METHODS AND APPLICATIONS IN MOLECULAR PHYLOGENETICS

EDITED BY: Juan Wang, Quan Zou and Qiguo Dai

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-344-3
DOI 10.3389/978-2-88976-344-3

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

METHODS AND APPLICATIONS IN MOLECULAR PHYLOGENETICS

Topic Editors:
Juan Wang, Inner Mongolia University, China
Quan Zou, University of Electronic Science and Technology of China, China
Qiguo Dai, Dailian Minzu University, China

Citation: Wang, J., Zou, Q., Dai, Q., eds. (2022). Methods and Applications in Molecular Phylogenetics. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-344-3

Table of Contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 14 July 2022
doi: 10.3389/fgene.2022.923409

Editorial: Methods and Applications in Molecular Phylogenetics

Juan Wang *

School of Computer Science, Inner Mongolia University, Hohhot, China

Keywords: molecular phylogenetics, whole genome sequences, protein, application, disease

Editorial on the ResearchTopic

Methods and Applications in Molecular Phylogenetics

The purpose of molecular phylogenetics is to infer the evolutionary history of organisms and gene sequences. In the early stages of research, molecular phylogenetics mainly considers the changes vertically, such as insertion, substitution, and deletion in loci (Siepel and Haussler, 2004). With the development of sequencing technologies, the whole genomes are available for more and more organisms and are used to analyze their phylogenetics (Henz et al., 2005; Birin et al., 2008). The evolutionary history of organisms at this stage is described as a phylogenetic tree (Bruno et al., 2000). Then, genes of genomes are rearranged under horizontal events, such as inversions, duplications, and transpositions, which change the content and order of genes. Many studies introduce computing methods of molecular phylogenetics for whole genomes (Greenman et al., 2012). Phylogenetic networks are used to describe the evolutionary history (Wang and Guo, 2019). Molecular phylogenetics has been applied in many areas, such as the analysis of proteins (Lv et al., 2020).

Traditional methods for molecular phylogenetics need to do the alignment for sequences. It is very time-consuming to process the alignment of whole genome sequences. Therefore, it is a hard issue to do phylogenetic analysis from whole genome sequences of organisms. Wu et al. introduce a metric called information-entropy position-weighted k-mer relative measure (IEPWRMkmer), which combines the position-weighted measure and the information entropy of frequency for k-mers. Accordingly, they denote the whole genomes as feature sequences and then use Manhattan distance to compute the distance between two whole genomes. Finally, they use the Neighbor-Joining method to construct the phylogenetic tree from distance matrices. The IEPWRMkmer is efficient and effective for extracting key information for evolutionary analysis, and it is free to align for whole genomes.

Many studies have been done in applications of molecular phylogenetics. A protein complex contains proteins that interact with each other in function due to the evolutionary relationship. Wang et al. used semantic information of GO terms and the topological information of PPI networks to propose a method called TSSN for constructing a weighted PPI network. They proposed a new algorithm (NNP) for recognizing protein complexes from the weighted PPI network. Experiments showed that the algorithm could identify more protein complexes more accurately. PredMHC, proposed by Chen et al., is used to predict major histocompatibility complex (MHC). The PredMHC extracts information on amino acid composition from proteins, which is different due to the evolution of coding genes. It uses the voting of the SGD, the SMO, and random forest to predict and achieve the best performance on both training and testing datasets than other methods.

Molecular phylogenetics is also applied in predicting disease-related proteins. Anti-inflammatory peptides (AIPs) are important to treat some inflammatory and autoimmune diseases. Zhao et al. introduced a model (called iAIPs) to identify AIPs. iAIPs extract features from AIPs based on the information of sequences changed in evolution and then use the random forest to train.

Edited and reviewed by:
Simon Charles Heath,
Center for Genomic Regulation (CRG),
Spain

***Correspondence:**
Juan Wang
wangjuan@imu.edu.cn

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: *19 April 2022*
Accepted: *27 May 2022*
Published: *14 July 2022*

Citation:
Wang J (2022) Editorial: Methods and
Applications in
Molecular Phylogenetics.
Front. Genet. 13:923409.
doi: 10.3389/fgene.2022.923409

Experimental results show that iAIPs can identify AIPs accurately. Cancer is a serious threat to human health and is one of the main causes of disease death. MultiGATAE, proposed by Zhang et al., can identify the cancer subtypes. It first constructs a similarity graph from multi-omics data (i.e., mRNA, miRNA, and DNA methylation) and then uses a deep learning method to learn embedding representation. It uses the K-means clustering method to identify cancer subtypes from embedding representation.

AUTHOR CONTRIBUTIONS

JW wrote the manuscript.

REFERENCES

Birin, H., Gal-Or, Z., Elias, I., and Tuller, T. (2008). Inferring Horizontal Transfers in the Presence of Rearrangements by the Minimum Evolution Criterion†. *Bioinformatics* 24 (6), 826–832. doi:10.1093/bioinformatics/btn024

Bruno, W. J., Socci, N. D., and Halpern, A. L. (2000). Weighted Neighbor Joining: a Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction. *Mol. Biol. Evol.* 17 (1), 189–197. doi:10.1093/oxfordjournals.molbev.a026231

Greenman, C. D., Pleasance, E. D., Newman, S., Yang, F., Fu, B., Nik-Zainal, S., et al. (2012). Estimation of Rearrangement Phylogeny for Cancer Genomes. *Genome Res.* 22 (2), 346–361. doi:10.1101/gr.118414.110

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., and Schuster, S. C. (2005). Whole-genome Prokaryotic Phylogeny. *Bioinformatics* 21 (10), 2329–2335. doi:10.1093/bioinformatics/bth324

Lv, Z., Wang, P., Zou, Q., and Jiang, Q. (2020). Identification of Sub-golgi Protein Localization by Use of Deep Representation Learning Features. *Bioinformatics* 36 (24), 5600–5609. doi:10.1093/bioinformatics/btaa1074

Siepel, A., and Haussler, D. (2004). Phylogenetic Estimation of Context-dependent Substitution Rates by Maximum Likelihood. *Mol. Biol. Evol.* 21, 468–488. doi:10.1093/molbev/msh039

Wang, J., and Guo, M. (2019). A Review of Metrics Measuring Dissimilarity for Rooted Phylogenetic Networks. *Briefings Bioinforma.* 20 (6), 1972–1980. doi:10.1093/bib/bby062

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 October 2021
doi: 10.3389/fgene.2021.768747

Hypertension-Related Drug Activity Identification Based on Novel Ensemble Method

Bin Yang[1], *Wenzheng Bao*[2]* *and Jinglong Wang*[3]

[1]School of Information Science and Engineering, Zaozhuang University, Zaozhuang, China, [2]School of Information and Electrical Engineering, Xuzhou University of Technology, Xuzhou, China, [3]College of Food Science and Pharmaceutical Engineering, Zaozhuang University, Zaozhuang, China

Hypertension is a chronic disease and major risk factor for cardiovascular and cerebrovascular diseases that often leads to damage to target organs. The prevention and treatment of hypertension is crucially important for human health. In this paper, a novel ensemble method based on a flexible neural tree (FNT) is proposed to identify hypertension-related active compounds. In the ensemble method, the base classifiers are Multi-Grained Cascade Forest (gcForest), support vector machines (SVM), random forest (RF), AdaBoost, decision tree (DT), Gradient Boosting Decision Tree (GBDT), KNN, logical regression, and naïve Bayes (NB). The classification results of nine classifiers are utilized as the input vector of FNT, which is utilized as a nonlinear ensemble method to identify hypertension-related drug compounds. The experiment data are extracted from hypertension-unrelated and hypertension-related compounds collected from the up-to-date literature. The results reveal that our proposed ensemble method performs better than other single classifiers in terms of ROC curve, AUC, TPR, FRP, Precision, Specificity, and F1. Our proposed method is also compared with the averaged and voting ensemble methods. The results reveal that our method could identify hypertension-related compounds more accurately than two classical ensemble methods.

Keywords: hypertension, flexible neural tree, ensemble, network pharmacology, machine learning

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Zhiwei Ji,
Nanjing Agricultural University, China
Chunhou Zheng,
Anhui University, China

****Correspondence:***
Wenzheng Bao
baowz55555@126.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *01 September 2021*
Accepted: *27 September 2021*
Published: *15 October 2021*

Citation:
Yang B, Bao W and Wang J (2021) Hypertension-Related Drug Activity Identification Based on Novel Ensemble Method.
Front. Genet. 12:768747.
doi: 10.3389/fgene.2021.768747

INTRODUCTION

Hypertensive disease is a frequent cardiovascular disease characterized by elevated arterial blood pressure and accompanied by the target organ injury or clinical diseases (Essiarab et al., 2011; Owlia and Bangalore, 2016). It is a risk factor leading to many serious complications such as stroke, hypertensive heart disease, renal failure, atherosclerosis, and so on (Sakai and Sigmund, 2005; Brinks and Eckhart, 2010). Due to the increasing pressure of work and life, many people do not develop good eating and living habits, and often stay up late. The age of hypertensive patients tends to be younger. Therefore, the prevention and treatment of hypertension has become very important for human health.

Network pharmacology (NP) could construct a multi-dimensional network based on “traditional Chinese medicine prescription-chemical component-targets-disease targets” to analyze the relationships between traditional Chinese medicine multi-components and activity, which could provide a theoretical basis for further experimental research on a pharmacodynamic material basis and action mechanism (Wang et al., 2018; Xu et al., 2018). In recent years, network pharmacology has revealed therapeutic targets for hypertension and become a research hotspot, as it has been clinically verified to be an effective method of drug screening (Chen et al., 2020). Chen et al. screened out the key compounds and targets of JiaWeiSiWu granule to reveal the mechanism of JiaWeiSiWu granule in treating hypertension by NP method (Chen et al., 2021a). By NP and molecular docking (MD) methods, Zhai et al. investigated the mechanism of Pinellia ternate in treating hypertension (Zhai et al., 2021). Chen et al. analyzed the network based on Guizhi decoction, active compounds, and targets, and found hypertension-related targets and key pathways (Chen et al., 2021b). Chen et al. utilized NP and MD to analyzed the genistein for treating pulmonary hypertension (PH) and provided new guidance for further PH-related research (Chen et al., 2019). Liu et al. explained the pharmacological mechanism of TaohongSiwu decoction in the treatment of essential hypertension (EH) by the NP method (Liu et al., 2020). Wang et al. utilized NP to analyze the mechanism of Yeju Jiangya decoction against hypertension (Wang et al., 2021).

In recent decades, many data mining methods have been applied to reveal the disease mechanism and medication law of many complex diseases, especially hypertension (Ji and Wang, 2014; Ji et al., 2015; Hwang et al., 2016; Hu et al., 2018; Liang et al., 2018; Amaratunga et al., 2020; Liu et al., 2021; Zhao et al., 2021). Zhang et al. utilized SPSS21.0 and Apriori algorithm to analyze the symptom/sign information of EH patients collected and gave their distribution law and correlation (Zhang et al., 2019a). Yuan and Chen proposed niche technology and an artificial bee colony algorithm to mine association rules from Traditional Chinese Medicine (TCM) cases for treating hypertension (Yuan and Chen, 2011). Ma et al. collected the new literature about hypertension and constructed the gene network by analysis (Ma et al., 2018). Ramezankhani et al. utilized a decision tree to predict the risk factors of hypertension incidence in data collected from Iranian adults (Ramezankhani et al., 2016). Aljumah et al. utilized a data mining method to predict the treatment of hypertension patients with different age groups (Aljumah et al., 2011). Fang et al. proposed a new model-based KNN and LightGBM to predict the risk of hypertension (Fang et al., 2021).

Few studies have involved the use of data mining methods to improve network pharmacology. In this paper, a novel ensemble method based on a flexible neural tree (FNT) is proposed to identify hypertension-related active compounds. In the ensemble method, the used base classifiers are Multi-Grained Cascade Forest, support vector machines, random forest, AdaBoost, decision tree, Gradient Boosting Decision Tree, KNN, logical regression, and naïve Bayes. The classification results of nine classifiers are input to the FNT model, which is trained to predict hypertension-related compounds. The data used in the experiment are from up-to-date literature collected about hypertension and network pharmacology. By analysis of the literature, hypertension-related compounds were collected as positive samples and the generated decoys were utilized as negative samples. The molecular descriptor of each compound is extracted as the feature vector.

METHODS

Classifiers

Assume that the training data is $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ containing n sample points. Sample point $x_i = \{x_i^1, x_i^2, \ldots x_i^m\}$ contains m features and category label $y_i = \{c_1, c_2\}$ contains two cases. The nine classifiers used are introduced in the following sections of the article.

Multi-Grained Cascade Forest

Multi-Grained Cascade Forest (gcForest) is a novel ensemble machine learning method, which utilizes the cascade forest (ensemble of decision trees) to learn and generate models (Zhou and Feng, 2017). The core of gcForest mainly includes two modules: multi-grained scanning and cascade forest. The flowchart of gcForest is depicted in **Figure 1**.

1) Multi-grained scanning

Multi granularity scanning is a technical means to enhance cascade forest and do more processing on features. Firstly, a complete m- dimensional sample is input, and then sliding sampling is carried out through the k_1-dimensional and k_2-dimensional sampling windows in order to obtain $s_1 = (m - k_1) + 1$ and $s_2 = (m - k_2) + 1$ feature subsample vectors, respectively. Each sub-sample is used for the training of completely random forest (A) and random forest (B). A probability vector with 2-dimension is obtained in each forest, so that two kinds of forests can produce $2s_1$ and $2s_2$ representation vectors, respectively. Finally, the results of all forests are spliced together to obtain the sample output.

2) Cascade forest

Cascade forest includes several layers, each layer is composed of many forests, and each forest is composed of many decision trees. Completely random forest (A) and random forest (B) in each layer ensure the diversity of the model. For a completely random forest, each tree in the forest randomly selects a feature as the splitting node of the splitting tree, which grows until each leaf node is subdivided into only one class. For random forest, each tree randomly selects $\sqrt{m}$candidate features, and the splitting nodes are filtered through the Gini coefficient. Each forest could generate a two-dimensional class vector. The two-dimensional class vectors of all forests are averaged to obtain the final two-dimensional class vector. Finally, the category with the maximum value in the final two-dimensional class vector is taken as the final classification result.

Support Vector Machines

Support vector machines (SVM) is a supervised learning algorithm based on statistical learning theory (Suykens and Vandewalle, 1999; Furey et al., 2000). With the sample set containing positive and negative samples, SVM could search a hyperplane that could segment the samples according to positive

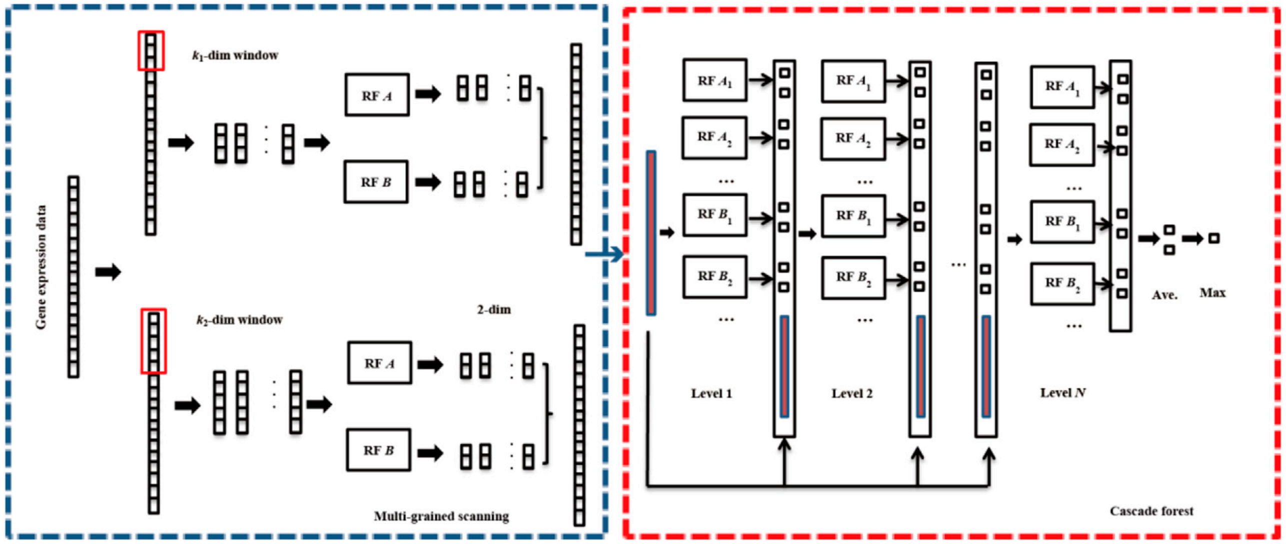

FIGURE 1 | The process of gcForest.

and negative classes. The classification hyperplane can be given as follows.

$$w^T x + b = 0. \quad (1)$$

Where x is the data point on the classification hyperplane, w is a vector perpendicular to the classification hyperplane, and b is the displacement.

Linear separated data can be distinguished by the optimal classification hyperplane. For non-linear separated data, SVM can be transformed into solving the following optimization problem by the soft interval optimization and kernel techniques.

$$\left\{ \min \phi(w, \varsigma) = \|w\|^2 + \frac{1}{2} C \sum_{i=1}^{n} \varsigma_i s.t. y_i [(w \cdot x_i + b)] \geq 1 - \varsigma_i. \right. \quad (2)$$

Where $C$is the penalty factor, ς_i is the relaxation variable, and x_i is mapped to a high-dimensional space by ϕ. SVM could find a hyperplane with the largest interval in this high-dimensional space to classify the data.

Random Forest

Random forest (RF) is a machine learning method based on an ensemble of decision trees for classification and regression (Breiman, 2001; Díaz-Uriarte and Alvarez de Andrés, 2006). Random forest is a combined classification model composed of many decision tree classification models. Each decision tree has the right to vote to determine the best classification result. In random forest, firstly, K sample sets are extracted from the original training set by bootstrap sampling method, and the size of each extracted sample set is the same as that of the original training set. Then, K decision tree models are established from K sample sets, respectively. And K trees will create K classification results. The random forest integrates all the classified results by voting method, and the category with the most votes is designated as the final classification result.

AdaBoost

AdaBoost is a dynamic ensemble classification algorithm, which is to reasonably combine multiple weak classifiers (single-layer decision tree) to make it a strong classifier (Morra et al., 2009; Cao et al., 2013). The detailed algorithm is given as follows.

1) Initialize the weight of each sample. Assuming that the dataset contains n samples, each training sample point is given the same weight ($\frac{1}{n}$) at the beginning.
2) Train weak classifiers. According to the samples, the weak classifiers are trained. If a sample has been accurately classified, its weight will be reduced in constructing the next training set. On the contrary, if a sample point is not accurately classified, its weight is increased. At the same time, according to the classification error of the weak classifier, its weight is calculated. Then, the sample set with updated weights is used to train the next classifier, and the whole training process goes on iteratively. T weak classifiers are obtained after T iterations.
3) The trained weak classifiers are combined into strong classifiers. Each weak classifier connects its respective weights through the classification function to form a strong classifier. After the training process of each weak classifier, the weight of the weak classifier with a smaller classification error rate is larger, which plays a greater decisive role in the final classification function, while the weight of the weak classifier with a larger classification error rate is smaller, which plays a smaller decisive role in the final classification function.

Decision Tree

A Decision Tree (DT) learning algorithm is usually a process of recursively selecting the optimal features and segmenting the training data according to the features so that each sub dataset has the best classification. The CART algorithm is one of the most common decision tree algorithms, which is mainly used for classification and regression (Breiman et al., 1984; Temkin et al., 1995). CART introduces the knowledge of probability theory and statistics into the research of decision tree. Different from the C4.5 algorithm, the CART algorithm could make a binary partition of the feature space and can split scalar attributes and continuous attributes. The specific algorithm is as follows:

1) Calculate the Gini index of the existing features. The feature with the smallest Gini index is selected as the splitting attribute of the root node. According to the optimal feature and cut point, two sub-nodes are generated from the current node, and the training dataset is allocated to the two sub-nodes according to the feature. According to an attribute value, a node is segmented to make the data in each descendant subset more "pure" than the data in its parent subset. Gini coefficient measures the impurity of sample division, and the smaller the impurity is, the higher the "purity" of the samples is.

For 2-class problems, the training set S is divided into two subsets S_1 and S_2 according to an attribute A. The Gini coefficient of the given division S is calculated as follows.

$$Gini_A(S) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2). \tag{3}$$

Where $|S|$ is the number of samples in set S, and $Gini(S_i)$ is the Gini coefficient of sample set S_i, which is calculated as follows:

$$Gini(S_i) = 1 - \sum_{k=1}^{2} \left(\frac{|C_k|}{|S_i|}\right)^2. \tag{4}$$

Where $|C_k|$ denotes the number of samples belonging to class k in the set S_i.

2) Step (1) is called recursively for two child nodes, and the iteration continues until the samples in all child nodes belong to the same category or no attributes can be selected as splitting attributes.
4) Prune the CART decision tree generated.

Gradient Boosting Decision Tree

Gradient Boosting Decision Tree (GBDT) is an integrated learning algorithm (Hu and Min, 2018; Zhang et al., 2019b). By boosting method, N weak learners are created, which are combined into a strong learner after many iterations. The performance of the strong learner is higher than any weak learner. In GBDT, the used weak learner is the CART regression tree. During each iteration of GBDT, the residual of the previous model is reduced, and a new model is trained and established in the gradient direction of residual reduction, to improve the performance of the classifier. The specific algorithm is shown as follows:

1) Initialize the weak learner.

$$f_0(x) = \text{argmin}_\kappa \sum_{i=1}^{n} L(y_i, \kappa). \tag{5}$$

Where L is the loss function.

2) For $t-th$ iteration $(t = 1, 2, \ldots, T)$

a) For $i-th$ sample, the residual reduction is calculated as follows.

$$r_{ti} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}. \tag{6}$$

Where $f_{t-1}(x)$ is the classifier during the $t-1-th$ iteration.

$$\kappa_{tj} = \text{argmin}_\kappa \sum_{x_i \in R_{tj}} L(y_i, f_{t-1}(x_i) + \kappa). \tag{7}$$

Where κ_{tj} is the value of the leaf node in the regression tree.

b) The calculated residues are used as new sample data, (x_i, r_{ti}) is utilized to fit a new CART regression tree and the probability of each category is calculated. The leaf node region of the CART regression tree R_{tj} $(j = 1, 2, \ldots, J)$ is obtained. J is the number of leaf nodes of the regression tree.
c) Calculate the optimal coefficient for the leaf area, which is given as follows.
d) The strong learner is updated with **Eq. 8**.

$$f_t(x) = f_{t-1}(x_i) + \sum_{j=1}^{J} \kappa_{tj} I(x \in R_{tj}). \tag{8}$$

When$x \in R_{tj}$ is true, I is equal to 1; otherwise, it is equal to 0.

3) The final strong learner $f(x)$ is obtained with **Eq. 9**.

$$f(x) = f_0(x) + \sum_{t=1}^{T} \sum_{j=1}^{J} c_{tj} I(x \in R_{tj}). \tag{9}$$

K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a classification algorithm based on supervised learning, which is to classify the data points according to the sample set with the known categories (Liao and Vemuri, 2002). Select the K neighbors with the smallest distance from the input data in the training set, and take the category with the most times among the K neighbors as the category of the classified data point. In the KNN algorithm, the selected neighbors are objects that have been correctly classified.

In the KNN method, the most commonly used measurement of distance is the Euclidean distance. The Euclidean distance of two variables (x_i and x_j) is defined as follows.

$$D\left(\left(x_i, x_j\right)\right) = \sqrt{\sum_{k=1}^{m}\left(x_i^k - x_j^k\right)^2}. \tag{10}$$

Logistic Regression

Logistic regression (LR) is utilized to deal with the regression problem, which obtains the minimum result of cost function by gradient descent method to obtain the better classification boundary (Maalouf, 2011; Munshi et al., 2014). LR maps the values of linear regression to the interval [0, 1] by Sigmoid function, which is defined as follows.

$$y_i = h_\theta\left(x_i\right) = \frac{1}{1 + e^{-\theta^T x_i}}. \tag{11}$$

Where $\theta^T x_i = \theta_0 + \theta_1 x_i^1 + \theta_1 x_i^2 + \ldots + \theta_m x_i^m$, θ_0 is a deviation parameter and θ_i represents the weight.

In order to solve the logistic regression model, the gradient descent algorithm is generally used to iteratively calculate the optimal parameters of the model.

Naïve Bayes

Naïve Bayes (NB) is one of the most widely utilized models in Bayesian classifiers, which is based on the assumption that the influence of an attribute value on the given class is independent of the values of other attributes (class conditional independence) (Rish, 2001; Li and Guo, 2005). The specific algorithm idea is as follows.

According to the joint probability and the prediction data x, the prediction category of x is defined as follows.

$$\arg\max p\left(y = c_k \middle| x\right). \tag{12}$$

According to the Bayesian theorem, $p\left(y = c_k|x\right)$ is calculated as follows.

$$p\left(y = c_k \middle| x\right) = \frac{p\left(x \middle| y = c_k\right) p\left(y = c_k\right)}{p(x)}. \tag{13}$$

Since the denominator is constant for all categories, just maximize the numerator, and **Eq. 12** could be defined as follows.

$$\arg\max p\left(x \middle| y = c_k\right) p\left(y = c_k\right). \tag{14}$$

Because each feature attribute is conditionally independent, $p\left(x|y = c_k\right)$ could be calculated as follows.

$$p\left(x \middle| y = c_k\right) = \prod_{i=1}^{m} p\left(x^i \middle| y = c_k\right) \tag{15}$$

According to **Eq. 15**, **Eq. 14** can be calculated as follows.

$$\arg\max p\left(y = c_k\right) \prod_{i=1}^{m} p\left(x^i \middle| y = c_k\right) \tag{16}$$

Select the category with the largest posteriori probability as the prediction category.

Ensemble Methods

To improve the classification performance of a single classifier, a novel ensemble method based on a flexible neural tree (FNT) is proposed. An example of our proposed ensemble method is depicted in **Figure 2**. From **Figure 2**, it could be seen that the used base classifiers are gcForest, SVM, RF, AdaBoost, decision tree, GBDT, KNN, logical regression, and naïve Baye, which are introduced in detail in *Classifiers*. Firstly according to the training data, these nine classifiers can output their corresponding confidence level set ($c = (c_1, c_2, \ldots, c_9)$), which is utilized as the input layer of the FNT model. The other hidden layers of the FNT model can be created randomly from operator set ($F = (+_2, +_3, \ldots, +_n)$) and variable set ($T = (c_1, c_2, \ldots, c_9)$) (Chen et al., 2006). $+_i$ denotes a flexible neuron operator, which can be calculated as follows:

$$\begin{cases} net_i = \sum_{j=1}^{i} w_j x_j, \\ o_i = f\left(a_i, b_i, net_i\right) = e^{-\left(\frac{net_i - a_i}{b_i}\right)^2}. \end{cases} \tag{17}$$

Where $f(\cdot)$ is an activation function, a_i and b_i are the parameters of function, x_j is the input variable and w_j is the corresponding weight of the input variable.

FNT is a kind of cross-layer neural network, so each hidden layer can contain both operator and variable nodes. Because the structure of the FNT model is not fixed and this model contains many parameters such as a_i, b_i, and w_j, many swarm algorithms have been proposed to search the optimal FNT model by iterations. In this paper, a hybrid evolutionary method based on genetic programming like structure optimization algorithm and simulated annealing was utilized for the training dataset. The detailed algorithms were introduced in another study (Yang et al., 2013).

Hypertension-Related Activity Drug Identification

In order to identify hypertension-related active compounds accurately, an ensemble method based on nine classifiers and a flexible neural tree is proposed. The process of hypertension-related active compounds identification is depicted in **Figure 3**. A total of 44 important studies were collected by querying the literature database according to two keywords: hypertension and network pharmacology. Through analyzing this literature, many important medicines such as Banxia Baizhu Tianma Tang, Chaihu Longgu Muli Decoction, compound reserpine and triamterene tablets, and Huanglian Jiedu Decoction, were collected and 88 hypertension-related compounds were searched. These important compounds were verified by biology experiments or molecular docking, which were used as positive samples in this paper. To obtain the negative samples, 20% of these compounds were randomly selected and input into the DUD•E website to generate decoys (Mysinger et al., 2012). In total, 264 decoys are selected randomly as negative samples.

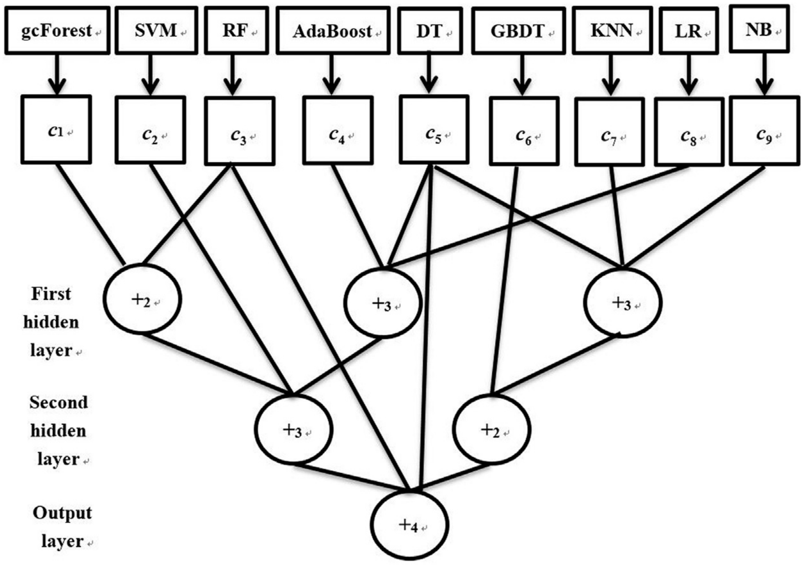

FIGURE 2 | The flowchart of our proposed ensemble method.

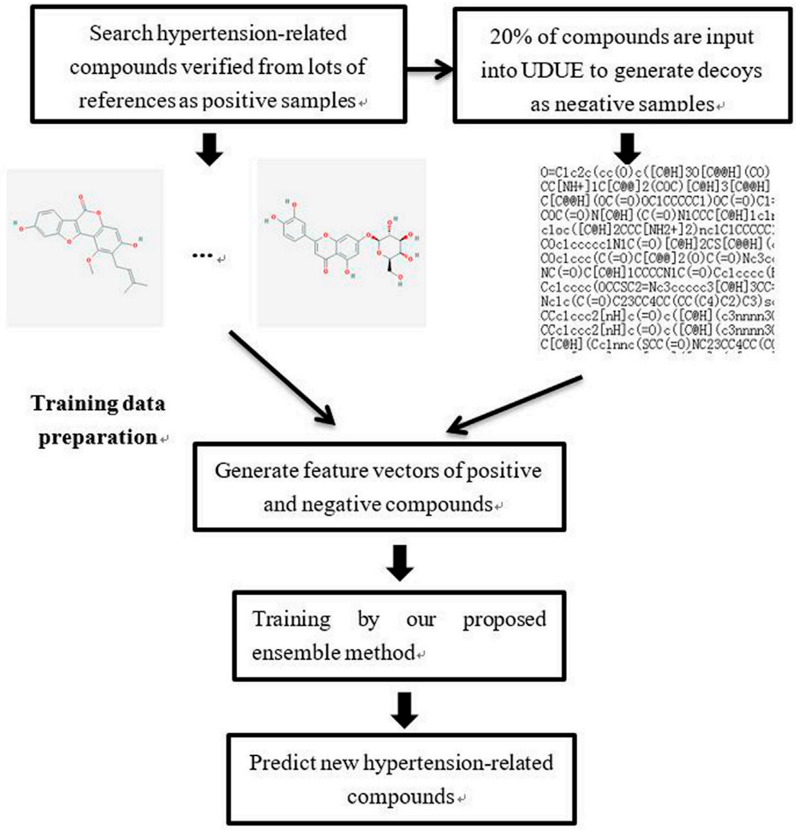
Search hypertension-related compounds verified from lots of references as positive samples
20% of compounds are input into UDUE to generate decoys as negative samples
…
O=C1c2c(cc(O)c([C@H]3O[C@@H](CO)
CC[NH+]1C[C@@]2(COC)[C@H]3[C@@H]
C[C@@H](OC(=O)OC1CCCCC1)OC(=O)C1
COC(=O)N[C@H](C(=O)N1CCC[C@H]1c1
c1oc([C@H]2CCC[NH2+]2)nc1C1CCCCC
COc1ccccc1N1C(=O)[C@H]2CS[C@@H](
COc1ccc(C(=O)C[C@@]2(O)C(=O)Nc3c
NC(=O)C[C@H]1CCCCN1C(=O)Cc1cccc(
Cc1cccc(OCCSC2=Nc3ccccc3[C@H]3CC
Nc1c(C(=O)C23CC4CC(CC(C4)C2)C3)s
CCc1ccc2[nH]c(=O)c([C@H](c3nnnn3(
CCc1ccc2[nH]c(=O)c([C@H](c3nnnn3(
C[C@H](Cc1nnc(SCC(=O)NC23CC4CC(C(
Training data preparation
Generate feature vectors of positive and negative compounds
Training by our proposed ensemble method
Predict new hypertension-related compounds

FIGURE 3 | The flowchart of hypertension-related active compound identification.

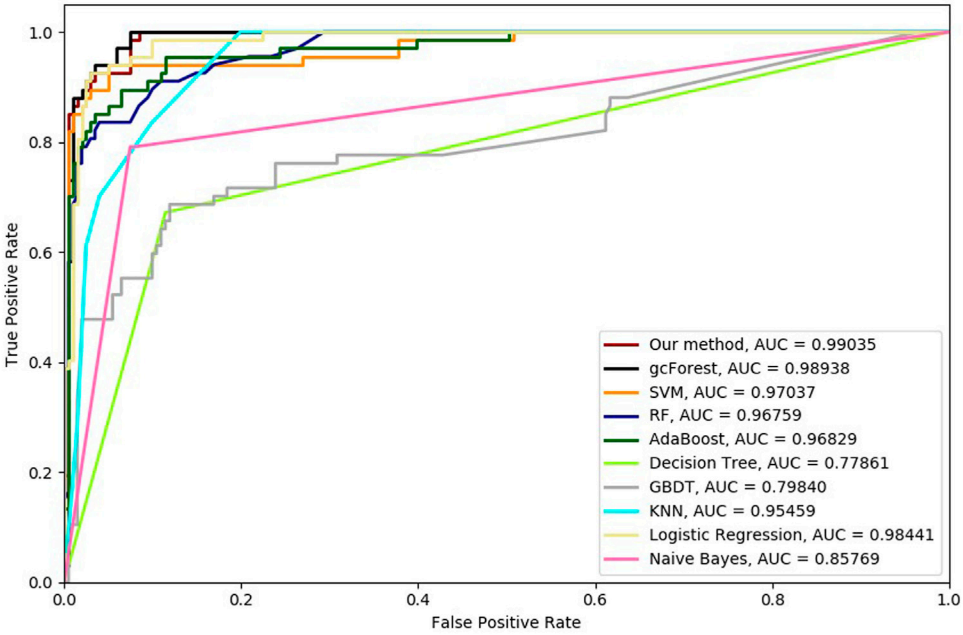

FIGURE 4 | Hypertension-related compound identification performances of ten methods with 2-cross validation methods.

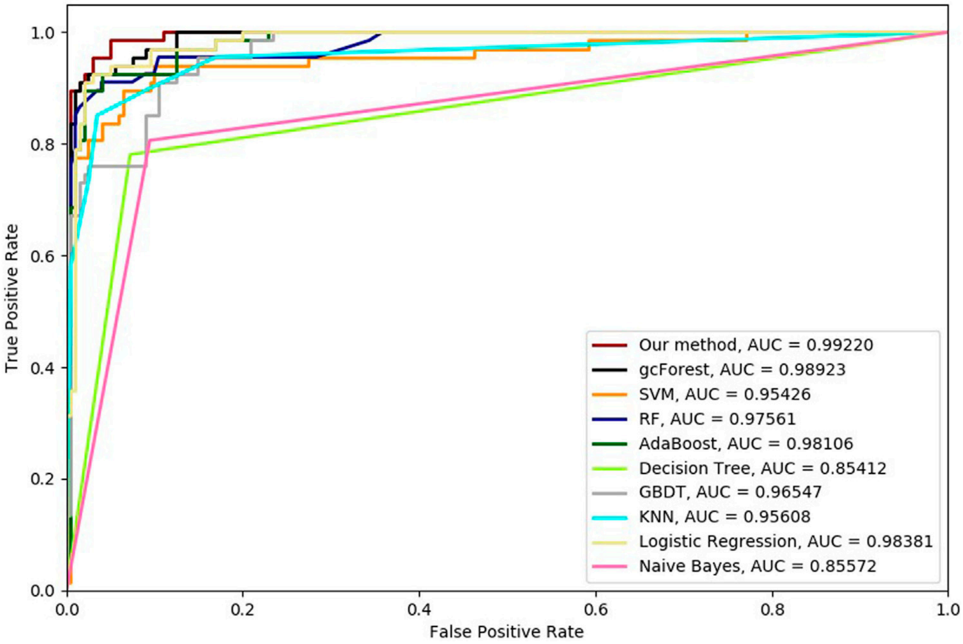

FIGURE 5 | Hypertension-related compound identification performances of ten methods with 4-cross validation methods.

The molecular descriptions of positive and negative compounds were extracted to constitute the hypertension-related dataset. With the collected dataset, our proposed ensemble method was fitted to predict other hypertension-related compounds.

EXPERIMENT RESULTS

In this part, the hypertension-related dataset collected is utilized, which contains 88 related compounds and 264 unrelated compounds. AUC, ROC curve, TPR, FRP, Precision,

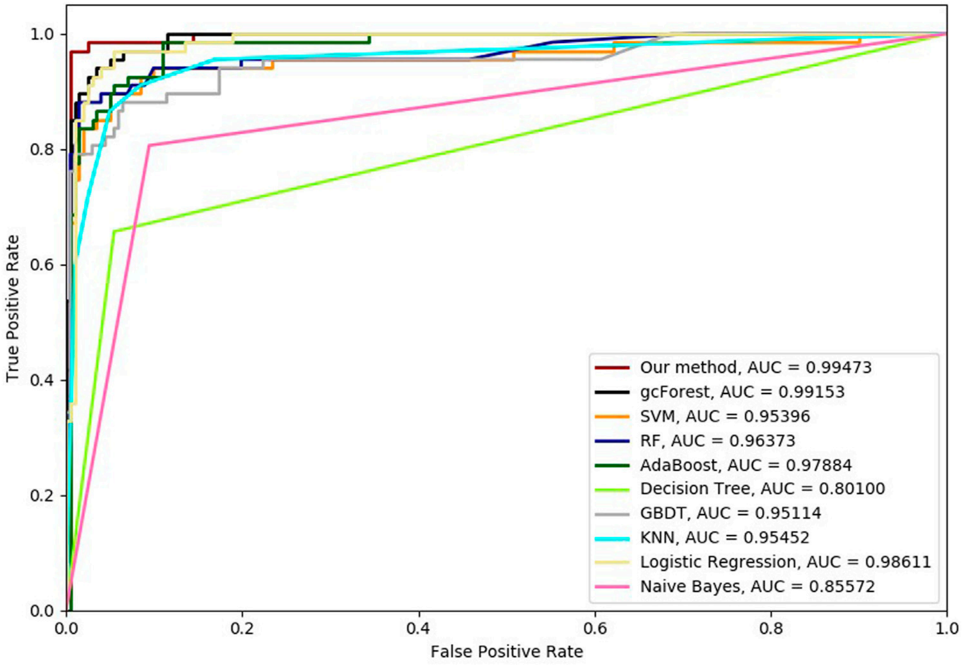

FIGURE 6 | Hypertension-related compound identification performances of ten methods with 6-cross validation methods.

