riedl, J. (1999). "Recommender systems in e-commerce," in *Proceedings of the* 1st ACM conference on Electronic commerce, (New York, NY: ACM), 158–166.

Stein, B., and Morrison, A. (2014). The enterprise data lake: Better integration and deeper analytics. *Technol. Forecast: Rethink. Integrat*. 2014, 1–9. doi: 10.1007/978-1-4842-3522-5_1

Suriarachchi, I., and Plale, B. (2016). "Crossing analytics systems: a case for integrated provenance in data lakes," in *Proceedings of the 12th IEEE International Conference on eScience (e-Science 2016)*, (Baltimore, MD: IEEE), 349–354.

UniProt (2020). http://www.uniprot.org/. (accessed July 19, 2020).

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. (2008). "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, (Helsinki).

Wang, J., Fujimaki, R., and Motohashi, Y. (2015). "Trading interpretability for accuracy: oblique treed sparse additive models," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY: ACM).

Wang, X., Chen, Y., Yang, J., Wu, L., Wu, Z., and Xie, X. (2018). "A reinforcement learning framework for explainable recommendation," in *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*, (Piscataway, NJ: IEEE), 587–596.

Ying, C., Argentinis, E., and Weber, G. (2016). IBM watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clin. Ther*. 38, 688–701. doi: 10.1016/j.clinthera.2015.12.001

Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., and Ma, S. (2014). "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *Proceedings of the* 37th international ACM SIGIR conference on Research & development in information retrieval, (New York. NY: Association for Computing Machinery), 83–92.

Conflict of Interest: HC was employed by Great Wall Motors company at the time of the study. The company was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

METHODS
published: 03 November 2020
doi: 10.3389/fbioe.2020.553847

Apathy Classification Based on Doppler Radar Image for the Elderly Person

Naoto Nojiri[1], Zelin Meng[2], Kenshi Saho[3], Yucong Duan[4], Kazuki Uemura[3], C. V. Aravinda[5], G. Amar Prabhu[5], Hiromitsu Shimakawa[1] and Lin Meng[2]**

[1] College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan, [2] College of Science and Engineering, Ritsumeikan University, Kusatsu, Japan, [3] Faculty of Engineering, Toyama Prefectural University, Imizu, Japan, [4] Data Science and Technology Department, Hainan University, Haikou, China, [5] Department of Computer Science and Engineering, NMAM Institute of Technology, NITTE, Karkala, India

Apathy is a disease characterized by diminished motivation not attributable to a diminished level of consciousness, cognitive impairment, or emotional distress. It is a serious problem facing the elderly in today's society. The diagnosis of apathy needs to be done at a clinic, which is particularly inconvenient and difficult for elderly patients. In this work, we examine the possibility of using doppler radar imaging for the classification of apathy in the elderly. We recruited 178 elderly participants to help create a dataset by having them fill out a questionnaire and submit to doppler radar imaging while performing a walking action. We selected walking because it is one of the most common actions in daily life and potentially contains a variety of useful health information. We used radar imaging rather than an RGB camera due to the greater privacy protection it affords. Seven machine learning models, including our proposed one, which uses a neural network, were applied to apathy classification using the walking doppler radar images of the elderly. Before classification, we perform a simple image pre-processing for feature extraction. This pre-processing separates every walking doppler radar image into four parts on the vertical and horizontal axes and the number of feature points is then counted in every separated part after binarization to create eight features. In this binarization, the optimized threshold is obtained by experimentally sliding the threshold. We found that our proposed neural network achieved an accuracy of more than 75% in apathy classification. This accuracy is not as high as that of other object classification methods in current use, but as an initial research in this area, it demonstrates the potential of apathy classification using doppler radar images for the elderly. We will examine ways of increasing the accuracy in future work.

Keywords: apathy classification, doppler radar image, the elderly person, machine learning, deep learning

OPEN ACCESS

Edited by:
Madhuchhanda Bhattacharjee, University of Hyderabad, India

Reviewed by:
Jianzhen Xu, Shantou University, China
Xinguo Lu, Hunan University, China

****Correspondence:***
Naoto Nojiri
ri0005ri0724@gmail.com
Lin Meng
menglin@fc.ritsumei.ac.jp

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *29 April 2020*
Accepted: *30 September 2020*
Published: *03 November 2020*

Citation:
Nojiri N, Meng Z, Saho K, Duan Y, Uemura K, Aravinda CV, Prabhu GA, Shimakawa H and Meng L (2020) Apathy Classification Based on Doppler Radar Image for the Elderly Person. Front. Bioeng. Biotechnol. 8:553847. doi: 10.3389/fbioe.2020.553847

1. INTRODUCTION

Apathy is a disease characterized by diminished motivation not attributable to a diminished level of consciousness, cognitive impairment, or emotional distress (Marin, 1990, 1991; Marin et al., 1991). It has a relationship with others diseases such as Parkinson's, Alzheimer's, and stroke, all of which tend to befall elderly people and threaten their health and well-being (Landes et al., 2001; Fuh et al., 2005; Caeiro et al., 2013; Pagonabarraga et al., 2015). Studies have shown that roughly

47% of patients with Alzheimer's disease also suffer from apathy (Fuh et al., 2005). However, to get an apathy diagnosis, elderly patients need to go to a clinic, which is both inconvenient for them and sometimes physically difficult. A computer vision system for assistance with apathy diagnosis in remote operation has been developed (Happy et al., 2019), but since it uses images of the patient's face, problems related to privacy protection arise. Another issue is that patients typically need to exhibit subjective symptoms before seeking a doctor, but apathy rarely has subjective symptoms, particularly among the elderly who often live in solitude. Hence, elderly people may delay getting diagnoses and miss out on the best treatment period.

Society is currently facing a rapid increase in the aging population—especially in Japan, where the percentage of the population aged 65 and over (elderly) is 28.1%. As of 2018, the population aged between 65 and 74 years was 13.9% and aged 75 years and over was 14.2%. By 2065, these numbers are expected to increase to 38.4% for ages 65+ and 25.6% for ages 75+ (CabinetOfficeJapan, 2019). Hence, developing a more convenient apathy assessment is becoming an important issue.

In this work, we examine the use of Doppler radar imaging for the classification of apathy in the elderly. Our objective is to encourage earlier access to apathy assessment. Doppler radar imaging is advantageous because it does not use face images, which helps with protecting privacy, and the equipment it uses is simple enough to set up that apathy checks can be performed routinely without any special preparation. Besides, because the Doppler radar directly measures the velocities, the accuracy of velocity measurement is better than other optical sensor techniques that mainly measures position information (Li et al., 2019). Furthermore, with its applicability to low-light conditions and to persons wearing ordinary clothes as its advantages, the Doppler radar has been investigated for using in home and hospital health monitoring applications in recent years (Seifert et al., 2019).

Unfortunately, as little research has been done in this area, it is not clear which action is best suited for apathy classification. Hence, we select one of the most normal actions in daily life: walking. Walking has a deep relationship with health condition and has been used since 1984 for clinical gait assessment in the neurologically impaired (Holden et al., 1984). It is easy to see how the action of walking relates to health condition; for example, stroke victims often have difficulty controlling their body when walking. Recently, researchers found that the action of walking can reveal a lot of a person's health information, including age (Handri et al., 2009; Makihara et al., 2011) and chronic illness (Pitta et al., 2005; Jehn et al., 2009; Rabinovich et al., 2013).

In this study, we created the Elderly Person Apathy Doppler Radar Image Dataset (EPADRI Dataset) with the help of elderly people aged 65 years or more. We had each participant fill out a questionnaire to determine if they had apathy or non-apathy and then perform a walking action under doppler radar to obtain experimental images. We then combined image processing and machine learning to perform apathy classification using the EPADRI Dataset.

As pre-processing, we utilize a simple image processing for extracting the features from the radar images. In this processing, a walking doppler radar image is separated into four parts by the vertical and horizontal axes and then binarization is applied to count the features of the eight parts for training and classification by machine learning. We apply four patterns for binarization—red channel, green channel, blue channel, and YUV—and slide the threshold of binarization from 50 to 220 to determine the optimized value. Finally, the number of feature points is used for the apathy classification.

As we know, machine learning models and Numerical Analysis methods are widely used in the Biology and Bioinformatics (Lu et al., 2015, 2019; Saho et al., 2020). In this work, we applied seven machine-learning models to a classification task: a support vector machine (SVM) (Vapnik, 1998), K-nearest neighbor (KNN) (Naomi, 1992), naive Bayes, decision tree (Quinlan, 1886), random forest (Breiman, 2001), an ensemble model (Opitz and Maclin, 1999), and our proposed neural network model (Homma et al., 1998).

This is the first paper to tackle apathy classification by using doppler radar images of walking action for the elderly. Our experimental results demonstrate the effectiveness of this approach.

The contributions of this work are as follows.