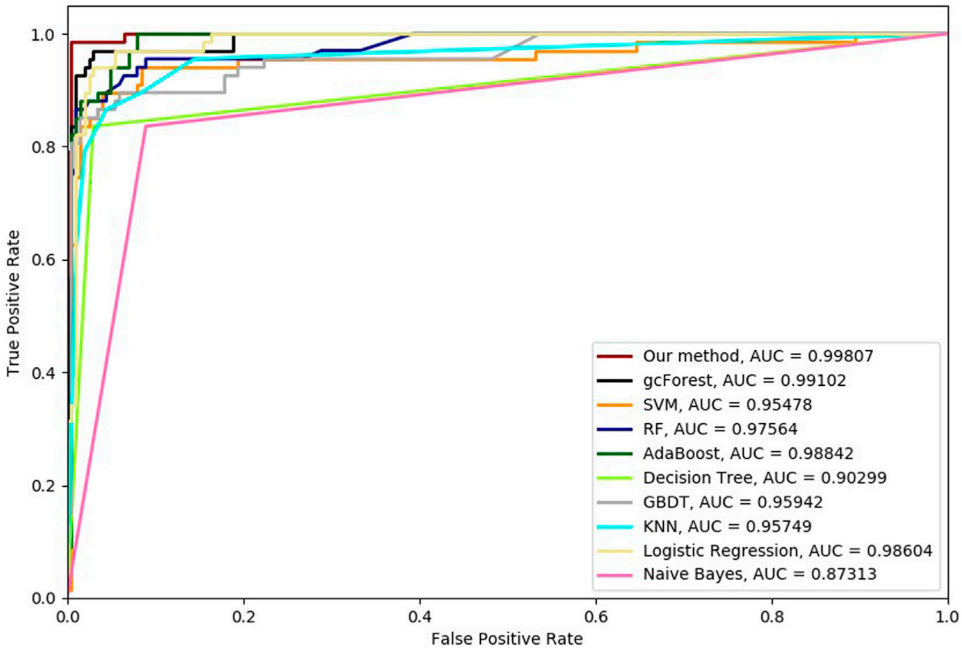

FIGURE 7 | Hypertension-related compound identification performances of ten methods with 8-cross validation methods.

Specificity, and F1 were used to test the performance of our proposed method. In our method, the parameters of nine classifiers were set by default. In FNT, the variable set is defined as $T = (c_1, c_2, \ldots, c_9)$ and the operator set is defined as $F = (+_2, +_3, +_4, +_5)$.

Six cross-validation methods were utilized to validate our proposed method. Nine classifiers were also utilized to identify hypertension-related compounds with the same dataset. The ROC curves and AUC performances with the different cross-validation methods are depicted in **Figures 4–9**, respectively.

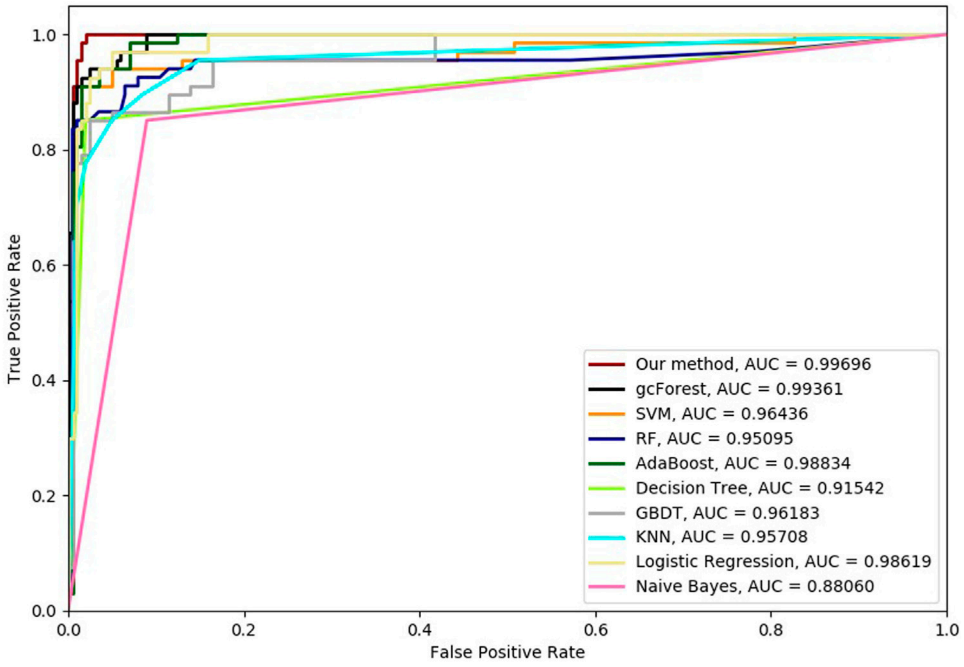

FIGURE 8 | Hypertension-related compound identification performances of ten methods with 10-cross validation methods.

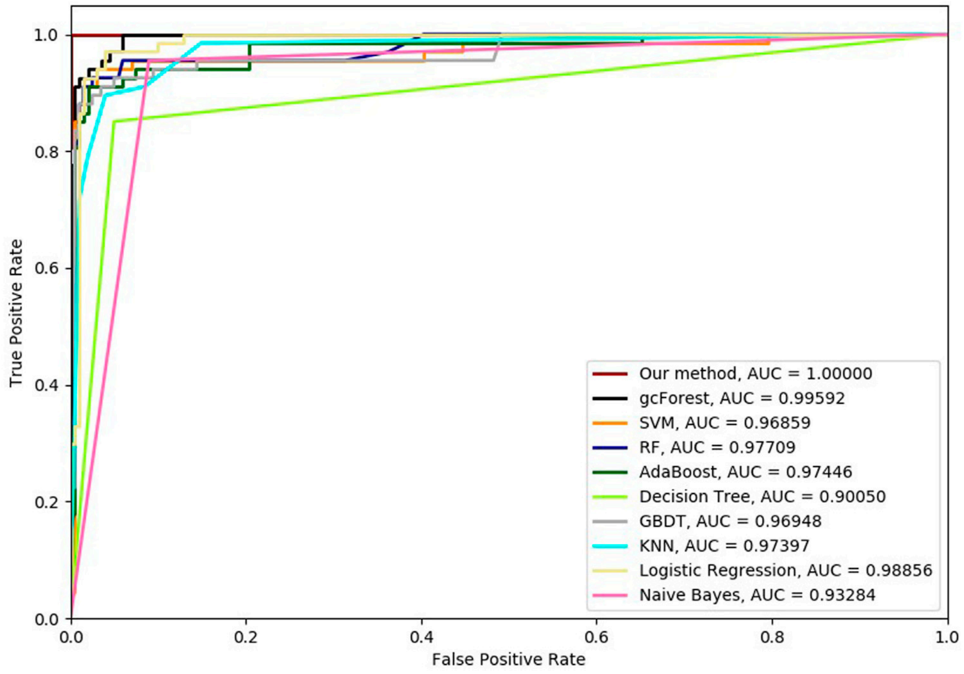

FIGURE 9 | Hypertension-related compound identification performances of ten methods with 15-cross validation methods.

From these results, it can be seen that gcForest has the best ROC curves and AUC values among the nine single classifiers. Our proposed ensemble method could perform better than gcForest in terms of ROC and AUC. With 2-cross, 4-cross, 6-cross, 8-cross, 10-cross, and 15-cross validation methods, in terms of AUC, our method is 0.1, 0.3, 0.3, 0.7, 0.3, and 0.4% higher than gcForest,

TABLE 1 | Classification performances of ten methods with 2-cross validation methods.

	TPR	FRP	Precision	Specificity	F1
Our method	0.880597	0.019900	0.936508	0.980100	0.907692
gcForest	0.940299	0.054726	0.851351	0.945274	0.893617
AdaBoost	0.791045	0.014925	0.946429	0.985075	0.861789
Decision Tree	0.671642	0.114428	0.661765	0.885572	0.666667
GBDT	0.61194	0.104478	0.66129	0.895522	0.635659
KNN	0.701493	0.039801	0.854545	0.960199	0.770492
LR	0.985075	0.199005	0.622642	0.800995	0.763006
Naive Bayes	0.791045	0.074627	0.779412	0.925373	0.785185
RF	0.671642	0.00995	0.957447	0.99005	0.789474
SVM	0.850746	0.00995	0.966102	0.99005	0.904762

TABLE 2 | Classification performances of ten methods with 4-cross validation methods.

	TPR	FRP	Precision	Specificity	F1
Our method	0.895522	0.014925	0.952381	0.985075	0.923077
gcForest	0.925373	0.039801	0.885714	0.960199	0.905109
AdaBoost	0.835821	0.0199	0.933333	0.9801	0.88189
Decision Tree	0.686567	0.039801	0.851852	0.960199	0.760331
GBDT	0.671642	0.00995	0.957447	0.99005	0.789474
KNN	0.850746	0.034826	0.890625	0.965174	0.870229
LR	0.940299	0.074627	0.807692	0.925373	0.868966
Naive Bayes	0.80597	0.094527	0.739726	0.905473	0.771429
RF	0.791045	0.00995	0.963636	0.99005	0.868852
SVM	0.776119	0.024876	0.912281	0.975124	0.83871

TABLE 3 | Classification performances of ten methods with 6-cross validation methods.

	TPR	FRP	Precision	Specificity	F1
Our method	0.955224	0.004975	0.984615	0.995025	0.969697
gcForest	0.925373	0.024876	0.925373	0.975124	0.925373
AdaBoost	0.835821	0.0199	0.933333	0.9801	0.88189
Decision Tree	0.656716	0.054726	0.8	0.945274	0.721311
GBDT	0.791045	0.00995	0.963636	0.99005	0.868852
KNN	0.865672	0.049751	0.852941	0.950249	0.859259
LR	0.940299	0.049751	0.863014	0.950249	0.9
Naive Bayes	0.80597	0.094527	0.739726	0.905473	0.771429
RF	0.820896	0.014925	0.948276	0.985075	0.88
SVM	0.791045	0.014925	0.946429	0.985075	0.861789

TABLE 4 | Classification performances of ten methods with 8-cross validation methods.

	TPR	FRP	Precision	Specificity	F1
Our method	0.970149	0.004975	0.984848	0.995025	0.977444
gcForest	0.940299	0.0199	0.940299	0.9801	0.940299
AdaBoost	0.850746	0.014925	0.95	0.985075	0.897638
Decision Tree	0.835821	0.029851	0.903226	0.970149	0.868217
GBDT	0.80597	0.004975	0.981818	0.995025	0.885246
KNN	0.865672	0.044776	0.865672	0.955224	0.865672
LR	0.940299	0.044776	0.875	0.955224	0.906475
Naive Bayes	0.835821	0.089552	0.756757	0.910448	0.794326
RF	0.835821	0.00995	0.965517	0.99005	0.896
SVM	0.791045	0.014925	0.946429	0.985075	0.861789

TABLE 5 | Classification performances of ten methods with 10-cross validation methods.

	TPR	FRP	Precision	Specificity	F1
Our method	0.955224	0.014925	0.955224	0.985075	0.955224
gcForest	0.925373	0.0199	0.939394	0.9801	0.932331
AdaBoost	0.850746	0.014925	0.95	0.985075	0.897638
Decision Tree	0.850746	0.0199	0.934426	0.9801	0.890625
GBDT	0.776119	0.014925	0.945455	0.985075	0.852459
KNN	0.850746	0.049751	0.850746	0.950249	0.850746
LR	0.940299	0.044776	0.875	0.955224	0.906475
Naive Bayes	0.850746	0.089552	0.76	0.910448	0.802817
RF	0.820896	0.004975	0.982143	0.995025	0.894309
SVM	0.880597	0.014925	0.951613	0.985075	0.914729

TABLE 6 | Classification performances of ten methods with 15-cross validation methods.

	TPR	FRP	Precision	Specificity	F1
Our method	0.955224	0	1	1	0.977099
gcForest	0.940299	0.0199	0.940299	0.9801	0.940299
AdaBoost	0.880597	0.0199	0.936508	0.9801	0.907692
Decision Tree	0.850746	0.049751	0.850746	0.950249	0.850746
GBDT	0.835821	0.00995	0.965517	0.99005	0.896
KNN	0.895522	0.039801	0.882353	0.960199	0.888889
LR	0.940299	0.034826	0.9	0.965174	0.919708
Naive Bayes	0.955224	0.089552	0.780488	0.910448	0.85906
RF	0.850746	0.00995	0.966102	0.99005	0.904762
SVM	0.880597	0.014925	0.951613	0.985075	0.914729

which reveals that our proposed method performs better than nine single classifiers for hypertension-related compound identification.

The TPR, FRP, Precision, Specificity, and F1 performances of the ten methods with the different cross-validation methods are listed in **Tables 1–6**, respectively. With 2-cross validation and 4-cross validation methods, LR could obtain the highest TPR performances, which shows that LR could identify more true hypertension-related compounds. For **Table 1**, RF and SVM have the best FPR performance, which shows that these two methods could identify less non-related compounds as related ones. SVM also has the highest Precision and Specificity performances among the ten methods. For **Table 2**, RF has the best FPR, Precision, and Specificity performances. Our method performed best in terms of F1, which reveals that it could identify hypertension-related compounds more accurately overall. With 6-cross validation, 8-cross validation, 10-cross validation, and 15-cross validation methods, our methods perform best among ten methods in terms of TPR, FRP, Precision, Specificity, and F1, except that RF has the lowest performance with 4-cross validation methods. The results show that our proposed ensemble method could identify more true hypertension-related and hypertension-unrelated compounds than the other nine single classifiers.

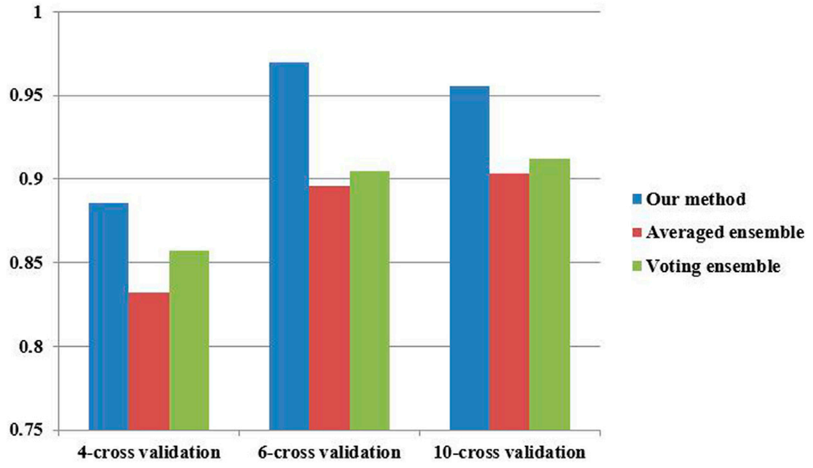

FIGURE 10 | F1 performances of hypertension-related compound by three ensemble methods.

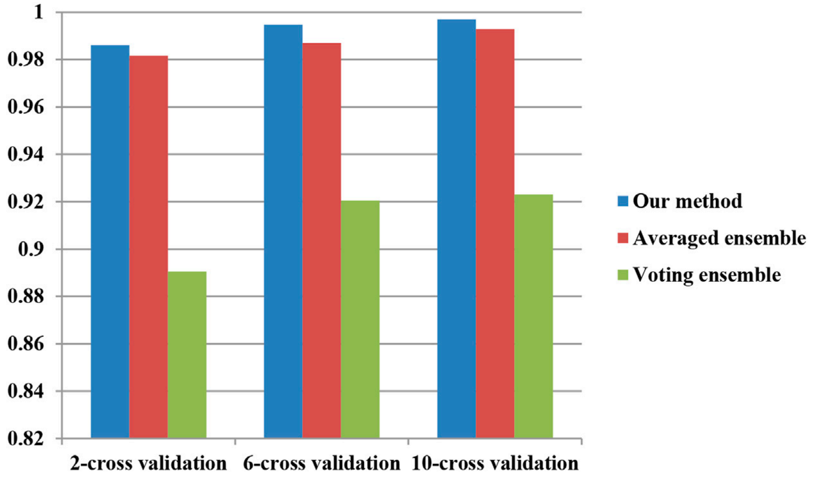

FIGURE 11 | AUC performances of hypertension-related compound by three ensemble methods.

DISCUSSION

To investigate the performance of our proposed ensemble further, two classical ensemble methods (averaged ensemble and voting ensemble) were also utilized to infer hypertension-related compounds. The F1 and AUC performances of the hypertension-related compounds by three ensemble methods are depicted in **Figure 10** and **Figure 11**, respectively. From **Figures 10**, **11**, it can be seen that our proposed ensemble method obtained better F1 and AUC performances than averaged and voting ensemble methods, which also shows that our method could identify hypertension-related compounds more accurately than the other two classical ensemble methods.

CONCLUSION

To identify hypertension-related closely active compounds, this paper proposed a novel ensemble method based on a flexible neural tree and nine classifiers. In our method, the classification results of nine single classifiers was utilized as the input vector of the flexible neural tree. An FNT model was utilized as a nonlinear ensemble method to identify hypertension-related drug activity. A hybrid evolutionary method based on genetic programming like structure optimization algorithm and simulated annealing is proposed to evolve the FNT model. In order to test the performance of our proposed ensemble method, data were extracted from hypertension-unrelated and hypertension-related compounds collected from up-to-date literature. By the different cross-validation methods, our proposed method obtained better ROC curves and AUC values than nine other single classifiers. Our proposed method also performs better than other single classifiers in terms of TPR, FRP, Precision, Specificity, and F1 in most cases. We also compare our proposed ensemble method with the averaged and voting ensemble methods. The results reveal that our method could identify hypertension-related compounds more accurately than the two classical ensemble methods.

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

WB collected and analyes the data of this work. BY and JW designed the model of this work.

FUNDING

This work was supported by the talent project of "Qingtan scholar" of Zaozhuang University, Jiangsu Provincial Natural Science Foundation, China (No. SBK2019040953), the Youth Innovation Team of Scientific Research Foundation of the Higher Education Institutions of Shandong Province, China (No. 2019KJM006), the Key Research Program of the Science Foundation of Shandong Province (ZR2020KE001), the Ph.D. research startup foundation of Zaozhuang University (No.2014BS13), and Zaozhuang University Foundation (No. 2015YY02).

REFERENCES

Aljumah, A., Ahamad, M., and Siddiqui, M. (2011). Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia. *Intell. Inf. Manag.* 3 (6), 252–261. doi:10.4236/iim.2011.36031

Amaratunga, D., Cabrera, J., Sargsyan, D., Kostis, J. B., Zinonos, S., and Kostis, W. J. (2020). Uses and Opportunities for Machine Learning in Hypertension Research. *Int. J. Cardiol. Hypertens.* 5, 100027. doi:10.1016/j.ijchy.2020.100027

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees (CART). *Biometrics*. Monterey, CA: Wadsworth.

Breiman, L. (2001). Random forest. *Machine Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Brinks, H. L., and Eckhart, A. D. (2010). Regulation of GPCR Signaling in Hypertension. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1802 (12), 1268–1275. doi:10.1016/j.bbadis.2010.01.005

Cao, Y., Miao, Q.-G., Liu, J.-C., and Gao, L. (2013). Advance and Prospects of AdaBoost Algorithm. *Acta Automatica Sinica* 39 (6), 745–758. doi:10.1016/s1874-1029(13)60052-x

Chen, J., Zhang, Y., Wang, Y., Jiang, P., Zhou, G., Li, Z., et al. (2021b). Potential Mechanisms of Guizhi Decoction against Hypertension Based on Network Pharmacology and Dahl Salt-Sensitive Rat Model. *Chin. Med.* 16 (1), 34. doi:10.1186/s13020-021-00446-x

Chen, L., Zhu, T., Qi, J., Zhang, Y., Zhang, Z., and Liu, H. (2021a). Pharmacological Mechanism of JiaWeiSiWu Granule in the Treatment of Hypertension Based on Network Pharmacology. *Ann. Palliat. Med.* 10 (7), 7486–7513. doi:10.21037/apm-21-1140

Chen, Y., Abraham, A., and Bo, Y. (2006). Feature Selection and Classification Using Flexible Neural Tree. *Neurocomputing* 70 (1/3), 305–313. doi:10.1016/j.neucom.2006.01.022

Chen, Y., Chen, D., Liu, S., Yuan, F., Guo, F., Fang, F., et al. (2019). Systematic Elucidation of the Mechanism of Genistein against Pulmonary Hypertension via Network Pharmacology Approach. *Int. J. Mol. Sci.* 20 (22), 5569. doi:10.3390/ijms20225569

Chen, Y., Yuan, T., Chen, D., Liu, S., Guo, J., Fang, L., et al. (2020). Systematic Analysis of Molecular Mechanism of Resveratrol for Treating Pulmonary Hypertension Based on Network Pharmacology Technology. *Eur. J. Pharmacol.* 888, 173466. doi:10.1016/j.ejphar.2020.173466

Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006). Gene Selection and Classification of Microarray Data Using Random forest. *BMC Bioinformatics* 7, 3. doi:10.1186/1471-2105-7-3

Essiarab, F., Taki, H., Malki, E. A., Hassar, M., Ghalim, N., Saile, R., et al. (2011). Cardiovascular Risk Factors Prevalence in a Moroccan Population. *Eur. J. Scientific Res.* 49 (4), 581–589.

Fang, M., Chen, Y., Xue, R., Wang, H., Chakraborty, N., Su, T., et al. (2021). A Hybrid Machine Learning Approach for Hypertension Risk Prediction. *Neural Comput. Applic.* doi:10.1007/s00521-021-06060-0

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. *Bioinformatics* 16 (10), 906–914. doi:10.1093/bioinformatics/16.10.906

Hu, J., and Min, J. (2018). Automated Detection of Driver Fatigue Based on EEG Signals Using Gradient Boosting Decision Tree Model. *Cogn. Neurodyn* 12 (4), 431–440. doi:10.1007/s11571-018-9485-1

Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., and Sun, Q. (2018). Deep Learning for Image-Based Cancer Detection and Diagnosis – A Survey. *Pattern Recognition* 83, 134–149. doi:10.1016/j.patcog.2018.05.014

Hwang, K.-Y., Lee, E.-S., Kim, G.-W., Hong, S.-O., Park, J.-S., Kwak, M.-S., et al. (2016). Developing Data Quality Management Algorithm for Hypertension Patients Accompanied with Diabetes Mellitus by Data Mining. *J. Digital Convergence* 14 (7), 309–319. doi:10.14400/jdc.2016.14.7.309

Ji, Z., and Wang, B. (2014). Identifying Potential Clinical Syndromes of Hepatocellular Carcinoma Using PSO-Based Hierarchical Feature Selection Algorithm. *Biomed. Res. Int.* 2014, 127572. doi:10.1155/2014/127572

Ji, Z., Wu, D., Zhao, W., Peng, H., Zhao, S., Huang, D., et al. (2015). Systemic Modeling Myeloma-Osteoclast Interactions under Normoxic/hypoxic Condition Using a Novel Computational Approach. *Sci. Rep.* 5, 13291. doi:10.1038/srep13291

Li, X., and Guo, Y. (2005). Naive Bayesian Classifier Based on Multiple Discriminant Analysis. *Inf. Control* 34 (5), 580–584.

Liang, Y., Chen, Z., Ward, R., and Elgendi, M. (2018). Photoplethysmography and Deep Learning: Enhancing Hypertension Risk Stratification. *Biosensors* 8 (4), 101. doi:10.3390/bios8040101

Liao, Y., and Vemuri, V. R. (2002). Use of K-Nearest Neighbor Classifier for Intrusion Detection. *Comput. Security* 21 (5), 439–448. doi:10.1016/s0167-4048(02)00514-x

Liu, T. H., Chen, W. H., Chen, X. D., Liang, Q. E., Tao, W. C., Jin, Z., et al. (2020). Network Pharmacology Identifies the Mechanisms of Action of TaohongSiwu Decoction Against Essential Hypertension. *Med. Sci. Monit.* 26, e920682. doi:10.12659/MSM.920682

Liu, X., Zhang, Y., Fu, C., Zhang, R., and Zhou, F. (2021). EnRank: An Ensemble Method to Detect Pulmonary Hypertension Biomarkers Based on Feature Selection and Machine Learning Models. *Front. Genet.* 12, 636429. doi:10.3389/fgene.2021.636429

Ma, X. L., Zhai, X., Liu, J. W., Xue, X. X., Guo, S. Z., Xie, H., et al. (2018). Study on the Biological Basis of Hypertension and Syndrome with Liver-Fire Hyperactivity Based on Data Mining Technology. *World J. Traditional Chin. Med.* 4 (4), 176–180. doi:10.4103/wjtcm.wjtcm_23_18

Maalouf, M. (2011). Logistic Regression in Data Analysis: an Overview. *Ijdats* 3 (3), 281–299. doi:10.1504/ijdats.2011.041335

Morra, J. H., Tu, Z., Apostolova, L. G., Green, A. E., Toga, A. W., and Thompson, P. M. (2009). Comparison of AdaBoost and Support Vector Machines for Detecting Alzheimer's Disease Through Automated Hippocampal Segmentation. *IEEE Trans. Med. Imaging* 29 (1), 30–43. doi:10.1109/TMI.2009.2021941

Munshi, T., Zuidgeet, M., Brussel, M., and Maarseveen, M. V. (2014). Logistic Regression and Cellular Automata-Based Modeling of Retail, Commercial and Residential Development in the City of Ahmedabad India. *Cities* 38 (2), 88–101. doi:10.1016/j.cities.2014.02.007

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* 55 (14), 6582–6594. doi:10.1021/jm300687e

Owlia, M., and Bangalore, S. (2016). In Hypertensive Patients with Elevated Risk of Cardiovascular Disease, Targeting Systolic Blood Pressure to Less Than 120 Mm Hg Significantly Reduces the Rate of Fatal and Non-fatal Cardiovascular Events as Well as Death from Any Cause. *Evid. Based Med.* 21 (3), 101. doi:10.1136/ebmed-2016-110397

Ramezankhani, A., Kabir, A., Pournik, O., Azizi, F., and Hadaegh, F. (2016). Classification-based Data Mining for Identification of Risk Patterns Associated with Hypertension in Middle Eastern Population. *Medicine* 95 (35), e4143. doi:10.1097/md.0000000000004143

Rish, I. (2001). "An Empirical Study of the Naive Bayes Classifier", in IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, August 4, 2001 3, 41–46.

Sakai, K., and Sigmund, C. D. (2005). Molecular Evidence of Tissue Renin-Angiotensin Systems: A Focus on the Brain. *Curr. Sci. Inc* 7 (2), 135–140. doi:10.1007/s11906-005-0088-y

Suykens, J. A. K., and Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* 9 (3), 293–300. doi:10.1023/a:1018628609742

Temkin, N. R., Holubkov, R., Machamer, J. E., Winn, H. R., and Dikmen, S. S. (1995). Classification and Regression Trees (CART) for Prediction of Function at 1 Year Following Head Trauma. *J. Neurosurg.* 82 (5), 764–771. doi:10.3171/jns.1995.82.5.0764

Wang, T., He, M., Du, Y., Chen, S., and Lv, G. (2021). Network Pharmacology Prediction and Pharmacological Verification Mechanism of Yeju Jiangya Decoction on Hypertension. *Evid. Based Complement. Alternat Med.* 2021, 5579129. doi:10.1155/2021/5579129

Wang, T., Zhang, Y., Lu, J., Chai, R., Chen, X., et al. (2018). Research on the Functional Mechanism of Shengmai Injection Based on Network Pharmacology. *J. Pharm. Res.* 37 (11), 621–624. doi:10.13506/j.cnki.jpr.2018.11.001

Xu, S. N., Li, Z., Zhai, Y. Y., Yao, W. F., Xu, J., Liu, Q., et al. (2018). Material Basis and Mechanism of Erzhi Pill for Preventing Osteoporosis Based on Network Pharmacology. *Chin. Pharm. J.* 53 (22), 1913–1920. doi:10.11669/cpj.2018.22.007

Yang, B., Chen, Y., and Jiang, M. (2013). Reverse Engineering of Gene Regulatory Networks Using Flexible Neural Tree Models. *Neurocomputing* 99, 458–466. doi:10.1016/j.neucom.2012.07.015

Yuan, F., and Chen, S. (2011). Model Construction on Efficient Mining Association Rules in Clinical Data of Hypertension. *Comput. Eng. Appl.* 47 (36), 226–229+233. doi:10.3778/j.issn.1002-8331.2011.36.062

Zhai, Z., Tao, X., Alami, M. M., Shu, S., and Wang, X. (2021). Network Pharmacology and Molecular Docking Combined to Analyze the Molecular and Pharmacological Mechanism of Pinellia Ternata in the Treatment of Hypertension. *Cimb* 43 (1), 65–78. doi:10.3390/cimb43010006

Zhang, B., Ren, J., Cheng, Y., Wang, B., and Wei, Z. (2019b). Health Data Driven on Continuous Blood Pressure Prediction Based on Gradient Boosting Decision Tree Algorithm. *IEEE Access* 7, 32423–32433. doi:10.1109/access.2019.2902217

Zhang, Y., Lei, L., and He, J. (2019a). Study on Distribution Rules of TCM Signs and Symptoms and Syndrome Elements in Essential Hypertension Based on Data Mining. *Chin. J. Inf. Traditional Chin. Med.* 26 (1), 99–104. doi:10.3969/j.issn.1005-5304.2019.01.023

Zhao, L., Wu, Y. J., and Zhang, M. Q. (2021). Research Progress of Data Mining in the Treatment of Hypertension by Traditional Chinese Medicine. *Food Ther. Health Care* 3 (2), 36–46. doi:10.12032/FTHC20210503

Zhou, Z. H., and Feng, J. (2017). Deep Forest: Towards An Alternative to Deep Neural Networks. Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, August 19-25, 2017, 3553–3559.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 22 October 2021
doi: 10.3389/fgene.2021.766496

An Information-Entropy Position-Weighted *K*-Mer Relative Measure for Whole Genome Phylogeny Reconstruction

Yao-Qun Wu[1,2], Zu-Guo Yu[1]*, Run-Bin Tang[1], Guo-Sheng Han[1] and Vo V. Anh[3]

[1]Hunan Key Laboratory for Computation and Simulation in Science and Engineering and Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Hunan, China, [2]Provincial Key Laboratory of Informational Service for Rural Area of Southwestern Hunan, Shaoyang University, Shaoyang, China, [3]Faculty of Science, Engineering and Technology, Swinburne University of Technology, Hawthorn, VIC, Australia

Alignment methods have faced disadvantages in sequence comparison and phylogeny reconstruction due to their high computational costs in handling time and space complexity. On the other hand, alignment-free methods incur low computational costs and have recently gained popularity in the field of bioinformatics. Here we propose a new alignment-free method for phylogenetic tree reconstruction based on whole genome sequences. A key component is a measure called *information-entropy position-weighted k-mer relative measure* (IEPWRMkmer), which combines the position-weighted measure of *k*-mers proposed by our group and the information entropy of frequency of *k*-mers. The Manhattan distance is used to calculate the pairwise distance between species. Finally, we use the Neighbor-Joining method to construct the phylogenetic tree. To evaluate the performance of this method, we perform phylogenetic analysis on two datasets used by other researchers. The results demonstrate that the *IEPWRMkmer* method is efficient and reliable. The source codes of our method are provided at https://github.com/wuyaoqun37/IEPWRMkmer.

Keywords: alignment-free method, k-mer relative distance, information entropy, phylogenetic analysis, genome

OPEN ACCESS

Edited by:
Juan Wang,
Inner Mongolia University, China

Reviewed by:
Liang Cheng,
Harbin Medical University, China
Yanjuan Li,
Quzhou University, China

****Correspondence:***
Zu-Guo Yu
yuzuguo@aliyun.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *29 August 2021*
Accepted: *29 September 2021*
Published: *22 October 2021*

Citation:
Wu Y-Q, Yu Z-G, Tang R-B, Han G-S and Anh VV (2021) An Information-Entropy Position-Weighted K-Mer Relative Measure for Whole Genome Phylogeny Reconstruction. Front. Genet. 12:766496. doi: 10.3389/fgene.2021.766496

INTRODUCTION

The reconstruction of a phylogenetic tree is a primary problem in evolutionary biology. Sequence alignment is a key step in the reconstruction, aiming to identify the homology of sequences and uncover phylogenetic relationships in sequences. Traditional sequence comparison is based on pairwise or multiple sequence alignment (Felsenstein and Felenstein, 2004; Morrison, 2006) and was implemented by software packages such as BLAST (Altschul et al., 1990), ClustalW (Thompson et al., 1994), and MrBayes (Ronquist et al., 2012). However, the methods based on sequence alignment have some disadvantages, including high computational cost in handling the time and space complexity of the algorithm. Therefore, alignment-free methods have been proposed to overcome these problems (Zielezinski et al., 2017). The computational cost of alignment-free methods is low because they are generally of linear complexity (Fox et al., 1977).

Several alignment-free methods for sequence comparison are based on word counts (Blaisdell, 1986; Höhl et al., 2006; Wang et al., 2016). A key idea is to use the close

TABLE 1 | Names, species, and accession numbers for mitochondrial genomes of 30 mammalian species.

No	Accession no	Species	Sequence name
1	AJ002189	*Sus scrofa*	Pig
2	AJ010957	*Homo sapiens*	*Hippopotamus*
3	AJ001588	*Pan troglodytes*	Rabbit
4	U96639	*Canis familiaris*	Dog
5	AF010406	*Ovis aries*	Sheep
6	V00662	*Homo sapiens*	Human
7	U20753	*Felis catus*	Cat
8	X72004	*Halichoerus grypus*	Gray seal
9	D38115	*Pongo pygmaeus*	Orangutan
10	V00654	*Bos taurus*	Cow
11	X97337	*Equus asinus*	Donkey
12	D38116	*Pan troglodytes*	Common chimpanzee
13	D38113	*Pan paniscus*	Pigmy chimpanzee
14	Z29573	*Didelphis virginiana*	Opossum
15	Y10524	*Macropus robustus*	Wallaroo
16	X99256	*Hylobates lar*	Gibbon
17	Y18001	*Papio hamadryas*	Baboon
18	X97336	*Rhinoceros unicornis*	Indian rhinoceros
19	Y07726	*Ceratotherium simum*	White rhinoceros
20	X63726	*Phoca vitulina*	Harbor seal
21	AJ238588	*Sciurus vulgaris*	Squirrel
22	AJ001562	*Glis glis*	Fat dormouse
23	AJ222767	*Cavia porcellus*	Guinea pig
24	X79547	*Equus caballus*	Horse
25	X14848	*Rattus norvegicus*	Rat
26	V00711	*Mus musculus*	Mouse
27	D38114	*Gorilla gorilla*	*Gorilla*
28	X61145	*Balenoptera physalus*	Fin whale
29	X72204	*Balenoptera musculus*	Blue whale
30	X83427	*Ornithorhyncus anatinus*	Platypus

TABLE 2 | Accession numbers, subtype, and area for 44 HIV-1.

No	Area	Accession no	Subtype
1	Belgium (DRC)	AF084936	G
2	Finland (Kenya)	AF061641	G
3	Sweden (DRC)	AF061642	G
4	Belgium	AF190128	H
5	Belgium	AF190127	H
6	Cent. Afr. Rep	AF005496	H
7	Tanzania	AF447763	CPZ
8	Cameroon	L20571	O
9	Senegal	AJ302647	O
10	Cameroon	L20587	O
11	Cameroon	AY169812	O
12	India	AF067155	C
13	South Africa	AY772699	C
14	Ethiopia	U46016	C
15	Brazil	U52953	C
16	Cameroon	AY371157	D
17	DRC	K03454	D
18	Uganda	U88824	D
19	Somalia	AF069670	A1
20	Uganda	AF484509	A1
21	Uganda	U51190	A1
22	Kenya	AF004885	A1
23	DRC	AF286238	A2
24	Cyprus	AF286237	A2
25	Sweden	AF082395	J
26	Sweden	AF082394	J
27	Cameroon	AJ249239	K
28	DRC	AJ249235	K
29	Cameroon	AJ249237	F2
30	Cameroon	AY371158	F2
31	Cameroon	AJ249236	F2
32	Cameroon	AF377956	F2
33	Finland	AF075703	F1
34	France	AJ249238	F1
35	Brazil	AF005494	F1
36	Belgium (DRC)	AF077336	F1
37	Cameroon	AJ271370	N
38	Cameroon	AY532635	N
39	Cameroon	AJ006022	N
40	Netherlands	AY423387	B
41	Thailand	AY173951	B
42	Australia	Gray seal	B
43	France	K03455	B
44	U.S.	AY331295	B

distribution of k-mers to imply the high correlation degree, hence the similarity of the sequences. The methods have been implemented in software tools, such as FFP (Sims et al., 2009), kWIP (Murray et al., 2017), CVtree (Qi et al., 2004), and DLtree (Wu et al., 2017). Many k-mer methods transform the input sequence into a frequency vector of k-mers, then define the distance of the sequences by that of the frequency vector of k-mers (Qi et al., 2004; Wu et al., 2017). To reduce the statistical dependence between adjacent word matches, Spaced-Words (Leimeister and Boden, 2014) proposed to use spaced words, which are defined by patterns of matches without reference to positions. Some alignment-free methods are based on match length, which defines the distance between sequences based on the length of substring matches between two sequences. These include the shortest unique substring method (Haubold et al., 2005), ACS (Ulitsky et al. 2006), UA (Comin and Verzotto, 2012), and ALFRED (Thankachan et al. 2016). In addition, graphical representation was used to construct the probability distribution of a DNA sequence (Yu et al., 2011). The chaos game representation transforms the distribution of characters in a DNA sequence into the distribution of nodes in a graph (Hoang et al. 2016; Yin, 2017; Mendizabal-Ruiz et al., 2018). Many researchers considered extracting the position information of a k-mer (Huang and Wang, 2011; Ding et al., 2013; Tang et al., 2014). Ding et al. (2013) used the average interval distance of normalized k-mers to capture evolutionary information for sequence comparison. Tang et al. (2014) presented the average relative distance of normalized k-mers to improve the method of Ding et al. (2013). Ma et al. (2020) proposed the *PWKmer* method, which combines the k-mer counts and k-mer position distributions for phylogenetic analysis.

In this work, we propose a new alignment-free method which combines the position-weighted measure of k-mers proposed by Ma et al. (2020) and the information entropy of frequency of k-mers to obtain phylogenetic information for sequence comparison. It is named *information-entropy position-weighted k-mer relative measure* (IEPWRMkmer). To evaluate the performance of this method, we carry out phylogenetic analysis on two data sets used by other researchers.

MATERIALS AND METHODS

Genomic Datasets

Dataset 1

The first dataset for analysis consists of the same whole genome DNA sequences of 30 mammalian species studied in Li et al. (2001), Otu and Sayood (2003), and Tang et al. (2014). The accession numbers, species, and species name are listed in **Table 1**. All sequences were downloaded from NCBI GenBank.

Dataset 2

The second dataset for analysis is the HIV-1 dataset studied in Ma et al. (2020). This dataset contains 43 HIV genome sequences used in Wu et al. (2007) and a controversial taxonomic sequence used in Chang et al. (2014). The dataset includes subtypes A, B, C, D, F, G, J, K, and H of the HIV-1 M, O, N groups and the CPZ sequence. The area, accession numbers, and subtypes are listed in **Table 2**. All these sequences were downloaded from NCBI GenBank.