- We constructed the Elderly Person Apathy Doppler Radar Image Dataset (EPADRI Dataset), which is the first dataset for apathy classification of the elderly by walking action.
- We demonstrate the effectiveness of using doppler radar images of walking actions for apathy classification and show that it both ensures privacy protection and is convenient to use.
- We propose image processing and machine learning for apathy classification of the elderly and describe the optimized threshold of binarization, color channel, and machine learning models.

Section 2 discusses related work on apathy classification and health care research on walking and doppler radar imaging. We present our dataset in Section 3. Section 4 introduces our approach, featuring the machine learning used in the experiments. The experimentation results on the apathy classification task are shown in section 5. Section 6 discusses the contributions of this work as well as the limitations. We conclude in section 7 with a brief summary and mention of future work.

2. RELATED WORK

2.1. Research on Apathy Classification

Apathy, which is derived from the Greek pathos, or passion, is conventionally defined as the absence or lack of feeling, emotions, interest, or concern (Marin, 1990). Robert et al. define apathy in clinical terms as including diminished motivation not attributable to a diminished level of consciousness, cognitive impairment, or emotional distress (Marin, 1990, 1991). Apathy occurs in several neurological and psychiatric disorders and seems to have a relationship with Parkinson's disease, Alzheimer's disease, stroke, etc., which often appear in the elderly (Landes et al., 2001; Caeiro et al., 2013; Pagonabarraga et al., 2015). Hence,

the assessment and early diagnosis of apathy is quite important, especially among the elderly.

Currently, patients need to go to a clinic for an apathy diagnosis, which usually entails medical personnel administering time-consuming clinical interviews and questionnaires. Such interviews, and getting to the clinic itself, are sometimes inconvenient and can be very hard on the elderly. This is unfortunate because if diagnosis is delayed, an elderly person will miss out on the best treatment period. Several researchers have examined the use of computer science-based methods such as computer vision and machine learning for apathy classification. Happy et al. classified apathetic and non-apathetic patients by machine learning in which the analysis target was facial dynamics entailing both emotion and facial movement (Happy et al., 2019). They administered apathy assessment interviews to 45 participants, which included short video clips with wide face pose variations, very low-intensity expressions, and insignificant inter-class variations, and reported the accuracy of 84%.

Liu et al. designed a system called ECOCAPTURE that assesses apathy in a quantitative and objective manner. It consists of observation of a patient's behavior in a multi-step scenario reproducing a brief, real-life situation by using a single 3D accelerometer under an ecological condition. An evaluation with 30 patients and 30 healthy individuals showed that ECOCAPTURE is a promising technique for more precise assessment of apathy (Liu et al., 2018).

2.2. Research on Walking in the Filed of Healthcare

Walking is one of the most common actions in daily life and can reveal abundant health information such as age and chronic illness. In 1984, walking ability was utilized for clinical gait assessment in the neurologically impaired (Holden et al., 1984). It is easy to see that the action of walking has some relationship with health condition; for example, stroke victims often have difficulty controlling their body when walking. More recently, researchers have found that measuring a patient's ability to walk is important in the diagnosis of chronic illness (Pitta et al., 2005; Jehn et al., 2009; Rabinovich et al., 2013). This has led to research into devices that protect patients by monitoring walking, such as a natural walking monitor for pulmonary patients used in conjunction with a mobile phone (Juen et al., 2015).

Researchers have also found that the walking action can be linked to an individual's age (Handri et al., 2009; Makihara et al., 2011). This has led to the development of devices like the walking-age analyzer for healthcare applications (Jin et al., 2014).

2.3. Research on Doppler Radar Imaging in Health Care Industry

Doppler radar imaging is a promising method in the e-health industry due to its assurance of privacy protection and the fact that it is non-wearable. Li et al. designed e-health applications by using passive doppler radar as a non-contact sensing method to capture human body movements, recognize respiration, and measure physical activities. Techniques related to health monitoring include micro doppler extraction for breathing detection and a support vector machine classifier utilized for physical activity recognition. Non-contact passive doppler radar has proven to be a complementary technology to meet the challenges of future healthcare applications (Li et al., 2018).

Chen et al. also applied radar imaging for classifying the six key activities of interest in the e-health area and found that it is effective for activity recognition (Chen et al., 2016).

Our motivation in the present study is to use radar images of walking action for apathy classification in the elderly. Our approach circumvents the issues in previous research because walking action is a normal daily action, which makes it simple to assess, and radar imaging protects privacy and is non-wearable. As such, we hope to make apathy assessment for the elderly simpler and more convenient. In this work, we examined seven machine-learning models for classification and a simple image processing method for feature extraction.

3. CREATION OF THE ELDERLY PERSON APATHY DOPPLER RADAR IMAGE DATASET (EPADRI DATASET)

We created the Elderly Person Apathy Doppler Radar Image Dataset (EPADRI Dataset) for training the machine-learning model and testing the accuracy of apathy classification.

We recruited 178 elderly people to help create the EPADRI Dataset. These individuals had previously filled out a Japanese version of a questionnaire known as Apathy Scale (Starkstein et al., 1992; Okada et al., 1997) we administered for apathy classification. The Apathy Scale is one of the generally used test to classify the Apathy in the field of physiotherapy and epidemiology and its effectiveness is validated in numerous studies (den Brok et al., 2015). Of the participants, 81 were between 65 and 75 years old and 98 were between 76 and 94 years old. All participants were Japanese and the questionnaire and answers were in Japanese.

The experimental protocol was approved by the local ethics committee (Toyama Prefectural University, approval no. H29-1). Participants were provided with written and verbal instructions of the testing procedures, and written consent was obtained from each participant prior to testing.

3.1. Questionnaire for Apathy Classification

Table 1 lists the apathy questionnaire items in the Apathy Scale (Starkstein et al., 1992). **Table 2** lists the responses and points. Points were tallied to judge the apathy situation. Participants with a score of 16 or more were judged to be apathetic people for the Japanese version of the Apathy Scale as verified in Okada et al. (1997).

3.2. Doppler Radar Image Creation

Figure 1 shows the creation of a radar image, where **Figure 1A** is an example of a doppler radar image with walking action and **Figure 1B** shows the walk process that is taken. **Figure 1C** shows the experimental environment, where the radar size is about 53 cm, the height is 62 cm, the start point is about 70 cm from the radar, and the walking distance is 100 cm.

TABLE 1 | Questionnaire items.

	Questions
Q1	Do you want to study something new?
Q2	Do you have any interests?
Q3	Are you interested in your health?
Q4	Can you focus on things?
Q5	Do you always want to do something?
Q6	Do you have plans or goals for the future?
Q7	Are you willing to try doing something new?
Q8	Do you spend time doing something every day?
Q9	Does someone have to tell you to do something every day?
Q10	Are you indifferent to anything?
Q11	Is there anything that interests you?
Q12	Do you do nothing unless someone tells you?
Q13	Do you ever feel not happy, not sad, but somewhere in the middle?
Q14	Do you think you are motivated?

TABLE 2 | Questionnaire responses and points.

Selection	Points
No	0
A little	1
Yes	2
Very	3

4. APATHY CLASSIFICATION BY MACHINE LEARNING

In this section, we propose our method for apathy classification that combines image processing with machine learning. **Figure 2A** shows the classification flow, which consists of feature extraction and classification.

We propose a simple image processing to extract the features from the walking radar image. Next, we apply seven machine-learning models, including an NN model we developed, to perform classification by using the extracted features.

Our objectives are two-fold. First, we want to demonstrate the possibility of performing apathy classification for the elderly by machine learning. Second, we want to determine the best model and best parameters by means of experimentation.

4.1. Feature Extraction

The feature extraction consists of binarization, image separation, and feature pixel counting, as shown in **Figure 2A**.

As discussed earlier, it is not clear which channel in an image is most suitable for apathy classification. We therefore focus on pixel configuration for the binarization and apply four kinds binarization: red channel, green channel, blue channel, and the Y of YUV. YUV is a color encoding system which encodes a color image taking human perception. The Y is defined as

$$Y = 0.299 * r[i][j] + 0.587 * g[i][j] + 0.114 * b[i][j]$$

(Charles, 2003), where the r, g, and b is the red, green and blue channel, i and j describe the coordinates of pixel.

Threshold is a key parameter in binarization as it may influence the classification accuracy. In our threshold decision, when a pixel ($P_{i,j}$) is more than the threshold, the pixel value is set to 255 (white pixel), and otherwise is set to 0. We set the pixel as one of four kinds (red channel, green channel, blue channel, and Y of YUV) and slide the threshold from 50 to 220 to determine the best value.

$$B_{i,j} = \begin{cases} 255 & (P_{i,j} >= threshold) \\ 0 & (otherwise) \end{cases}$$

After the binarization, every image is separated into four parts by the vertical and horizontal axes. An example of a separated image is shown in **Figure 2**, which includes lists items from *a* to *h*. The white pixel numbers of eight parts in the binarized image are counted. Finally, the eight numbers are decided as the features for apathy classification by the following machine-learning models.