We use two approaches to validate the method. First, we use the Robinson-Foulds (RF) distance to compare our method with other alignment-free methods. Second, we use the bootstrap method to construct consensus trees and show the stability of the trees obtained by our method.

METHODS

Let $S = s_1 s_2 \cdots s_L$ be a DNA sequence with length L, $a_1 a_2 \cdots a_k$is a k-mer, where $a_i \in$(A,T,C,G). If the k-mer $a_1 a_2 \cdots a_k$ occurs in S, we denote by $p_{a_1 a_2 \cdots a_k}$ the vector composed of the positions of $a_1 a_2 \cdots a_k$ in this given sequence and by $p_{a_1 a_2 \cdots a_k}(i)$ its ith element. If the k-mer $a_1 a_2 \cdots a_k$ does not occur in S, we set $p_{a_1 a_2 \cdots a_k}$=(0). For example, for the DNA sequence GTAACCTGAACGTACTTGGA with length 20, we list all 2-mer position vectors:

P_{AA}=(3,9); P_{AC}=(4,10,14); P_{AG}= (0); P_{AT}= (0); P_{CA}=(0); P_{CC}=(5); P_{CG}=(11); P_{CT}=(6,15); P_{GA}=(8,19); P_{GC}=(0); P_{GG}=(18); P_{GT}=(1,12); P_{TA}=(2,13); P_{TC} = 0; P_{TG}=(7,17); P_{TT}=(16).

In this example, the 2-mers AG, AT, CA, GC, and TC do not appear. For each k-mer, its position vector provides its position distribution information in the sequence. One can use the k-mer position vectors to reconstruct the DNA sequence (Ma et al., 2020).

Ma et al. (2020) defined the position-weighted measure $D(a_1 a_2 \cdots a_k)$ of $a_1 a_2 \cdots a_k$ based on its position in the sequence as

$$D(a_1 a_2 \cdots a_k) = \begin{cases} \dfrac{\sum_{i=1}^{n} p_{a_1 a_2 \cdots a_k}(i)}{L(L-k+1)}, & n \neq 0, \\ 0, & n = 0, \end{cases} \qquad (1)$$

where n is the length of the vector $p_{a_1 a_2 \cdots a_k}$. Actually $p_{a_1 a_2 \cdots a_k}(i)/L$ means the position weight of $a_1 a_2 \cdots a_k$ in the given sequence with length L.

We denote by N the number of sequences in a dataset. In order to characterize the importance of k-mers in the whole dataset, we count the number m of the sequences that contain a k-mer $a_1 a_2 \cdots a_k$. Then the occurrence frequency $F(a_1 a_2 \cdots a_k)$ of this k-mer in the whole dataset is defined as m/N. We introduce the Shannon entropy $H(a_1 a_2 \cdots a_k)$ of frequency $F(a_1 a_2 \cdots a_k)$ defined by Murray et al. (2017) as

$$H(a_1 a_2 \cdots a_k) = -\left(F \log_2(F) + (1-F)\log_2(1-F)\right), \qquad (2)$$

where F stands for $F\,(a_1 a_2 \cdots a_k)$.

In this study, we aim to get more DNA phylogenetic information by combining the above two methods and defining

$$E\,(a_1 a_2 \cdots a_k) = D\,(a_1 a_2 \cdots a_k) \times H\,(a_1 a_2 \cdots a_k) \qquad (3)$$

Here, we regard Shannon entropy $H\,(a_1 a_2 \cdots a_k)$ as another weight.

For a fixed K, there are 4^K k-mers. For each k-mer $a_1 a_2 \cdots a_k$, we can calculate the corresponding $E\,(a_1 a_2 \cdots a_k)$, then arrange 4^K of these $E\,(a_1 a_2 \cdots a_k)$ to get a feature representation vector $(E_1, E_2, \cdots, E_{4^K})$ according to the alphabet order of the 4^K k-mers for each genome.

For two given genome sequences A and B, we can obtain $E_\mathrm{A} = (E_1^A, E_2^A, \cdots, E_{4^K}^A)$ and $E_\mathrm{B} = (E_1^B, E_2^B, \cdots, E_{4^K}^B)$ by the method. We use the Manhattan distance to calculate the pairwise distance between these two genome sequences:

.

$$D\,(A,B) = \sum\nolimits_{i}^{4^K} \left|\left(E_i^A - E_i^B\right)\right| \qquad (4)$$

For a given dataset, we can derive a distance matrix by **Eq. 4**. This distance matrix contains the sequence similarity information. After obtaining the distance matrix, we insert it into the mega 7.0 software (Sudhir et al., 2016) and use Neighbor-Joining (NJ) program (Saitou et al. 1987) to construct the phylogenetic tree.

Robinson-Foulds Distance and the Bootstrap Method

We use the Robinson-Foulds (RF) distance (Robinson and Foulds 1981) to judge the quality of the method. A smaller RF value means a closer distance between the phylogenetic tree and the reference tree.

(Yu et al., 2010) proposed a modified version of the bootstrap method to evaluate the reliability of the constructed phylogenetic tree. We also use this method in the present work. Its workflow is as follows: Each row is the feature vector $(E_1, E_2, \cdots, E_{4^K})$ of a species, and each column is the feature value of all genome sequences based on the same k-mer. Through random sampling of all columns, in which some columns may be selected many times, while some columns may not be selected at all, we randomly select one column. After 4^K times of selection, a new N×4^K feature matrix is constructed. Using the new feature matrix, the Manhattan distance of any two rows is calculated to get a new distance matrix. Then we use the NJ method to construct a phylogenetic tree and repeat the above steps 100 times. Finally, a consensus tree is drawn by using consense. exe in

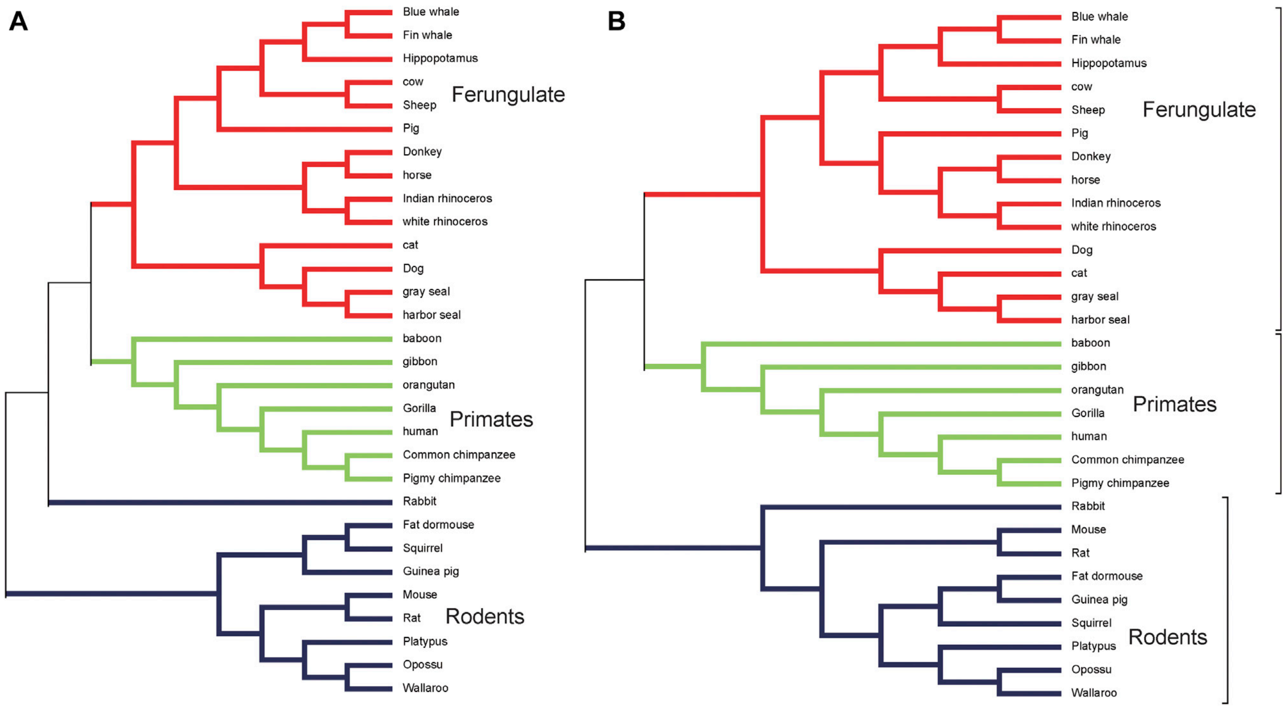

FIGURE 1 | (A) The phylogenetic tree of 30 mammalian species reconstructed by ClustalX. **(B)** The phylogenetic tree of 30 mammalian species at *K* = 8 based on our method.

TABLE 3 | The RF distance between the phylogenetic tree conducted by our method at K = 5,6,7,8,9 and the reference tree conducted by ClustalX.

K	5	6	7	8	9
RF distance	38	28	22	8	10

the Phylip package. The frequency of a particular branch of a phylogenetic tree can be used as a measure of the stability of this branch.

RESULTS

Experiment 1

We use the genomes of 30 mammalian species in dataset 1 to construct a phylogenetic tree using ClustalX (Larkin et al. 2007) as the reference tree. ClustalX is one of the widely used multiple alignment programs. The result is shown in **Figure 1A**. It is seen that rabbit, fat dormouse, squirrel, guinea pig, mouse, rat, platypus, opossum, and wallaroo belong to the rodents group; human, baboon, orangutan, gibbon, gorilla, pigmy chimpanzee, and common chimpanzee belong to the primates group; blue whale, fin whale, hippopotamus, cow, sheep, pig, donkey, horse, Indian-rhinoceros, white rhinoceros, cat, dog, gray seal, and harbor seal belong to the ferungulates group. When *K* < 5, it is not feasible to construct a phylogenetic tree using our method. When *K* = 5, 6, the 30 mammals cannot be divided into three groups in our tree. When *K* = 7, it can be divided into three groups, but the relationship between guinea pig and fat dormouse is not correct. When *K* = 8, 9, the branches of the tree become correct. We list the RF distances between the phylogenetic tree constructed by our method at *K* = 5, 6, 7, 8, 9 and the reference tree constructed by ClustalX in **Table 3**. From **Table 3**, we can see that the RF distance reaches the minimum when *K* = 8. We show the phylogenetic tree of *K* = 8 constructed by our method in **Figure 1B**. From **Figure 1B**, we can see that the species in the three main categories are grouped correctly. Primates and ferungulates are closer, and this relationship is consistent with that in **Figure 1A**. In terms of branches, monotremes (platypus), marsupials (wallaroo, opossum), murid rodents (mouse, rat), non-murid rodents (guinea pig, squirrel, fat dormouse, rabbit), perissodactyls (white rhinoceros, horse, Indian rhinoceros, donkey), carnivores (harbor seal, dog, gray seal, cat), artiodactyls (sheep, cow, hippopotamus, pig), primates (human, pigmy chimpanzee, common chimpanzee, gorilla, baboon, gibbon, orangutan), and cetaceans (blue whale, fin whale) are grouped into respective taxonomic classes accurately.

Figure 2 shows the RF distance between the reference tree constructed by ClustalX and the phylogenetic tree constructed by our method, Tang's method, PWKmer, DLtree, and CVtree on dataset 1. Using our method, when *K* = 8, the RF distance is 8. The shortest RF distance of DLtree (*K* = 9) is 10, the shortest distance of CVtree (*K* = 9) is 16, the shortest distance of Tang's method (*K* = 7) is 16, and the shortest distance of *PWKmer* (*K* = 9) is 10. Therefore, the results of our method are closer to those of

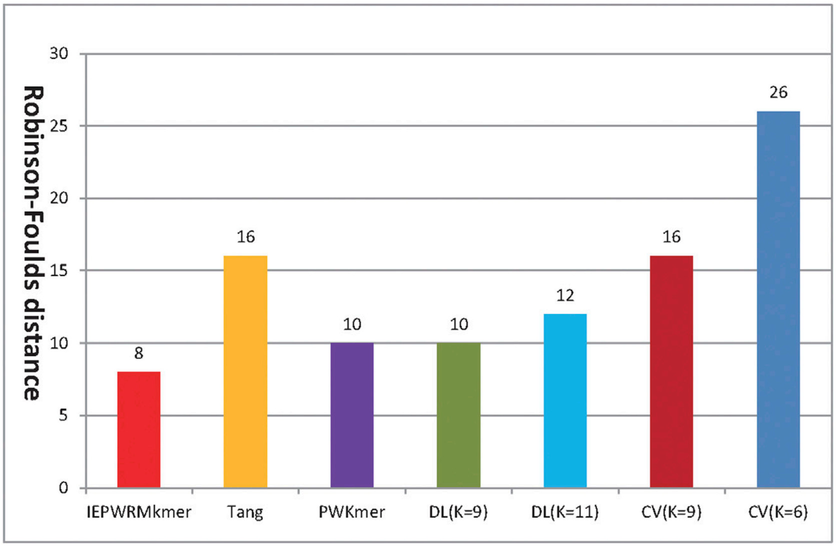

FIGURE 2 | The Robinson–Foulds distance between the tree reconstructed by ClustalX method and the phylogenetic trees reconstructed by our method (IEPWRMkmer K = 8), the CVTree method, the DLTree method, Tang's method (K = 7), and the PWKmer method (K = 9) on dataset 1 (we used the optimal tree by CVTree and DLTree).

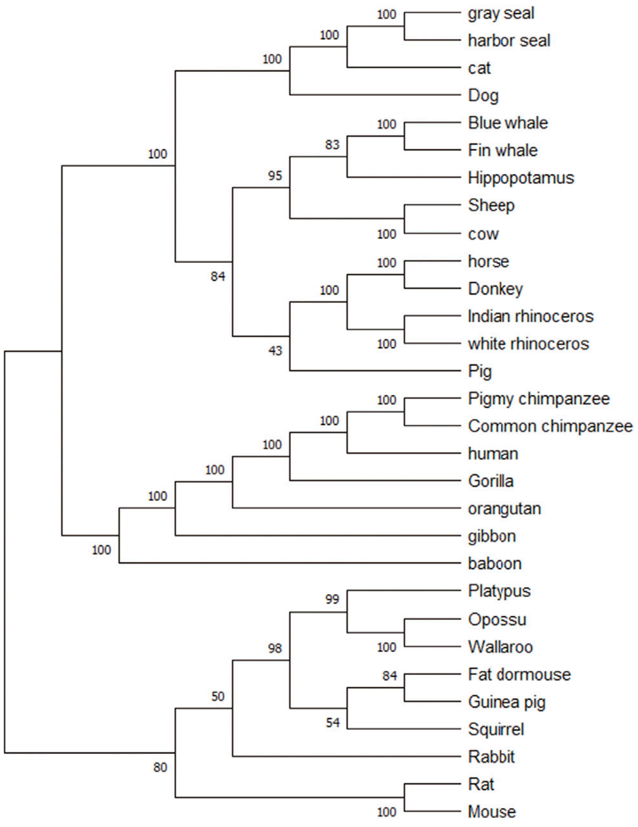

FIGURE 3 | The modified bootstrap consensus tree for **Figure 1B** based on 100 replicates.

ClustalX than those of the other methods, which indicates that our method is effective.

Figure 3 shows the consensus tree of 30 mammalian species based on our method. Compared with **Figure 1B**, 30 mammalian species are divided into the rodents group, the ferungulates group, and the primates group correctly. The support rate is 80% for the rodents group and 100% for both ferungulates and primates groups. Among the branches, marsupials (opossum, wallaroo), carnivores (dog, cat, harbor seal, gray seal), murid roots (rat, mouse), and cetaceans (fin whale, blue whale) are all supported by a 100% rate. In the artiodactyls group (cow, sheep, pig, hippopotamus), pig is separated out of the artiodactyls group, but the support rate is low at 43%. It indicates that the phylogenetic tree constructed by our method is quite robust.

Experiment 2

The human immunodeficiency viruses (HIV) represent a group of retroviruses, which are not presumed to have originated from human cellular DNA sequences, hence are distinct from endogenous retroviruses (Wu et al., 2007). HIV-1 can be classified into three major phylogenetic groups, namely M (major), N (new), and O (others). Group M is responsible for the HIV pandemic, it is divided into nine subtypes, namely A, B, C, D, F, G, J, K, and H. Based on differential phylogenetic clustering, the subtypes A and F are further divided into sub-subtypes (A1, A2) and (F1, F2), respectively. Groups N and O are derived from other primates and then infect humans. CPZ is a non-human primate virus isolated from chimpanzees, which is closest to human-to-human transmission of HIV.

We performed the phylogenetic analysis of 44 HIV-1 complete genome sequences in dataset 2 using ClustalX and our method.

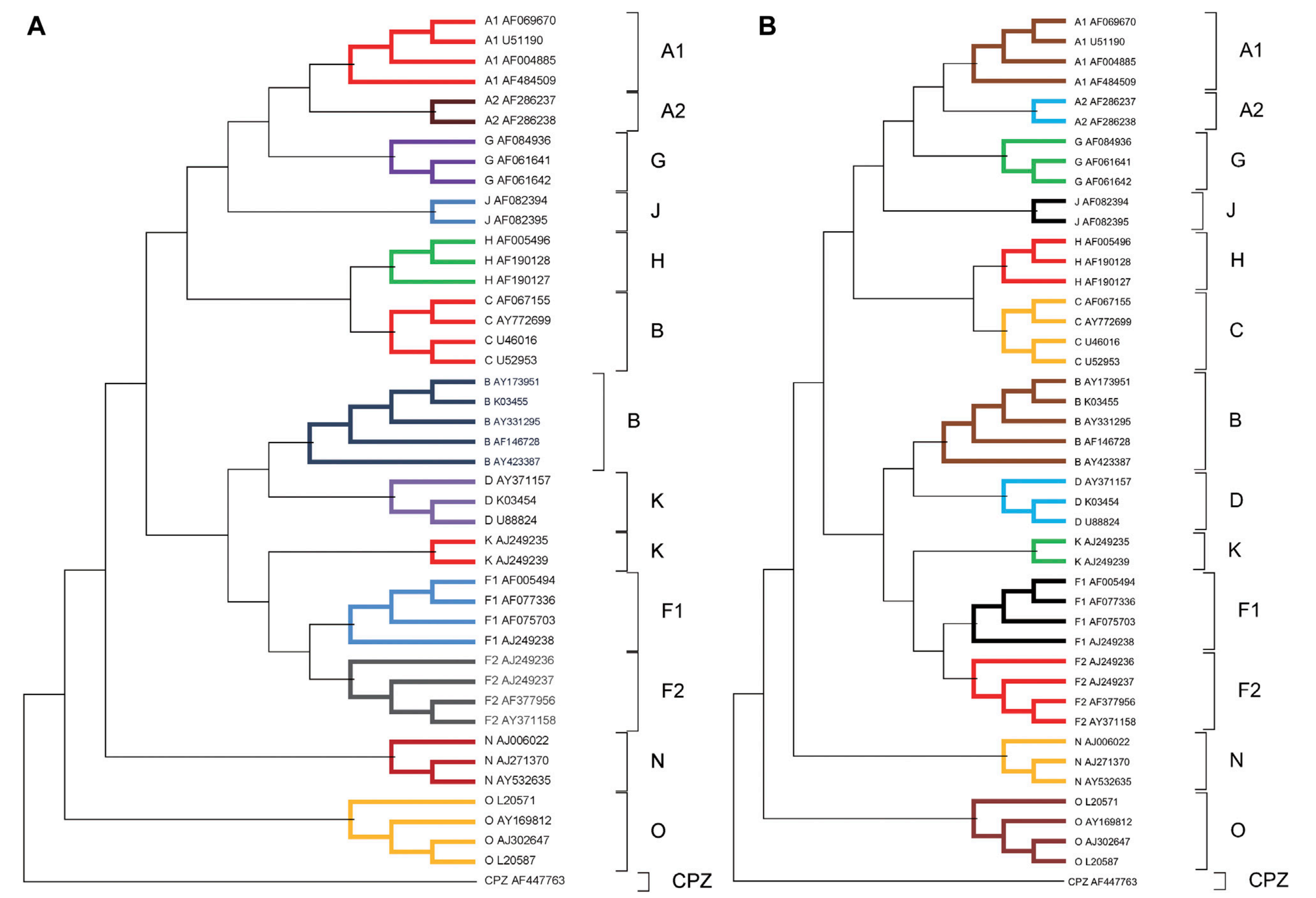

FIGURE 4 | (A) The phylogenetic tree of 44 HIV-1 genomes reconstructed by ClustalX. **(B)** The phylogenetic tree of 44 HIV-1 genomes reconstructed by our method ($K = 7$).

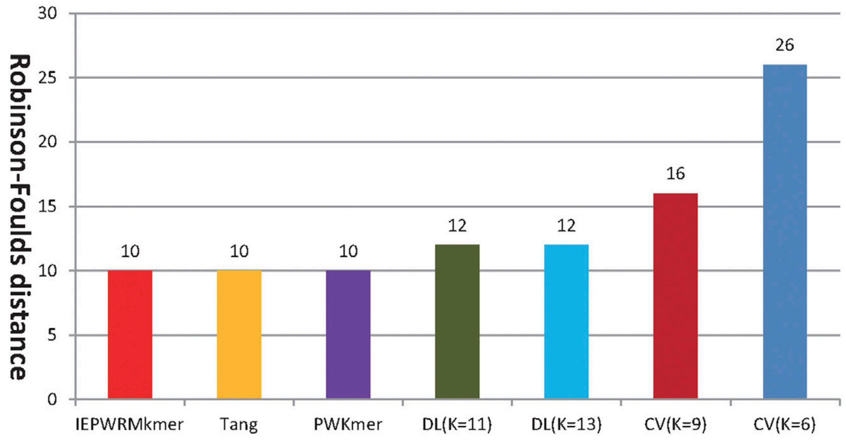

FIGURE 5 | The RF distance between the reference tree constructed by Clustalx and the phylogenetic trees constructed by our method (IEPWRMkmer, $K = 7$), Tang's method ($K = 8$), the PWKmer method ($K = 9$), the DLtree method, and the CVtree method. (For the PWKmer method, the DLtree method, and the CVtree method, we chose their optimal classification tree).

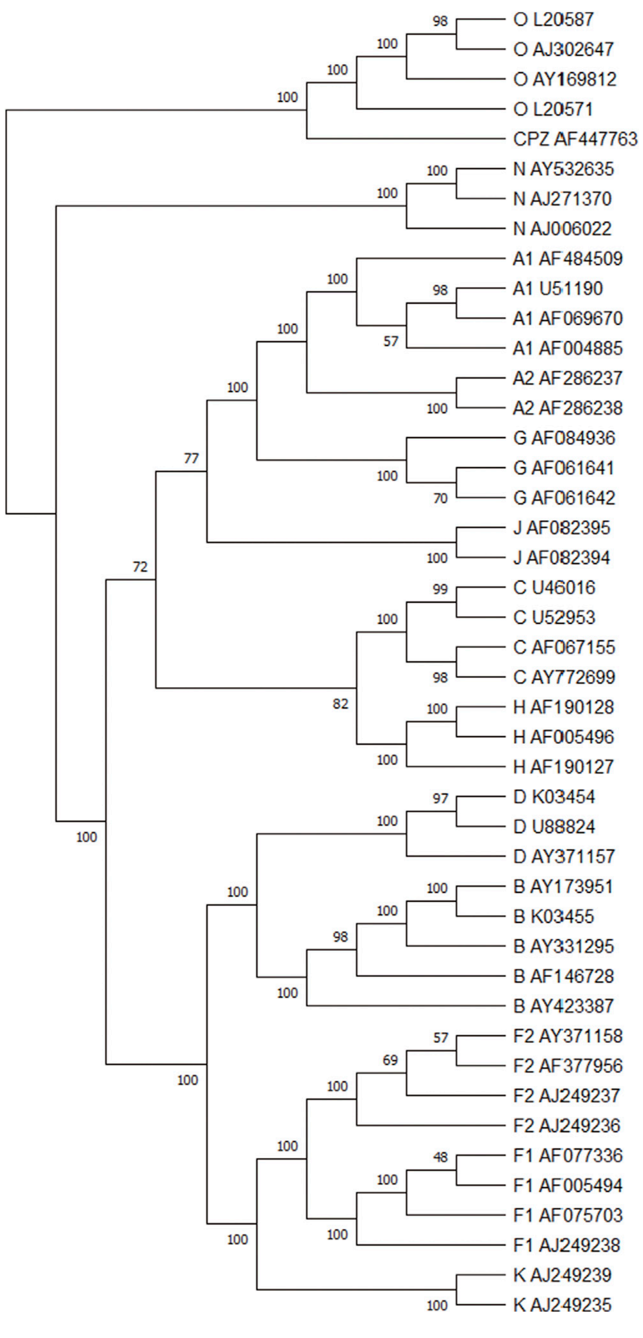

FIGURE 6 | The modified bootstrap consensus tree for **Figure 4B** based on 100 replicates.

The phylogenetic trees reconstructed by ClustalX and our method ($K = 7$) are shown in **Figure 4A** and **Figure 4B**, respectively. From **Figure 4B**, we can see that the species from all subtypes can be correctly classified into their groups (A, B, C, D, F, G, J, K, H, O, and M), and CPZ as the reference sequence is separated into the outermost. From the internal branches, both F and A contain two subtypes (F1 and F2) and (A1 and A2), respectively. Our method can separate the two subtypes, and in the branches, both F and A subtypes can be closely grouped together.

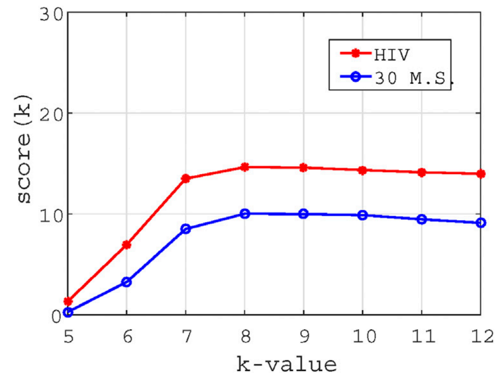

FIGURE 7 | The trend chart of K value vs scoring measure score(K). The red circles represent the scores of the dataset of 30 mammalian species for different K values, and the blue dots represent the scores of the HIV dataset for different K values.

Figure 5 shows the RF distances between the reference tree constructed by ClustalX and the phylogenetic trees constructed by our method, Tang's method, PWKmer, DLtree, and CVtree. Using our method, when $K = 7$, the RF distance is 10. The shortest RF distance of the DLtree ($K = 11$) is 12, the shortest distance of the CVtree ($K = 9$) is 16, the shortest distance of the PWKmer ($K = 9$) is 10, and the shortest distance of Tang's method ($K = 9$) is 10. Therefore, our method performs better than the DLtree and the CVtree on dataset 2 and has the same performance as Tang's method and PWKmer. The results indicate that our method is quite effective again.

Figure 6 shows the consensus tree of 44 HIV-1 based on our method. Comparing with **Figure 4B**, all HIV-1 sequences are divided into the M, N, O, and CPZ groups, whose support rate is 100%. From the branch point of view, in group M, the branch support rate of all subtypes is 100%. For subtypes A and F, the subtypes (A1, A2) and (F1 and F2) are clustered with 100% support. It again indicates that the phylogenetic tree constructed by our method is quite robust.

Estimate of the Optimal Parameter *K*

Different lengths of k-mers contain different phylogenetic information. Short k-mers may not contain sufficient DNA sequence information. Long k-mers contain sufficient phylogenetic information, but it needs large memory and takes a long time to calculate the distance based on information on long k-mers. Therefore, it is also very important to estimate an optimal value of K as heralded in (Yu et al., 2010) for the DLTree method and (Qi et al., 2004) for the CVTree method.

In this paper, we propose to use the Shannon entropy of the feature matrix to determine the optimal value of K. Using **Eq. 3**, we can obtain an $N \times 4^K$ feature matrix for a dataset with N genomes. Then, we propose to define a scoring strategy as

$$\mathrm{score}(K) = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{4^K}\left(E_{ij}\log_2 E_{ij} + \left(1 - E_{ij}\right)\log_2\left(1 - E_{ij}\right)\right). \tag{5}$$

The optimal K is the value at which $\mathrm{score}(K)$ reaches its maximum.

We use **Eq. 5** to calculate score(K) on datasets 1 and 2 for different K. The relationship between score(K) and K is shown in **Figure 7** for these two datasets. It is seen that score(K) reaches the largest value when $K = 8$ on the two datasets. Considering that the larger K is, the more memory resources are consumed, we only consider the values near $K = 8$ (e.g., $K = 7, 8, 9$). For the 30 mammalian species dataset, we have seen that the phylogenetic tree for $K = 8$ constructed by our method is closest to the reference tree. The same happened for the HIV-1 dataset with $K = 7$. The outcomes indicate that score(K) can provide an effective means to estimate the optimal value of K.

CONCLUSION

In this paper, a new alignment-free method is proposed for phylogenetic analysis and sequence comparison based on whole genome sequences. Our method combines the position-weighted measure of *k*-mers and the information entropy of frequency of *k*-mers. We used the Manhattan metric to measure the distance between a pair of sequences and the NJ method to construct the phylogenetic tree. In order to test the effectiveness and reliability of our method, we applied it on two datasets of 30 mammalian species and 44 HIV-1 genomes. The results demonstrated that the present method is efficient and reliable. A suitable K value is important to capture rich phylogenetic information of DNA sequences. In order to choose an optimal K value, we proposed a scoring measure based on the information entropy. The obtained results on two real datasets support that the method can capture the *k*-mer distribution information and is effective for whole genome sequence comparison and phylogenetic analysis.

Remark: The method of this paper is derived from the two studies Ma et al. (2020) and Murray et al. (2017). There are differences between this work and previous works: Tang et al. presented the average relative distance for normalized *k*-mers. PWKmer uses the counts and position distributions of *k*-mers to capture more evolutionary information. KWIP (Murray et al. 2017) uses information entropy to weight the inner product (Si∗Sj), while we use information entropy to weight the relative positions of *k*-mers. KWIP uses a kernel function to calculate the distance, while we use the Manhattan metric to calculate the pairwise distance between species. Here, we claimed that the results obtained by the IEPWRMkmer method are close to those by ClustalX and the IEPWRMkmer is superior to the other distance metrics. We used the phylogenetic tree constructed by ClustalX as the reference tree or standard tree, hence we cannot claim that our method is superior to the ClustalX method.

DATA AVAILABILITY STATEMENT

The genome datasets analyzed for this study can be found in the GenBank https://www.ncbi.nlm.nih.gov/

AUTHOR CONTRIBUTIONS

Y-QW contributed to the conception and design of the study, developed the method, and wrote the manuscript. Z-GY gave the ideas and supervised the project. All authors discussed the results and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by funds from the National Natural Science Foundation of China (grant numbers: 11871061 and 12026213); The National Key Research and Development Program of China (grant number: 2020YFC0832405); Innovation Foundation of Qian Xuesen Laboratory of Space Technology.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.* 215 (3), 403–410. doi:10.1016/S0022-2836(05)80360-2

Blaisdell, B. E. (1986). A Measure of the Similarity of Sets of Sequences Not Requiring Sequence Alignment. *Proc. Natl. Acad. Sci.* 83 (14), 5155–5159. doi:10.1073/pnas.83.14.5155

Chang, G., Wang, H., and Zhang, T. (2014). A Novel Alignment-free Method for Whole Genome Analysis: Application to HIV-1 Subtyping and HEV Genotyping. *Inf. Sci.* 279, 776–784. doi:10.1016/j.ins.2014.04.029

Comin, M., and Verzotto, D. (2012). Alignment-free Phylogeny of Whole Genomes Using Underlying Subwords. *Algorithms Mol. Biol.* 7 (1), 1–12. doi:10.1186/1748-7188-7-34

Ding, S., Li, Y., Yang, X., and Wang, T. (2013). A Simple *K*-word Interval Method for Phylogenetic Analysis of DNA Sequences. *J. Theor. Biol.* 317, 192–199. doi:10.1016/j.jtbi.2012.10.010

Felsenstein, J., and Felenstein, J. (2004). *Inferring Phylogenies*. (Sunderland, MA: Sinauer Associates). doi:10.1086/383584

Fox, G. E., Magrum, L. J., Balch, W. E., Wolfe, R. S., and Woese, C. R. (1977). Classification of Methanogenic Bacteria by 16S Ribosomal RNA Characterization. *Proc. Natl. Acad. Sci.* 74 (10), 4537–4541. doi:10.1073/pnas.74.10.4537

Haubold, B., Pierstorff, N., Möller, F., and Wiehe, T. (2005). Genome Comparison without Alignment Using Shortest Unique Substrings. *BMC Bioinformatics* 6 (1), 123–211. doi:10.1186/1471-2105-6-123

Hoang, T., Yin, C., and Yau, S. S.-T. (2016). Numerical Encoding of DNA Sequences by Chaos Game Representation with Application in Similarity Comparison. *Genomics* 108, 134–142. doi:10.1016/j.ygeno.2016.08.002

Höhl, M., Rigoutsos, I., and Ragan, M. A. (2006). Pattern-based Phylogenetic Distance Estimation and Tree Reconstruction. *Evol. Bioinformatics* 2, 359–375. doi:10.2174/157489306775330570

Huang, Y., and Wang, T. (2011). Phylogenetic Analysis of DNA Sequences with a Novel Characteristic Vector. *J. Math. Chem.* 49 (8), 1479–1492. doi:10.1007/s10910-011-9811-x

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874. doi:10.1093/molbev/msw054

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X Version 2.0. *Bioinformatics* 23 (21), 2947–2948. doi:10.1093/bioinformatics/btm404

Leimeister, C.-A., Boden, M., Horwege, S., Lindner, S., and Morgenstern, B. (2014). Fast Alignment-free Sequence Comparison Using Spaced-word Frequencies. *Bioinformatics* 30, 1991–1999. doi:10.1093/bioinformatics/btu177

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P., and Zhang, H. (2001). An Information-Based Sequence Distance and its Application to Whole Mitochondrial Genome Phylogeny. *Bioinformatics* 17 (2), 149–154. doi:10.1093/bioinformatics/17.2.149

Ma, Y., Yu, Z., Tang, R., Xie, X., Han, G., and Anh, V. V. (2020). Phylogenetic Analysis of HIV-1 Genomes Based on the Position-Weighted *K*-Mers Method. *Entropy* 22 (2), 255. doi:10.3390/e22020255

Mendizabal-Ruiz, G., Román-Godínez, I., Torres-Ramos, S., Salido-Ruiz, R. A., Vélez-Pérez, H., and Morales, J. A. (2018). Genomic Signal Processing for DNA Sequence Clustering. *PeerJ* 6 (3), e4264. doi:10.7717/peerj.4264

Morrison, D. A. (2006). Multiple Sequence Alignment for Phylogenetic Purposes. *Aust. Syst. Bot.* 19 (6), 479–539. doi:10.1071/sb06020

Murray, K. D., Webers, C., Ong, C. S., Borevitz, J., and Warthmann, N. (2017). KWIP: The *K*-Mer Weighted Inner Product, a De Novo Estimator of Genetic Similarity. *Plos Comput. Biol.* 13 (9), e1005727. doi:10.1371/journal.pcbi.1005727

Otu, H. H., and Sayood, K. (2003). A New Sequence Distance Measure for Phylogenetic Tree Construction. *Bioinformatics* 19 (16), 2122–2130. doi:10.1093/bioinformatics/btg295

Qi, J., Luo, H., and Hao, B. (2004). CVTree: a Phylogenetic Tree Reconstruction Tool Based on Whole Genomes. *Nucleic Acids Res.* 32 (Suppl. l_2), W45–W47. doi:10.1093/nar/gkh362

Robinson, D. F., and Foulds, L. R. (1981). Comparison of Phylogenetic Trees. *Math. Biosciences* 53 (1-2), 131–147. doi:10.1016/0025-5564(81)90043-2

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Syst. Biol.* 61 (3), 539–542. doi:10.1093/sysbio/sys029

Saitou, N., and Nei, M. (1987). The Neighbor-Joining Method: a New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* 4 (4), 406–425. doi:10.1093/oxfordjournals.molbev.a040454

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009). Alignment-free Genome Comparison with Feature Frequency Profiles (FFP) and Optimal Resolutions. *Pnas* 106 (8), 2677–2682. doi:10.1073/pnas.0813249106

Tang, J., Hua, K., Chen, M., Zhang, R., and Xie, X. (2014). A Novel *K*-word Relative Measure for Sequence Comparison. *Comput. Biol. Chem.* 53, 331–338. doi:10.1016/j.compbiolchem.2014.10.007

Thankachan, S. V., Chockalingam, S. P., Liu, Y., Apostolico, A., and Aluru, S. (2016). ALFRED: a Practical Method for Alignment-free Distance Computation. *J. Comput. Biol.* 23 (6), 452–460. doi:10.1089/cmb.2015.0217

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-specific gap Penalties and Weight Matrix Choice. *Nucl. Acids Res.* 22 (22), 4673–4680. doi:10.1093/nar/22.22.4673

Ulitsky, I., Burstein, D., Tuller, T., and Chor, B. (2006). The Average Common Substring Approach to Phylogenomic Reconstruction. *J. Comput. Biol.* 13 (2), 336–350. doi:10.1089/cmb.2006.13.336

Wang, Y., Lei, X., Wang, S., Wang, Z., Song, N., Zeng, F., et al. (2016). Effect of K-Tuple Length on Sample-Comparison with High-Throughput Sequencing Data. *Biochem. Biophysical Res. Commun.* 469 (4), 1021–1027. doi:10.1016/j.bbrc.2015.11.094

Wu, Q., Yu, Z.-G., and Yang, J. (2017). DLTree: Efficient and Accurate Phylogeny Reconstruction Using the Dynamical Language Method. *Bioinformatics* 33 (14), 2214–2215. doi:10.1093/bioinformatics/btx158

Wu, X., Cai, Z., Wan, X.-F., Hoang, T., Goebel, R., and Lin, G. (2007). Nucleotide Composition String Selection in HIV-1 Subtyping Using Whole Genomes. *Bioinformatics* 23 (14), 1744–1752. doi:10.1093/bioinformatics/btm248

Yin, C. (2019). Encoding and Decoding DNA Sequences by Integer Chaos Game Representation. *J. Comput. Biol.* 26 (2), 143–151. doi:10.1089/cmb.2018.0173

Yu, C., Deng, M., and Yau, S. S.-T. (2011). DNA Sequence Comparison by a Novel Probabilistic Method. *Inf. Sci.* 181 (8), 1484–1492. doi:10.1016/j.ins.2010.12.010

Yu, Z.-G., Chu, K. H., Li, C. P., Anh, V., Zhou, L.-Q., and Wang, R. W. (2010). Whole-proteome Phylogeny of Large dsDNA Viruses and Parvoviruses through a Composition Vector Method Related to Dynamical Language Model. *BMC Evol. Biol.* 10 (1), 1–11. doi:10.1186/1471-2148-10-192

Yu, Z.-G., Zhan, X.-W., Han, G.-S., Wang, R. W., Anh, V., and Chu, K. H. (2010). Proper Distance Metrics for Phylogenetic Analysis Using Complete Genomes without Sequence Alignment. *Ijms* 11 (3), 1141–1154. doi:10.3390/ijms11031141

Zielezinski, A., Vinga, S., Almeida, J., and Karlowski, W. M. (2017). Alignment-free Sequence Comparison: Benefits, Applications, and Tools. *Genome Biol.* 18 (1), 1–17. doi:10.1186/s13059-017-1319-7

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 November 2021
doi: 10.3389/fgene.2021.773202

iAIPs: Identifying Anti-Inflammatory Peptides Using Random Forest

Dongxu Zhao[1], Zhixia Teng[1], Yanjuan Li[2] and Dong Chen[2]*

[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2]College of Electrical and Information Engineering, Quzhou University, Quzhou, China

Recently, several anti-inflammatory peptides (AIPs) have been found in the process of the inflammatory response, and these peptides have been used to treat some inflammatory and autoimmune diseases. Therefore, identifying AIPs accurately from a given amino acid sequences is critical for the discovery of novel and efficient anti-inflammatory peptide-based therapeutics and the acceleration of their application in therapy. In this paper, a random forest-based model called iAIPs for identifying AIPs is proposed. First, the original samples were encoded with three feature extraction methods, including g-gap dipeptide composition (GDC), dipeptide deviation from the expected mean (DDE), and amino acid composition (AAC). Second, the optimal feature subset is generated by a two-step feature selection method, in which the feature is ranked by the analysis of variance (ANOVA) method, and the optimal feature subset is generated by the incremental feature selection strategy. Finally, the optimal feature subset is inputted into the random forest classifier, and the identification model is constructed. Experiment results showed that iAIPs achieved an AUC value of 0.822 on an independent test dataset, which indicated that our proposed model has better performance than the existing methods. Furthermore, the extraction of features for peptide sequences provides the basis for evolutionary analysis. The study of peptide identification is helpful to understand the diversity of species and analyze the evolutionary history of species.

Keywords: anti-inflammatory peptides, random forest, feature extraction, evolutionary information, evolutionary analysis

OPEN ACCESS

Edited by:
Juan Wang,
Inner Mongolia University, China

Reviewed by:
Haiying Zhang,
Xiamen University, China
Jiajie Peng,
Northwestern Polytechnical University, China

****Correspondence:***
Zhixia Teng
tengzhixia@nefu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *09 September 2021*
Accepted: *08 October 2021*
Published: *30 November 2021*

Citation:
Zhao D, Teng Z, Li Y and Chen D (2021) iAIPs: Identifying Anti-Inflammatory Peptides Using Random Forest. Front. Genet. 12:773202. doi: 10.3389/fgene.2021.773202

1 INTRODUCTION

As a part of the nonspecific immune response, inflammation response usually occurs in response to any type of bodily injury (Ferrero-Miliani et al., 2007). When the inflammatory response occurs in the condition of no obvious infection, or when the response continues despite the resolution of the initial insult, the process may be pathological and leads to chronic inflammation (Patterson et al., 2014). At present, the therapy for inflammatory and autoimmune diseases usually uses nonspecific anti-inflammatory drugs or other immunosuppressants, which may produce some side effects (Tabas and Glass, 2013; Yu et al., 2021). Several endogenous peptides found in the process of inflammatory response have become anti-inflammatory agents and can be used as new therapies for autoimmune diseases and inflammatory disorders (Gonzalez-Rey et al., 2007; Yu et al., 2020a). Compared with small-molecule drugs, the therapy based on peptides has minimal toxicity and high specificity under normal conditions, which is a better choice for inflammatory and autoimmune disorders and has been widely used in treatment (de la Fuente-Núñez et al., 2017; Shang et al., 2021).

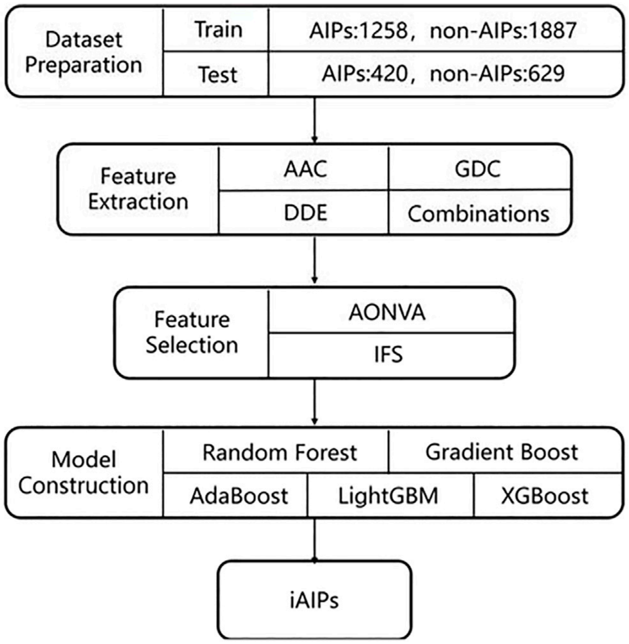

FIGURE 1 | The framework of iAIPs.

Due to the biological importance of AIPs, many biochemical experimental methods have been developed for identifying AIPs. However, these biochemical methods usually need a long experimental cycle and have a high experimental cost. In recent years, machine learning has increasingly become the most popular tool in the field of bioinformatics (Zhao et al., 2017; Liu et al., 2020; Luo et al., 2020; Sun et al., 2020; Zhao et al., 2020; Jin et al., 2021; Wang et al., 2021a). Many researchers have tried to adopt machine learning algorithms to identify AIPs only based on peptide amino acid sequence information. In 2017, Gupta et al. proposed a predictor of AIPs based on the machine learning method. They constructed the combined features and inputted them in the SVM classifier to construct the prediction model (Gupta et al., 2017).

In 2018, Manavalan et al. proposed a novel prediction model called AIPpred. They encoded the original peptide sequence by the dipeptide composition (DPC) feature representation method, and then, they developed a random forest-based model to identify AIPs (Manavalan et al., 2018). AIEpred is a novel prediction model and is proposed by Zhang et al. AIEpred encodes peptide sequences based on three feature representations. Based on various feature representations, it constructed many base classifiers, which are the basis of ensemble classifier (Zhang et al., 2020a).

In this paper, we proposed a novel identification model of AIPs for further improving the identification ability. First, we encoded the samples with multiple features consisting of AAC, DDE, and GDC. It has been proven that multiple features can effectively discriminate positive instances from negative ones in various biological problems. Second, we selected the optimal features based on a feature selection strategy, which has achieved better performance in many biological problems. Finally, we used the random forest classifier to construct an identification model based on the optimal features. The experimental result shows that our proposed method in this paper has better performance than the existing methods.

2 MATERIALS AND METHODS

Figure 1 gives the general framework of iAIPs proposed in this paper. The framework consists of four steps as follows: 1) Dataset preparation—It collects the data required for the experiment. 2) Feature extraction—It converts the collected sequence data from step 1 into numerical features. 3) Feature selection—removes redundant features from a feature set. 4) Prediction model construction. Each step of the framework will be described as follows.

2.1 Dataset Preparation

A high-quality dataset is critical to construct an effective and reliable prediction model. To measure the performance of our model by comparing it with other existing machine learning-based prediction models, we used the dataset with no change proposed in AIPpred (Manavalan et al., 2018). The dataset was first retrieved from the IEDB database (Kim et al., 2012; Vita et al., 2019), and then the samples with sequence identity >80% (Zou et al., 2020) are excluded by using CD-HIT (Huang et al., 2010). The dataset contains 1,678 AIPs and 2,516 non-AIPs. For this dataset, it is randomly selected as the training dataset, which is inputted into the classifier and used to construct the identification model. The training dataset is also used to measure the cross-validation performance of our model. The remaining dataset is used as an independent dataset, which will be used to evaluate the generalization capability of our identification model. In detail, the training dataset consists of 1,258 AIPs and 1,887 non-AIPs, and the independent dataset consists of 420 AIPs and 629 non-AIPs.

2.2 Feature Extraction Methods

In the process of peptide identification, finding an effective feature extraction method is the most important step (Liu, 2019; Fu et al., 2020; Cai et al., 2021). In this study, we tried a variety of feature extraction methods and used the random forest classifier to evaluate the performance of those methods. Finally, we chose three efficient feature extraction methods to encode peptide amino acid sequences, including amino acid composition, dipeptide deviation from expected mean, and g-gap dipeptide composition. The details of each feature extraction method are described as follows.

2.2.1 Amino Acid Composition

Different peptide sequences consist of different amino acid sequences. AAC tried to count the composition information of peptides. In detail, AAC calculates the frequency of occurrence of each amino acid type (Wei et al., 2018a; Liu et al., 2019; Ning et al., 2020; Yang et al., 2020; Zhang and Zou, 2020; Wu and Yu, 2021). The computation formula of AAC is as follows:

$$\mathrm{AAC}(\mathrm{j}) = \frac{N(\mathrm{j})}{L}, \quad \mathrm{j} \in \{A, C, D, E, F, \ldots, Y\}$$

where L denotes the length of the peptide, which is the number of characters in the peptide, AAC (j) denotes the percentage of amino acid j, N (j) denotes the total number of amino acid j. The dimension of AAC is 20.

2.2.2 Dipeptide Deviation From the Expected Mean

According to the dipeptide composition information, DDE computes deviation frequencies from expected mean values (Saravanan and Gautham, 2015). The feature vector extracted by DDE is generated by three parameters: theoretical variance (TV), dipeptide composition (DC), and theoretical mean (TM). The formulas of the three parameters are as follows:

$$D_C(\mathrm{j}) = \frac{n_j}{L-1}$$

where n_j denotes the occurred frequency of dipeptide j, and L denotes the length of peptide sequences.

$$T_M(j) = \frac{C_{j1}}{C_N} \times \frac{C_{j2}}{C_N}$$

C_{j1} denotes the number of codons that encode for the first amino acid, and C_{j2} denotes the number of codons that encode for the second amino acid in the dipeptide j. CN denotes the total number of possible codons.

$$T_V(j) = \frac{T_M(j)(1 - T_M(j))}{L-1}$$

The formula of DDE(i) is as follows.

$$DDE(j) = \frac{D_C(j) - T_M(j)}{\sqrt{T_V(j)}}$$

2.2.3 G-Gap Dipeptide Composition

GDC is used to measure the correlation of two non-adjacent residues; its dimension is 400 (Wei et al., 2018b). GDC can be represented as follows:

$$GDC(g) = (f_1^g, f_2^g, \ldots, f_{400}^g)$$

where f_v^g is the frequency of v (v = 1,2, . . ., 400), and it can be calculated as:

$$f_v^g = \frac{N_v^g}{\sum_{v=1}^{400} N_v^g}$$

where N_v^g denotes the number of the v-th g-gap dipeptide in a given peptide. In this study, every peptide has a different length; the minimum length is 5. Therefore, we set the range of g from 1 to 4. For the different values of g, we represent the feature as GDC-gap1, GDC-gap2, GDC-gap3, and GDC-gap4.

2.3 Feature Selection

In the *Feature extraction methods* section, we introduced the feature extraction method used in this paper. However, like other feature representation methods, our feature representation may also produce many noises (Wei et al., 2014; Wang et al., 2020a; Li et al., 2020; Tang et al., 2020; Wang et al., 2021b). Recently, many feature selection methods for eliminating noise has been used to solve many bioinformatics problems (He et al., 2020), such as TATA-binding protein prediction (Zou et al., 2016), DNA 4mc site prediction (Manavalan et al., 2019), antihypertensive peptide prediction (Manayalan et al., 2019), drug-induced hepatotoxicity prediction (Su et al., 2019), and enhance-promoter interaction prediction (Hong et al., 2020; Min et al., 2021).

Likewise, we will use a two-step feature selection method to solve the noise of features. In detail, the feature is first ranked based on the ANOVA score. Then, based on the orderly features, we use the incremental feature selection (IFS) strategy to generate different feature subsets, the feature subset with optimal performance is selected as the optimal feature subset. In the *Result and discussion* section, we will give the experiments about feature extraction, in which we will verify the effectiveness of our feature representation.

2.3.1 Analysis of Variance

In this work, the feature is first ranked based on the ANOVA score. For every feature, ANOVA calculated the ratio of the variance between groups and the variance within groups, which can test the mean difference between groups effectively (Ding et al., 2014). The score is calculated as follows:

$$S(t) = \frac{S_B^2(t)}{S_W^2(t)}$$

where S (t) is the score of the feature t, $S_B^2(t)$ is the variance between groups, and $S_W^2(t)$ is the variance within groups. The formula of $S_B^2(t)$ and $S_W^2(t)$ is as follows:

$$S_B^2(t) = \frac{1}{K-1}\sum_{i=1}^{K} m_i \left(\frac{\sum_{j=1}^{m_i} f_t(i,j)}{m_i} - \frac{\sum_{i=1}^{K}\sum_{j=1}^{m_i} f_t(i,j)}{\sum_{i=1}^{K} m_i} \right)^2$$

$$S_w^2(t) = \frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{m_i} \left(f_t(i,j) - \frac{\sum_{j=1}^{m_i} f_t(i,j)}{m_i} \right)^2$$

where K denotes the number of groups, and N denotes the total number of instances; $f_t(i, j)$ denote the value of the j-th sample in the i-th group of the feature t.

2.3.2 Incremental Feature Selection

Based on the orderly features, we use the incremental feature selection strategy to generate different feature subsets; the feature subset with optimal performance is selected as the optimal feature subset. In the incremental feature selection method, the feature set is constructed as empty at first, and then the feature vector is added one by one from the ranked feature set. Meanwhile, the new feature set is inputted into a classifier, and then a prediction model is constructed. We evaluate the performance of the model according to some indicators. Finally, the feature subset with the optimal performance is considered as the optimal feature set.

2.4 Machine Learning Methods

In this paper, we utilized various ensemble learning classification algorithms to develop identification models, which contain random forest (Ru et al., 2019; Wang et al., 2020b; Ao et al., 2021), AdaBoost, Gradient Boost Decision Tree (Yu et al., 2020b), LightGBM, and XGBoost. In addition, we also tried some traditional machine learning classification algorithms, such as logistic regression and Naïve Bayes. The description of these methods is as follows.

2.4.1 Random Forest

As one of the most powerful ensemble learning methods, random forest was proposed by Breiman (2001). Due to its effectiveness, random forest has been widely used in bioinformatics areas. Random forest can solve regression and classification tasks. To solve the problem, random forest uses the random feature selection method to construct hundreds or thousands of decision trees (Akbar et al., 2020). By voting on these decision trees, the final identification result is obtained. The random forest algorithm used in this paper is from WEKA (Hall et al., 2008), and all parameters are default.

2.4.2 AdaBoost

The AdaBoost algorithm is an iterative algorithm, which was proposed by Freund (1990). For a benchmark dataset, AdaBoost will train various weak classifiers and combine these weak classifiers by sample weight to construct a stronger final classifier. Among samples, low weights are assigned to easy samples that are classified correctly by the weak learner, while high weights are for the hard or misclassified samples. By constantly adjusting the weight of samples, AdaBoost will focus more on the samples that are classified incorrectly.

2.4.3 Gradient Boost Decision Tree

Similar to AdaBoost, Gradient Boost Decision Tree (GBDT) also combines weak learners to construct a prediction model (Friedman, 2001). Different from AdaBoost, GBDT will constantly adapt to the new model when the weak learners are learned. In detail, based on the negative gradient information of the loss function of the current model, the new weak classifier is trained. The training result is accumulated into the existing model to improve its performance (Basith et al., 2018).

2.4.4 LightGBM and XGBoost

Both LightGBM and XGBoost are improved algorithms based on GBDT. LightGBM is mainly optimized in three aspects. The histogram algorithm is used to convert continuous features into discrete features, the gradient-based one-side sampling (GOSS) method is used to adjust the sample distribution and reduce the numbers of samples, and the exclusive feature bundling (EFB) is used to merge multiple independent features. XGBoost adds the second-order Taylor expansion and regularization term to the loss function.

2.4.5 Na ve Bayes

Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem, which assumes that the features are independent of each other. According to this theorem, the probability of a given sample classified into class k can be calculated as

$$\mathrm{P}(C_k|X) = \frac{P(C_k)P(X|C_k)}{P(X)}$$

where the sample has the expression formula of {X, C}.

2.4.6 Other Machine Learning Methods

Other traditional machine learning methods used for performance comparison include J48, logistic, SMO, and SGD. J48 is a decision tree algorithm provided in Weka, which is implemented based on the C4.5 idea. Logistic is a probability-based classification algorithm. Based on linear regression, Logistic introduces sigmoid function to limit the output value to [0,1] interval. SMO and SGD are optimization algorithms provided in Weka. SMO (sequential minimal optimization) is based on support vector machine (SVM), and SGD is based on linear regression.

2.5 Performance Evaluation

To measure the performance of our proposed model, we chose four commonly used measurements: SN, SP, ACC, and MCC (Jiang et al., 2013; Wei et al., 2017a; Ding et al., 2019; Shen et al., 2019; Huang et al., 2020). These measurements are calculated as follows.

$$SN = \frac{TP}{TP + FN}$$
$$SP = \frac{TN}{TN + FP}$$
$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$
$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

where *FP*, *FN*, *TN*, and *TP* show the number of false-positive, false-negative, true-negative, and true-positive, respectively. These are widely used in bioinformatics studies, such as protein fold recognition (Shao et al., 2021), DNA-binding protein prediction (Wei et al., 2017b), protein–protein interaction prediction (Wei et al., 2017c), and drug–target interaction identification (Ding et al., 2020; Ding and JijunGuo, 2020).

Furthermore, we also used the receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982; Fushing and Turnbull, 1996) to evaluate the performance of our proposed model. ROC computes the true-positive rate and low false-positive rate by setting various possible thresholds (Gribskov and Robinson, 1996). The area under the ROC curve (AUC) also shows the performance of the proposed model, which is more accurate in the aspect of evaluating the performance of the prediction model constructed by an imbalanced dataset.

TABLE 1 | Performance comparison of various single features.

Feature	SN	SP	ACC	MCC	AUC
Amino acid composition (AAC)	0.529	0.845	0.719	0.398	0.760
Dipeptide deviation for the expected mean (DDE)	0.589	0.854	0.748	0.464	0.784
G-gap dipeptide composition (GDC)-gap1	0.456	0.862	0.700	0.353	0.764
GDC-gap2	0.466	0.852	0.697	0.348	0.751
GDC-gap3	0.454	0.869	0.703	0.361	0.741
GDC-gap4	0.449	0.853	0.692	0.335	0.733
CKSAAGP	0.477	0.861	0.707	0.371	0.732
CTriad	0.215	0.897	0.624	0.155	0.668
GAAC	0.533	0.750	0.663	0.288	0.679
GDPC	0.525	0.826	0.706	0.370	0.727
GTPC	0.470	0.855	0.701	0.357	0.742
TPC	0.304	0.910	0.668	0.277	0.739

TABLE 2 | Performance comparison of various combined features of fivefold cross-validation on the training dataset.

Feature	SN	SP	ACC	MCC	AUC
AAC+DDE	0.582	0.857	0.747	0.461	0.784
AAC+GDC-gap1	0.483	0.870	0.715	0.388	0.770
AAC+GDC-gap2	0.453	0.871	0.704	0.363	0.773
AAC+GDC-gap3	0.435	0.866	0.694	0.339	0.759
AAC+GDC-gap4	0.447	0.873	0.703	0.360	0.760
DDE+GDC-gap1	0.586	0.858	0.749	0.466	0.790
DDE+GDC-gap2	0.588	0.854	0.748	0.464	0.791
DDE+GDC-gap3	0.583	0.860	0.749	0.466	0.785
DDE+GDC-gap4	0.587	0.851	0.746	0.459	0.784
AAC+DDE+GDC-gap1	0.585	0.860	0.750	0.468	0.794
AAC+DDE+GDC-gap2	0.584	0.852	0.745	0.457	0.790
AAC+DDE+GDC-gap3	0.593	0.857	0.751	0.471	0.784
AAC+DDE+GDC-gap4	0.587	0.855	0.748	0.464	0.785

TABLE 3 | Performance comparison of various combined features on the independent dataset.

Feature	SN	SP	ACC	MCC	AUC
AAC+DDE	0.564	0.860	0.742	0.450	0.808
AAC+GDC-gap1	0.488	0.884	0.725	0.413	0.799
AAC+GDC-gap2	0.455	0.878	0.708	0.373	0.787
AAC+GDC-gap3	0.448	0.881	0.707	0.371	0.795
AAC+GDC-gap4	0.462	0.865	0.704	0.362	0.783
DDE+GDC-gap1	0.569	0.857	0.742	0.450	0.812
DDE+GDC-gap2	0.560	0.854	0.736	0.437	0.805
DDE+GDC-gap3	0.576	0.857	0.745	0.456	0.808
DDE+GDC-gap4	0.569	0.857	0.742	0.450	0.801
AAC+DDE+GDC-gap1	0.56	0.859	0.739	0.443	0.806
AAC+DDE+GDC-gap2	0.557	0.855	0.736	0.437	0.805
AAC+DDE+GDC-gap3	0.552	0.855	0.734	0.433	0.806
AAC+DDE+GDC-gap4	0.567	0.859	0.742	0.450	0.801

3 RESULTS AND DISCUSSION

To verify the effectiveness of our proposed model, we will measure the performance of our model from different perspectives. The detailed process of these experiments is presented as follows.

3.1 Performance of Different Features

In this study, we use a variety of feature extraction methods and their combinations to encode peptide sequences. At first, we measure the effectiveness of single features. The comparison results of the fivefold cross-validation on the training dataset are shown in **Table 1**.

Table 1 shows that DDE is much better than other features according to the indicators of AUC, MCC, ACC, SP, and SN. In detail, the AUC value reaches 0.784, which is 2%–11.6% higher than other features. Based on the indicator of AUC, the features of DDE, GDC-gap1, and AAC have the best performance.

To achieve better performance, we further test the performance of multiple features on the basis of DDE, GDC, and AAC. In detail, the GDC feature adopts four different parameters, that is, gap1, gap2, gap3, and gap4. The corresponding feature is GDC-gap1, GDC-gap2, GDC-gap3, and GDC-gap4. The performance comparison of the fivefold cross-validation on the training dataset is shown in **Table 2**.

According to **Table 2**, the multiple features of AAC + DDE + GDC-gap1 has the best performance. Its value of SN, SP, ACC, MCC, and AUC are 0.585, 0.860, 0.750, 0.468, and 0.794, respectively.

To verify the performance of these combined features, we tested them on the independent test set. **Table 3** shows the experimental results on the independent dataset. The results show that the combined features of AAC + DDE + GDC-gap1 have the best performance on the independent dataset.

3.2 Performance of Different Classifiers

In this study, we chose the random forest algorithm to construct the classifier. To verify the effectiveness of the random forest classifier, we compared its performance with other classifiers. We chose several ensemble classifiers that are similar to the random forest classifier, including AdaBoost, GBDT, LightGBM, and XGBoost. In addition, we also chose some machine learning classifiers, including J48, Logistic, SMO, SGD, and Naïve Bayes.

Based on the best feature combination, which is obtained from previous experiments, we constructed different identification models using different classifiers. The performance of these classifiers on the training dataset is shown in **Table 4**.

TABLE 4 | Performance of various classifiers utilizing AAC-DDE-GDC-gap1 feature and fivefold cross-validation on the training dataset.

Classifier	SN	SP	ACC	MCC	AUC
Random forest	0.585	0.860	0.750	0.468	0.794
AdaBoost	0.579	0.743	0.678	0.324	0.661
Gradient Boost Decision Tree (GBDT)	0.583	0.788	0.706	0.379	0.686
LightGBM	0.564	0.754	0.678	0.321	0.659
XGBoost	0.576	0.757	0.684	0.336	0.666
J48	0.552	0.737	0.663	0.292	0.647
Logistic	0.497	0.677	0.605	0.175	0.624
Sequential minimal optimization (SMO)	0.476	0.725	0.626	0.206	0.601
SGD	0.491	0.689	0.610	0.182	0.590
Naïve Bayes	0.483	0.684	0.603	0.168	0.604

TABLE 5 | Performance of various classifiers based on AAC-DDE-GDC-gap1 feature on the independent dataset.

Classifier	SN	SP	ACC	MCC	AUC
Random forest	0.560	0.859	0.739	0.443	0.806
AdaBoost	0.607	0.809	0.728	0.426	0.708
GBDT	0.640	0.798	0.735	0.443	0.719
LightGBM	0.538	0.859	0.730	0.424	0.698
XGBoost	0.579	0.847	0.740	0.446	0.713
J48	0.524	0.738	0.652	0.266	0.621
Logistic	0.498	0.658	0.594	0.156	0.615
SMO	0.442	0.701	0.598	0.147	0.572
SGD	0.493	0.679	0.604	0.173	0.586
Naïve Bayes	0.486	0.676	0.600	0.162	0.602

The results in **Table 4** show that the performance of the random forest classifier is the best, and its AUC value is 10.8%–20.4% higher than other classifiers. To further compare the generalization ability of these classifiers, we test those models on the independent dataset. **Table 5** shows the experimental results. The results showed that the random forest classifier is also better than other classifiers on the independent dataset.

3.3 The Analysis of Feature Selection

In the extracted features, some feature vectors may be noisy or redundant. To further improve the identification performance, we try to find optimal features by feature selection methods in this section. In this paper, the two-step feature selection strategy is used as the feature selection strategy to eliminate noise. In detail, we first used the ANOVA method to rank feature vectors, and then we used the IFS strategy to filter the optimal feature set.

The comparison of performance before and after dimensionality reduction is shown in **Figure 2**. All indicators of the selected features have higher values than the original ones. The results suggest that the optimal feature set can improve the overall performance of our identification model and our fewer selected features can still accurately describe AIPs.

3.4 Comparison With Existing Methods

Independent dataset test plays an important role in testing the generalization ability of the identification model. Therefore, the independent dataset was used to measure our identification model; the performance of our identification model was

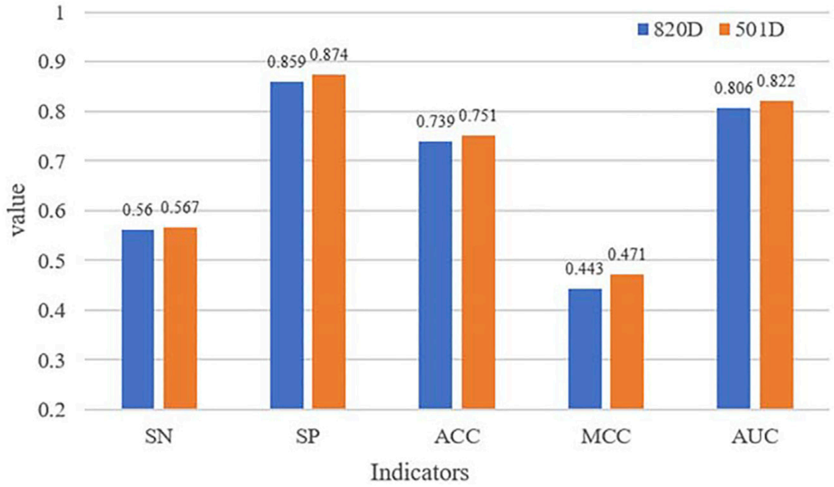

FIGURE 2 | Comparison of identification performance before and after dimensionality reduction.

TABLE 6 | Performance of different identification models on the independent dataset.

Method	SN	SP	ACC	MCC	AUC
AntiInflam (LA)	0.258	0.892	0.638	0.197	0.647
AntiInflam (MA)	0.786	0.417	0.565	0.210	0.706
AIEpred	0.555	0.899	0.762	0.495	0.767
AIPpred	0.741	0.746	0.744	0.479	0.813
iAIPs (our work)	0.567	0.874	0.751	0.471	0.822

compared with existing methods, which contains AntiInflam (Ferrero-Miliani et al., 2007), AIPpred, and AIEpred. **Table 6** shows the detailed results of the different methods for identifying AIPs, where the results are ranked according to AUC.

As shown in **Table 6**, the value of our proposed identification model iAIPs in SN, SP, ACC, AUC, and MCC are 0.567, 0.874, 0.751, 0.822, and 0.471, respectively. Furthermore, the same independent dataset-based experimental results showed that the ACC of iAIPs was 0.007–0.186 higher than that of AntiInflam and AIPpred, which is similar to AIEpred. Moreover, according to AUC, our performance is better than the other methods, which is 0.009–0.175 higher than the others. The results indicate that our method has better performance than other existing prediction models.

4 CONCLUSION

In this paper, an identifying AIP model based on peptide sequence is proposed. We tried various features and their combinations, utilized various commonly used ensemble learning classification algorithms and the two-step feature selection strategy. After trying a large number of experiments, we finally constructed an effective AIP prediction model. By conducting a large number of experiments on the training dataset and independent dataset, we verified that our proposed prediction model iAIPs could efficiently identify AIPs from the newly synthesized and discovered peptide sequences, which is better than the existing AIP prediction models.

In the future, the optimization of the feature representation method is a research direction. Especially, the research on a new feature representation method that can adaptively encode peptide sequences is of great significance. Furthermore, other optimization methods and computational intelligence models will be considered for identifying anti-inflammatory peptides. Deep learning (Lv et al., 2019; Zeng et al., 2020a; Zeng et al., 2020b; Zhang et al., 2020b; Du et al., 2020; Pang and Liu, 2020), unsupervised learning (Zeng et al., 2020c), and ensemble learning (Sultana et al., 2020; Zhong et al., 2020; Li et al., 2021; Niu et al., 2021; Shao and Liu, 2021) will be employed when the dataset is large enough.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: http://www.thegleelab.org/AIPpred/.

AUTHOR CONTRIBUTIONS

DZ and ZT conceptualized the study. DZ and YL formulated the methodology. DZ validated the study and wrote the original draft. DC and YL reviewed and edited the manuscript. ZT supervised the study and acquired the funding. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work is supported by the Fundamental Research Funds for the Central Universities (2572018BH05, 2572017CB33), the National Natural Science Foundation of China (61901103, 28961671189), and the Natural Science Foundation of Heilongjiang Province (LH2019F002).

REFERENCES

Akbar, S., Ateeq Ur, R., Maqsood, H., and Mohammad, S. (2020). cACP: Classifying Anticancer Peptides Using Discriminative Intelligent Model via Chou's 5-step Rules and General Pseudo Components. *Chemometrics Intell. Lab. Syst.* 196, 103912. doi:10.1016/j.chemolab.2019.103912

Ao, C., Zou, Q., and Yu, L. (2021). *RFhy-m2G: Identification of RNA N2-Methylguanosine Modification Sites Based on Random forest and Hybrid Features*. Methods (San Diego, Calif.): Elsevier.

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

Breiman, L. (2001). Random Forests. *Machine Learn.* 45 (1), 5–32. doi:10.1023/a:1010933404324

Cai, L., Wang, L., Fu, X., Xia, C., Zeng, X., and Zou, Q. (2021). ITP-pred: an Interpretable Method for Predicting, Therapeutic Peptides with Fused Features Low-Dimension Representation. *Brief Bioinform* 22 (4), bbaa367. doi:10.1093/bib/bbaa367

de la Fuente-Núñez, C., Silva, O. N., Lu, T. K., and Franco, O. L. (2017). Antimicrobial Peptides: Role in Human Disease and Potential as Immunotherapies. *Pharmacol. Ther.* 178, 132–140. doi:10.1016/j.pharmthera.2017.04.002

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 (8), 2229–2235. doi:10.1039/c4mb00316k

Ding, Y. T., and JijunGuo, F. (2020). Identification of Drug-Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion. *Knowledge-Based Syst.*, 204. doi:10.1016/j.knosys.2020.106254

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

Ding, Y., Tang, J., and Guo, F. (2020). Identification of Drug-Target Interactions via Fuzzy Bipartite Local Model. *Neural Comput. Applic* 32, 10303–10319. doi:10.1007/s00521-019-04569-z

Du, Z., Xiao, X., and Uversky, V. N. (2020). Classification of Chromosomal DNA Sequences Using Hybrid Deep Learning Architectures. *Curr. Bioinformatics* 15 (10), 1130–1136. doi:10.2174/1574893615666200224095531

Ferrero-Miliani, L., Nielsen, O. H., Andersen, P. S., and Girardin, S. E. (2007). Chronic Inflammation: Importance of NOD2 and NALP3 in Interleukin-1beta

Generation. *Clin. Exp. Immunol.* 147 (2), 227–235. doi:10.1111/j.1365-2249.2006.03261.x

Freund, Y. (1990). Boosting a Weak Learning Algorithm by Majority. *Inf. Comput.* 121 (2), 256–285. doi:10.1006/inco.1995.1136

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* 29 (5), 1189–1232. doi:10.1214/aos/1013203451

Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). StackCPPred: a Stacking and Pairwise Energy Content-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency. *Bioinformatics* 36 (10), 3028–3034. doi:10.1093/bioinformatics/btaa131

Fushing, H., and Turnbull, B. W. (1996). Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. *Ann. Stat.* 24 (1), 25–40. doi:10.1214/aos/1033066197

Gonzalez-Rey, E., Anderson, P., and Delgado, M. (2007). Emerging Roles of Vasoactive Intestinal Peptide: a New Approach for Autoimmune Therapy. *Ann. Rheum. Dis.* 66 (3), iii70–6. doi:10.1136/ard.2007.078519

Gribskov, M., and Robinson, N. L. (1996). Use of Receiver Operating Characteristic (ROC) Analysis to Evaluate Sequence Matching. *Comput. Chem.* 20 (1), 25–33. doi:10.1016/s0097-8485(96)80004-0

Gupta, S., Sharma, A. K., Shastri, V., Madhu, M. K., and Sharma, V. K. (2017). Prediction of Anti-inflammatory Proteins/peptides: an Insilico Approach. *J. Transl Med.* 15 (1), 7. doi:10.1186/s12967-016-1103-6

Hall, M., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., and Witten, I. H. (2008). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsl.* 11 (1), 10–18. doi:10.1145/1656274.1656278

Hanley, J. A., and McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 143 (1), 29–36. doi:10.1148/radiology.143.1.7063747

He, S., Fei, G., Quan, Z., and Hui, D. (2020). MRMD2.0: A Python Tool for Machine Learning with Feature Ranking and Reduction. *Curr. Bioinformatics* 15 (10), 1213–1221. doi:10.2174/1574893615999200503030350

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 (4), 1037–1043. doi:10.1093/bioinformatics/btz694

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a Web Server for Clustering and Comparing Biological Sequences. *Bioinformatics* 26 (5), 680–682. doi:10.1093/bioinformatics/btq003

Huang, Y., Zhou, D., Wang, Y., Zhang, X., Su, M., Wang, C., et al. (2020). Prediction of Transcription Factors Binding Events Based on Epigenetic Modifications in Different Human Cells. *Epigenomics* 12 (16), 1443–1456. doi:10.2217/epi-2019-0321

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting Human microRNA-Disease Associations Based on Support Vector Machine. *Ijdmb* 8 (3), 282–293. doi:10.1504/ijdmb.2013.056078

Jin, S., Zeng, X., Xia, F., Huang, W., and Liu, X. (2021). Application of Deep Learning Methods in Biological Networks. *Brief. Bioinform.* 22 (2), 1902–1917. doi:10.1093/bib/bbaa043

Kim, Y., Ponomarenko, J., Zhu, Z., Tamang, D., Wang, P., Greenbaum, J., et al. (2012). Immune Epitope Database Analysis Resource. *Nucleic Acids Res.* 40, W525–W530. doi:10.1093/nar/gks438

Li, J., Pu, Y., Tang, J., Zou, Q., and Guo, F. (2020). DeepATT: a Hybrid Category Attention Neural Network for Identifying Functional Effects of DNA Sequences. *Brief Bioinform* 21, 8. doi:10.1093/bib/bbaa159

Li, J., Wei, L., Guo, F., and Zou, Q. (2021). EP3: An Ensemble Predictor that Accurately Identifies Type III Secreted Effectors. *Brief. Bioinform.* 22 (2), 1918–1928. doi:10.1093/bib/bbaa008

Liu, B. (2019). BioSeq-Analysis: a Platform for DNA, RNA and Protein Sequence Analysis Based on Machine Learning Approaches. *Brief. Bioinform.* 20 (4), 1280–1294. doi:10.1093/bib/bbx165

Liu, B., Gao, X., and Zhang, H. (2019). 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(20): p. e127. doi:10.1093/nar/gkz740

Liu, Y., Yalin, H., Guohua, W., and Yadong, W. (2020). A Deep Learning Approach for Filtering Structural Variants in Short Read Sequencing Data. *Brief Bioinform* 22 (4). doi:10.1093/bib/bbaa370

Luo, X., Wang, F., Wang, G., and Zhao, Y. (2020). Identification of Methylation States of DNA Regions for Illumina Methylation BeadChip. *BMC Genomics* 21(Suppl. 1): p. 672. doi:10.1186/s12864-019-6019-0

Lv, Z., Ao, C., and Zou, Q. (2019). Protein Function Prediction: From Traditional Classifier to Deep Learning. *Proteomics* 19 (14), e1900119. doi:10.1002/pmic.201900119

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). 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., Kim, M. O., and Lee, G. (2018). AIPpred: Sequence-Based Prediction of Anti-inflammatory Peptides Using Random Forest. *Front. Pharmacol.* 9, 276. doi:10.3389/fphar.2018.00276

Manayalan, B., Shaherin, B., Tae Hwan, S., Leyi, W., and Gwang, L. (2019). mAHTPred: a Sequence-Based Meta-Predictor for Improving the Prediction of Anti-hypertensive Peptides Using Effective Feature Representation. *Bioinformatics* 35 (16), 2757–2765. doi:10.1093/bioinformatics/bty1047

Min, X., Ye, C., Liu, X., and Zeng, X. (2021). Predicting Enhancer-Promoter Interactions by Deep Learning and Matching Heuristic. *Brief. Bioinform.* 22. doi:10.1093/bib/bbaa254

Ning, L., Huang, J., He, B., and Kang, J. (2020). An In Silico Immunogenicity Analysis for PbHRH: An Antiangiogenic Peptibody by Fusing HRH Peptide and Human IgG1 Fc Fragment. *Cbio* 15 (6), 547–553. doi:10.2174/1574893614666190730104348

Niu, M., Lin, Y., and Zou, Q. (2021). sgRNACNN: Identifying sgRNA On-Target Activity in Four Crops Using Ensembles of Convolutional Neural Networks. *Plant Mol. Biol.* 105 (4-5), 483–495. doi:10.1007/s11103-020-01102-y

Pang, Y., and Liu, B., (2020). SelfAT-Fold: Protein Fold Recognition Based on Residue-Based and Motif-Based Self-Attention Networks. *Ieee/acm Trans. Comput. Biol. Bioinf.* 1, 1. doi:10.1109/TCBB.2020.3031888

Patterson, H., Nibbs, R., McInnes, I., and Siebert, S. (2014). Protein Kinase Inhibitors in the Treatment of Inflammatory and Autoimmune Diseases. *Clin. Exp. Immunol.* 176 (1), 1–10. doi:10.1111/cei.12248

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 (7), 2931–2939. doi:10.1021/acs.jproteome.9b00250

Saravanan, V., and Gautham, N. (2015). Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. *OMICS: A J. Integr. Biol.* 19 (10), 648–658. doi:10.1089/omi.2015.0095

Shang, Y., Gao, L., Zou, Q., and Yu, L. (2021). Prediction of Drug-Target Interactions Based on Multi-Layer Network Representation Learning. *Neurocomputing* 434, 80–89. doi:10.1016/j.neucom.2020.12.068

Shao, J., and Liu, B. (2021). ProtFold-DFG: Protein Fold Recognition by Combining Directed Fusion Graph and PageRank Algorithm. *Brief. Bioinform.* 22, 32–40. doi:10.1093/bib/bbaa192