4.2. Classification of Machine-Learning Models

This subsection introduces the seven machine-learning models we examined to determine which one was most suitable for apathy classification: a support vector machine (SVM) (Vapnik, 1998), k-nearest neighbor (KNN), naive Bayes, decision tree, random forest, an ensemble model, and our proposed neural network (NN).

4.2.1. SVM

An SVM is a supervised learning model for the boundary decision and classification of data by maximum-margin hyperplane. The most basic idea is classification using linear separability. Data are defined as $Data = \{X, Y\}$, where $X = \{X_1, \ldots X_N\}$ is the feature of the input data and $Y = \{y_1, \ldots y_N\}$ is the class label of each input data. The boundary decision is defined as $w^T X + b = 0$, where w is the normal vector to a hyperplane and b is the intercept. The constraint condition is $y_i(w^t X_i + b) >= 1$, which is used for the boundary decision and classification.

4.2.2. KNN

k-nearest neighbor is a basic classifier that calculates the k closest training in the feature space (Naomi, 1992). Data are usually defined as $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_N\}$. Absolute distance measuring, Euclidean distance measuring, or some other distance function is used for calculating the minimum distance. In this study, we define two k: one for the classification of apathy and the other for non-apathy.

4.2.3. Naive Bayes

Naive Bayes is a simple technique for constructing classifiers. In abstract terms, naive Bayes is a conditional probability model: when given a problem instance to be classified, represented by a vector $x = \{x_1, \ldots, x_n\}$ representing some n features, it assigns to this instance probabilities $p(C_k \mid x_1, \ldots, x_n)$ for each of K possible outcomes or classes C_k.

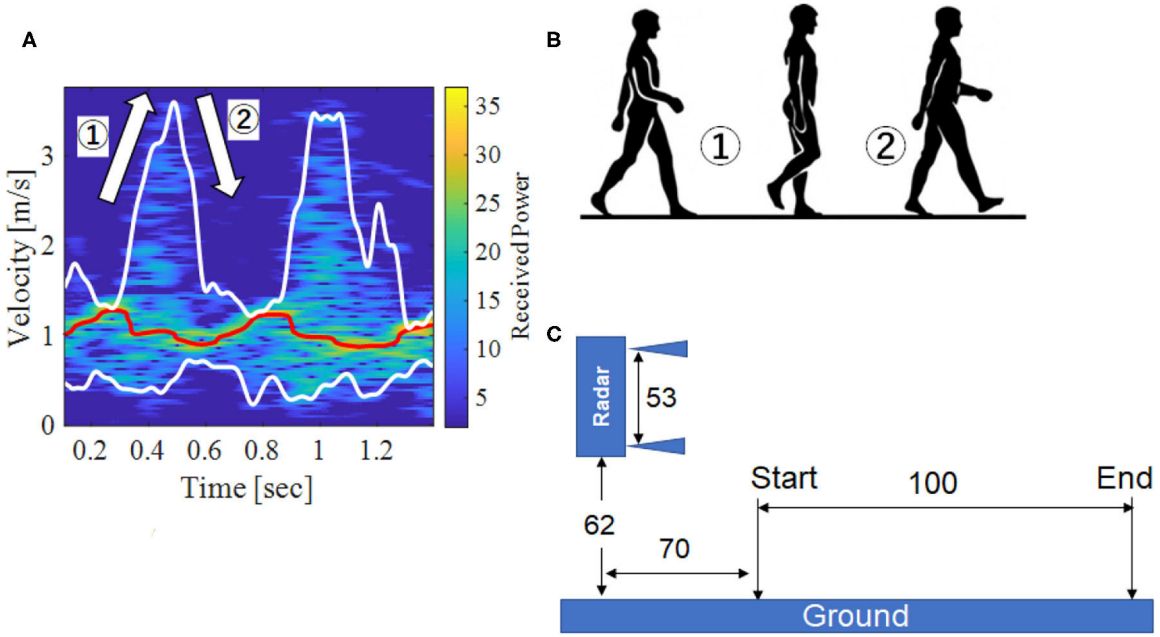


FIGURE 1 | Doppler radar image and experimental environment. **(A)** Original image, **(B)** Walk process, **(C)** Experimental environment.

The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k \mid x) = p(C_k)p(x \mid C) \div P(x).$$

It can also be

$$posterior = prior \times likelihood \div evidence.$$

4.2.4. Decision Tree

Decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is typically used to display an algorithm that only contains conditional control statements (Quinlan, 1886).

In the decision tree (two-class) model, the correct decision tree for class $Data = X, Y$, number of classes P is p, and the number of the another class N is n. Any correct decision tree for $Data$ will classify objects in the same proportion as their representation in $Data$. An arbitrary object will be determined to belong to the class P with probability $p(p+n)$ and the class N with probability $n \div (p+n)$.

To classify an object, the expected information is generated by

$$I(p,n) = -p \div (p+N)log_2 p \div (p+n) \\ -n \div (p+n)log_2 n \div (p+n).$$

The expected information required for the tree with A as a root is then obtained as the weighted average

$$E(A) = \sum_{i=1}^{v}(P_i + n_i) \div (p-n)I(P_i, n_i),$$

where the weight for the ith branch is the proportion of the objects in C that belong to X_i. The information gained by branching on A is therefore $gain(A) = I(p,n) - E(A)$.

4.2.5. Random Forest

Random forest is a combination of tree predictors in which each tree depends on the values of a random vector sampled independently and where all trees in the forest have the same distribution (Breiman, 2001).

The point is to create a group of decision trees with low correlation by using randomly sampled training data and randomly selected explanatory variables.

First, m training sets are generated by a bootstrap model. Then, for each training set, a decision tree is constructed. When a node searches for a feature and splits it, this is not to find the feature that can maximize the index (such as information gain) but to randomly extract various features and find the optimal solution among them, which is then applied to the node and split again. The random forest model uses the idea of bagging, that is, integrating, so is actually equivalent to sampling samples and features, which means it can avoid overfitting. The prediction stage includes the bagging strategy, classified voting, and regression of mean value.

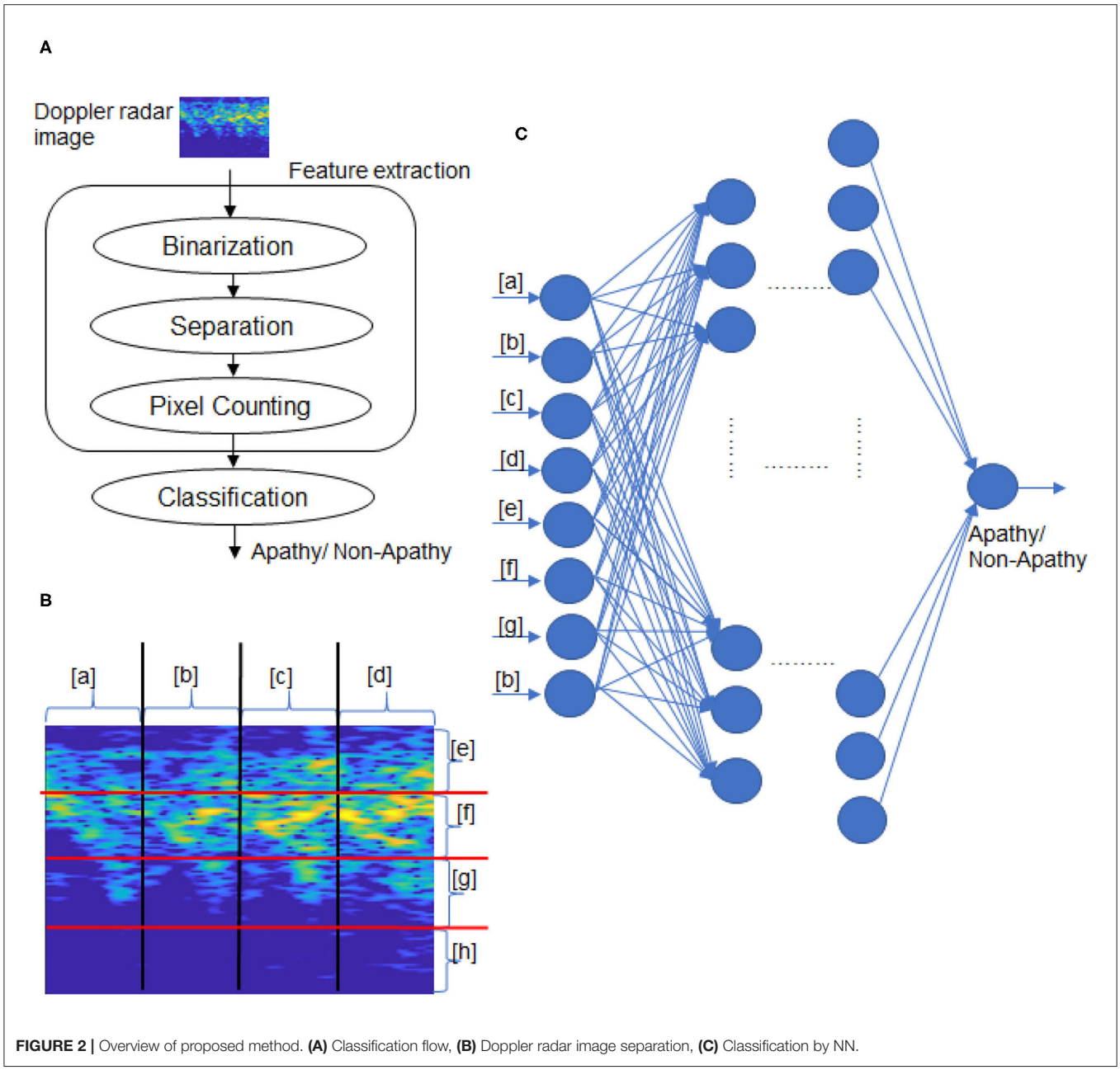


FIGURE 2 | Overview of proposed method. **(A)** Classification flow, **(B)** Doppler radar image separation, **(C)** Classification by NN.