Shao, J., Yan, K., and Liu, B. (2021). FoldRec-C2C: Protein Fold Recognition by Combining Cluster-To-Cluster Model and Protein Similarity Network. *Brief. Bioinform.* 22, 32–40. doi:10.1093/bib/bbaa144

Shen, Y., Tang, 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

Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2019). Developing a Multi-Dose Computational Model for Drug-Induced Hepatotoxicity Prediction Based on Toxicogenomics Data. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16 (4), 1231–1239. doi:10.1109/tcbb.2018.2858756

Sultana, N., Sharma, N., Sharma, K. P., and Verma, S. (2020). A Sequential Ensemble Model for Communicable Disease Forecasting. *Cbio* 15 (4), 309–317. doi:10.2174/1574893614666191202153824

Sun, S., Lei, X., Quan, Z., and Guohua, W. (2020). BP4RNAseq: a Babysitter Package for Retrospective and Newly Generated RNA-Seq Data Analyses Using Both Alignment-Based and Alignment-free Quantification Method. *Bioinformatics* 37 (9), 1319–1321. doi:10.1093/bioinformatics/btaa832

Tabas, I., and Glass, C. K. (2013). Anti-inflammatory Therapy in Chronic Disease: Challenges and Opportunities. *Science* 339 (6116), 166–172. doi:10.1126/science.1230720

Tang, Y.-J., Pang, Y.-H., Liu, B., and Idp-Seq2Seq (2020). IDP-Seq2Seq: Identification of Intrinsically Disordered Regions Based on Sequence to Sequence Learning. *Bioinformaitcs* 36 (21), 5177–5186. doi:10.1093/bioinformatics/btaa667

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2019). The Immune Epitope Database (IEDB): 2018 Update. *Nucleic Acids Res.* 47 (D1), D339–D343. doi:10.1093/nar/gky1006

Wang, C., Zhang, Y., and Han, S. (2020). Its2vec: Fungal Species Identification Using Sequence Embedding and Random Forest Classification. *Biomed. Res. Int.* 2020, 1–11. doi:10.1155/2020/2468789

Wang, H. D., Tang, J., Zou, Q., and Guo, F. (2021). Identify RNA-Associated Subcellular Localizations Based on Multi-Label Learning Using Chou's 5-steps Rule. *BMC Genomics* 22(56): p. 1-1.doi:10.1186/s12864-020-07347-7

Wang, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of Membrane Protein Types via Multivariate Information Fusion with Hilbert-Schmidt Independence Criterion. *Neurocomputing* 383, 257–269. doi:10.1016/j.neucom.2019.11.103

Wang, X., Yang, Y., Jian, L., and Guohua, W. (2021). The Stacking Strategy-Based Hybrid Framework for Identifying Non-coding RNAs. *Brief Bioinform* 22 (5), 32–40. doi:10.1093/bib/bbab023

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 (23), 4007–4016. doi:10.1093/bioinformatics/bty451

Wei, L., Jie, H., Fuyi, Li., Jiangning, S., Ran, S., and Quan, Z. (2018). *Comparative Analysis and Prediction of Quorum-Sensing Peptides Using Feature Representation Learning and Machine Learning Algorithms*, 21. Brief Bioinform, 106–119. doi:10.1093/bib/bby107

Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. *Ieee/acm Trans. Comput. Biol. Bioinf.* 11 (1), 192–201. doi:10.1109/tcbb.2013.146

Wei, L., Tang, J., and Zou, Q. (2017). Local-DPP: An Improved DNA-Binding Protein Prediction Method by Exploring Local Evolutionary Information. *Inf. Sci.* 384, 135–144. doi:10.1016/j.ins.2016.06.026

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017). A Novel Hierarchical Selective Ensemble Classifier with Bioinformatics Application. *Artif. Intelligence Med.* 83, 82–90. doi:10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved Prediction of Protein-Protein Interactions Using Novel Negative Samples, Features, and an Ensemble Classifier. *Artif. Intelligence Med.* 83, 67–74. doi:10.1016/j.artmed.2017.03.001

Wu, X., and Yu, L. (2021). *EPSOL: Sequence-Based Protein Solubility Prediction Using Multidimensional Embedding*. Oxford, England): Bioinformatics. doi:10.1093/bioinformatics/btab463

Zhong, Y., Lin, W., and Zou, Q. (2020). Predicting Disease-Associated Circular RNAs Using Deep Forests Combined with Positive-Unlabeled Learning Methods. *Brief. Bioinformatics* 21 (4), 1425–1436. doi:10.1093/bib/bbz080

Yang, L., Gao, H., Wu, K., Zhang, H., Li, C., and Tang, L. (2020). Identification of Cancerlectins by Using Cascade Linear Discriminant Analysis and Optimal G-gap Tripeptide Composition. *Cbio* 15 (6), 528–537. doi:10.2174/1574893614666190730103156

Yu, L., Wang, M., Yang, Y., Xu, F., Zhang, X., Xie, F., et al. (2021). Predicting Therapeutic Drugs for Hepatocellular Carcinoma Based on Tissue-specific Pathways. *Plos Comput. Biol.* 17 (2), e1008696. doi:10.1371/journal.pcbi.1008696

Yu, L., Shi, Q., Wang, S., Zheng, L., and Gao, L. (2020). Exploring Drug Treatment Patterns Based on the Action of Drug and Multilayer Network Model. *Ijms* 21 (14), 5014. doi:10.3390/ijms21145014

Yu, X., Jianguo, Z., Mingming, Z., Chao, Y., Qing, D., Wei, Z., et al. (2020). Exploiting XGBoost for Predicting Enhancer-Promoter Interactions. *Curr. Bioinformatics* 15 (9), 1036–1045. doi:10.2174/1574893615666200120103948

Zeng, X., Wang, W., Chen, C., and Yen, G. G. (2020). A Consensus Community-Based Particle Swarm Optimization for Dynamic Community Detection. *IEEE Trans. Cybern.* 50 (6), 2502–2513. doi:10.1109/tcyb.2019.2938895

Zeng, X., Yinglai, L., Yuying, H., Linyuan, L., Xiaoping, M., and Rodriguez-Paton, A. (2020). Deep Collaborative Filtering for Prediction of Disease Genes. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17 (5), 1639–1647. doi:10.1109/tcbb.2019.2907536

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020). Target Identification Among Known Drugs by Deep Learning from Heterogeneous Networks. *Chem. Sci.* 11 (7), 1775–1797. doi:10.1039/c9sc04336e

Zhang, J., Zehua, Z., Lianrong, P., Jijun, T., and Fei, G. (2020). *AIEpred: An Ensemble Predictive Model of Classifier Chain to Identify Anti-inflammatory Peptides*, 18. IEEE/ACM Trans Comput Biol Bioinform, 1831–1840. doi:10.1109/tcbb.2020.2968419

Zhang, Y., Jianrong, Y., Siyu, C., Meiqin, G., Dongrui, G., Min, Z., et al. (2020). Review of the Applications of Deep Learning in Bioinformatics. *Curr. Bioinformatics* 15 (8), 898–911. doi:10.2174/1574893615999200711165743

Zhang, Y. P., and Zou, Q. (2020). PPTPP: A Novel Therapeutic Peptide Prediction Method Using Physicochemical Property Encoding and Adaptive Feature Representation Learning. *Bioinformatics* 36 (13), 3982–3987. doi:10.1093/bioinformatics/btaa275

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an Ensemble Classifier-Based Feature Selection for Differential Expression Analysis on Expression Profiles. *BMC Bioinformatics* 21 (1), 43. doi:10.1186/s12859-020-3388-y

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

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016). 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., Gang, L., Xingpeng, J., Xiangrong, L., and Xiangxiang, Z. (2020). Sequence Clustering in Bioinformatics: an Empirical Study. *Brief. Bioinform.* 21 (1), 1–10. doi:10.1093/bib/bby090

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 December 2021
doi: 10.3389/fgene.2021.792265

A New Method for Recognizing Protein Complexes Based on Protein Interaction Networks and GO Terms

*Xiaoting Wang, Nan Zhang, Yulan Zhao and Juan Wang**

School of Computer Science, Inner Mongolia University, and with Ecological Big Data Engineering Research Center of the Ministry of Education, Hohhot, China

Motivation: A protein complex is the combination of proteins which interact with each other. Protein–protein interaction (PPI) networks are composed of multiple protein complexes. It is very difficult to recognize protein complexes from PPI data due to the noise of PPI.

Results: We proposed a new method, called Topology and Semantic Similarity Network (TSSN), based on topological structure characteristics and biological characteristics to construct the PPI. Experiments show that the TSSN can filter the noise of PPI data. We proposed a new algorithm, called Neighbor Nodes of Proteins (NNP), for recognizing protein complexes by considering their topology information. Experiments show that the algorithm can identify more protein complexes and more accurately. The recognition of protein complexes is vital in research on evolution analysis.

Availability and implementation: https://github.com/bioinformatical-code/NNP.

Keywords: protein interaction network, protein complex, GO terms, NNP, function of proteins

OPEN ACCESS

Edited by:
Jijun Tang,
University of South Carolina,
United States

Reviewed by:
Zhanchao Li,
Guangdong Pharmaceutical
University, China
Ruofan Xia,
Facebook (United States),
United States

****Correspondence:***
Juan Wang
wangjuan@imu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *10 October 2021*
Accepted: *10 November 2021*
Published: *13 December 2021*

Citation:
Wang X, Zhang N, Zhao Y and Wang J (2021) A New Method for Recognizing Protein Complexes Based on Protein Interaction Networks and GO Terms. Front. Genet. 12:792265. doi: 10.3389/fgene.2021.792265

INTRODUCTION

The recognition for protein complexes based on the PPI network has become one of the most important channels in current research. Detection of protein complexes from PPI networks is an important work in the understanding of biological processes. It is also of great significance for researching mechanisms and developing new drugs. Researchers have put forward a variety of effective methods to recognize protein complexes. The MCODE algorithm chooses a vertex with the maximum weight as the initial cluster, and then recursively searches for the vertices that meet a threshold value to add to the cluster (Bader and Hogue, 2003). The DPClus is a modified algorithm that chooses the vertices with high connectivity with the present cluster iteratively (Altaf-Ul-Amin et al., 2006). Jerarca uses the hierarchical cluster to partition the complexes based on the distance among proteins (Aldecoa and Marín, 2010). RNSC divides the complexes by means of a cost function (King et al., 2004). MCL (Enright et al., 2002) simulates network flow by constructing a similarity matrix, alternately performs expansion and inflation operations, and achieves clustering effect after multiple iterations. But the method is difficult to identify the complexes with little overlap. After that, an improved method was proposed which measured the reliability of PPI based on the annotations of protein function (Cho et al., 2007). SCI-BN and ClusterM combine topology of PPI and biological information of sequences to identify complexes (Qi et al., 2008; Wang et al., 2020).

Although these methods can effectively identify functional modules of proteins, they all ignore the internal structure of the modules. The basic structure of a protein complex is composed of the

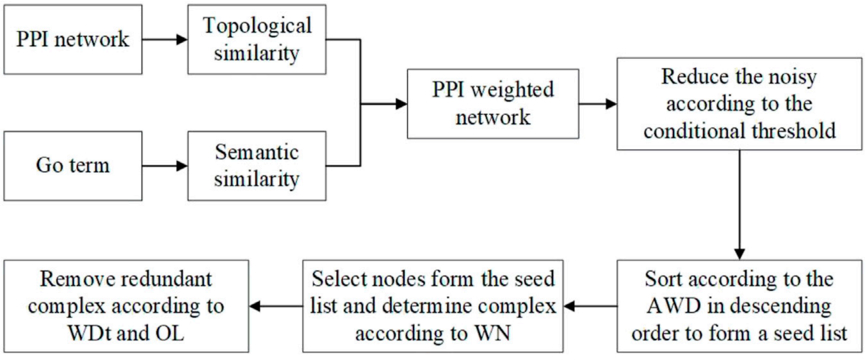
PPI network
Topological similarity
Go term
Semantic similarity
PPI weighted network
Reduce the noisy according to the conditional threshold
Sort according to the AWD in descending order to form a seed list
Select nodes form the seed list and determine complex according to WN
Remove redundant complex according to WDt and OL

FIGURE 1 | Workflow of the NNP.

TABLE 1 | Results of methods are used in the unweighted networks and weighted networks computed by the TSSN.

Metrics Method	*R*	*P*	F1
ClusterOne-u	0.32	0.415	0.361
ClusterOne-T	**0.34**	**0.43**	**0.38**
MCODE-u	0.21	0.49	0.294
MCODE-T	**0.23**	**0.51**	**0.317**
MCL-u	0.58	0.21	0.308
MCL-T	**0.605**	**0.228**	**0.331**

Bold values represents the experimental results on ClusterOne, MCode and MCL weighted by the TSSN method.

nucleus of a protein complex and all its subordinate proteins (Gavin et al., 2006). So, a protein complex can be regarded as a subgraph with a nucleus and its subordinate proteins for assisting the nucleus to play a specific role. COACH (Wu et al., 2009) and CORE (Leung et al., 2009) are proposed based on the idea. The F-MCL algorithm combines firefly algorithm and MCL (Lei et al., 2016). ClusterONE is a clustering algorithm guided by cohesion which can identify subgraphs of dense substructure (Nepusz et al., 2012). However, the cohesion formula may lead to deviation in the clustering process. EA (Halim et al., 2015) uses multi-population evolutionary algorithm to cluster the probability map. MNC is a novel clustering model based on multi networks which combines the shared clustering structure in PPI and domain–domain interaction (DDI) networks in order to improve the accuracy of identification (Ou-Yang et al., 2017). IdenPC-CAP recognizes protein complexes from the interaction networks consisting of RNA–RNA interactions, RNA–protein interactions, and PPIs (Wu et al., 2021). CSC uses both topological and biological characteristics to identify protein complexes (Liu et al., 2018; Sharma et al., 2018). DPCMNE detects protein complexes *via* multilevel network embedding (Meng et al., 2021). PC2P formalizes protein complexes as biclique spanned subgraphs and converts the problem of detecting protein complex to coherent partition (Omranian et al., 2021). A semi-supervised model based on non-negative matrix tri-factorization is also used to detect protein complex (Liu et al., 2021). In the FCAN-PCI, the semantic similarity of proteins and the topology of PPI network are integrated into a fuzzy clustering model (Pan et al., 2021). GECA proposes a model based on the gene expression and core-attachment (Noori et al., 2021). The idenPC-MIIP method modifies the weights of original network by defining mutually important neighbors on the weighted network and then identifies protein complexes using a greedy algorithm (Wu et al., 2021)

METHODS

For a PPI network *N*, TSSN computes the edge aggregation coefficient as the topology characteristics of *N*, makes use of the GO annotation as the biological characteristics of *N*, and then constructs a weighted network. NNP identifies protein complexes based on this weighted network.

TSSN

A PPI network can be seen as an undirected graph *G*= (*V*, *E*), and each protein is a node in *V*. Two proteins interact with each other if and only if there is an edge between the two nodes representing two proteins. In order to describe the structural similarity among proteins in the PPI network, Jaccard coefficient between two nodes *u* and *v* in *G*= (*V*, *E*) is defined as follows:

$$J(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}, \tag{1}$$

where *N*(*u*) [or *N*(*v*)] represents the set of all neighbor nodes of protein *u* (or *v*) in the network.

We adopted the simGIC method (Tian and Guo, 2017), which is an improved method from the GIC (Pesquita et al., 2007) to calculate semantic similarity between proteins. Assuming that proteins *u* and *v* are annotated by term sets $A_= \{T_1, T_2, \cdots, T_m\}$ and $B_= \{S_1, S_2, \cdots, S_n\}$ respectively, the semantic similarity between *u* and *v* is defined as follows:

$$se(u,v) = \frac{\sum_{T_i \in A \cap B} - \log p(T_i)}{\max\{IC(A), IC(B)\}}, \tag{2}$$

TABLE 2 | F1 values of NNP on different thresholds of WNT.

t	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
F1	0.4	0.41	**0.42**	0.41	0.4	0.39	0.395	0.37	0.3	0.2	0.13

Bold values shows that when the threshold t is 0.2, the value of F1 reaches a maximum of 0.42.

TABLE 3 | Precision values of NNP on different thresholds of WNT.

t	0.2	0.21	0.22	0.23	0.24	0.25
Precision	0.491	0.492	**0.5**	0.495	0.493	0.493

Bold values shows that when the threshold t is 0.5, the precision value reaches the maximum 0.5.

TABLE 4 | Each algorithm identifies the cluster information.

No.	Algorithm	Number	Average	Coverage
1	CYC2008	408	4.71	1,628
2	CFinder	178	11.31	2,147
3	ClusterONE	413	5	1898
4	MCODE	110	6.5	1,299
5	NNP	538	4.54	1937
6	MCL	623	6.57	4096
7	EA	398	13.5	2,661
8	PC2P	434	4.50	1953

Where $IC(A)$ is the set of $\{-\log(T_1), -\log(T_2),..., -\log(T_m)\}$, and $p(T_i)$ represents the times that GO terms or single function of protein appear in the specified term data.

Here, the similarity between two proteins u and v is defined as the average between their topological similarity and semantic similarity, that is,

$$s(u,v) = \frac{\sum\limits_{u_1 \in N(u), v_1 \in N(v)} \left(J(u_1,v_1) + se(u_1,v_1)\right)}{2}, \tag{3}$$

where the value of $s(u,v)$ is [0,1].

NNP

Given a weighted network $G = (V, E, W)$, where $V = \{v_1, v_2, \cdots, v_m\}$, $E = \{e_1, e_2, \cdots, e_n\}$, $W = \{w(e_1), w(e_2), \cdots, w(e_n)\}$, and $w(e_i)$ represents the weight of the edge e_i. The distance between the nodes v_i and v_j is the minimum among all lengths of paths. V_j is denoted as the set of nodes with the distance 2 between v_j, which is referred to as the set of second-order neighbor nodes between *vj*. The network $G_j = (V_j, E_j, W_j)$ is derived by V_j. The weighed degree of v_j in G is defined as follows:

$$WD\left(v_j, G\right) = \sum_{i=1}^{n} w\left(v_j, v_i\right), \tag{4}$$

where $(v_j, v_i) \in E$ and $w(v_j, v_i)$ indicates the weight of the edge between node *j* and node *i*. The average weighted degree of v_j in G is computed by the following equation:

$$AWD\left(v_j, G\right) = \sum_{i=1}^{n} w\left(v_j, v_i\right) / \left|V\right|. \tag{5}$$

The weighted neighbor ratio is defined as follows:

$$WN\left(v_j, G\right) = \frac{WD\left(v_j, G\right)}{WD\left(v_j, G\right) + WD\left(v_j, G_j\right)}. \tag{6}$$

In order to assess complexes, we compute the tightness degree of a complex $G = (V, E, W)$ as follows:

$$WDt(G) = 2\sum_{i=1}^{n} w(e_i) / (|V| \times (|V| - 1)). \tag{7}$$

For two complexes C1 and C2, the overlap ratio (OL) between them is defined as follows:

$$OL(C_1, C_2) = \frac{|C_1 \cap C_2|^2}{|C_1| \cdot |C_2|}. \tag{8}$$

NNP identifies complexes by four main steps. First, the NNP uses the TSSN method to compute the similarity among proteins, and then builds a PPI weighted network and neighbor networks. Second, it calculates a conditional threshold in order to reduce the noise, and then the network is transformed into a matrix, which is arranged in descending order according to the average weighted degree (AWD) of nodes to form a seed list. Third, it selects nodes from the seed list iteratively as the initial complex to cluster, and then removes or retains the node according to the weighted neighbor ratio (WN) until all nodes list are solved. Finally, it calculates the OL among protein complexes and judges whether the complexes are retained or discarded through the network tightness (WDt). Finally, the complex set was obtained. **Figure 1** shows the workflow of NNP. The pseudo code can be seen in the Algorithm.

RESULTS AND DISCUSSION

In order to assess the TSSN method, we compare the protein complexes identified by three classical methods, that is, ClusterONE, MCODE, and MCL, respectively, based on the PPI networks with the weight computed by TSSN and the PPI networks without weight. We compare the results of protein complexes predicted by CFinder, ClusterONE, MCODE, MCL, EA, and NNP methods.

Datasets

In all experiments, we use the PPI data of yeast downloaded from the DIP database (https://dip.doe-mbi.ucla.edu/dip/Download.cgi?SM=7&TX=4932), version 20170205. In order to reduce the noise of data, we delete the repeated interactions and the

TABLE 5 | Three complexes identified by methods were analyzed from the DIP.

Algorithm Protein complex	CFinder (%)	Cluster -ONE	MCODE (%)	NNP (%)	MCL (%)	EA (%)	PC2P (%)
CFI	100	100%	100	100	100	100	83.3
NEC	83.3	64.1%	91.7	100	100	91.7	83.3
DRC	56.3	100%	61.4	91.7	67.5	83.3	53.3

TABLE 6 | Results of protein complexes recognized by algorithms.

Metrics method	*R*	*P*	F1
CFinder	0.3408	0.2698	0.3012
ClusterONE	0.4068	0.3554	0.3794
MCODE	0.2293	0.501	0.3146
NNP	**0.3515**	**0.5107**	**0.4164**
MCL	0.3326	0.4093	0.367
EA	0.34	0.383	0.3602
PC2P	0.4340	0.1935	0.2677

Bold values show that the experimental results of the NNP method are optimal.

TABLE 7 | Numbers of protein complexes perfectly matched by each algorithm for DIP data set.

Algorithm	Perfect matching
CFinder	11
ClusterONE	10
MCODE	6
NNP	**17**
MCL	15
EA	14
PC2P	0

Bold values show that the experimental results of the NNP method are optimal.

TABLE 8 | Protein complexes with lower *p*-value identified by the algorithm on the DIP.

GO term	OL (%)	*p*-value
mRNA processing	96	1.54E-36
Small nuclear ribonucleo protein complex	86.1	2.73E-58
mRNA splicing, *via* spliceosome	95.7	4.48E-38
Transferase activity, transferring glycosyl groups	89.59	1.81E-76
Ribosomal small subunit biogenesis	88.2	2.45E-48
Transporter activity	94.38	6.84E-100

circle of a node to itself. Then the PPI network contains 5,115 nodes and 22,552 edges. GO annotations and ontology data of yeast are downloaded from the website (http://www.geneontology.org/).

Reference Sets

Here, two standard sets, namely, CYC2008 (Pu et al., 2009) and NewMIPS (Friedel et al., 2008), are used in the experiments, where CYC2008 is downloaded from (http://wodaklab.org/cyc2008/downloads). These data are predicted by biological methods, including 408 complexes and 1,628 proteins. The NewMIPS is a set of protein complexes, including 428 complexes and 1,171 proteins.

Metrics

For a prediction algorithm, its effectiveness is measured by four indexes: recall, precision, F1, and overlap ratio. The recall value *R* is the ratio of the number of complexes which are identified by methods and matched with the complexes in the standard set to the number of complexes in the standard set; the precision value *P* is the ratio of the number of complexes which are identified by methods and matched with the complexes in the standard set to the number of all complexes identified by the algorithm. F1 is the harmonic average of *P* and *R*, that is,

$$F1 = \frac{2 \times R \times P}{R + P}. \tag{9}$$

To judge the biological significance of complexes, a functional enrichment analysis is used to analyze the gene annotation information in the GO database, that is, *p*-value. The calculation method is given as follows:

$$p - value = 1 - \sum_{i=0}^{m-1} \frac{\binom{|F|}{i}\binom{|V| - |F|}{|C| - i}}{\binom{|V|}{|C|}}, \tag{10}$$

where *m* is the number of identified complexes that are the same as those in the standard data set, *F* the complexes in the standard data set, *V* the number of proteins contained in the PPI network, and *C* the number of identified complexes. Here, if *p-value* is less than 0.01, the complex is regarded with biological significance.

RESULTS

In all recorded experimental results, we use CYC2008 as the standard set and set the threshold of OL as 0.2. OL represents the overlap rate between the two complexes. The value of OL being 0.2 indicates that the identified complex is considered correct when the OL with the standard complex reaches 0.2.

Table 1 shows the results. For each method in **Table 1**, u represents the methods that are used to identify the complexes from the unweighted networks and T represents the methods that are used to identify the complexes from the weighted networks computed by the TSSN. From **Table 1**, we can see that the precision values for the weighted networks

TABLE 9 | Algorithm perfectly matches the protein complex on the DIP.

GO term	OL (%)	*p*-value
mRNA metabolic process	100	7.37E-27
Anaphase-promoting complex–dependent catabolic process	100	4.68E-24
Polyadenylation-dependent snoRNA 3′-end processing	100	1.45E-32

Algorithm | detecting protein complexes.

```
input: an unweighted PPI network $G$ ($V$, $E$) and the annotations of proteins
output: all protein complexes
$C = \varnothing$;
calculate the similarity between the two nodes of each edge and obtain a
weighted PPI network $G$ ($V$, $E$, $W$) by formula (3);
for each node $v \in V$ do
    obtain the first-order neighbor graph $G'$ ($V'$, $E'$, $W'$) of $v$;
    compute $AWD$ ($v$, $G'$) by formula (5);
    if $AWD$ ($v$, $G'$) = 0 then
        delete $v$ from $V$;
    end if
end for
arrange nodes in $V$ by descending $AWD$ values to form the seed set $S$;
for $s \in S$ do
    add the first-order neighbor graph $G'$ ($V'$, $E'$, $W'$) of $s$ as a complex $C_0$ to
    $C$;
    for $v \in V'$
        if $WN$ ($v$, $G'$) < $WNT$ then
        $v$ is marked as disposed and removed from $C_0$;
        end if
    end for
end for
for every disposed node $v$ do
    obtain the first-order neighbor graph $G'$ ($V'$, $E'$, $W'$) of $v$;
    for each complex $C_0$ in $C$ do
        if $AWD$ ($v$, $C_0$) > $AWD$ ($v$, $G'$) then
            add $v$ to $C_0$;
        end if
    end for
end for
for every two complexes $C_1$ and $C_2$ in $C$ do
    if $OL$ ($C_1$, $C_2$) $\geq$ 0.2 then
        if $WDt$ ($C_1$) < $WDt$ ($C_2$) then
            delete $C_1$ from $C$;
        end if
    end if
end for
return $C$;
```

computed by the TSSN method are higher than those for the unweighted networks. So the TSSN method is efficient for computing the weigh values of networks.

The precision results of the NNP algorithm depend on the thresholds of weighted neighbor ratio (WNT). **Table 2** shows that F1 values gradually increase with the increase in *t* values if the thresholds of WNT is (0,0.2), and F1 gradually decreases as a whole if the t values of WNT continue to increase from 0.2. So F1 can reach the maximum 0.42 if values of WNT are (0.2, 0.25). **Table 3** shows the precision values of NNP on different thresholds of WNT. When the WNT value is 0.22, the precision is 0.5, which is slightly higher than the other five values. Therefore, it is reasonable for the NNP algorithm to set the threshold of the WNT as 0.22.

Table 4 lists the comparison of the cluster information identified by the six algorithms compared with CYC2008. CYC2008 is selected as the benchmark, and its average size is 4.71; the closer the average size of the cluster identified by a method is to 4.71, the more accurate the method is. Among the six algorithms, the average size of clusters identified by the NNP is 4.54, which is closest to the size of clusters in the standard data. So the recognition result of NNP has high theoretical reliability.

Table 5 shows the results identified by the CFinder, ClusterONE, MCODE, MCL, EA, NNP, and PC2P methods for three complexes randomly selected from DIP. CFI is the mRNA cleavage factor complex with size 5; NEC is the nuclear exosome complex with size 12, and DRC is the DNA-directed RNA polymerase II complex. The table shows that six methods recognize the same proteins as the CYC2008 for the CFI, that is, OL 100%, OL of NNP, and MCL is both 100% for NEC. The OL of PC2P is 83.3%. The OL of EA and that of MCODE are the same, which is 91.7%, ranking second. There is one missed protein: YHR081W. CFinder has two missed proteins and the OL is 84%. The OL of PC2P is 83.3%. So, the accuracy of ClusterONE is low. For DRC, the performance of NNP and ClusterONE is better, while the OL value of EA is 83.3%. There are many omissive and wrong proteins detected by CFinder, MCODE, MCL, and PC2P. The OL of CFinder is 56.3%. The OL of PC2P is only 53.3%.

Table 6 shows the results of six methods. In terms of precision, the value of CFinder is lowest, which is only 26.98%, and the value of NNP is largest compared with other algorithms, reaching 51.07%. The precision of MCODE lists second, reaching 50.1%. Although the precision of MCODE is high, the recall is low, which leads to the low F1 value. From the table, it is obvious that the F1 of NNP is max among all other methods. So NNP has better accuracy in identifying protein complexes than other methods.

Table 7 lists the number of protein complexes identified by CFinder, ClusterONE, MCODE, MCL, EA, NNP, and PC2P from DIP data set, matched with CYC2008. As shown in **Table 7**, the protein complexes identified by NNP based on the DIP data set are perfectly matched with 17 protein complexes. The MCODE only has six complexes perfectly matched with the standard set. The PC2P has no perfectly matched complex with the standard set. Therefore, compared with other algorithms, the NNP algorithm can accurately and perfectly match more protein complexes on the DIP data set.

Table 8 lists some protein complexes with low *p-values* identified by the NNP algorithm on the DIP, which can show that the protein complexes identified by the NNP algorithm have significant biological significance. **Table 9** lists three protein complexes perfectly matched with DIP and NewMIPS identified by the NNP method.

CONCLUSION

Considering the topological structure of the PPI network, it introduces the gene ontology in biological information. We propose the methods for computing weight of protein interaction network and the recognizing of protein complexes on the weighted network. By comparing with other algorithms, the TSSN method based on topological features and GO term similarity can filter the noise, which can reduce the impact of noise data. The NNP algorithm can identify the protein complexes. The experimental results show that the NNP is superior to other classical algorithms.

In the future, we will adopt new technologies to detect false-positive edges and predict false-negative edges in the PPI network, thus improving the quality of the PPI network. Machine learning methods will be used to detect protein complexes based on their biological characteristics. Finally, since static PPI networks only contain the interaction between proteins and cannot reflect the dynamic characteristics of proteins interactions over time, we will study how to build a dynamic PPI network and identify protein complexes in the dynamic network.

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

XW, NZ, and JW proposed and designed the method. XW and NZ performed the experiments. All authors wrote the manuscript.

FUNDING

This work has been supported by the National Natural Science Foundations of China (62002181, 62061035) and the Self-topic/Open Project of Ecological Big Data Engineering Research Center of the Ministry of Education.

REFERENCES

Aldecoa, R., and Marín, I. (2010). Jerarca: Efficient Analysis of Complex Networks Using Hierarchical Clustering. *PLoS ONE* 5 (7), e11585. doi:10.1371/journal.pone.0011585

Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., and Kanaya, S. (2006). Development and Implementation of an Algorithm for Detection of Protein Complexes in Large Interaction Networks. *BMC bioinformatics* 7 (1), 1–13. doi:10.1186/1471-2105-7-207

Bader, G. D., and Hogue, C. W. (2003). An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC bioinformatics* 4 (1), 2–27. doi:10.1186/1471-2105-4-2

Cho, Y.-R., Hwang, W., Ramanathan, M., and Zhang, A. (2007). Semantic Integration to Identify Overlapping Functional Modules in Protein Interaction Networks. *BMC bioinformatics* 8 (1), 1–13. doi:10.1186/1471-2105-8-265

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An Efficient Algorithm for Large-Scale Detection of Protein Families. *Nucleic Acids Res.* 30 (7), 1575–1584. doi:10.1093/nar/30.7.1575

Friedel, C. C., Krumsiek, J., and Zimmer, R. (2009). "Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast," in *Annual International Conference on Research in Computational Molecular Biology*, 16, 971–987. doi:10.1089/cmb.2009.0023*J. Comput. Biol.*

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., et al. (2006). Proteome Survey Reveals Modularity of the Yeast Cell Machinery. *Nature* 440 (7084), 631–636. doi:10.1038/nature04532

Halim, Z., Waqas, M., and Hussain, S. F. (2015). Clustering Large Probabilistic Graphs Using Multi-Population Evolutionary Algorithm. *Inf. Sci.* 317, 78–95. doi:10.1016/j.ins.2015.04.043

King, A. D., Przulj, N., and Jurisica, I. (2004). Protein Complex Prediction via Cost-Based Clustering. *Bioinformatics* 20 (17), 3013–3020. doi:10.1093/bioinformatics/bth351

Lei, X., Wang, F., Wu, F.-X., Zhang, A., and Pedrycz, W. (2016). Protein Complex Identification through Markov Clustering with Firefly Algorithm on Dynamic Protein-Protein Interaction Networks. *Inf. Sci.* 329, 303–316. doi:10.1016/j.ins.2015.09.028

Leung, H. C. M., Xiang, Q., Yiu, S. M., and Chin, F. Y. L. (2009). Predicting Protein Complexes from Ppi Data: a Core-Attachment Approach. *J. Comput. Biol.* 16 (2), 133–144. doi:10.1089/cmb.2008.01TT

Liu, G., Liu, B., Li, A., Wang, X., Yu, J., and Zhou, X. (2021). Identifying Protein Complexes with Clear Module Structure Using Pairwise Constraints in Protein Interaction Networks. *Front. Genet.* 12, 786. doi:10.3389/fgene.2021.664786

Liu, W., Ma, L., Jeon, B., Chen, L., and Chen, B. (2018). A Network Hierarchy-Based Method for Functional Module Detection in Protein-Protein Interaction Networks. *J. Theor. Biol.* 455, 26–38. doi:10.1016/j.jtbi.2018.06.026

Meng, X., Xiang, J., Zheng, R., Wu, F., and Li, M. (2021). DPCMNE: Detecting Protein Complexes from Protein-Protein Interaction Networks via Multi-Level Network Embedding. *Ieee/acm Trans. Comput. Biol. Bioinf.*, 1. doi:10.1109/TCBB.2021.3050102

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting Overlapping Protein Complexes in Protein-Protein Interaction Networks. *Nat. Methods* 9 (5), 471–472. doi:10.1038/nmeth.1938

Noori, S., Al-A'Araji, N., and Al-Shamery, E. (2021). Identifying Protein Complexes from Protein-Protein Interaction Networks Based on the Gene Expression Profile and Core-Attachment Approach. *J. Bioinform. Comput. Biol.* 19 (3), 2150009. doi:10.1142/S0219720021500098

Omranian, S., Angeleska, A., and Nikoloski, Z. (2021). PC2P: Parameter-free Network-Based Prediction of Protein Complexes. *Bioinformatics* 37 (1), 73–81. doi:10.1093/bioinformatics/btaa1089

Ou-Yang, L., Yan, H., and Zhang, X.-F. (2017). A Multi-Network Clustering Method for Detecting Protein Complexes from Multiple Heterogeneous Networks. *BMC bioinformatics* 18 (13), 23–34. doi:10.1186/s12859-017-1877-4

Pan, X., Hu, L., Hu, P., and You, Z.-H. (2021). Identifying Protein Complexes from Protein-Protein Interaction Networks Based on Fuzzy Clustering and GO Semantic Information. *Ieee/acm Trans. Comput. Biol. Bioinf.* 14 (8), 1. doi:10.1109/TCBB.2021.3095947

Pesquita, C., Faria, D., Bastos, H., Falcao, A., and Couto, F. (2007). July)Evaluating Go-Based Semantic Similarity Measures. *Proc. 10th Annu. Bio-Ontologies Meet.* 37 (40), 38.

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009). Up-to-date Catalogues of Yeast Protein Complexes. *Nucleic Acids Res.* 37 (3), 825–831. doi:10.1093/nar/gkn1005

Qi, Y., Balem, F., Faloutsos, C., Klein-Seetharaman, J., and Bar-Joseph, Z. (2008). Protein Complex Identification by Supervised Graph Local Clustering. *Bioinformatics* 24 (13), i250–i268. doi:10.1093/bioinformatics/btn164

Sharma, P., Bhattacharyya, D. K., and Kalita, J. K. (2018). Detecting Protein Complexes Based on a Combination of Topological and Biological Properties in Protein-Protein Interaction Network. *J. Genet. Eng. Biotechnol.* 16 (1), 217–226. doi:10.1016/j.jgeb.2017.11.005

Tian, Z., and Guo, M. Z. (2017). An Improved Method for Measuring the Functional Similarity of Genes. *Intell. Comp. Appl.* 7 (5), 123–126. doi:10.3969/j.issn.2095-2163.2017.05.034

Wang, Y., Jeong, H., Yoon, B.-J., and Qian, X. (2020). ClusterM: a Scalable Algorithm for Computational Prediction of Conserved Protein Complexes across Multiple Protein Interaction Networks. *BMC genomics* 21 (10), 1–14. doi:10.1186/s12864-020-07010-1

Wu, M., Li, X., Kwoh, C.-K., and Ng, S.-K. (2009). A Core-Attachment Based Method to Detect Protein Complexes in Ppi Networks. *BMC bioinformatics* 10 (1), 1–16. doi:10.1186/1471-2105-10-169

Wu, Z., Liao, Q., Fan, S., and Liu, B. (2021). idenPC-CAP: Identify Protein Complexes from Weighted RNA-Protein Heterogeneous Interaction Networks Using Co-assemble Partner Relation. *Brief. Bioinform.* 22 (4), bbaa372. doi:10.1093/bib/bbaa372

Wu, Z., Liao, Q., and Liu, B. (2021). idenPC-MIIP: Identify Protein Complexes from Weighted PPI Networks Using Mutual Important Interacting Partner Relation. *Brief. Bioinformatics* 22 (2), 1972–1983. doi:10.1093/bib/bbaa016

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 11 January 2022
doi: 10.3389/fgene.2021.801728

Autoregressive Modeling and Prediction of the Activity of Antihypertensive Peptides

Xufen Xie[1], *Chuanchuan Zhu*[1], *Di Wu*[2,3] *and Ming Du*[2,3]*

[1]*School of Information Science and Engineering, Dalian Polytechnic University, Dalian, China,* [2]*School of Food Science and Technology, Dalian Polytechnic University, Dalian, China,* [3]*National Engineering Technology Research Center of Seafood, Dalian Polytechnic University, Dalian, China*

Naturally derived bioactive peptides with antihypertensive activities serve as promising alternatives to pharmaceutical drugs. There are few relevant reports on the mapping relationship between the EC_{50} value of antihypertensive peptide activity (AHTPA-EC_{50}) and its corresponding amino acid sequence (AAS) at present. In this paper, we have constructed two group series based on sorting natural logarithm of AHTPA-EC_{50} or sorting its corresponding AAS encoding number. One group possesses two series, and we find that there must be a random number series in any group series. The random number series manifests fractal characteristics, and the constructed series of sorting natural logarithm of AHTPA-EC_{50} shows good autocorrelation characteristics. Therefore, two non-linear autoregressive models with exogenous input (NARXs) were established to describe the two series. A prediction method is further designed for AHTPA-EC_{50} prediction based on the proposed model. Two dynamic neural networks for NARXs (NARXNNs) are designed to verify the two series characteristics. Dipeptides and tripeptides are used to verify the proposed prediction method. The results show that the mean square error (MSE) of prediction is about 0.5589 for AHTPA-EC_{50} prediction when the classification of AAS is correct. The proposed method provides a solution for AHTPA-EC_{50} prediction.

Keywords: antihypertensive peptides, NARXNN, fractal characteristics, EC_{50} prediction, machine learning

OPEN ACCESS

Edited by:
Juan Wang,
Inner Mongolia University, China

Reviewed by:
Liang Cheng,
Harbin Medical University, China
Yanjuan Li,
Quzhou University, China

Correspondence:
Ming Du
duming@dlpu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *25 October 2021*
Accepted: *09 December 2021*
Published: *11 January 2022*

Citation:
Xie X, Zhu C, Wu D and Du M (2022) Autoregressive Modeling and Prediction of the Activity of Antihypertensive Peptides. Front. Genet. 12:801728. doi: 10.3389/fgene.2021.801728

1 INTRODUCTION

Hypertension is a clinical syndrome characterized by increased systemic arterial blood pressure, which can be accompanied by functional or organic damage of the heart, brain, kidney, and other organs. The renin–angiotensin system (RAS) controls blood pressure by regulating the volume of blood in blood vessels. The angiotensin-converting enzyme (ACE) is the core component of the RAS. The ACE can convert inactive angiotensin I into angiotensin II with vasoconstriction, which indirectly increases blood pressure (Zhang et al., 2000). Therefore, ACE inhibitors are widely used as drugs for the treatment of cardiovascular diseases (Stone, 2018). Antihypertensive active peptide is an effective ACE inhibitor (Tu et al, 2018a; Tu et al, 2018b; Wu et al, 2019), which has attracted great attention in the treatment and prevention of hypertension. The EC_{50} value (sample concentration when the ACE inhibition rate is 50%) describes the activity of antihypertensive peptide, which is the most important index to select antihypertensive active peptide. Some research studies focus on feature representation (Tong,

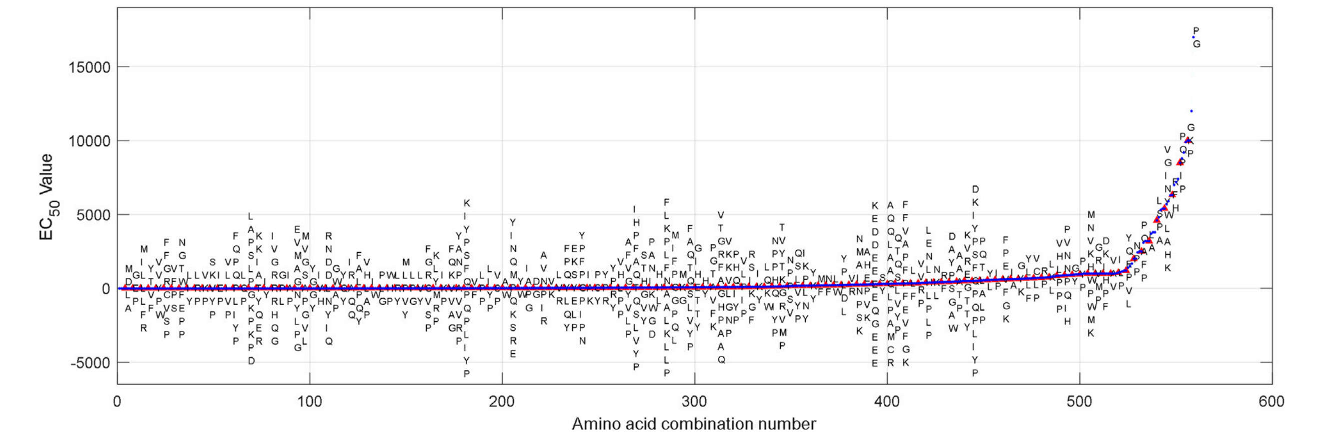

FIGURE 1 | Constructed time series of natural logarithm of AHTPA-EC$_{50}$ and its corresponding amino acid combination.

et al, 2008; Manavalan et al, 2019), and some research studies focus on identification (Majumder, and Wu, 2010). Machine learning (ML) approaches are becoming more and more popular in bioinformatics (Baldi et al., 2001; Libbrecht and Noble, 2015; Zou and Qiliu, 2019; Yang et al., 2020; Zhang et al., 2021). Some research studies are associated with classification, and some are associated with regression. In 2015, Kumar et al. developed four different model types for predicting AHTPs with varied lengths using ML approaches (Kumar et al., 2015a; Kumar et al., 2015b). Another paper on AHTP prediction used random forest (RF) approaches (Win et al., 2018). However, there is great uncertainty in the relationship between the AAS of antihypertensive peptides and its corresponding AHTPA-EC$_{50}$. So far, the mapping relationship between AHTPA-EC$_{50}$ and its corresponding AAS has not been reported. The existing published data show that AHTPA-EC$_{50}$ has multi-scale characteristics. It is difficult to establish a deterministic model between the AAS and AHTPA-EC$_{50}$ directly.

Fractal phenomena generally exist in nature. Fractal data have the characteristics of instability, self-similarity, and multi-scale (Ruderman, 1996; Ghosh and Somvanshi, 2008; Al-Hamdan, et al, 2010; Al-Hamdan et al, 2012). The spectrum of fractal data is consistent (Pentland, 1984; Nill and Bouzas, 1992; Wornell and Oppenheim, 1992). These characteristics can be used to describe physical phenomena with statistical fractal. Fractional Brownian motion (FBM) (Chow, 2011; Kim and Kim, 2004; Fouché and Mukeru, 2013) is more universal than ordinary Brownian motion, and it can better describe the fractal phenomena in nature. FBM can be modeled and described by the time series of dynamic system, and time-series analysis is an important method of system identification and analysis. Yule first proposed the autoregressive (AR) model to predict the law of market change in 1927. In the 1960s, time-series analysis made a great progress in spectral analysis and estimation. The research of linear time-series model has been greatly developed from the AR model to autoregressive moving average (ARMA) modeling theory. Engle and Granger developed estimation procedures, tests, and empirical examples for the relationship between co-integration and error correction models (Engle and Granger 1987), and Hannan and Deistler proposed the multivariable VARMA model and VARMAX model (Hannan and Deistler, 1988). However, Moran proposed the limitations of linear model in the 1950s (Moran, 1953). The non-linear time-series model follows to become an attracting research topic until the late 1970s and early 1980s. These research studies include the threshold autoregressive model, exponential autoregressive model, bilinear model, non-linear autoregressive model, and state-dependent model. Tong et al. gave the threshold autoregressive model (Tong, 1983), and Ozaki proposed an exponential autoregressive model (Ozaki, 1980). The system identification is generally based on the complete clarity of input–output causality. In practical application, the system output can be measured, but the input of some specific systems is difficult to observe and measure. In that situation, it is not easy to determine the causal relationship between input and output. In that case, the traditional system identification method is difficult to apply. Although the system's input cannot always be determined, it is certain that there is a relationship between some known parameters or data and the system output. These known parameters or data can directly or indirectly affect the system output. If the relevant data are also regarded as the system input, then the time-series model with exogenous input is determined. Tong analyzed the non-linear time series with exogenous input, established the relationship between non-linear time series and non-linear dynamic system (chaos), and studied the prediction based on non-linear time series (Tong, 1990).

In this paper, a kind of time series construction method on AHTPA-EC$_{50}$ and its corresponding AAS is proposed firstly. We can find a lot of fractal characteristics from the two group time series. Then, the two groups of constructed series are modeled as two different NARX time-series models. Furthermore, two NARXNNs are used to perform the

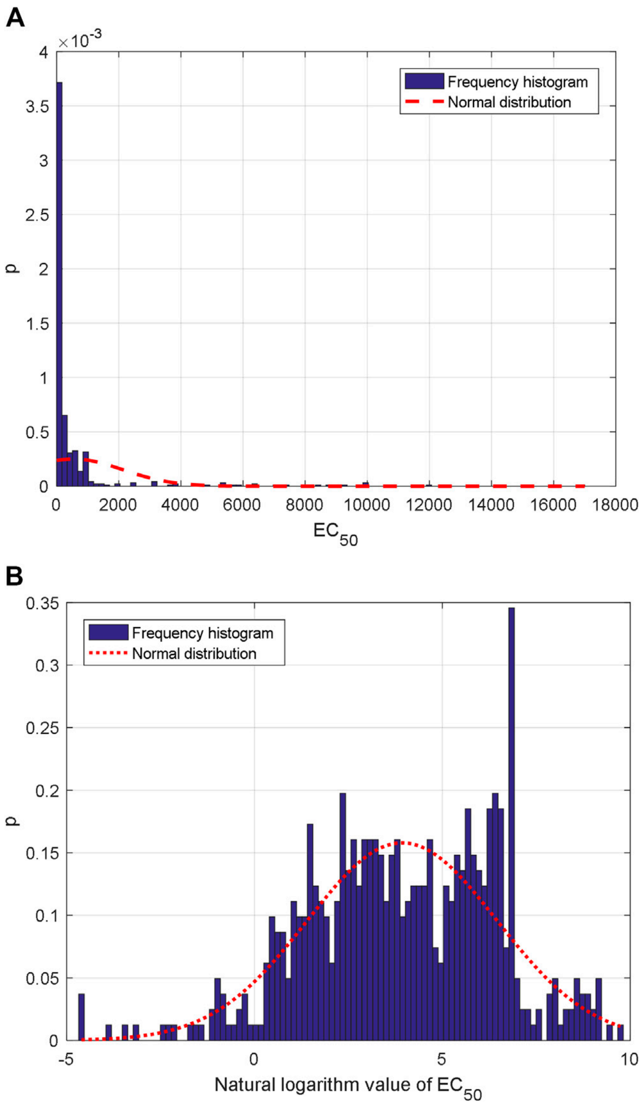

FIGURE 2 | Statistical histogram: frequency histogram of **(A)** AHTPA-EC_{50} and **(B)** natural logarithm of AHTPA-EC_{50}.

proposed model. And then we further proposed a prediction method for AHTPA-EC_{50} based on two NARXNNs and ML classification algorithms. The model and prediction method are useful and meaningful on antihypertensive active peptide research, drug design, and industrial production.

2 MATERIALS AND METHODS

2.1 Analysis of AHTPA-EC_{50} and Its Corresponding AAS

2.1.1 Statistical Analysis of AHTPA-EC_{50}

559 group AHTPA-EC_{50} data and their corresponding AAS are shown in **Figure 1**. Due to the difficulty of display, **Supplementary Material** marks the corresponding AAS every four EC_{50} values (interval = 3). The statistical histogram is analyzed, and histogram analysis of AHTPA-EC_{50} is shown in **Figure 2A**. We can see that AHTPA-EC_{50} is concentrated on the right side of the longitudinal axis of the coordinate and there is some very large AHTPA-EC_{50} value in these data. The characteristics of large distribution span and asymmetry appear in AHTPA-EC_{50} data. Comparing with the normal distribution data with the same mean and variance, it can be seen that AHTPA-EC_{50} data deviate very far from the normal distribution. In order to reduce the scale of AHTPA-EC_{50}, the natural logarithm of AHTPA-EC_{50} data is calculated. The distribution of natural logarithm of AHTPA-EC_{50} is further analyzed, and the histogram distribution is shown in **Figure 2B**. Compared with the normal distribution of the same mean and variance, the natural logarithm histogram of AHTPA-EC_{50} cut off more slowly in the tail, and it shows the characteristics of a long tail. This is an important feature of fractal data.

TABLE 1 | Numerical definitions of amino acids.

Amino acids	A	C	D	E	F	G	H	I	K	L
Numerical definitions	1	2	3	4	5	6	7	8	9	10
Amino acids	M	N	P	Q	R	S	T	V	W	Y
Numerical definitions	11	12	13	14	15	16	17	18	19	20

2.1.2 Encoding for AAS

The expression of amino acid is different from the digital number, and it is a symbolic quantity that cannot be directly quantified. In order to analyze the relationship between the AAS and its corresponding AHTPA-EC_{50}, it is necessary to encode for the AAS. The numerical definitions of different amino acids are shown in **Table 1**. The AAS is digitally encoded in a 21 base system. Because the number 0 cannot appear in the first place of the combined code, the number 0 is not defined here.

2.1.3 Constructed Time Series and Its Time–Frequency Characteristics

(1) Constructed time series based on sorting code of AAS

As mentioned above, the AAS can be converted to decimal digit by numerical definitions of amino acids. After sorting the natural logarithm of coding numbers from small to large, the natural logarithm of AHTPA-EC_{50} can be constructed. The constructed time series is shown in **Figure 3A**. Multi-scale wavelet transform is performed to the constructed AHTPA-EC_{50} time series, and the time–frequency distribution is shown in **Figure 3B**. There is also no obvious law between high-energy data and series number and frequency in **Figure 3B**, and different time–frequency relationships show similar patterns.

(2) Constructed time series based on sorting AHTPA-EC_{50}

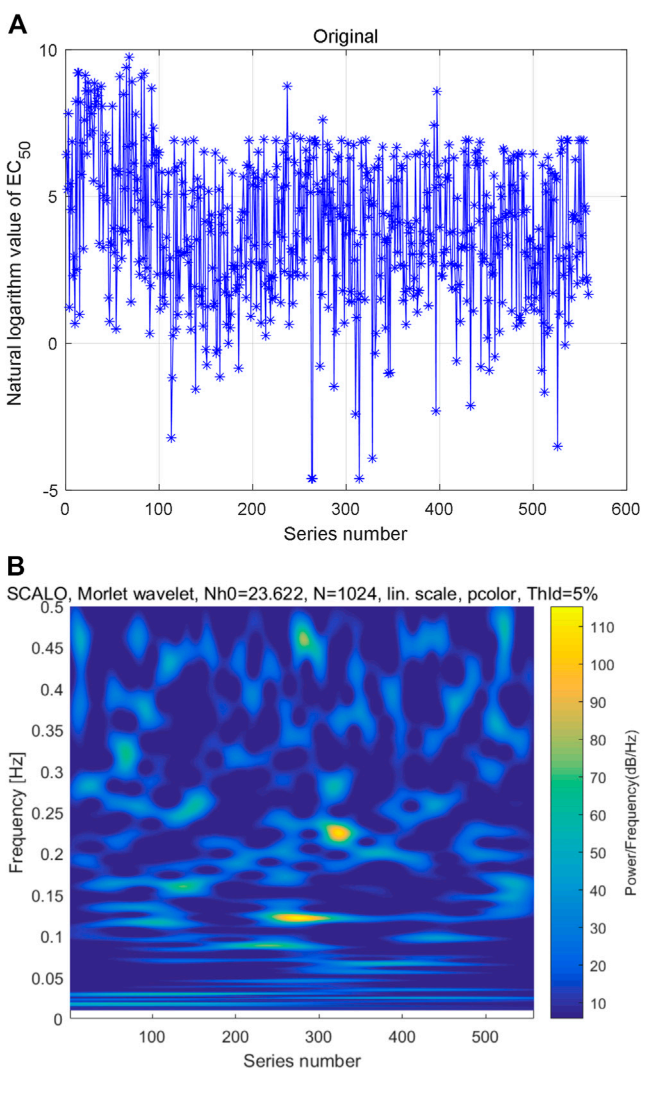

FIGURE 3 | Constructed first time series and its multi-scale wavelet transform: **(A)** time series of natural logarithm of AHTPA-EC_{50} and **(B)** time–frequency distribution of multi-scale wavelet transform.

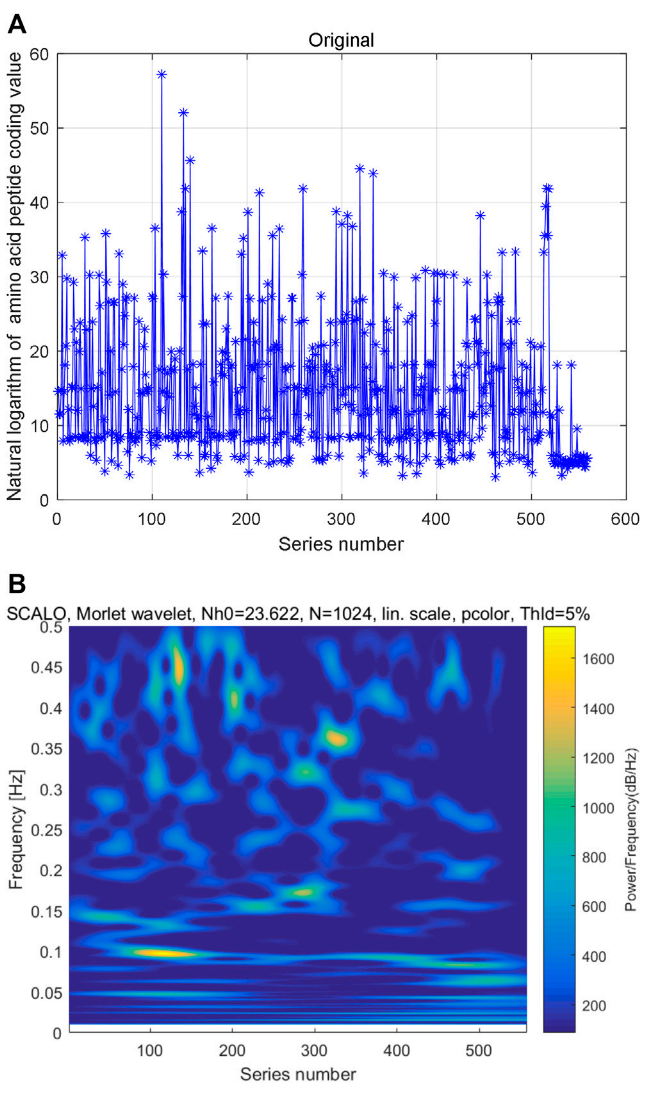

FIGURE 4 | Constructed second time series and its multi-scale wavelet transform: **(A)** time series of natural logarithm of coding AAS and **(B)** time–frequency distribution of multi-scale wavelet transform.

We also constructed natural logarithm of AHTPA-EC_{50} time series by sorting the data from small to large. The AAS is converted to decimal digit by numerical definitions of amino acids. After sorting the natural logarithm of AHTPA-EC_{50} from small to large, the time series of natural logarithm of coding value of AAS is also constructed. The constructed time series is shown in **Figure 4A**. Multi-scale wavelet transform is performed to the natural logarithm of coding value of AAS. The constructed time series of AAS and its time–frequency distribution are shown in **Figure 4B**. And there is no obvious law between high-energy data and series number and frequency. However, different time–frequency relationships show similar patterns.

In summary, the relationship between the natural logarithm of AHTPA-EC_{50} and its corresponding natural logarithm of coding AAS is special. If one of the series is sorted, the other will be a random number series. We deduce that there is not a direct regression modeling for their relationship.

The Haar wavelet is further used to decompose the reconstructed time series to analyze fractal characteristics (data in **Figure 3A**) in multiple scales. The low-frequency data of different scales are shown in **Figures 5A,B,C,D**. The Hurst index of the time series is estimated by multi-scale wavelet transform data, as shown in **Figure 6A**, in which the wavelet transform scales are 1–9. The estimated Hurst index is used to generate FBM, and the empirical probability distribution of the generated FBM data is shown in **Figure 6B**. 10,000 FBM data are generated by the Monte Carlo method here. The probability distribution data corresponding to the constructed natural logarithm of

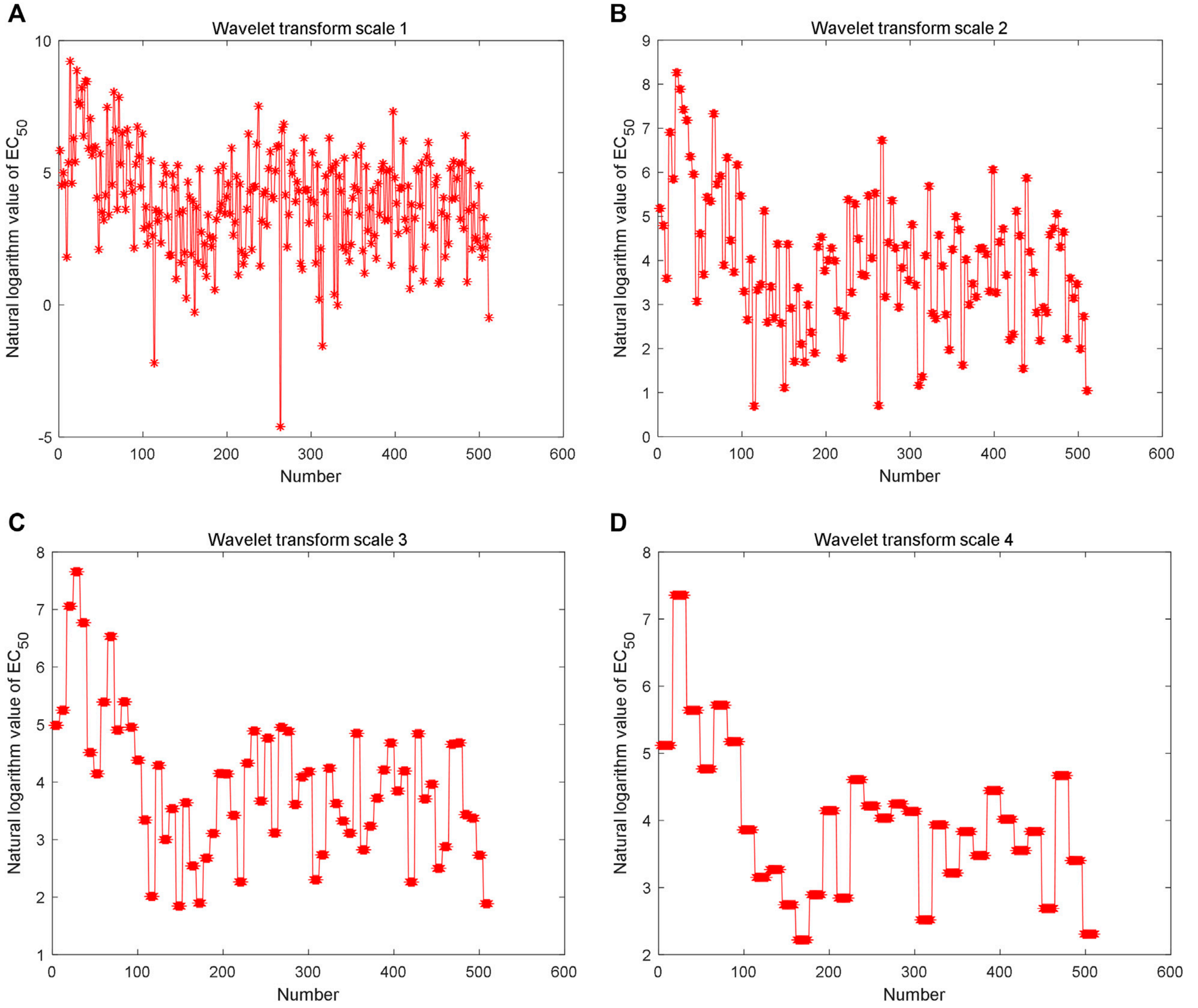

FIGURE 5 | Multi-scale wavelet decomposition of constructed time series: low frequency data of **(A)** level 1 wavelet transform, **(B)** level 2 wavelet transform, **(C)** level 3 wavelet transform, and **(D)** level 4 wavelet transform.

AHTPA-EC$_{50}$ are represented in red, and the curve closest to the constructed natural logarithm of AHTPA-EC$_{50}$ is shown in blue. It can be seen that the constructed AHTPA-EC$_{50}$ is very close to the FBM time series.

2.2 Non-Linear Autoregressive Time-Series Modeling and Its Implementation

2.1.4 Correlation Analysis

Although the constructed series shows fractal characteristics, the relationship between the natural logarithm of coding value of AAS and its corresponding natural logarithm of AHTPA-EC$_{50}$ still needs to be analyzed. **Figure 7A** shows the cross-correlation analysis for the first group of constructed time series, and it shows weak correlation between the two time series. **Figure 7B** shows the autocorrelation analysis for sorting natural logarithm of AHTPA-EC$_{50}$. We can see that the sorting natural logarithm of AHTPA-EC$_{50}$ showed weak autocorrelation. **Figure 8A** shows the cross-correlation analysis for the second group of time series, and it shows weak correlation between the two time series. **Figure 8B** shows the autocorrelation analysis for constructed natural logarithm of AHTPA-EC$_{50}$, and the natural logarithm of AHTPA-EC$_{50}$ based on the coding value AAS showed obvious autocorrelation.

2.1.5 Non-Linear Autoregressive Model With Exogenous Input

According to the above analysis, the two groups' constructed AHTPA-EC$_{50}$ data are modeled as an autoregressive time series, and the natural logarithm of coding AAS is used as the exogenous input parameter. The non-linear autoregressive model with

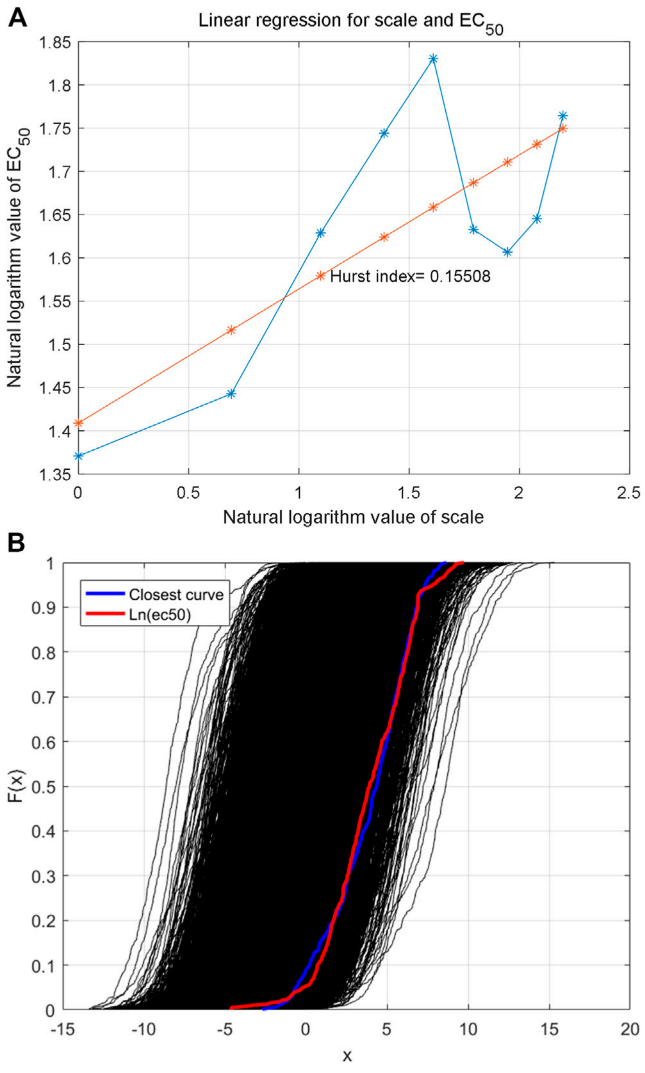

FIGURE 6 | Estimation of Hurst index of the time series **(A)** and empirical probability distribution of FBM with the same Hurst index **(B)**.

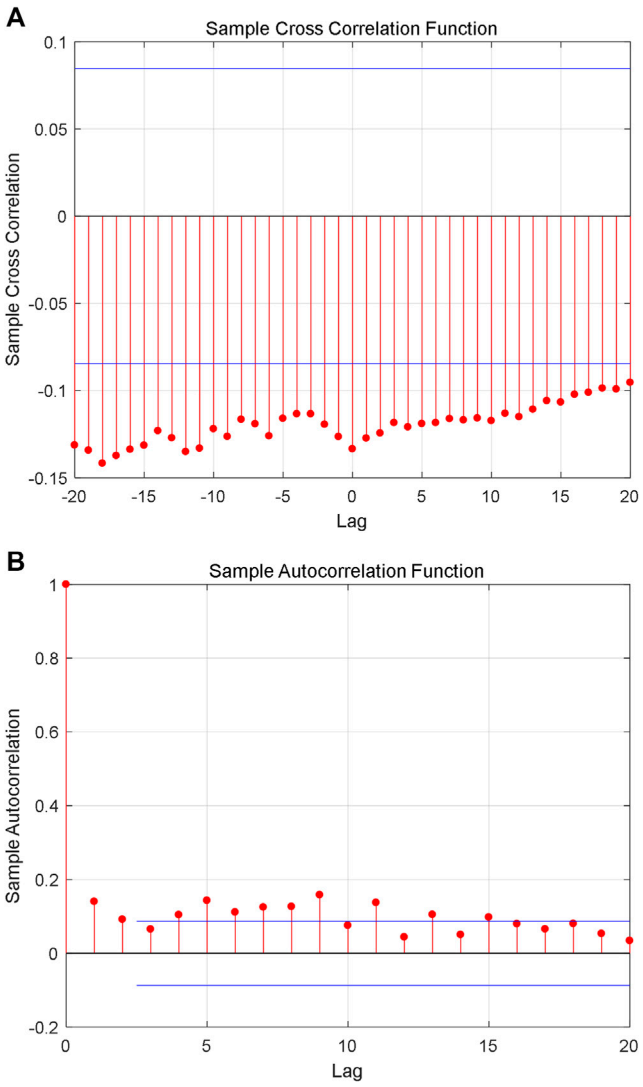

FIGURE 7 | Correlation analysis of the second group time series. **(A)** Cross-correlation with the sorting natural logarithm of coding AAS. **(B)** Autocorrelation of natural logarithm of AHTPA-EC$_{50}$.

exogenous input is established to describe the relationship between the AAS and its corresponding AHTPA-EC$_{50}$, and this relationship is described as

$$y(t) = f\begin{bmatrix} y(t-1), y(t-2), \ldots, y(t-n_y) \\ u(t-1), u(t-2), \ldots, u(t-n_u) \end{bmatrix}, \tag{1}$$

where $y(t), y(t-1), y(t-2), \ldots, y(t-n_y)$ represent time series at different time and $u(t-1), u(t-2), \ldots, u(t-n_u)$ represent exogenous inputs at different time, y denotes the natural logarithm of AHTPA-EC$_{50}$, and u denotes the natural logarithm of coding AAS value. According to the characteristics of AAS and its corresponding AHTPA-EC$_{50}$, the AAS is defined as the input parameter affecting AHTPA-EC$_{50}$ here.

2.1.6 Neural Network Implementation of Model

The NARX model of AHTPA-EC$_{50}$ and AAS was realized by the NARXNN. This neural network was performed in Matlab. The two neural network structures are shown in **Figure 9**. The mean square error (MSE) is selected as the performance function of NARXNN. The Levenberg–Marquardt algorithm is used for net training. The division ratio of training set, verification set, and test set in neural network learning samples is 0.7:0.15:0.15. The delay corresponding to the two constructed series is 1:3 and 1:2, respectively, and the hidden layer has 10 neurons.

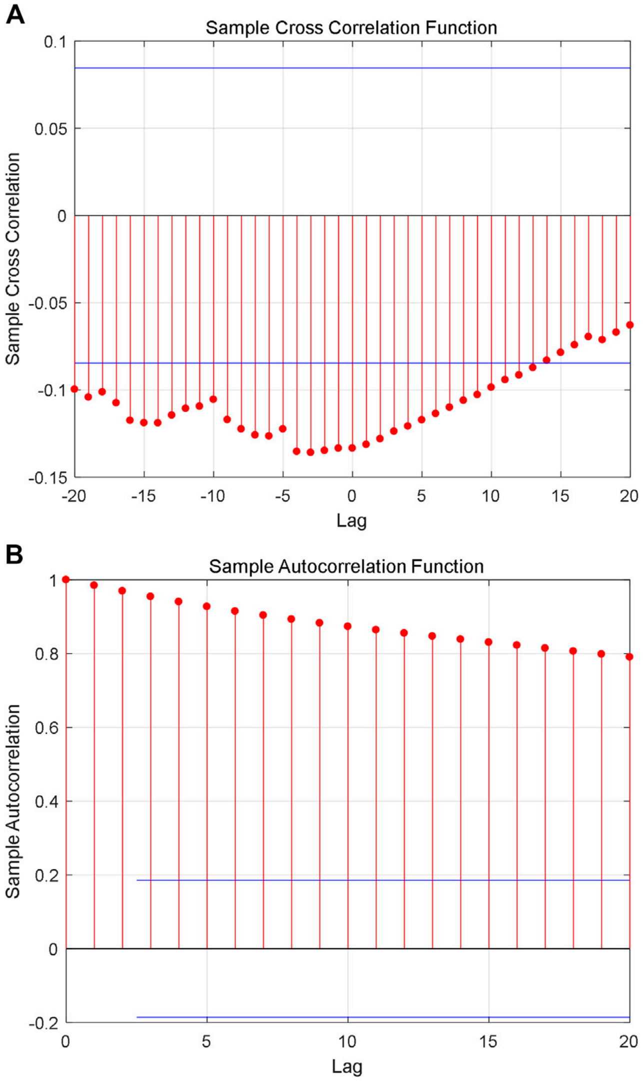

FIGURE 8 | Correlation analysis of the first group time series. **(A)** Cross-correlation with the natural logarithm of coding value of AAS. **(B)** Autocorrelation of sorting natural logarithm of AHTPA-EC_{50}.

2.1.7 Prediction Method for AHTPA-EC_{50}

We further proposed a method for AHTPA-EC_{50} prediction. This method includes two parts: classification and AHTPA-EC_{50} prediction. The ML algorithm is used to classify the AAS. The classification corresponds to different digital segments of AHTPA-$EC_{50.}$ The feature representation is necessary in this process. This prediction method is described in **Figure 10**. Support vector machine (SVM) is used for classification in this research.

3 Results

3.1 Prediction Results of the Proposed Model

As mentioned above, there are 559 groups of samples in total. However, these data include different antihypertensive peptides, whose length is from 2 to 20. We select the samples of AAS, whose length is fixed. There are 231 samples of dipeptides and tripeptides in our dataset. They are larger than other peptides. These data are used to verify the proposed model and prediction model. We also constructed two series according to the above method. The first 200 groups in the first series of samples are used for training, and the last 31 data are used for validation and testing. The training results of the constructed series are shown in **Figure 11**.

For the first NARXNN corresponding to the first group series, the training error is 4.895193, the validation error is 4.636605, and the testing error is 3.546904. For the second NARXNN corresponding to the second group series, the training error is 0.001881, the validation error is 0.124045, and the testing error is 0.010165. The second NARXNN has high accuracy; however, it needs the sorting number, and it cannot be used for prediction alone. The classification of the proposed prediction method can provide a rough location in the series. The first NARXNN also gives an original estimation value of AHTPA-EC_{50}. The AHTPA-EC_{50} will be predicted in the segment of the second series, and two known term AASs help in prediction. The known AASs are selected by the rough location and original estimation value. The second NARXNN is trained every time; therefore, the output will be changed slightly. The first and second NARXNNs are trained in **Figures 11A,B**.

The AHTPA-EC_{50} of dipeptides and tripeptides is used to verify the prediction method. The first 200 groups in the first series of samples are used for training, and the last 31 data in the first series are used for testing. The proposed method demands classification, and we assume that the classification is correct here; thus, we input the AAS in segments. And the classification is designed as three classification. AHTPA-EC_{50} = 1, and median values of the series are designed as segment points. The results of prediction are shown in **Figure 12**. Therefore, when the classification is correct, the MSE is 0.5589. We also designed a backpropagation neural network (BPNN) for comparison. The network structure is designed as 3–10–1. The mean square error (MSE) is selected as a performance function. The Levenberg–Marquardt algorithm is used for net training. The logsig function is set as the input function, and the pure linear function is used in the second layer. The number of iterations is set to 1000, the learning rate is 0.1, and the learning target is 0.00001. The results are shown in **Figure 13**, where test samples are randomly selected 100 times. The results reveal that the proposed method has better accuracy than the BPNN.

3.2 Classification of AAS for AHTPA-EC_{50}

As mentioned above, the proposed prediction method demands a rough position which is used in NARX2 prediction. Two classification and three classification are designed for the proposed prediction method here. SVM is used for the classification of AHTPA-EC_{50} and its corresponding AAS here. We classify the AAS whose length is less than three amino acids. 231 samples of dipeptides and

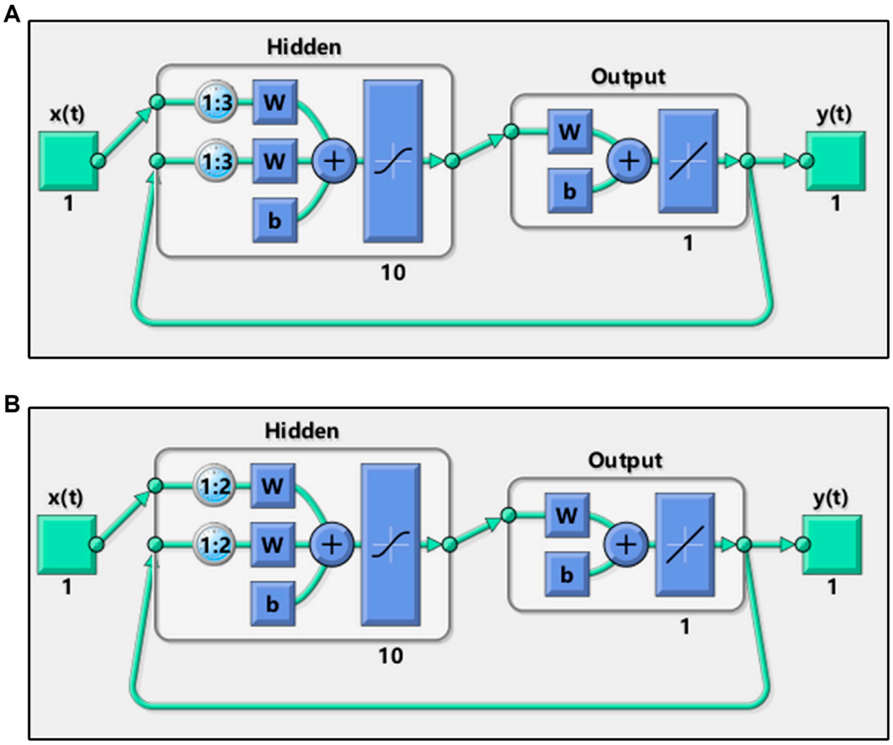

FIGURE 9 | Structures of the neural network for the **(A)** first series and **(B)** second series.

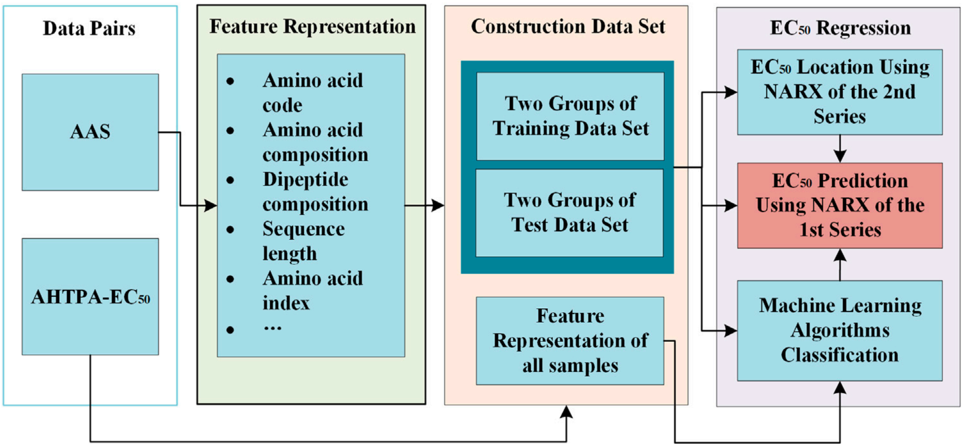
Data Pairs
AAS
AHTPA-EC50
Feature Representation
• Amino acid code
• Amino acid composition
• Dipeptide composition
• Sequence length
• Amino acid index
• …
Construction Data Set
Two Groups of Training Data Set
Two Groups of Test Data Set
Feature Representation of all samples
EC50 Regression
EC50 Location Using NARX of the 2nd Series
EC50 Prediction Using NARX of the 1st Series
Machine Learning Algorithms Classification

FIGURE 10 | Prediction method for AHTPA-EC$_{50}$–based NARX.

tripeptides are classified here. For three classification, AHTPA-EC$_{50}$ = 1 and median values of the series are designed as segment points. For two classification, the median value of the series is designed as the segment point. The label design is shown in **Figure 14**.