4.2.6. Neural Network (NN)

An NN is a mathematical model that mimics the network structure of nerve cells (neurons) in the brain (Cun, 1989; Homma et al., 1998). It builds multiple layers of interconnected nodes for training data and is typically used for pattern recognition, data classification, and future prediction.

In this work, we propose an NN model that consists of five layers, as follows.

- Input layer: 8-node, activation is relu.
- Second layer: 16-node, activation is relu.
- Third layer: 32-node, activation is relu.
- Fourth layer: 64-node, activation is relu.
- Output layer: 1-output, activation is sigmoid

The training epoch is set to 50. The confidence is set to 0.5, which means when the confidence of the apathy classification is >0.5, the prediction result is judged as apathy, and otherwise as non-apathy.

4.2.7. Ensemble Model

The ensemble model performs predictions by means of a combination of several basic prediction models. The key idea is to generate a final prediction result based on the principle of majority voting with respect to the prediction results of all the models (Opitz and Maclin, 1999; Polikar, 2006).

In this method, SVM, random forest, NN, and KNN are used as the basic models. *apVote* and *noapVote* are calculated by counting the number of apathy and non-apathy prediction results, respectively. The final result is then predicted by

$$Finalresult = \begin{cases} Apathy & (apVote >= noapVote) \\ NonaApathy & (otherwise) \end{cases}$$

5. EXPERIMENTATION

5.1. Experimental Conditions

We used 178 walking radar images of 178 elderly participants in our experiment. A total of 150 images were used for training (48 apathy, 102 non-apathy), and the remaining 28 images were used for testing (eight apathy, 20 non-apathy). Each participant had one doppler radar image of a walking action.

Python 3.7 was used for programming the feature exaction and machine learning design. Anaconda was used as the standard platform. The hardware environment was a CPU (core i7 8th Gen, memory: 32 GB).

5.2. Overview of Accuracy

Figures 3–6 show the apathy classification accuracy when using only red channel, blue channel, green channel, and Y of YUV, respectively.

In these figures, the horizontal axis shows the threshold of binarization and the vertical axis shows the accuracy. Apathy-C denotes the correct classification rate of Apathy, Apathy-M the incorrect classification rate of apathy, Non-Apathy-C the correct classification rate of Non-Apathy, and Non-Apathy-M the incorrect classification rate of Non-Apathy. We should point out that this dataset was somewhat limited, as about 28.5% of the testing data was apathy data.

The seven sub-figures in each figure depict the respective apathy classification accuracy of each of the seven machine-learning models.

5.2.1. Red Channel

Figure 3 shows the apathy classification accuracy of using only the red channel. We found that SVM, decision tree, random forest, and ensemble performed poorly in Apathy-C. Naive Bayes achieved a good accuracy in Apathy-C, but its accuracy in Non-Apathy-C was very low. On the other hand, Naive Bayes achieved slight improvement in Non-Apathy-C during the threshold is 70 to 50. However, accuracy in Apathy-C is very low in these thresholds. Hence, its total accuracy (Non-Apathy-C + Apathy-C) was low.

In KNN, the total accuracy (Non-Apathy-C + Apathy-C) was more than 64% in the case of the threshold from 140 to 220, and in NN, the total accuracy was more than 75% in the case of the threshold from 150 to 220.

5.2.2. Green Channel

Figure 4 shows the apathy classification accuracy of using only the green channel. As with the experiment using the red channel, SVM, decision tree, random forest, and ensemble performed poorly in Apathy-C, and naive Bayes performed poorly in Non-Apathy-C.

In contrast to the results for the red channel, here KNN had a total accuracy (Non-Apathy-C + Apathy-C) of more than 65% in the case of the threshold from 130 to 170, and from 50 to 70. NN achieved a total accuracy of more than 75% in the case of the threshold from 160 to 200.

5.2.3. Blue Channel

Figure 5 shows the apathy classification accuracy of using only the blue channel. The same as when using the red and green channels, SVM, decision tree, random forest, and ensemble performed poorly in Apathy-C, and naive Bayes performed poorly in Non-Apathy-C.

NN also performed poorly here, and missed almost all of the apathy images. KNN did not achieve an accuracy of more than 71% in total. These results demonstrate that using only the blue channel degrades the accuracy.

5.2.4. Y of YUV

Figure 6 shows the apathy classification accuracy of using only the Y of YUV. As with the experiments with the red, green, and blue channels, SVM, decision tree, random forest, and ensemble performed poorly in Apathy-C, and naive Bayes performed poorly in Non-Apathy-C.

In addition, as in the experiment with the blue channel, KNN did not achieve an accuracy of more than 71% in total. As for NN, the total accuracy was more than 75% in the case of the threshold from 150 to 190.

5.3. Conclusion on Experimental Results

The results of the above experiments demonstrate that SVM, decision tree, random forest, and ensemble are not appropriate for use as machine-learning models for apathy classification of the elderly using doppler radar imaging. We conclude that KNN and NN are better models.

In terms of color channel, we found that the blue channel is not effective. Also, the Y of YUV is no better than the red or green channels, as Y is calculated using the blue channel. The accuracy of using Y is also just as bad as when using the blue channel, as only the slightest coefficient (0.114) is used for calculating Y.

When comparing all of the models and all of the thresholds, the proposed NN performed the best, with a total accuracy of more than 75%. The optimal threshold is from 150 to 190 when using red channel, green channel, and Y of YUV.

For giving more accurate analysis about NN, we list the experimental results about the accuracy of red channel, green channel, and Y of YUV during the threshold from 150 to 190 in **Figure 7**. The experimental results show the three channels achieve the same accuracy in NN, especially in the threshold from 160 to 180. (Note: Almost all of the Apathy can not be recognized correctly in blue channel by NN which was shown in **Figure 5**.)

We performed additional experiments to see if we could further improve the performance of NN by changing the number of layers, activation functions, and epochs, but no improvements were observed. Hence, we consider the optimal

FIGURE 3 | Experimental results: using red channel. **(A)** SVM, **(B)** KNN, **(C)** Naive Bayes, **(D)** decision tree, **(E)** random forest, **(F)** neural network, **(G)** ensemble.

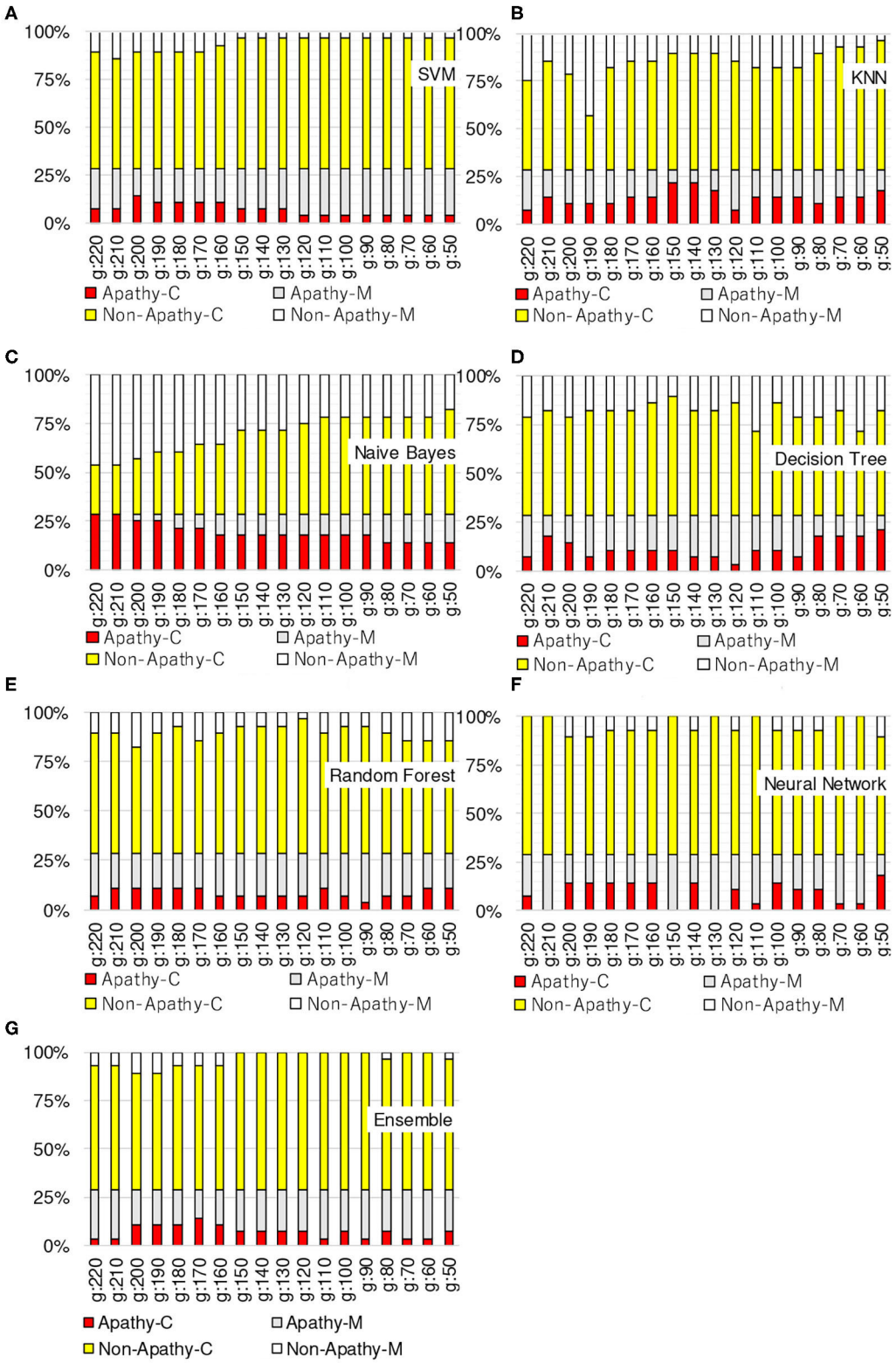


FIGURE 4 | Experimental results: using green channel. **(A)** SVM, **(B)** KNN, **(C)** Naive Bayes, **(D)** decision tree, **(E)** random forest, **(F)** neural network, **(G)** ensemble.

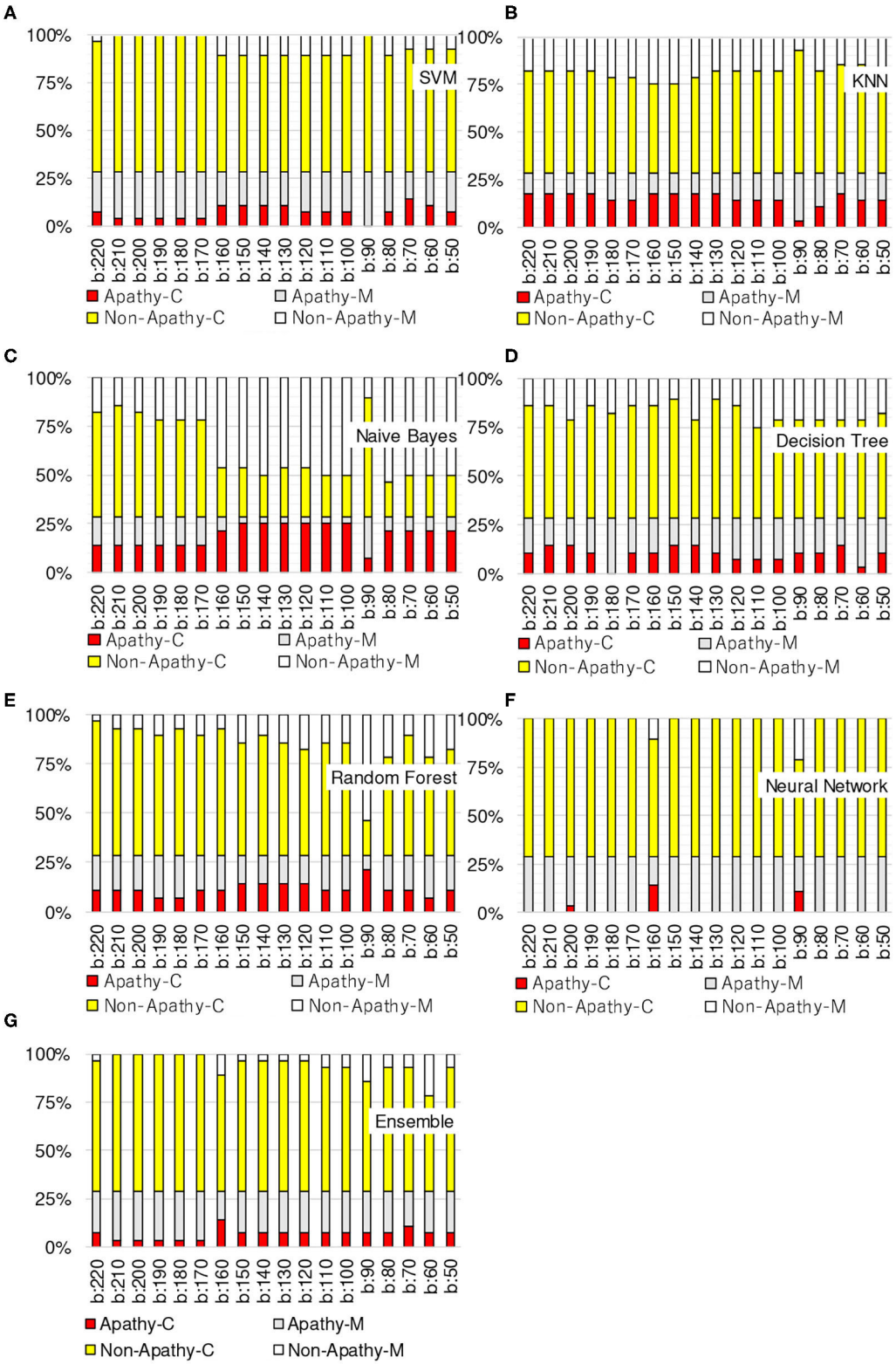


FIGURE 5 | Experimental results: using blue channel. **(A)** SVM, **(B)** KNN, **(C)** Naive Bayes, **(D)** decision tree, **(E)** random forest, **(F)** neural network, **(G)** ensemble.