For two classification, there are 161 training data pairs and 70 testing data pairs which are used for classification. And eight feature descriptors are extracted from the peptide sequence. They are the amino acid composition, the digital description of AAS, the peptide sequence code, and the length of peptide sequence.

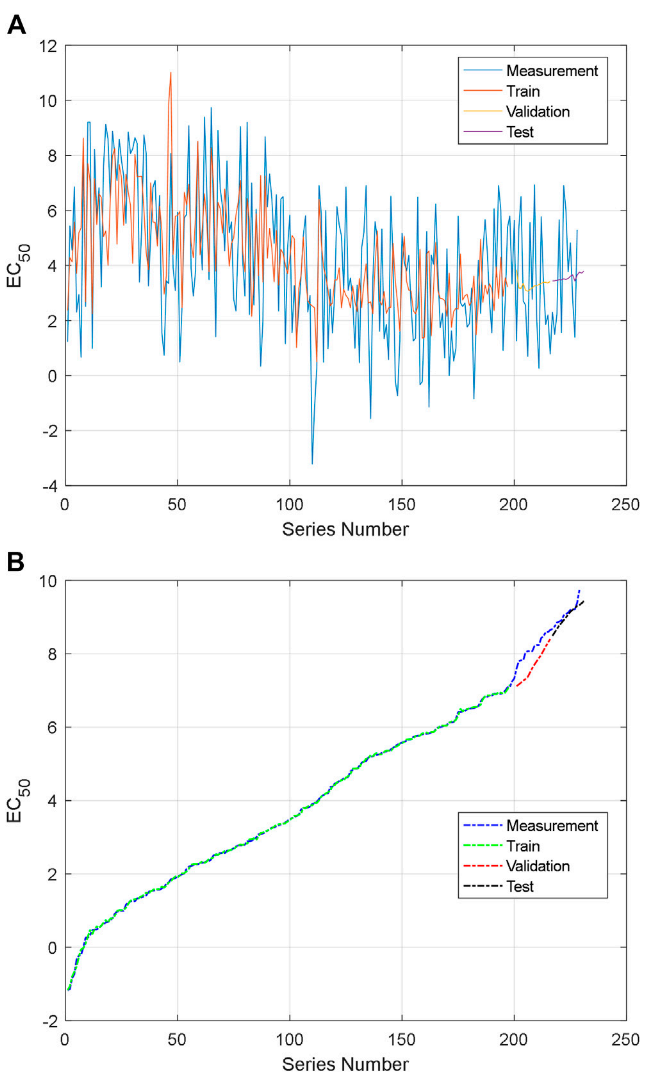

FIGURE 11 | Training and testing data: **(A)** the first NARXNN prediction for the first group series and **(B)** the second NARXNN prediction for the second group series.

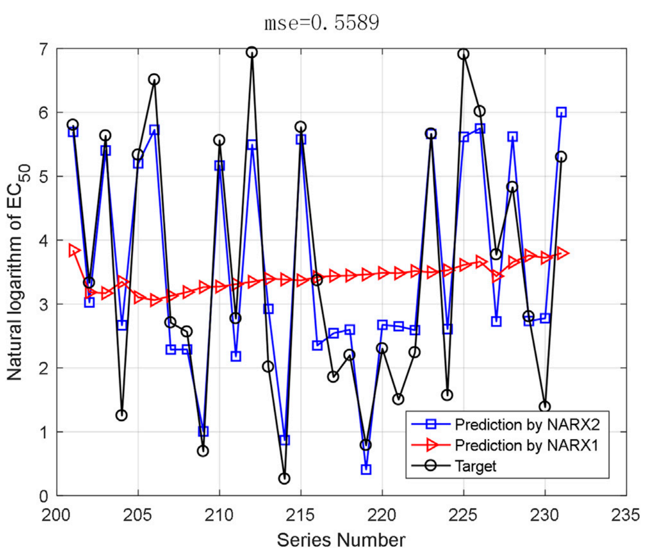

FIGURE 12 | Prediction results by the proposed method.

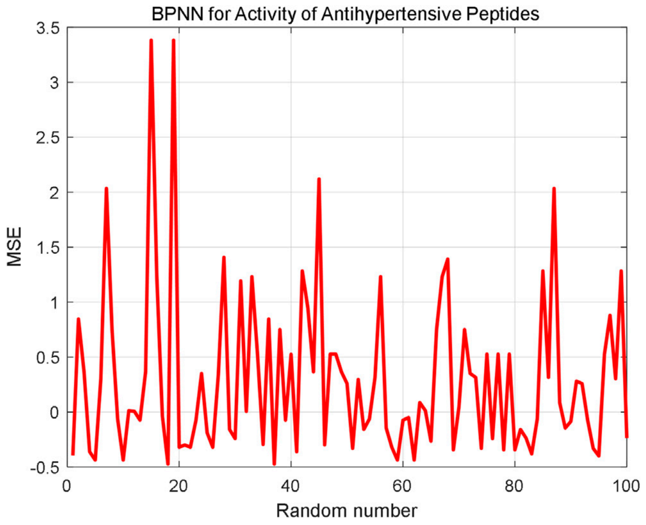

FIGURE 13 | Prediction results by the BPNN.

The classification results are shown in **Figure 15**. We can see that the two classification accuracy is 68.57% and the three classification accuracy is 60.00%. Due to the limitations in training, the effect of three classification is not very good. If the quantity of training sample increases and other ML algorithms are also used, we think the accuracy can be improved.

4 Conclusion

In this paper, the statistical distribution of AHTPA-EC_{50} is analyzed. Two group time series are constructed between AHTPA-EC_{50} and its corresponding AAS. According to the characteristics of constructed time series, AHTPA-EC_{50} is modeled by the NARX model. Then, a prediction method of AHTPA-EC_{50} is proposed. Dipeptides and tripeptides are used to verify the proposed model and prediction method. The results show that the MSE is 0.5589 when the classification is correct. Finally, we tried to classify the dipeptide and tripeptide data by SVM. Although the accuracy of classification is not very high, it is still feasible. The proposed model and prediction method provide a solution for AHTPA-EC_{50} prediction, and they are useful and meaningful on antihypertensive active peptide research, drug design, and industrial production (Chen et al., 2020; Granger and Joyeux 1980).

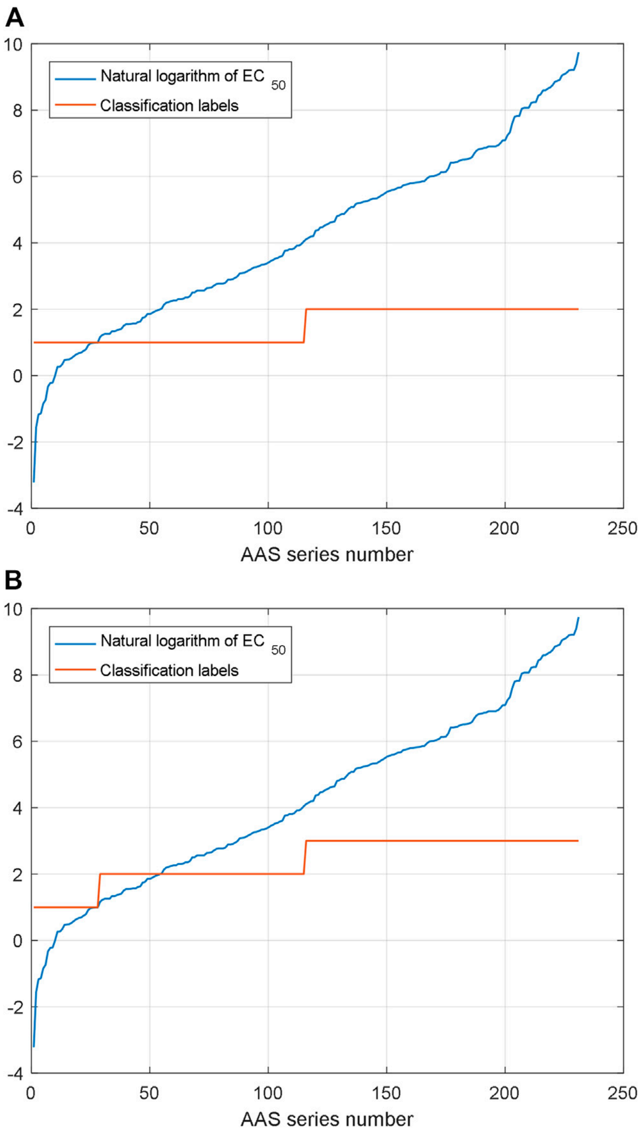

FIGURE 14 | Classification by label design: **(A)** two classification and **(B)** three classification.

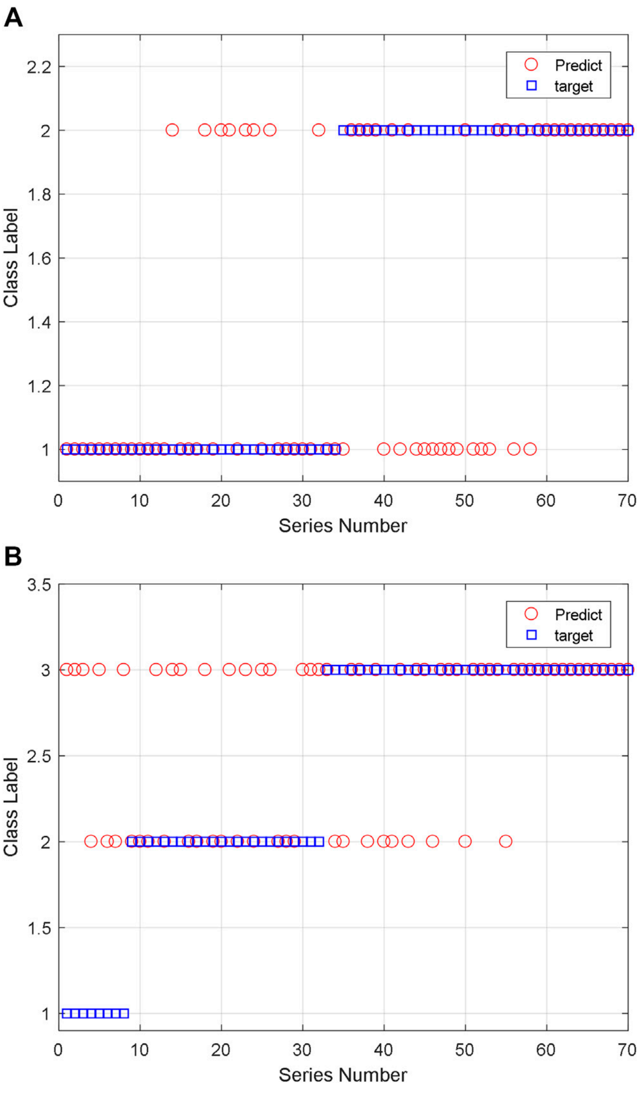

FIGURE 15 | Classification results of the AAS using SVM: **(A)** two classification and **(B)** three classification.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, and further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

XX and CZ designed the algorithm. DW and MD proposed the problem, pointed research direction, and provided the dataset. XX wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 31730069 and 31371805).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.801728/full#supplementary-material

REFERENCES

Al-Hamdan, M., Cruise, J., Rickman, D., and Quattrochi, D. (2010). Effects of Spatial and Spectral Resolutions on Fractal Dimensions in Forested Landscapes. *Remote Sensing* 2, 611–640. doi:10.3390/rs2030611

Al-Hamdan, M. Z., Cruise, J. F., Rickman, D. L., and Quattrochi, D. A. (2012). Characterization of Forested Landscapes from Remotely Sensed Data Using Fractals and Spatial Autocorrelation. *Adv. Civil Eng.* 2012, 1–14. doi:10.1155/2012/945613

Baldi, P., Brunak, S., and Bach, F. (2001). *Bioinformatics: The Machine Learning Approach*. Cambridge: Mass MIT Press.

Chen, Z., Pang, M., Zhao, Z., Li, S., Miao, R., Zhang, Y., et al. (2020). Feature Selection May Improve Deep Neural Networks for the Bioinformatics Problems. *Bioinformatics* 36 (5), 1542–1552. doi:10.1093/bioinformatics/btz763

Chow, W. C. (2011). Fractal (Fractional) Brownian Motion. *Wires Comp. Stat.* 3 (2), 149–162. doi:10.1002/wics.142

Engle, R. F., and Granger, C. W. J. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica* 55 (2), 251–276. doi:10.2307/191325310.2307/1913236

Fouché, W. L., and Mukeru, S. (2013). On the Fourier Structure of the Zero Set of Fractional Brownian Motion. *Stat. Probab. Lett.* 83, 459–466. doi:10.1016/j.spl.2012.10015

Ghosh, J. K., and Somvanshi, A. (2008). Fractal-based Dimensionality Reduction of Hyperspectral Images. *J. Indian Soc. Remote Sens* 36, 235–241. doi:10.1007/s12524-008-0024-0

Granger, C. W. J., and Joyeux, R. (1980). An Introduction to Long-Memory Time Series Models and Fractional Differencing. *J. Time Ser. Anal.* 1, 15–29. doi:10.1017/CBO9780511753978.01810.1111/j.1467-9892.1980.tb00297.x

Hannan, E. J., and Deistler, M. (1988). *The Statistical Theory of Linear Systems,*. New York: Wiley, 5–48.

Kim, T. S., and Kim, S. (2004). Singularity Spectra of Fractional Brownian Motions as a Multi-Fractal. *Chaos, Solitons & Fractals* 19, 613–619. doi:10.1016/S0960-0779(03)00187-5

Kumar, R., Chaudhary, K., Sharma, M., Nagpal, G., Chauhan, J. S., Singh, S., et al. (2015a). AHTPDB: a Comprehensive Platform for Analysis and Presentation of Antihypertensive Peptides. *Nucleic Acids Res.* 43, D956–D962. doi:10.1093/nar/gku1141

Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., et al. (2015b). An In Silico Platform for Predicting, Screening and Designing of Antihypertensive Peptides. *Sci. Rep.* 5, 12512. doi:10.1038/srep12512

Libbrecht, M. W., and Noble, W. S. (2015). Machine Learning Applications in Genetics and Genomics. *Nat. Rev. Genet.* 16 (6), 321–332. doi:10.1038/nrg3920

Majumder, K., and Wu, J. (2010). A New Approach for Identification of Novel Antihypertensive Peptides from Egg Proteins by QSAR and Bioinformatics. *Food Res. Int.* 43, 1371–1378. doi:10.1016/j.foodres.2010.04.027

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). mAHTPred: a Sequence-Based Meta-Predictor for Improving the Prediction of Anti-hypertensive Peptides Using Effective Feature Representation. *Bioinformatics* 35 (16), 2757–2765. doi:10.1093/bioinformatics/bty1047

Moran, P. (1953). The Statistical Analysis of the Canadian lynx Cycle. *Aust. J. Zool.* 1, 291–298. doi:10.1071/ZO9530291

Nill, N. B., and Bouzas, B. H. (1992). Objective Image Quality Measure Derived from Digital Image Power Spectra. *Opt. Eng.* 31, 813–825. doi:10.1117/12.56114

Ozaki, T. (1980). Non-linear Time Series Models for Non-linear Random Vibrations. *J. Appl. Probab.* 17, 84–93. doi:10.1017/S0021900200046829

Pentland, A. P. (1984). Fractal-Based Description of Natural Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-6, 661–674. doi:10.1109/TPAMI.1984.4767591

Ruderman, D. L. (1996). Origins of Scaling in Natural Images. *Vis. Res.* 37, 3385–3398. doi:10.1117/12.238707

Stone, M. E. (20182018). *Kaplan's Essentials of Cardiac Anesthesia*. ISBN 978-0-323-49798-5. Elsevier. Mechanisms of Action: ACE inhibitors act by inhibiting one of several proteases responsible for cleaving the decapeptide Ang I to form the octapeptide Ang II. doi:10.1016/c2012-0-06151-0

Tong, H. (1990). *Nonlinear Time Series: A Dynamical Systems Approach*. Oxford: Oxford University Press, 14–37.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*. New York: Springer-Verlag, 7–34.

Tong, J., Liu, S., Zhou, P., Wu, B., and Li, Z. (2008). A Novel Descriptor of Amino Acids and its Application in Peptide QSAR. *J. Theor. Biol.* 253, 90–97. doi:10.1016/j.jtbi.2008.02.030

Tu, M., Cheng, S., Lu, W., and Du, M. (2018a). Advancement and Prospects of Bioinformatics Analysis for Studying Bioactive Peptides from Food-Derived Protein: Sequence, Structure, and Functions. *Trac Trends Anal. Chem.* 105, 7–17. doi:10.1016/j.trac.2018.04.005

Tu, M., Wang, C., Chen, C., Zhang, R., Liu, H., Lu, W., et al. (2018b). Identification of a Novel ACE-Inhibitory Peptide from Casein and Evaluation of the Inhibitory Mechanisms. *Food Chem.* 256, 98–104. doi:10.1016/j.foodchem.2018.02.107

Win, T. S., Schaduangrat, N., Prachayasittikul, V., Nantasenamat, C., and Shoombuatong, W. (2018). PAAP: a Web Server for Predicting Antihypertensive Activity of Peptides. *Future Med. Chem.* 10, 1749–1767. doi:10.4155/fmc-2017-0300

Wornell, G. W., and Oppenheim, A. V. (1992). Estimation of Fractal Signals from Noisy Measurements Using Wavelets. *IEEE Trans. Signal. Process.* 40, 611–623. doi:10.1109/78.120804

Wu, D., Tu, M., Wang, Z., Wu, C., Yu, C., Battino, M., et al. (2020). Biological and Conventional Food Processing Modifications on Food Proteins: Structure, Functionality, and Bioactivity. *Biotechnol. Adv.* 40, 107491. doi:10.1016/j.biotechadv.2019.107491

Yang, Z., Wang, J., Yang, J., Qi, Z., He, J., and He, J. (2020). Recognizing Proteins with Binding Function in Elymus Nutans Based on Machine Learning Methods. *Comb. Chem. High Throughput Screen.* 23 (6), 554–562. doi:10.2174/1386207323666200330120154

Zhang, R.-z., Xu, X.-h., Chen, T.-b., Li, L., and Rao, P.-f. (2000). An Assay for Angiotensin-Converting Enzyme Using Capillary Zone Electrophoresis. *Anal. Biochem.* 280 (2), 286–290. doi:10.1006/abio.2000.4535

Zhang, Z., Wang, J., and Liu, J. (2021). DeepRTCP: Predicting ATP-Binding Cassette Transporters Based on 1-Dimensional Convolutional Network. *Front. Cel Dev. Biol.* 8, 614080. doi:10.3389/fcell.2020.614080

Zou, Q., and LiuLiu, Q. (2019). Advanced Machine Learning Techniques for Bioinformatics. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16 (4), 1182–1183. doi:10.1109/TCBB.2019.2919039

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

METHODS
published: 28 January 2022
doi: 10.3389/fgene.2021.821996

Identify DNA-Binding Proteins Through the Extreme Gradient Boosting Algorithm

Ziye Zhao[1†], Wen Yang[2†], Yixiao Zhai[1], Yingjian Liang[3] and Yuming Zhao[1]**

[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2]International Medical Center, Shenzhen University General Hospital, Shenzhen, China, [3]Department of Obstetrics and Gynecology, The First Affiliated Hospital of Harbin Medical University, Harbin, China

The exploration of DNA-binding proteins (DBPs) is an important aspect of studying biological life activities. Research on life activities requires the support of scientific research results on DBPs. The decline in many life activities is closely related to DBPs. Generally, the detection method for identifying DBPs is achieved through biochemical experiments. This method is inefficient and requires considerable manpower, material resources and time. At present, several computational approaches have been developed to detect DBPs, among which machine learning (ML) algorithm-based computational techniques have shown excellent performance. In our experiments, our method uses fewer features and simpler recognition methods than other methods and simultaneously obtains satisfactory results. First, we use six feature extraction methods to extract sequence features from the same group of DBPs. Then, this feature information is spliced together, and the data are standardized. Finally, the extreme gradient boosting (XGBoost) model is used to construct an effective predictive model. Compared with other excellent methods, our proposed method has achieved better results. The accuracy achieved by our method is 78.26% for PDB2272 and 85.48% for PDB186. The accuracy of the experimental results achieved by our strategy is similar to that of previous detection methods.

Keywords: DNA-binding protein prediction, machine learning, feature extraction, dimensionality reduction, XGBoost model

OPEN ACCESS

Edited by:
Juan Wang,
Inner Mongolia University, China

Reviewed by:
Wei Lan,
Guangxi University, China
Junwei Luo,
Henan Polytechnic University, China

Correspondence:
Yingjian Liang
genomeliang@hotmail.com
Yuming Zhao
zym@nefu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *25 November 2021*
Accepted: *07 December 2021*
Published: *28 January 2022*

Citation:
Zhao Z, Yang W, Zhai Y, Liang Y and Zhao Y (2022) Identify DNA-Binding Proteins Through the Extreme Gradient Boosting Algorithm.
Front. Genet. 12:821996.
doi: 10.3389/fgene.2021.821996

INTRODUCTION

Organisms contain many macromolecular substances, such as DNA and proteins, which contain the genetic information of organisms and are important components of all cells and tissues that make up an organism. To study the life activities of cells, it is necessary to study DNA and proteins and the interaction between them. Research on DBPs has an extremely important status and significance in related life sciences and plays an important role in DNA replication and recombination, virus infection and proliferation. It is necessary to study the combination of DNA and protein to study the gene expression of organisms at the molecular level. Researchers are paying increasing attention to DBP studies. DBPs are a kind of protein that binds to DNA, and it is critical to determine which of the numerous proteins can attach to DNA (Liu et al., 2019a; Li et al., 2019; Li et al., 2020) However, the traditional use of biochemical methods to find DBP consumes considerable time and money. Based on the above requirements and the development of computer science and ML(Zheng et al., 2019; Zheng et al., 2020; Wang et al., 2021a), relevant researchers have developed many detection methods based on ML algorithms in the hopes of improving the efficiency of detecting DBP and saving manpower and material resources.

ML is frequently utilized in the fields of computational biology (Jiang et al., 2013a; Cheng et al., 2019a; Liu et al., 2019b; Wang et al., 2019; Liu et al., 2020a; Tao et al., 2020a; Wang et al., 2020a; Zhang et al., 2020a; Zhao et al., 2020a; Zhu et al., 2020; Wang et al., 2021b; Wang et al., 2021c; Dao et al., 2021; Yu et al., 2021) to analyze brain disease (Liu et al., 2018a; Cheng et al., 2019b; Bi et al., 2020; Iqubal et al., 2020; Zhang et al., 2021a), lncRNA-miRNA interactions (Cheng et al., 2016; Liu et al., 2020b; Han et al., 2021), protein remote homology (Hong et al., 2020), protein functions (Wei et al., 2018a; Shen et al., 2019a; Shen et al., 2019b; Ding et al., 2019; Wang et al., 2020b; Shen et al., 2020; Tang et al., 2020; Wang et al., 2021d; Shang et al., 2021; Shao and Liu, 2021; Zhao et al., 2021), electron transport proteins (Ru et al., 2019), differential expression (Yu et al., 2020a; Zhao et al., 2020b; Zhai et al., 2020) and protein-protein interconnections (Ding et al., 2016a; Ding et al., 2016b; Yu et al., 2020b).

The protein sequence is very sizeable, and its number far exceeds the number of structures known to researchers (Zuo et al., 2017). Therefore, ML is used in various computer programs that predict DBP. The model IDNA-Prot|dis (Liu et al., 2014) was proposed by Liu et al. and is used to detect DBP based on the pseudo amino acid composition (PseAAC), and it can accurately extract the characteristics of DNA binding proteins. There are two models that use PseACC and physical-chemical distance transformation and support vector machine (SVM) algorithms, named PseDNA-Pro (Liu et al., 2015a) and iDNAPro-PseAAC (Liu et al., 2015b). Lin et al. developed the IDNA-Prot (Lin et al., 2011) prediction model based on the random forest (RF) algorithm through the PseACC feature. Kummar et al. developed two models based on RF and SVM classifiers called DNA-Prot (Kumar et al., 2009) and DNAbinder (Kumar et al., 2007). Dong et al. proposed the Kmer1+ACC (Liu et al., 2016) model based on the SVM algorithms Kmer composition and autocross covariance transformation. The position-specific scoring matrix (PSSM) can be obtained by calculating the protein sequence's position frequency matrix, which has evolutionary information on the protein (Shao et al., 2021). The Local-DPP (Wei et al., 2017) uses the local pseudo position-specific scoring matrix (Pse-PSSM) and random forest algorithm to detect DBPs. Multiple kernel SVM is a DBP predictor from heuristically kernel alignment, and it is also named MKSVM-HKA (Ding et al., 2020a), which includes a variety of characteristics and was developed by Ding et al. The MSFBinder (Liu et al., 2018b) model proposed by Liu et al. is based on multiview features as well as classifiers. DPP-PseAAC (Rahman et al., 2018) is a model based on Chou's general PseAAC, and it is used to detect DBPs. Methods have also been developed that combine multiscale features and deep neural networks to predict DBPs, such as MsDBP (Du et al., 2019).Adilina et al. (2019) analyzed protein sequence characteristics and implemented two different feature selection methods to build a DBP predictor.

In recent years, an increasing number of researchers have adopted complex feature extraction methods (Fu et al., 2020; Jin et al., 2021) and classification models to identify DBPs. It is critical to develop a method that uses as few DBP features as possible and includes a simple classification model while also ensuring a good ability to detect DPB. According to previous work, we proposed a DBP identification method based on the XGBoost model. First, several features were extracted from the protein sequence. Second, the features of these sequences were spliced. Third, the dimension of the data was standardized and reduced. Finally, the XGBoost model was used to detect DBPs. We have evaluated the effectiveness of our method on some benchmark data sets. Compared with some current experimental methods, our method achieves a better Matthew's correlation coefficient (MCC), with a value of 0.713 for PDB186 and 0.5652 for PDB2272.

METHODS

Identifying DBPs is a common dichotomy problem. First, we used six different feature extraction models for DBPs sequences to extract the corresponding sequence feature information. Then, the sequence feature information was spliced. Next, dimensionality reduction was performed on the spliced sequence feature information. Finally, the XGBoost model was utilized to identify DBPs. **Figure 1** depicts the flowchart of our adopted technique.

Extracting Features

To recognize DBPs, the corresponding features must be extracted. We adopt six feature extraction methods to obtain sequence information: global encoding, GE (Li et al., 2009); multi-scale continuous as well as discontinuous descriptor, MCD (You et al., 2014); normalized Moreau-Broto auto correlation, NMBAC (Ding et al., 2016b; Feng and Zhang, 2000); position specific scoring matrix-based average blocks, PSSM-AB (Jeong et al., 2011; Zhu et al., 2019); PSSM-based discrete cosine transform, PSSM-DCT (Huang et al., 2015); and PSSM-based discrete wavelet transform, PSSM-DWT (Nanni et al., 2012). The abovementioned feature extraction models are all well-known protein sequence extraction algorithm s and commonly used, which could be described in related works (Zou et al., 2021). **Table 1** shows the feature dimensions derived by various feature extraction methods. After completing the above work, we used MATLAB to horizontally stitch together (Ding et al., 2020c; Ding et al., 2020d; Yang et al., 2021a) the features extracted from the same protein sequence using different feature extraction methods. The spliced features are represented by $\boldsymbol{Z}^{*}$. After splicing, the dimensions of PDB14189 and PDB2272 are 2692, and the dimensions of PDB1075 and PDB186 are 3092.

Standardize the Data

To make the data more standardized and unified and to strengthen the relationship between the characteristics of the data and the labels of the data, we use Z-score standardization to process the data.

Z-score standardization is defined as follows:

$$\mathbf{M}* = \frac{\boldsymbol{Z}_i^{*} - \bar{\boldsymbol{Z}}}{\sigma} \tag{1A}$$

$$\bar{\boldsymbol{Z}} = \frac{\sum_{i=0}^{N} \boldsymbol{Z}_i^{*}}{N} \tag{1B}$$

$$\sigma = \sqrt{\frac{\sum_{i=0}^{N} \left(\boldsymbol{Z}_i^{*} - \bar{\boldsymbol{Z}}\right)^2}{N}} \tag{1C}$$

$$i = 1, 2, \ldots, \mathrm{N} \tag{1D}$$

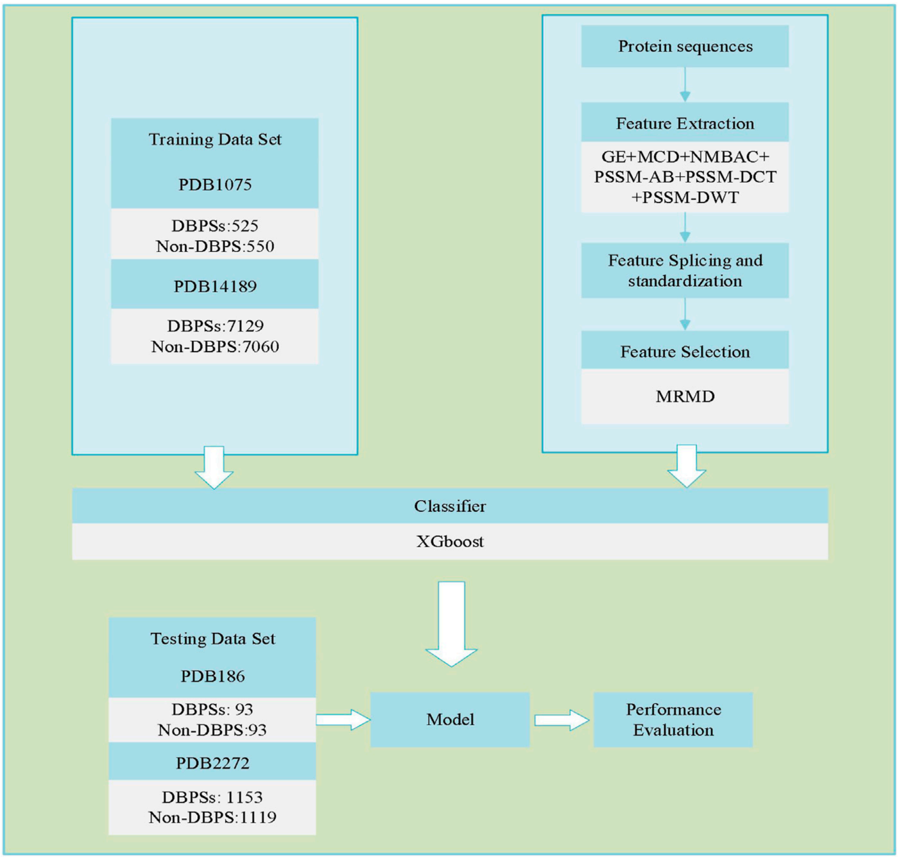

FIGURE 1 | Process of predicting DBPs.

TABLE 1 | Dimensional information about the features.

Model	Dimensionality
GE	150
MCD	882
MNBAC	200
PSSM-AB	200
PSSM-DCT	399
PSSM-DWT	1,040

where N is the total number of samples and σ is the standard deviation.

The DBP sequence was processed in three stages: feature extraction, feature information splicing, and data standardization. Following the aforementioned three stages, we can obtain the sequence feature information $\mathbf{M}^*$.

Dimensionality Reduction by Max-Relevance-Max-Distance

Zou et al. (Quan et al., 2016; Niu et al., 2020) developed a dimensionality reduction method in 2015 named Max-Relevance-Max-Distance (MRMD), and the user guide and complete runtime program can be obtained and downloaded from the following URL: https://github.com/heshida01/MRMD3.0. It judges data independence through a distance function and completes the dimensionality reduction operation in three steps (Tao et al., 2020b). It first evaluates each feature's contribution to the classification and then quantifies each feature's contribution to the classification. Second, the weights of different features are calculated for classification and the selected features are sorted accordingly. Third, the different numbers of features are filtered and classified and the results are recorded. We analyze and compare the results of the previous step to select the most effective group and use the sequence features chosen from this group as the result of dimensionality reduction.

The maximum correlation and the maximum distance are the main bases for the MRMD algorithm to judge the weight of each feature to the prediction result. The Pearson correlation coefficient can be used to quantify the degree of correlation between features and cases, and it can be calculated by the maximum relevance (MR).

The Pearson correlation coefficient is defined as follows:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

The i_{th} characteristic from the sequence and the category label to which those sequences belong make up the vectors X and Y.

The maximum distance (MD) is used to assess feature redundancy. We calculate the three indices between characteristics in total.

$$\mathrm{ED}(X,Y)=\sqrt{\sum_{i=0}^{N}(x_i-y_i)^2}\,(i=1,2,\ldots,\mathrm{N}) \tag{3A}$$

$$\cos(X,Y)=\frac{X\cdot Y}{\|X\|\|Y\|} \tag{3B}$$

$$TC(X,Y)=\frac{X\cdot Y}{\|X\|^2+\|Y\|^2-X\cdot Y} \tag{3C}$$

Equations 3A, **E3B**, **E3C** represent Euclidean distance, cosine similarity and Tanimoto coefficient, respectively. We can obtain the MD value by calculating the three indicators. Finally, the classification contribution value of each feature is calculated by combining MR and MD in a specific ratio.

After dimensionality reduction, the dimensions of PDB14189 and PDB2272 are 379, and the dimensions of PDB1075 and PDB186 are 1460.

Based on the three steps of feature extraction and splicing, data standardization and dimensionality reduction operations, we obtain the final sequence features.

Extreme Gradient Boosting Algorithm

In 2011, Tianqi Chen and Carlos Guestrin (Chen and Guestrin, 2016) first proposed the XGBoost algorithm, or the extreme gradient boosting algorithm. It is a machine learning model that achieves a stronger learning effect by integrating multiple weak learners. The XGBoost model has many advantages, such as strong flexibility and scalability (Yang et al., 2021b; Zhang et al., 2021b).

Generally, most boosting tree models have difficulty implementing distributed training because when training n_{th} trees, they will be affected by the residuals of the first *n-1* trees and only use first-order derivative information. The XGBoost model is different. It performs a second-order Taylor expansion of the loss function and uses a variety of methods to prevent overfitting as much as possible. XGBoost can also automatically use the CPU's multithreaded parallel computing to speed up the running speed. This feature represents a great advantage of XGBoost over other methods. XGBoost has improved significantly in terms of effect and performance.

The XGBoost algorithm is described in detail as follows:

$$\hat{y}_i=\sum_{m=1}^{M}f_m(x_i),\, f_m\in F \tag{4}$$

where M is the number of trees and F represents the basic model of the trees.

The objective function is defined as follows:

$$L=\sum_i l(\hat{y}_i,y_i)+\sum_m \Omega(f_m) \tag{5}$$

The error between the predicted value and the true value is represented by the loss function l, and the regularized function $\mathbf{\Omega}$ to prevent overfitting is defined as follows:

$$\mathbf{\Omega}(f)=\boldsymbol{\gamma}T+\frac{1}{2}\lambda\|w\|^2 \tag{6}$$

where the weight and number of leaves of each tree are represented by w and T, respectively.

After performing the quadratic Taylor expansion on the objective function, the information gain generated after each split of the objective function can be expressed as follows:

$$Gain=\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L}g_i\right)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{\left(\sum_{i\in I_R}g_i\right)^2}{\sum_{i\in I_R}h_i+\lambda}+\frac{\left(\sum_{i\in I}g_i\right)^2}{\sum_{i\in I}h_i+\lambda}\right]-\gamma \tag{7}$$

We can see that the split threshold γ is added to **Eq. 7** to prevent overfitting and inhibit the overgrowth of the tree. Only when the information gain is greater than γ is the leaf node allowed to split. It can optimize the objective function at the same time because the tree is prepriced.

XGBoost also has the following two features:

1. Splitting stops when the threshold is greater than the weight of all samples on the leaf node too prevent the model from learning special training samples.
2. The features are randomly sampled when constructing each tree.

These features can prevent the XGBoost model from overfitting during the experiment.

EXPERIMENTAL RESULTS

In this chapter, we obtain experimental results through experiments on four benchmark data sets, evaluate our methods of identifying DBP and compare our experimental results with that of other methods.

Data Sets

The four benchmark data sets are PDB1075, PDB186, PDB14189, and PDB2272. Liu et al. (2015a) and Lou et al. (2014) provided PDB1075 (training set) and PDB186 (independent testing set), respectively, and Du et al. (2019) provided PDB14189 (training set) and PDB2272 (independent testing set). These data sets are from the Protein Data Bank (PDB), and **Table 2** shows the results of their detailed information.

Measurement Standard

In this research, the following coefficients are used to evaluate our method: specificity (SP), sensitivity (SN), Matthew correlation coefficient (MCC), accuracy (ACC) and area under the ROC curve (AUC) (Jiang et al., 2013b; Wei et al., 2014; Wei et al., 2018a; Wei et al., 2018b; Cheng et al., 2018; Jin et al., 2019; Zhang et al., 2020b; Cheng et al., 2020; Liu et al., 2020c; Wang et al., 2020c; Guo et al., 2020; Huang et al., 2020; Wei et al., 2020; Zeng et al., 2020; Zhai et al., 2020). The calculation formulas for these coefficients are as follows:

TABLE 2 | Basic information about four standard data sets.

Data sets	The number of negative	The number of positive	The total numbers
PDB14189	7,060	7,129	14,189
PDB1075	550	525	1,075
PDB2272	1,119	1,153	2,272
PDB186	93	93	186

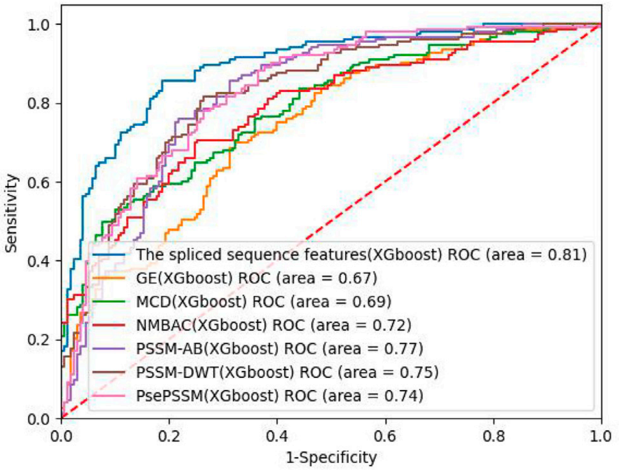

FIGURE 2 | ROC curves of different feature extraction methods on PDB1075 data.

$$Spec = \frac{TN}{TN + FP} \tag{8A}$$

$$SN = \frac{TP}{TP + FN} \tag{8B}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{8C}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{8D}$$

Among them, TN, TP, FP and FN reflect the values of true negatives, true positives, false positives, and false negatives, respectively.

Performance Analysis

On the PDB 1075 data set, the performance of the spliced sequence features and single sequence features is evaluated by randomly extracting 30% of the data as a test set. **Figure 2**; **Table 3** depict the experimental outcomes. PSSM-DWT (MCC: 0.4981) achieved better performance than other single sequence features. The spliced sequence features perform better than the single sequence feature on all parameters. The spliced sequence feature (ROC: 0.81) also gained the best ROC performance.