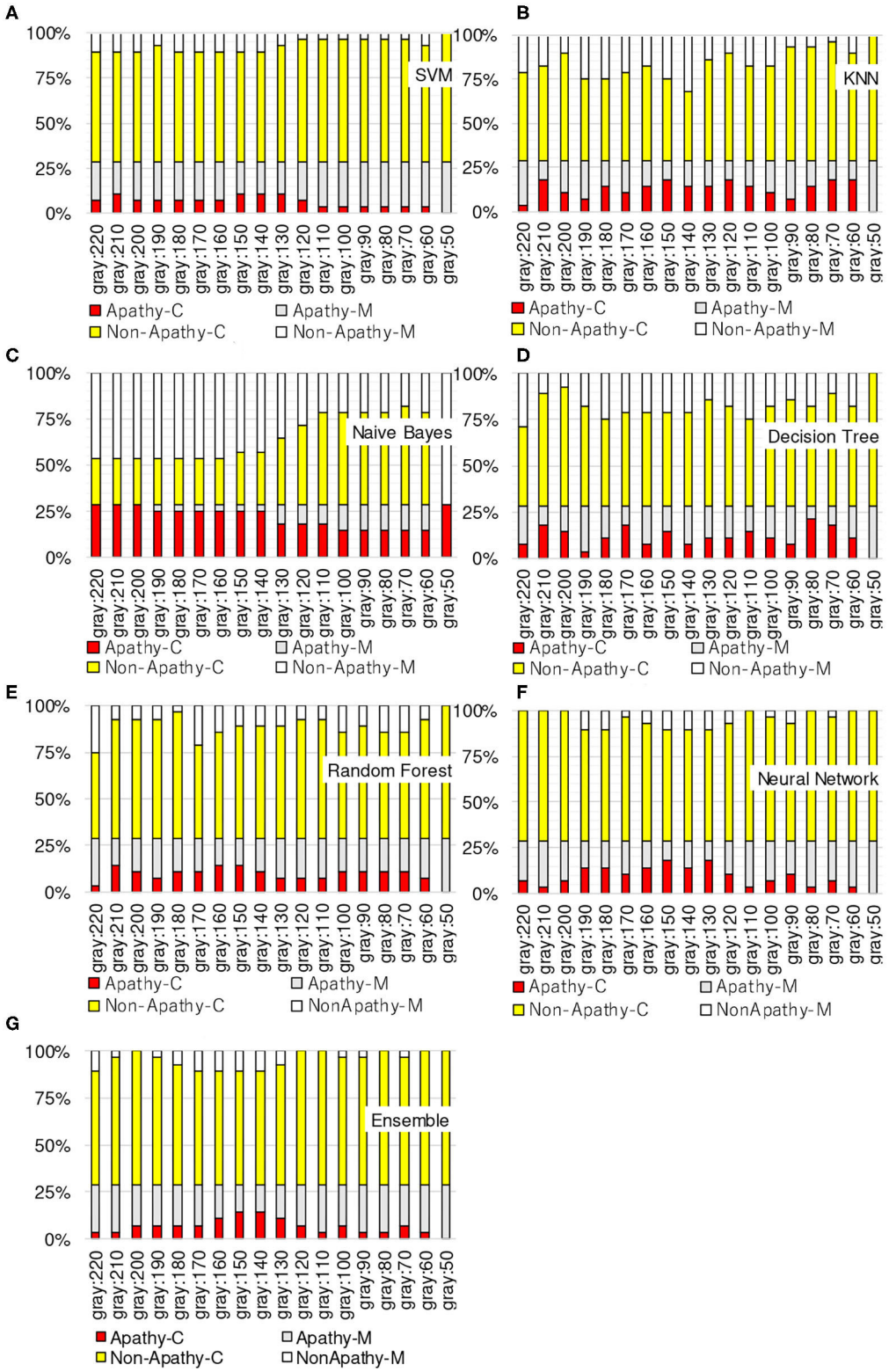


FIGURE 6 | Experimental results: using Y of YUV. **(A)** SVM, **(B)** KNN, **(C)** Naive Bayes, **(D)** decision tree, **(E)** random forest, **(F)** neural network, **(G)** ensemble.

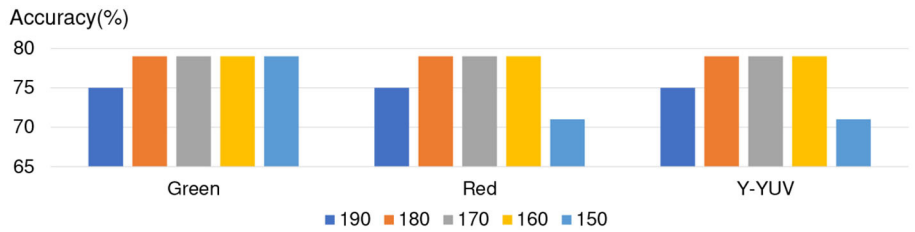


FIGURE 7 | Channels discussion about NN.

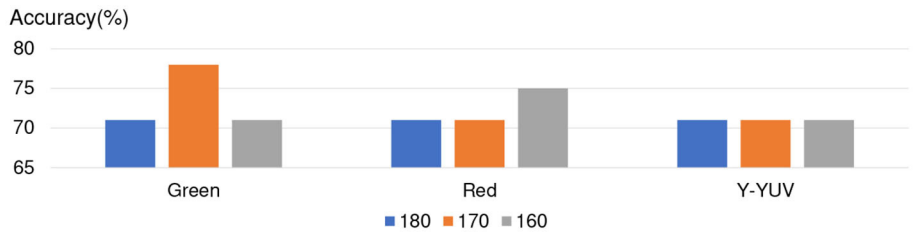


FIGURE 8 | Discussion about Ensemble model.

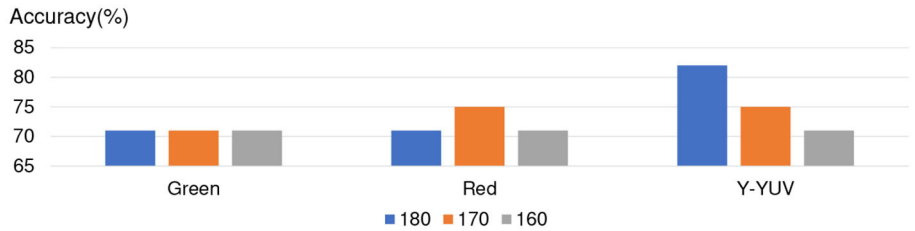


FIGURE 9 | Discussion on experimental results of NN (validation = 0.2).

model to be the proposed five-layer NN model (described in section 4).

Furthermore, for considering that Naive Bayes achieved better accuracy in Apathy-C and Neural Network achieved better accuracy in Non-Apathy-C. We only combines Naive Bayes and NN method in ensemble model for ensuring better accuracy. In term of the color channel and threshold, the red channel, green channel, and Y of YUV during the threshold from 150 to 190, are decided as the experimental condition. In these conditions, Naive Bayes achieved better accuracy in Apathy-C and Neural Network achieved better accuracy in Non-Apathy-C, by using the single model, respectively. The experimental results of Naive Bayes and NN combined ensemble model are listed in **Figure 8**, show that almost all of the cases only achieved 71% accuracy and were not better than NN.

In conclusion, five-layer slight Neural Network achieved better apathy classification accuracy based on Doppler Radar Image by using the red channel, green channel, and Y of YUV during the threshold from 160 to 180. For proving credibility of the conclusion, we separated 20% of training data as the validation data, and trained the NN again. The experimental results are shown in **Figure 9**, and achieved the similar results as without validation data.

5.4. Optimization in Image Separation

The 4 × 4 image separation presented in **Figure 2** considers the physical features of walking expressed on the radar images. Each separated image (**Figure 2D**) in [a]–[d] corresponds to the motion of each one step. The image [e] expresses the legs' motion in the stance phase of walking, [f] corresponds to body motion, and [g] expresses the legs' motion in the swing phase. The image [h] includes slight information on relatively large velocities of motions of toes or arms.

For proving the optimization of 4 × 4 image separation, we also added two separated method experimentation, including 4 × 5 and 5 × 5. The results of accuracy of 5 × 5 is shown in **Figure 10**, and 4 × 5 is shown in **Figure 11**. The experimental results show that the additional experimentation can not achieve better accuracy than the 4 × 4 image separation.

Hence, the 4 × 4 image separation is an optimized method which can be proved by the characters of image and the experimentation results.

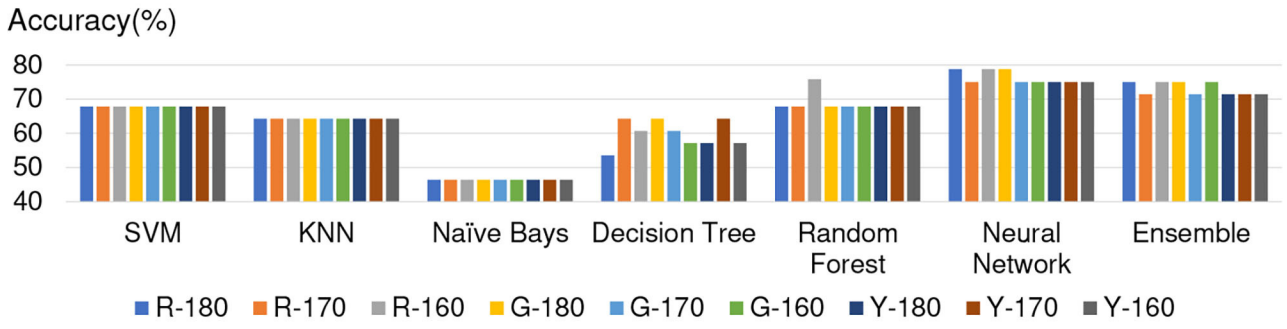


FIGURE 10 | Experimental results of the 5 × 5 image separation.

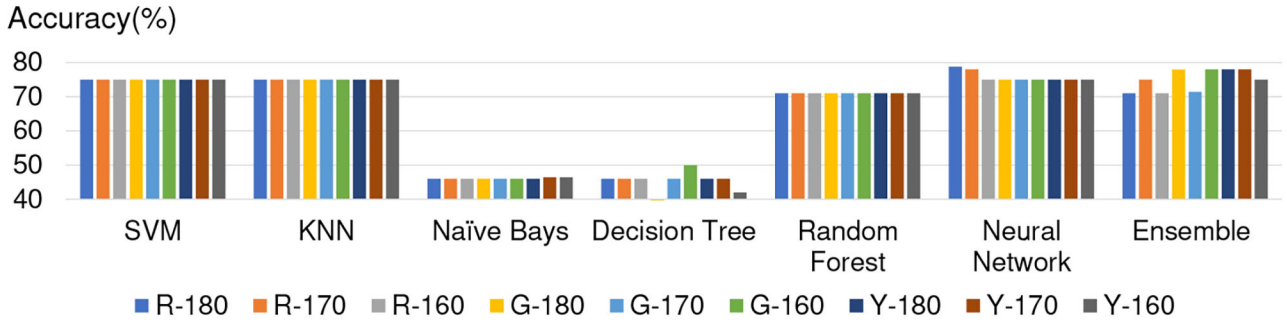


FIGURE 11 | Experimental results of the 4 × 5 image separation.