Independent Data Set of PDB186

In this experiment, different sequence features have different prediction performances. We use PDB1075 as the training set and PDB186 as the test set to evaluate our experimental method and compared the experimental findings of our approach to those of 13 other methods. **Table 4** clearly shows the complete experimental outcomes.

The MCC values of the five methods are all above 0.6 for MSDBP, MSFBinder, Local-DPP MKSVM-HKA, and Adilina's work (0.606, 0.616, 0.625, 0.648 and 0.670, respectively). Thus, these methods have excellent performance. Although Adilina's work (SN: 95.0%) performs best in terms of the value of SN, the results of XGBoost achieve optimal ACC (85.48%), MCC (0.713) and Spec (80.6%). On PDB1075 and PDB186, XGBoost outperforms the other methods.

Independent Data Set of PDB2272

Du et al. (2019) removed proteins in PDB2272 that shared more than 40% of their sequence with PDB14189 to avoid homology bias between the two data sets. We conducted experiments on Du's data set to verify the performance of the XGBoost model. PDB14189 is the training set, and PDB2272 is the test set. We independently tested XGBoost on PDB2272, used PDB14189 as the training set and compared it with five other classification methods. The detailed experimental results can be seen in **Table 5**. The results clearly show that XGBoost achieves the best ACC, MCC and Spec values of 78.26%, 0.5652 and 76.05%, respectively, compared with the other methods. For PDB2272, XGBoost presents a superior performance relative to the other classification methods.

Experimental Results With PDB2272 and PDB186 as Test Set

We combined PDB14189 and PDB1075 as the training set, and combined PDB2272 and PDB186 as the test set. After normalization and dimensionality reduction operations, we got an accuracy of 79.09% and the MCC value was 0.5818. It can be seen that this result is between the previous two experimental results.

DISCUSSION AND CONCLUSION

This paper proposes a method of predicting DBPs using the XGBoost algorithm and by splicing sequence feature information. The final sequence feature is built from multiple sequence features and spliced by MATLAB. To make the data more standardized and strengthen the relationship between data characteristics and data tags, the data are processed using Z-Score standardization. During the experiment, we used MRMD to reduce the dimensionality of the data and thus reduce the characteristics of the data. We

TABLE 3 | Performance of PDB1075 using different feature extraction methods in XGBoost.

Model name	Feature extraction method	ACC (%)	SN (%)	MCC	Spec (%)
XGboost	GE	66.87	71.17	0.3342	62.09
	MCD	69.04	70.00	0.3975	67.97
	NMBAC	72.14	75.29	0.4404	68.62
	PSSM-AB	76.47	75.29	0.5300	77.77
	PSSM-Pse	74.30	75.88	0.4845	72.54
	PSSM-DWT	74.92	74.70	0.4981	75.16
	The spliced sequence feature	**81.42**	**84.11**	**0.6272**	**78.43**

Bold indicates that their experimental results are the best and the experimental values are the highest.

TABLE 4 | Comparison between the XGBoost model and other methods on the PDB186 data set.

Models	ACC (%)	SN (%)	Spec (%)	MCC
IDNA-Prot|dis	72.0	79.5	64.5	0.445
IDNA-Prot	67.2	67.7	66.7	0.344
DNA-Prot	61.8	69.9	53.8	0.240
DNAbinder	60.8	57.0	64.5	0.216
DBPPre	76.9	79.6	74.2	0.538
IDNAPro-PseAAC	71.5	82.8	60.2	0.442
Kmerl + ACC	71.0	82.8	59.1	0.431
Local-DPP	79.0	92.5	65.6	0.625
DPP-PseAAC	77.4	83.0	70.9	0.550
MSFBinder	79.6	93.6	65.6	0.616
MsDBP	80.1	86.0	74.2	0.606
MKSVM-HKA	81.2	94.6	67.7	0.648
Adilina’s work	82.3	**95.0**	69.9	0.670
XGboost	**85.48**	90.3	**80.6**	**0.713**

Bold indicates that their experimental results are the best and the experimental values are the highest.

[a]*The experimental results of other methods come from (Wei et al., 2017).*

TABLE 5 | Experimental findings for the independent data set PDB2272 using the XGBoost algorithm and other models.

Methods	ACC (%)	MCC	SN (%)	Spec (%)
MK-FSVM-SVDD	76.12	0.5476	**91.50**	60.41
DPP-PseAAC	58.10	0.1625	56.63	59.61
PseDNA-Pro	61.88	0.2430	75.28	48.08
MK-SVM	75.00	0.5264	91.41	58.09
MsDBP	66.99	0.3397	70.69	63.18
XGboost	**78.26**	**0.5652**	80.39	**76.05**

Bold indicates that their experimental results are the best and the experimental values are the highest.

[a]*The experimental results of other methods come from (Du et al., 2019; Zou et al., 2021).*

performed experiments and compared the performance of XGBoost in terms of single sequence feature information and spliced sequence feature information. On the PDB 1075 data set, performance of the spliced sequence feature (MCC: 0.7272) is obviously better than that of the single sequence feature. To further assess our method, we applied the XGBoost model to the PDB186 and PDB2272 data sets. XGBoost produced superior results for PDB186 (MCC: 0.713) and PDB2272 (MCC: 0.5652) compared to available methods.

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 authors.

AUTHOR CONTRIBUTIONS

ZZ and WY designed, planned and implemented the experiment. ZZ also wrote the main part of the article, and YXZ wrote other parts of the article. YL and YMZ participated in the coordination of the study and reviewed the article. All authors read and approved the final article.

FUNDING

This work was supported by the National Natural Science Foundation of China (61971119), The Heilongjiang Postdoctoral Fund (LBH-Q20135).The National Natural Science Foundation of China (NSFC)is a sub ministerial institution in charge of NSFC. NSFC operates relatively independently, and is responsible for the organization and implementation of funding plans, project setting and evaluation, project approval, supervision, etc.

REFERENCES

Adilina, S., Farid, D. M., and Shatabda, S. (2019). Effective DNA Binding Protein Prediction by Using Key Features via Chou’s General PseAAC. *J. Theor. Biol.* 460, 64–78. doi:10.1016/j.jtbi.2018.10.027

Bi, X.-a., Liu, Y., Xie, Y., Hu, X., and Jiang, Q. (2020). Morbigenous Brain Region and Gene Detection with a Genetically Evolved Random Neural Network Cluster Approach in Late Mild Cognitive Impairment. *Bioinformatics* 36 (8), 2561–2568. doi:10.1093/bioinformatics/btz967

Chen, T., and Guestrin, C. (2016). “XGBoost: A Scalable Tree Boosting System,” in The 22nd ACM SIGKDD International Conference.

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a Comprehensive Web-Based Bioinformatics Toolkit for Exploring Disease Associations and ncRNA Function. *Bioinformatics* 34 (11), 1953–1956. doi:10.1093/bioinformatics/bty002

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a Comprehensive Database for Dysbiosis of the Gut Microbiota in Disorders and Interventions. *Nucleic Acids Res.* 48 (D1), D554–D560. doi:10.1093/nar/gkz843

Cheng, L., Shi, H., Wang, Z., Hu, Y., Yang, H., Zhou, C., et al. (2016). IntNetLncSim: an Integrative Network Analysis Method to Infer Human lncRNA Functional Similarity. *Oncotarget* 7 (30), 47864–47874. doi:10.18632/oncotarget.10012

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 (D1), D140–D144. doi:10.1093/nar/gky1051

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational Methods for Identifying Similar Diseases. *Mol. Ther. - Nucleic Acids* 18, 590–604. doi:10.1016/j.omtn.2019.09.019

Dao, F. Y., Lv, H., Su, W., Sun, Z-J., Huang, Q-L., and Lin, H. (2021). iDHS-Deep: an Integrated Tool for Predicting DNase I Hypersensitive Sites by Deep Neural Network. *Brief Bioinform* 22, bbab047. doi:10.1093/bib/bbab047

Ding, Y., Chen, F., Guo, X., Tang, J., and Wu, H. (2020). Identification of DNA-Binding Proteins by Multiple Kernel Support Vector Machine and Sequence Information. *Current Proteomics* 17 (4), 302–310. doi:10.2174/1570164616666190417100509

Ding, Y., Tang, J., and Guo, F. (2020). Human Protein Subcellular Localization Identification via Fuzzy Model on Kernelized Neighborhood Representation. *Appl. Soft Comput.* 96, 106596. doi:10.1016/j.asoc.2020.106596

Ding, Y., Tang, J., and Guo, F. (2020). Identification of Drug-Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion. *Knowledge-Based Syst.* 204, 106254. doi:10.1016/j.knosys.2020.106254

Ding, Y., Tang, J., and Guo, F. (2020). Identification of Drug–Target Interactions via Fuzzy Bipartite Local Model. *Neural Comput. Appl.* 32 (D1), 1–17. doi:10.1007/s00521-019-04569-z

Ding, Y., Tang, J., and Guo, F. (2016). Identification of Protein-Protein Interactions via a Novel Matrix-Based Sequence Representation Model with Amino Acid Contact Information. *Int. J. Mol. Sci.* 17 (10), 1623. doi:10.3390/ijms17101623

Ding, Y., Tang, J., and Guo, F. (2016). Predicting Protein-Protein Interactions via Multivariate Mutual Information of Protein Sequences. *Bmc Bioinformatics* 17 (1), 398. doi:10.1186/s12859-016-1253-9

Ding, Y., Tang, J., and Guo, F. (2019). Protein Crystallization Identification via Fuzzy Model on Linear Neighborhood Representation. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 18, 1986. doi:10.1109/TCBB.2019.2954826

Du, X., Diao, Y., Liu, H., and Li, S. (2019). MsDBP: Exploring DNA-Binding Proteins by Integrating Multiscale Sequence Information via Chou's Five-step Rule. *J. Proteome Res.* 18 (8), 3119–3132. doi:10.1021/acs.jproteome.9b00226

Feng, Z.-P., and Zhang, C.-T. (2000). Prediction of Membrane Protein Types Based on the Hydrophobic index of Amino Acids. *J. Protein Chem.* 19 (4), 269–275. doi:10.1023/a:1007091128394

Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). StackCPPred: a Stacking and Pairwise Energy Content-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency. *Bioinformatics* 36 (10), 3028–3034. doi:10.1093/bioinformatics/btaa131

Guo, Z., Wang, P., Liu, Z., and Zhao, Y. (2020). Discrimination of Thermophilic Proteins and Non-thermophilic Proteins Using Feature Dimension Reduction. *Front. Bioeng. Biotechnol.* 8, 584807. doi:10.3389/fbioe.2020.584807

Han, X., Kong, Q., Liu, C., Cheng, L., and Han, J. (2021). SubtypeDrug: a Software Package for Prioritization of Candidate Cancer Subtype-specific Drugs. *Bioinformatics* 2021, btab011. doi:10.1093/bioinformatics/btab011

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 (4), 1037–1043. doi:10.1093/bioinformatics/btz694

Huang, Y. A., You, Z. H., Gao, X., Wong, L., and Wang, L. (2015). Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence. *Biomed. Res. Int.* 2015, 902198. doi:10.1155/2015/902198

Huang, Y., Zhou, D., Wang, Y., Zhang, X., Su, M., Wang, C., et al. (2020). Prediction of Transcription Factors Binding Events Based on Epigenetic Modifications in Different Human Cells. *Epigenomics* 12 (16), 1443–1456. doi:10.2217/epi-2019-0321

Iqubal, A., Iqubal, M. K., Khan, A., Ali, J., Baboota, S., and Haque, S. E. (2020). Gene Therapy, A Novel Therapeutic Tool for Neurological Disorders: Current Progress, Challenges and Future Prospective. *Curr. Gene Ther.* 20 (3), 184–194. doi:10.2174/1566523220999200716111502

Jeong, J. C., Lin, X., and Chen, X.-W. (2011). On Position-specific Scoring Matrix for Protein Function Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics (Tcbb)* 8 (2), 308. doi:10.1109/tcbb.2010.93

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting Human microRNA-Disease Associations Based on Support Vector Machine. *Int. J. Data Min Bioinform* 8 (3), 282–293. doi:10.1504/ijdmb.2013.056078

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting Human microRNA-Disease Associations Based on Support Vector Machine. *Int. J. Data Min Bioinform* 8 (3), 282–293. doi:10.1504/ijdmb.2013.056078

Jin, S., Zeng, X., Fang, J., Lin, J., Chan, S. Y., Erzurum, S. C., et al. (2019). A Network-Based Approach to Uncover microRNA-Mediated Disease Comorbidities and Potential Pathobiological Implications. *NPJ Syst. Biol. Appl.* 5 (1), 41–11. doi:10.1038/s41540-019-0115-2

Jin, S., Zeng, X., Xia, F., Huang, W., and Liu, X. (2021). Application of Deep Learning Methods in Biological Networks. *Brief. Bioinform.* 22 (2), 1902–1917. doi:10.1093/bib/bbaa043

Kumar, K. K., Pugalenthi, G., and Suganthan, P. N. (2009). DNA-prot: Identification of DNA Binding Proteins from Protein Sequence Information Using Random Forest. *J. Biomol. Struct. Dyn.* 26 (6), 679–686. doi:10.1080/07391102.2009.10507281

Kumar, M., Gromiha, M. M., and Raghava, G. P. (2007). Identification of DNA-Binding Proteins Using Support Vector Machines and Evolutionary Profiles. *Bmc Bioinformatics* 8, 463. doi:10.1186/1471-2105-8-463

Li, H., Long, C., Xiang, J., Liang, P., Li, X., and Zuo, Y. (2020). Dppa2/4 as a Trigger of Signaling Pathways to Promote Zygote Genome Activation by Binding to CG-Rich Region. *Brief Bioinform* 22, bbaa342. doi:10.1093/bib/bbaa342

Li, H., Ta, N., Long, C., Zhang, Q., Li, S., liu, S., et al. (2019). The Spatial Binding Model of the pioneer Factor Oct4 with its Target Genes during Cell Reprogramming. *Comput. Struct. Biotechnol. J.* 17, 1226–1233. doi:10.1016/j.csbj.2019.09.002

Li, X., Liao, B., Shu, Y., Zeng, Q., and Luo, J. (2009). Protein Functional Class Prediction Using Global Encoding of Amino Acid Sequence. *J. Theor. Biol.* 261 (2), 290–293. doi:10.1016/j.jtbi.2009.07.017

Lin, W. Z., Fang, J. A., Xiao, X., and Chou, K. C. (2011). iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *Plos One* 6 (9), e24756. doi:10.1371/journal.pone.0024756

Liu, B., Wang, S., and Wang, X. (2015). DNA Binding Protein Identification by Combining Pseudo Amino Acid Composition and Profile-Based Protein Representation. *Sci. Rep.* 5, 15479. doi:10.1038/srep15479

Liu, B., Gao, X., and Zhang, H. (2019). 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 (20), e127. doi:10.1093/nar/gkz740

Liu, B., Wang, S., Dong, Q., Li, S., and Liu, X. (2016). Identification of DNA-Binding Proteins by Combining Auto-Cross Covariance Transformation and Ensemble Learning. *IEEE Trans.on Nanobioscience* 15 (4), 328–334. doi:10.1109/tnb.2016.2555951

Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., et al. (2014). iDNA-Prot Vertical Bar 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 (9), e106691. doi:10.1371/journal.pone.0106691

Liu, B., Xu, J., Fan, S., Xu, R., Zhou, J., and Wang, X. (2015). PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation. *Mol. Inf.* 34 (1), 8–17. doi:10.1002/minf.201400025

Liu, D., Li, G., and Zuo, Y. (2019). Function Determinants of TET Proteins: the Arrangements of Sequence Motifs with Specific Codes. *Brief Bioinform* 20 (5), 1826–1835. doi:10.1093/bib/bby053

Liu, G., Jin, S., Hu, Y., and Jiang, Q. (2018). Disease Status Affects the Association between Rs4813620 and the Expression of Alzheimer's Disease Susceptibility geneTRIB3. *Proc. Natl. Acad. Sci. USA* 115 (45), E10519–E10520. doi:10.1073/pnas.1812975115

Liu, H., Ren, G., Chen, H., Liu, Q., Yang, Y., and Zhao, Q. (2020). Predicting lncRNA-miRNA Interactions Based on Logistic Matrix Factorization with

Neighborhood Regularized. *Knowledge-Based Syst.* 191, 105261. doi:10.1016/j.knosys.2019.105261

Liu, X. J., Gong, X. J., Yu, H., and Xu, J. H. (2018). A Model Stacking Framework for Identifying DNA Binding Proteins by Orchestrating Multi-View Features and Classifiers. *Genes (Basel)* 9 (8). doi:10.3390/genes9080394

Liu, Y., Huang, Y., Wang, G., and Wang, Y. (2020). A Deep Learning Approach for Filtering Structural Variants in Short Read Sequencing Data. *Brief Bioinform* 22, bbaa370. doi:10.1093/bib/bbaa370

Liu, Y., Zhang, X., Zou, Q., and Zeng, X. (2020). Minirmd: Accurate and Fast Duplicate Removal Tool for Short Reads via Multiple Minimizers. *Bioinformatics* 37, 1604–1606. doi:10.1093/bioinformatics/btaa915

Lou, W., Wang, X., Chen, F., Chen, Y., Jiang, B., and Zhang, H. (2014). Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naive Bayes. *Plos One* 9 (1), 86703. doi:10.1371/journal.pone.0086703

Nanni, L., Brahnam, S., and Lumini, A. (2012). Wavelet Images and Chou's Pseudo Amino Acid Composition for Protein Classification. *Amino Acids* 43 (2), 657–665. doi:10.1007/s00726-011-1114-9

Niu, M., Zhang, J., Li, Y., Wang, C., Liu, Z., Ding, H., et al. (2020). CirRNAPL: A Web Server for the Identification of circRNA Based on Extreme Learning Machine. *Comput. Struct. Biotechnol. J.* 18, 834–842. doi:10.1016/j.csbj.2020.03.028

Quan, Z., Zenga, J., Caoa, L., and Jia, 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

Rahman, M. S., Shatabda, S., Saha, S., Kaykobad, M., and Rahman, M. S. (2018). DPP-PseAAC: A DNA-Binding Protein Prediction Model Using Chou's General PseAAC. *J. Theor. Biol.* 452, 22–34. doi:10.1016/j.jtbi.2018.05.006

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 (7), 2931–2939. doi:10.1021/acs.jproteome.9b00250

Shang, Y., Gao, L., Zou, Q., and Yu, L. (2021). Prediction of Drug-Target Interactions Based on Multi-Layer Network Representation Learning. *Neurocomputing* 434, 80–89. doi:10.1016/j.neucom.2020.12.068

Shao, J., and Liu, B. (2021). ProtFold-DFG: Protein Fold Recognition by Combining Directed Fusion Graph and PageRank Algorithm. *Brief. Bioinform.* 22, bbaa192. doi:10.1093/bib/bbaa192

Shao, J., Yan, K., and Liu, B. (2021). FoldRec-C2C: Protein Fold Recognition by Combining Cluster-To-Cluster Model and Protein Similarity Network. *Brief. Bioinform.* 22, bbaa144. doi:10.1093/bib/bbaa144

Shen, Y., Ding, Y., Tang, J., Zou, Q., and Guo, F. (2019). Critical Evaluation of Web-Based Prediction Tools for Human Protein Subcellular Localization. *Brief. Bioinformatics* 21, 1628. doi:10.1093/bib/bbz106

Shen, Y., Ding, Y., Tang, J., Zou, Q., and Guo, F. (2020). Critical Evaluation of Web-Based Prediction Tools for Human Protein Subcellular Localization. *Brief. Bioinform.* 21 (5), 1628–1640. doi:10.1093/bib/bbz106

Shen, Y., Tang, 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

Tang, Y.-J., Pang, Y.-H., and Liu, B. (2020). IDP-Seq2Seq: Identification of Intrinsically Disordered Regions Based on Sequence to Sequence Learning. *Bioinformaitcs* 36 (21), 5177–5186. doi:10.1093/bioinformatics/btaa667

Tao, Z., Li, Y., Teng, Z., and Zhao, Y. (2020). A Method for Identifying Vesicle Transport Proteins Based on LibSVM and MRMD. *Comput. Math. Methods Med.* 2020, 8926750. doi:10.1155/2020/8926750

Tao, Z., Li, Y., Teng, Z., and Zhao, Y. (2020). A Method for Identifying Vesicle Transport Proteins Based on LibSVM and MRMD. *Comput. Math. Methods Med.* 2020, 8926750. doi:10.1155/2020/8926750

Wang, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of Membrane Protein Types via Multivariate Information Fusion with Hilbert-Schmidt Independence Criterion. *Neurocomputing* 383, 257–269. doi:10.1016/j.neucom.2019.11.103

Wang, H., Jijun, T., Ding, Y., and Guo, F. (2021). Exploring Associations of Non-coding RNAs in Human Diseases via Three-Matrix Factorization with Hypergraph-Regular Terms on center Kernel Alignment. *Brief. Bioinform.* 22, bbaa409. doi:10.1093/bib/bbaa409

Wang, H., Liang, P., Zheng, L., Long, C. S., Li, H. S., Zuo, Y., et al. (2021). eHSCPr Discriminating the Cell Identity Involved in Endothelial to Hematopoietic Transition. *Bioinformatics* 37, 2157. doi:10.1093/bioinformatics/btab071

Wang, H., Yijie, D., Tang, J., Zou, Q., and Guo, F. (2021). Identify RNA-Associated Subcellular Localizations Based on Multi-Label Learning Using Chou's 5-steps Rule. *BMC Genomics* 22 (56), 1. doi:10.1186/s12864-020-07347-7

Wang, J., Wang, H., Wang, X., and Chang, H. (2020). Predicting Drug-Target Interactions via FM-DNN Learning. *Curr. Bioinformatics* 15 (1), 68–76. doi:10.2174/1574893614666190227160538

Wang, S., Wang, Y., Yu, C., Cao, Y., Yu, Y., Pan, Y., et al. (2020). Characterization of the Relationship between FLI1 and Immune Infiltrate Level in Tumour Immune Microenvironment for Breast Cancer. *J. Cel Mol Med* 24 (10), 5501–5514. doi:10.1111/jcmm.15205

Wang, Y., Ding, Y., Tang, J., Dai, Y., and Guo, F. (2021). CrystalM: A Multi-View Fusion Approach for Protein Crystallization Prediction. *Ieee/acm Trans. Comput. Biol. Bioinform* 18 (1), 325–335. doi:10.1109/TCBB.2019.2912173

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 (4), 282–294. doi:10.2174/1574893614666190304125221

Wei, L., Chen, H., and Su, R. (2018). 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. (2018). Prediction of Human Protein Subcellular Localization Using Deep Learning. *J. Parallel Distributed Comput.* 117, 212–217. doi:10.1016/j.jpdc.2017.08.009

Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2020). Comparative Analysis and Prediction of Quorum-sensing Peptides Using Feature Representation Learning and Machine Learning Algorithms. *Brief. Bioinform.* 21 (1), 106–119. doi:10.1093/bib/bby107

Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. *Ieee/acm Trans. Comput. Biol. Bioinf.* 11 (1), 192–201. doi:10.1109/tcbb.2013.146

Wei, L., Tang, J., and Zou, Q. (2017). Local-DPP: An Improved DNA-Binding Protein Prediction Method by Exploring Local Evolutionary Information. *Inf. Sci.* 384, 135–144. doi:10.1016/j.ins.2016.06.026

Yang, C., Ding, Y., Meng, Q., Tang, J., and Guo, F. (2021). Granular Multiple Kernel Learning for Identifying RNA-Binding Protein Residues via Integrating Sequence and Structure Information. *Neural Comput. Appl.* 33, 11387. doi:10.1007/s00521-020-05573-4

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big Data Mining with Fusion of Multifarious Physical Examination Indicators. *Inf. Fusion* 75, 140–149. doi:10.1016/j.inffus.2021.02.015

You, Z. H., Zhu, L., Zheng, C. H., Yu, H. J., Deng, S. P., and Ji, Z. (2014). Prediction of Protein-Protein Interactions from Amino Acid Sequences Using a Novel Multi-Scale Continuous and Discontinuous Feature Set. *Bmc Bioinformatics* 15 (Suppl. 15), S9. doi:10.1186/1471-2105-15-S15-S9

Yu, L., Shi, Y., Zou, Q., Wang, S., Zheng, L., and Gao, L. (2020). Exploring Drug Treatment Patterns Based on the Action of Drug and Multilayer Network Model. *Int. J. Mol. Sci.* 21 (14), 5014. doi:10.3390/ijms21145014

Yu, L., Wang, M., Yang, Y., Xu, F., Zhang, X., Xie, F., et al. (2021). Predicting Therapeutic Drugs for Hepatocellular Carcinoma Based on Tissue-specific Pathways. *Plos Comput. Biol.* 17 (2), e1008696. doi:10.1371/journal.pcbi.1008696

Yu, L., Zhou, D., Gao, L., and Zha, Y. (2020). Prediction of Drug Response in Multilayer Networks Based on Fusion of Multiomics Data. *Methods* 192, 85. doi:10.1016/j.ymeth.2020.08.006

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020). Target Identification Among Known Drugs by Deep Learning from Heterogeneous Networks. *Chem. Sci.* 11 (7), 1775–1797. doi:10.1039/c9sc04336e

Zhai, Y., Chen, Y., Teng, Z., and Zhao, Y. (2020). Identifying Antioxidant Proteins by Using Amino Acid Composition and Protein-Protein Interactions. *Front. Cel Dev. Biol.* 8, 591487. doi:10.3389/fcell.2020.591487

Zhang, C.-H., Li, M., Lin, Y.-P., and Gao, Q. (2020). Systemic Therapy for Hepatocellular Carcinoma: Advances and Hopes. *Curr. Gene Ther.* 20 (2), 84–99. doi:10.2174/1566523220666200628014530

Zhang, D., Chen, H. D., Zulfiqar, H., Yuan, S. S., Huang, Q. L., Zhang, Z. Y., et al. (2021). iBLP: An XGBoost-Based Predictor for Identifying Bioluminescent Proteins. *Comput. Math. Methods Med.* 2021, 6664362. doi:10.1155/2021/6664362

Zhang, J., Zhang, Z., Pu, L., Tang, J., and Guo, F. (2020). AIEpred: an Ensemble Predictive Model of Classifier Chain to Identify Anti-inflammatory Peptides. *Ieee/acm Trans. Comput. Biol. Bioinform* 18, 1831. doi:10.1109/TCBB.2020.2968419

Zhang, Z., Ding, J., Xu, J., Tang, J., and Guo, F. (2021). Multi-Scale Time-Series Kernel-Based Learning Method for Brain Disease Diagnosis. *IEEE J. Biomed. Health Inform.* 25 (1), 209–217. doi:10.1109/jbhi.2020.2983456

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36, 4466. doi:10.1093/bioinformatics/btaa428

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an Ensemble Classifier-Based Feature Selection for Differential Expression Analysis on Expression Profiles. *BMC Bioinformatics* 21 (1), 43. doi:10.1186/s12859-020-3388-y

Zhao, X., Wang, H., Li, H., Wu, Y., and Wang, G. (2021). Identifying Plant Pentatricopeptide Repeat Proteins Using a Variable Selection Method. *Front. Plant Sci.* 12, 506681. doi:10.3389/fpls.2021.506681

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 (Oxford)* 2019, baz131. doi:10.1093/database/baz131

Zheng, L., Liu, D., Yang, W., Yang, L., and Zuo, Y. (2020). RaacLogo: a New Sequence Logo Generator by Using Reduced Amino Acid Clusters. *Brief Bioinform* 22, bbaa096. doi:10.1093/bib/bbaa096

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. *Knowledge-Based Syst.* 163, 787–793. doi:10.1016/j.knosys.2018.10.007

Zhu, Y., Li, F., Xiang, D., Akutsu, T., Song, J., and Jia, C. (2020). Computational Identification of Eukaryotic Promoters Based on Cascaded Deep Capsule Neural Networks. *Brief. Bioinform.* 22, bbaa299. doi:10.1093/bib/bbaa299

Zou, Y., Wu, H., Guo, X., Peng, L., Ding, Y., Tang, J., et al. (2021). MK-FSVM-SVDD: A Multiple Kernel-Based Fuzzy SVM Model for Predicting DNA-Binding Proteins via Support Vector Data Description. *Curr. Bioinformatics* 16 (2), 274–283. doi:10.2174/1574893615999200607173829

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 (1), 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 February 2022
doi: 10.3389/fgene.2022.816825

Evaluation of CircRNA Sequence Assembly Methods Using Long Reads

Jingjing Zhang[1,2], Md. Tofazzal Hossain[1,2], Weiguo Liu[3], Yin Peng[4], Yi Pan[2] and Yanjie Wei[2,5]**

[1]University of Chinese Academy of Sciences, Beijing, China, [2]Centre for High Performance Computing, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, [3]School of Software, Shandong University, Jinan, China, [4]Guangdong Key Laboratory for Genome Stability and Disease Prevention and Regional Immunity and Diseases, Department of Pathology, Shenzhen University School of Medicine, Shenzhen, China, [5]CAS Key Laboratory of Health Informatics, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Hongmin Cai,
South China University of Technology, China
Lei Deng,
Central South University, China

****Correspondence:***
Yin Peng
ypeng@szu.edu.cn
Yanjie Wei
yj.wei@siat.ac.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *17 November 2021*
Accepted: *17 January 2022*
Published: *14 February 2022*

Citation:
Zhang J, Hossain MT, Liu W, Peng Y, Pan Y and Wei Y (2022) Evaluation of CircRNA Sequence Assembly Methods Using Long Reads. Front. Genet. 13:816825. doi: 10.3389/fgene.2022.816825

The functional study on circRNAs has been increasing in the past decade due to its important roles in micro RNA sponge, protein coding, the initiation, and progression of diseases. The study of circRNA functions depends on the full-length sequences of circRNA, and current sequence assembly methods based on short reads face challenges due to the existence of linear transcript. Long reads produced by long-read sequencing techniques such as Nanopore technology can cover full-length sequences of circRNA and therefore can be used to evaluate the correctness and completeness of circRNA full sequences assembled from short reads of the same sample. Using long reads of the same samples, one from human and the other from mouse, we have comprehensively evaluated the performance of several well-known circRNA sequence assembly algorithms based on short reads, including circseq_cup, CIRI_full, and CircAST. Based on the F1 score, the performance of CIRI-full was better in human datasets, whereas in mouse datasets CircAST was better. In general, each algorithm was developed to handle special situations or circumstances. Our results indicated that no single assembly algorithm generated better performance in all cases. Therefore, these assembly algorithms should be used together for reliable full-length circRNA sequence reconstruction. After analyzing the results, we have introduced a screening protocol that selects out exonic circRNAs with full-length sequences consisting of all exons between back splice sites as the final result. After screening, CIRI-full showed better performance for both human and mouse datasets. The average F1 score of CIRI-full over four circRNA identification algorithms increased from 0.4788 to 0.5069 in human datasets, and it increased from 0.2995 to 0.4223 in mouse datasets.

Keywords: circRNA, full-length sequences, short reads, long reads, assembly

INTRODUCTION

Only recently has circular RNA (circRNA) appeared as a hot research topic since it was first discovered in the 1970s (Sanger et al., 1976; Arnberg et al., 1980; Kos et al., 1986). Different from linear RNAs, the special covalent circular structure of circRNA is formed by back splicing (Jeck et al., 2013). Identifying the back splice sites is the most important factor for circRNA identification from the sequencing reads (Kristensen et al., 2019). Based on sequencing data, various identification algorithms were developed, such as find_circ (Memczak et al., 2013), KNIFE (Szabo et al., 2015), CIRI (Y. Gao et al., 2015), and PCirc (Yin et al., 2021), some of which require annotation information of genome sequences to improve identification sensitivity and reduce the false discovery rate (FDR) (Memczak et al., 2013; Baruzzo et al., 2017).

As more and more circRNAs were discovered in animals and plants in recent years (Glažar et al., 2014; J.; Zhang et al., 2020a), new functions of circRNAs in the organism have also been discovered. Acting as micro RNA (miRNA) sponge is mostly studied for circRNAs, and circRNAs regulate expression of miRNA target gene indirectly (Piwecka et al., 2017). Hansen *et al* found that exonic circRNA CDR1as can bind with miR-671, which can degrade CDR1as mediated by AGO (Hansen et al., 2013), and the binding sites are highly conserved. In addition, circRNAs can also interact with RNA binding proteins as endogenous competitive RNA (S. Zheng et al., 2021). The gene *muscleblind* (*MBL*) of Drosophila can encode MBL protein as a transcript factor, and MBL regulates the dynamic balance of circular transcript (circRNA circMbl) and linear transcript (Ashwal-Fluss et al., 2014). Although circRNAs were considered to be noncoding RNAs (Qu et al., 2015), some circRNAs have been found to translate proteins (Shi et al., 2020). For example, circRNA circPINT can translate into protein PINT-87aa for inhibiting malignant glioma (M. Zhang et al., 2018). Another circRNA, circE7, derived from oncogenic human papilloma viruses (HPVs), is found to produce E7 oncoprotein with modified N6-methyladenosine (m6A) (Zhao et al., 2019).

For the study of circRNA functions, sequence information is vital. Due to its special structure, it is difficult to obtain correct and complete sequences of circRNAs (full-length sequences) directly. Reconstruction of circRNAs full-length sequences was effected by linear transcripts (Szabo & Salzman, 2016). Computational tools such as circseq_cup (Ye et al., 2017), CIRI-full (Y. Zheng et al., 2019), and CircAST (Wu et al., 2019) were developed to assemble full-length sequences for circRNAs according to short reads (next-generation sequencing data and RNA-Seq data).

circseq_cup predicts circRNAs and constructs full-length sequences based on paired-end (PE) short reads. This method first relies on an alignment software (TopHat-Fusion, STAR-Fusion, or segemehl (Kim & Salzberg, 2011; Dobin et al., 2013; Hoffmann et al., 2014)) to identify fusion junction sites. The construction of the virtual reference sequence concatenates sequences between fusion junction sites. Full-length sequences of circRNAs were assembled by PE reads that could align to the middle of virtual reference sequences. Then, some criteria were used to filter out false-positive circRNAs, such as sequences supported by less than two pairs of PE reads. CIRI-full introduces a new feature named reverse overlap (RO) for assembling candidate circRNA sequences. Back-splice junctions (BSJs) are PE reads that are aligned to back splice sites which support the identification of circRNA. If RO reads or BSJ reads can cover all cirexons (circRNA's exon) between back splice sites, the complete sequences of circRNA can be assembled by connecting the cirexons. Otherwise, a combined strategy based on both RO reads and BSJ reads were used to reconstruct circRNA full-length sequences. Performance improvement of CIRI-full relies on longer reads, such as longer than 250 bp. CircAST assembles circRNA full-length sequences with mapped fragments using a multiple splice graph model. Each transcript was represented by a directed acyclic graph (DAG), exons between back splice sites represent the nodes on the graph, and directed edges on the graph indicate the mapped reads mapped on these two different exons. Source node and sink node should be the exons mapped by the fragments of back splice reads of circular transcript. In addition, CircAST is an annotated-based method and shows better performance on shorter read lengths (from 75 bp to 125 bp). For all the software/methods, the correctness and completeness of the constructed circRNA sequences are difficult to evaluate. Assembly software based on short reads could only reconstruct full-length sequences for some circRNAs due to the interference of linear transcripts, and some assembled circRNA full-length sequences are false positive due to the same reason (X. Li et al., 2020).

Long-read sequencing, such as Nanopore sequencing, is capable of generating longer lengths, between 5,000 and 30,000 base pairs (van Dijk et al., 2018). Long reads have a higher error rate (10–15%), but these sequencing errors are randomly distributed; the rates can therefore be greatly reduced through the use of circular consensus sequencing (Larsen et al., 2014). This makes direct sequencing the full-length sequences of circRNAs possible since the length of most circRNAs under study is shorter than 5,000 bp (Z. Gao et al., 2019; J. Zhang et al., 2020b). Thus, by using long-read sequencing results of a sample, it is possible to evaluate the quality of assembled circRNA full-length sequences based on the short read sequencing results of the same sample.

In this study, we used three evaluation strategies (read alignment, CIRI-long, and isoCirc; see in Method) based on long reads to verify the quality of full-length sequences assembled based on short reads. In our results, each assembly algorithm showed its own advantage; in CircAST and circseq_cup, the precision was high but the sensitivity was low, whereas in CIRI-full, the precision was low but the sensitivity was high. CIRI-full performed better (F1 score, read alignment: 0.6348, CIRI-long: 0.4093, isoCirc: 0.5965) in *Homo sapiens* (human) datasets, while CircAST was the better performer in *Mus musculus* (mouse) datasets (F1 score, read alignment: 0.4112, CIRI-long: 0.4733, isoCirc: 0.3212). Among these assembly tools, CIRI-full assembled more circRNA full-length sequences with less than 57% of precision in human datasets, while circseq_cup and CircAST assembled few circRNAs full-length sequences with about 80% of precision in human datasets. After careful analysis, we have introduced a screening protocol that selects out exonic circRNAs with full-length sequences consisting of all exons between back splice sites as the final result. After screening, CIRI-full showed the best performance for both human and mouse datasets.

MATERIALS AND METHODS

Data Collection

RNA-seq libraries (short reads; next-generation sequencing data) were downloaded from the Sequence Reads Archive (accession

ID: SRR10612068, SRR10612069, and SRR10612070) and the National Genomics Data Center (https://bigd.big.ac.cn/gsa) (accession ID: CRR194214 and CRR194215). Nanopore libraries (long reads; third-generation sequencing data) were downloaded from the Sequence Reads Archive (accession ID: SRR10612050, SRR10612051, SRR10612052, SRR10612053, SRR10612054, and SRR10612055) and the National Genomics Data Center (accession ID: CRR194190, CRR194191, CRR194194, and CRR194195). Short reads and long reads from the same database were derived from the same experiment samples. Sequencing data downloaded from the SRA were all derived from the cultured HEK293 cells, and data downloaded from the NGDC were derived from adult mice. Table S1 provides a summary of all of the datasets. The reference genomes of human (GRCh38/hg38) and mouse (GRCm38/mm10) were downloaded from UCSC.

Identification of circRNA and Recontruction of circRNA Full-Length Sequence Based on Short Reads

For analysis of short reads, sequencing reads were mapped to the genome using BWA (H. Li & Durbin, 2009), STAR (Dobin et al., 2013), and Tophat2 (Kim et al., 2013) with default parameters. Four tools, including CIRI2 (v2.0.6) (Y. Gao et al., 2018), CIRCexplorer2 (v2.3.5) (X. O. Zhang et al., 2014), circRNA_finder (v1.1) (Westholm et al., 2014), and find_circ (v1.2) (Memczak et al., 2013), were used for circRNA identification following the instructions of the software documentation. The identified circRNAs were selected with at least two back splice reads which were aligned to the circRNA junction sites.

Three pieces of software, circseq_cup, CIRI-full, and CircAST, were used for reconstruction of full-length sequences of circRNA with default parameters. Among them, CIRI-full and CircAST both require information of identified circRNA and sequencing reads as input, while circseq_cup only needs sequencing reads as input. Thus, for each short reads sequencing data, nine different results of full-length sequences are generated using different strategies, due to different combinations of identification algorithms and assembly algorithms.

Evaluation of circRNA Full-Length Sequences Using Long Reads

Long reads data are a cluster of long-read sequences, most of which are longer than the full sequences of circRNA. One could assess whether circRNAs full-length sequences (most of their length <1,000 bp) that were reconstructed based on short reads are correct according to long-read sequences, given that both short reads and long reads are derived from the same samples.

In this study, we have used three strategies based on long reads to evaluate the assembled circRNA full