5.5. Experiment Using Deep-Learning Models

As various deep-learning models are proposed and used in the field of classification such as animal classification, characters classification etc. (Meng et al., 2018a,b, 2019), which achieved good accuracy. These models include LeNet (Lecun et al., 1998), AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015), VGG16, VGG19 (Simonyan and Zisserman, 2015), ResNet152V2 (He et al., 2015), Inception (Szegedy et al., 2016b), InceptionResNetV2 (Szegedy et al., 2016a), Xception (Chollet, 2017), and MobilNet (Howard et al., 2017) etc.

We also applied these models for measuring the accuracy of Apathy classification. **Figure 12** shows the experimental results of 11 state-of-the-art deep learning models for Apathy classification. The accuracy and the loss are listed. The results show that few of these models converged well such as ResNet, Inception, InceptionResNet, Mobile Net, VGG. Furthermore, the other models do not achieve better accuracy than the Machine learning models. Hence, the results demonstrate the difficulty of applying current deep-learning models to apathy classification using walking doppler radar images.

6. DISCUSSION

6.1. Effectiveness and Significance of Research

In terms of feature extraction, we applied binarization using only the red channel, green channel, blue channel, and Y of YUV, and slid the threshold from 50 to 220. We found that all of the machine learning models could not achieve high accuracy when using only the blue channel. We discussed seven machine-learning models for apathy classification and showed that in the red channel, green channel, and Y of YUV, the threshold from 150 to 190 resulted in the accuracy of more than 75%. This result demonstrates the effectiveness and significance of this research.

We feel this accuracy should be improved in the future, but even so, our findings here demonstrate the possibility of achieving an apathy classification method for the elderly that is both convenient and protects their privacy.

6.2. Limitations

The limitation of this research is that the dataset is small, only 178 elderly person are participants help for creating the dataset. Even if it is very hard to realizing the current dataset, and the some current research only uses Dozens of participants such as paper (Happy et al., 2019) has 45 participants, and paper (Liu et al., 2018) has 30 patients. For improving the accuracy and realizing the Practical, the dataset set should be increased.

Another limitation is the set place of the drop radar and the walking action. As this is an initial study, we kept things simple by setting the drop radar in front of the participants and having them perform the walking action on command. For practical use in production and diagnosis, these limitations need to be considered.

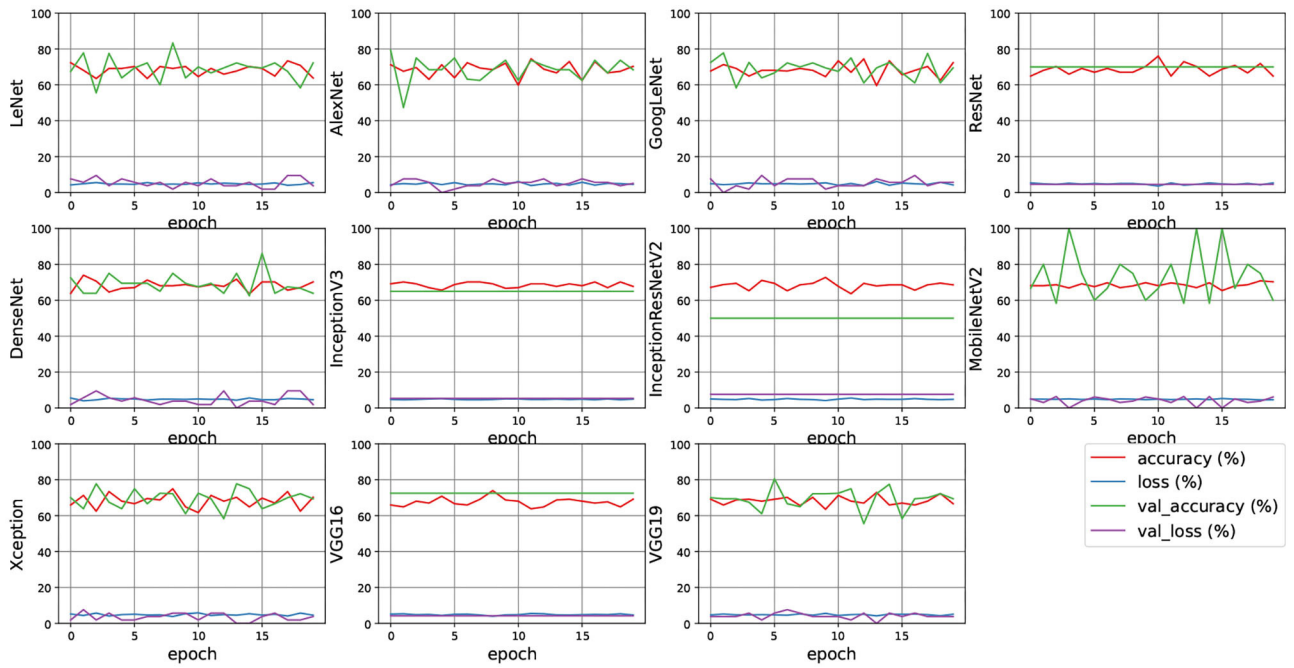


FIGURE 12 | Experimental results of deep learning models.

7. CONCLUSION

In this paper, we have examined using a walking action doppler radar image for the classification of apathy in the elderly. Walking is a common action in daily life and radar imaging is a good method in terms of privacy protection, so using a walking action doppler radar image may help us to achieve a diagnostic method that is both convenient and protects privacy. For the apathy classification, we proposed a method that combines image processing with machine learning. We had 168 elderly people help create a dataset by filling out a questionnaire to determine if they exhibited apathy or non-apathy and then used the results to train and test seven machine-learning models. The image processing consists of binarization, image separation, and feature pixel counting to extract features. We focused on pixel configuration for the binarization and slid the threshold from 50 to 220 to determine the optimized value. We then applied seven machine-learning models including our proposed NN model to a classification task by using the extracted features. We found that, in the red channel, green channel, and Y of YUV, the threshold from 150 to 190 resulted in an accuracy of more than 75%. This demonstrates the effectiveness of our approach and suggests its potential for achieving an apathy classification method for the elderly that is both convenient and protects their privacy. Further the EPADRI Dataset and the classification code are opened in our Lab website for reproducible study [http://www.ihpc.se.ritsumei.ac.jp/Publication.html: Apathy Dataset and Classification Code(2020)]. In future work, we will improve the accuracy further by increasing the size of the dataset.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The experimental protocol was approved by the local ethics committee (Toyama Prefectural University, approval no. H29-1). Participants were provided with written and verbal instructions of the testing procedures, and written consent was obtained from each participant prior to testing.

AUTHOR CONTRIBUTIONS

NN and ZM have the same contribution on experimentation and paper writing. KS and KU created the dataset and give advice on the data analysis. YD gave the advice on the data analysis on bio fields. CA and GA share the algorithms. HS and LM are supervisors of this project research. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by JSPS KAKENHI Grant Number 18K18337 and AMED.

REFERENCES

Breiman, L. (2001). Random forests. *Mach Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

CabinetOfficeJapan (2019). *Annual Report on the Ageing Society fy* 2019. Available online at: https://www8.cao.go.jp/kourei/english/annualreport/2019/pdf/2019.pdf

Caeiro, L., Ferro, J. M., and Costa, J. (2013). Apathy secondary to stroke: a systematic review and meta-analysis. *Cerebrovasc. Dis*. 35, 23–39. doi: 10.1159/000346076

Charles, P. (2003). *Digital Video and HDTV Algorithms and Inter-Faces*. San Francisco, CA: Morgan Kaufmann.

Chen, Q., Tan, B., Chetty, K., and Woodbridge, K. (2016). "Activity recognition based on micro-dopplersignature with in-home WI-FI," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)* (Munich). doi: 10.1109/HealthCom.2016.7749457

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolution," in *IEEE Conference on Pattern Recognition and Computer Vision, PRCV 2017* (Hawaii, HI). doi: 10.1109/CVPR.2017.195

Cun, Y. L. (1989). *Generalization and network design strategies, Technical Report CRG-TR-98-4*. University of Toronto Connectionist Research Group.

den Brok, M. G., van Dalen, J. W., van Gool, W. A., van Charante, E. P. M., de Bie, R. M., and Richard, E. (2015). Apathy in Parkinson's disease: a systematic review and meta-analysis. *Mov. Disord*. 30, 759–769. doi: 10.1002/mds.26208

Fuh, J., Wang, S., and JL., J. C. (2005). Neuropsychiatric profiles in patients with Alzheimer's disease and vascular dementia. *Neurol. Neurosurg. Psychiatry* 76, 1337–1041. doi: 10.1136/jnnp.2004.056408

Handri, S., Nakamura, K., and Nomura, S. (2009). Gender and age classification based on pattern of human motion using choquet integral agent networks. *J. Adv. Comput. Intell. Intell. Informat*. 13, 481–488. doi: 10.20965/jaciii.2009.p0481

Happy, S. L., Dantcheva, A., Das, A., Zeghari, R., Robert, P., and Bremond, F. (2019). "Characterizing the state of apathy with facial expression and motion analysis," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)* (Lille). doi: 10.1109/FG.2019.8756545

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv:1512.03385*. doi: 10.1109/CVPR.2016.90

Holden, M. K., Gill, K. M., Magliozzi, M. R., Nathan, J., and Piehl-Baker, L. (1984). Clinical gait assessment in the neurologically impaired. reliability and meaningfulness. *Phys. Ther*. 60, 35–40. doi: 10.1093/ptj/64.1.35

Homma, T., Atlas, L., and Marks, R. (1998). "An artificial neural network for Spatio-temporal bipolar patters: application to phoneme classification," in *Advances in Neural Information Processing Systems* (Denver, CO), 31–40.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Wey, T., et al. (2017). Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*

Jehn, M., Schmidt-Trucksäess, A., Schuster, T., H. Hanssen, M. W., Halle, M., and Koehler, F. (2009). Accelerometer-based quantification of 6-minute walk test performance in patients with chronic heart failure: applicability in telemedicine. *J. Cardiac Failure* 15, 334–340. doi: 10.1016/j.cardfail.2008.11.011

Jin, B., Thu, T. H., Baek, E., Sakong, S. H., Xiao, J., Mondal, T., et al. (2014). Walking-age analyzer for healthcare applications. *IEEE J. Biomed. Health Inform*. 18, 1034–1042. doi: 10.1109/JBHI.2013.2296873

Juen, J., Cheng, Q., and Schatz, B. (2015). A natural walking monitor for pulmonary patients using mobile phonese. *IEEE J. Biomed. Health Inform*. 19, 1399–1405. doi: 10.1109/JBHI.2015.2427511

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems, NIPS 2012*. (Lake Tahoe, NV)

Landes, A., Sperry, S., Strauss, M., and Geldmacher (2001). Apathy in Alzheimer's disease. *J. Am. Geriatr. Soc*. 49, 1700–1707. doi: 10.1046/j.1532-5415.2001.49282.x

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Li, W., Tan, B., and Piechocki, R. (2018). Passive radar for opportunistic monitoring in E-health applications. *IEEE J. Transl. Eng. Health Med*. 6, 1–10. doi: 10.1109/JTEHM.2018.2791609

Li, X., He, Y., and Jing, X. (2019). A survey of deep learning-based human activity recognition in radar. *Remote Sens*. 11. doi: 10.3390/rs11091068

Liu, Y., Batrancourt, B., Marin, F., and Levy, R. (2018). "Evaluation of apathy by single 3D accelerometer in ecological condition-case of patients with behavioral variant of fronto-temporal dementia," in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)* (Bogotá). doi: 10.1109/HealthCom.2018.8531167

Lu, X., Qian, X., Li, X., Miao, Q., and Peng, S. (2019). DMCM: a data-adaptive mutation clustering method to identify cancer-related mutation clusters. *Bioinformatics* 35, 389–397. doi: 10.1093/bioinformatics/bty624

Lu, X., Wang, X., Ding, L., Li, J., Gao, Y., and He, K. (2015). frDriver: A functional region driver identification for protein sequence. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 14. doi: 10.1109/TCBB.2020.3020096

Makihara, Y., Mannami, H., and Yagi, Y. (2011). Gait analysis of gender and age using a large-scale multi-view gait database. *Proc. Asian Conf. Comput. Vis*. 6493, 440–451. doi: 10.1007/978-3-642-19309-5_34

Marin, R. S. (1990). Differential diagnosis and classification of apathy. *Am. J. Psychiatry* 147, 22–30. doi: 10.1176/ajp.147.1.22

Marin, R. S. (1991). Apathy: a neuropsychiatric syndrome. *J. Neuropsychiat. Clin. Neurosci*. 3, 243–254. doi: 10.1176/jnp.3.3.243

Marin, R. S., Biedrzycki, R., and Firinciogullari, S. (1991). Reliability and validity of the apathy evaluation scale. *Psychiatr. Res*. 38, 143–162. doi: 10.1016/0165-1781(91)90040-V

Meng L., Aravinda, C. V., Uday Kumar Reddy, K. R., Izumi, T., and Yamazaki, K. (2018a). "Ancient Asian character recognition for literature preservation and understanding," in *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*. EuroMed 2018. Lecture Notes in Computer Science, Vol. 11196, ed Ioannides M, et al. (Cham: Springer), 741–751. doi: 10.1007/978-3-030-01762-0_66

Meng, L., Hirayama, T., and Oyanagig, S. (2018b). Underwater-drone with panoramic camera for automatic fish recognition based on deep learning. *IEEE Access*. 6, 17880–17886. doi: 10.1109/ACCESS.2018.2820326

Meng, L., Lyu, B., Zhang, Z., Aravinda, C., Kamitoku, N., and Yamazaki, K. (2019). "Ocrale bone inscription detector based on SSD," in *New Trends in Image Analysis and Processing' ICIAP 2019*, Lecture Notes in Computer Science, Vol. 11808 (Trento), 126–136. doi: 10.1007/978-3-030-30754-7_13

Naomi, S. A. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat*. 46, 175–185. doi: 10.1080/00031305.1992.10475879

Okada, K., Kobayashi, S., Yamagata, S., Takahashi, K., and Yamaguchi, S. (1997). Poststroke apathy and regional cerebral blood flow. *Stroke* 28, 2437–2441. doi: 10.1161/01.STR.28.12.2437

Opitz, D., and Maclin, R. (1999). Popular ensemble methods: an empirical study. *J. Artif. Intell. Res*. 11, 169–198. doi: 10.1613/jair.614

Pagonabarraga, J., Kulisevsky, J., Strafella, A., and Krack, P. (2015). Apathy in Parkinson's disease: clinical features, neural substrates, diagnosis, and treatment. *Lancet Neurol*. 14, 518–531. doi: 10.1016/S1474-4422(15)00019-8

Pitta, F., Troosters, T., Spruit, M. A., Probst, V. S., Decramer, M., and Gosselink, R. (2005). Characteristics of physical activities in daily life in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med*. 171, 972–977. doi: 10.1164/rccm.200407-855OC

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits Syst. Mag*. 6, 21–45. doi: 10.1109/MCAS.2006.1688199

Quinlan, J. (1886). Induction of decision trees. *Mach. Learn*. 1, 81–106. doi: 10.1007/BF00116251

Rabinovich, R. A., Louvaris, Z., Raste, Y., D. Langer, H. V. R., Giavedoni, S., Burtin, C., et al. (2013). Validity of physical activity monitors during daily life in patients with COPD. *Eur. Respir. J*. 42, 1205–1215. doi: 10.1183/09031936.00134312

Saho, K., Uemura, K., Fujimoto, M., and Matsumoto, M. (2020). Evaluation of higher-level instrumental activities of daily living via micro-doppler radar sensing of sit-to-stand-to-sit movement. *IEEE J. Transl. Eng. Health Med*. 8:2100211. doi: 10.1109/JTEHM.2020.2964209

Seifert, A. K., Amin, M., and Zoubir, A. M. (2019). Toward unobtrusive in-home gait analysis based on radar micro-doppler signatures. *IEEE Trans. Biomed. Eng*. 66, 2629–2640 doi: 10.1109/TBME.2019.2893528

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in *Advances in Neural Information Processing Systems, NIPS 2015* (Montreal, QC).

Starkstein, S. E., Mayberg, H. S., Preziosi, T., Andrezejewski, P., Leiguarda, R., and Robinson, P. G. (1992). Reliability, validity, and clinical correlates of apathy in Parkinson's disease. *J. Neuropsychiatry Clin. Neurosci.* 4, 134–139. doi: 10.1176/jnp.4.2.134

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2016a). Inception-v4, inception-ResNet and the impact of residual connections on learning. *arXiv:1602.07261*

Szegedy, C., Liu, W., Jia, Y. Q., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015* (Boston, MA). doi: 10.1109/CVPR.2015.7298594

Szegedy, C., Liu, W., Jia, Y. Q., Sermanet, P., Reed, S., Anguelov, D., et al. (2016b). "Rethinking the inception architecture for computer vision," in *IEEE Conference on Pattern Recognition and Computer Vision, PRCV 2016* (Las Vegas, NV).

Vapnik, V. (1998). *Statistical Learning Theory*. Physical Therapy Reviews New York, NY: Wiley.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership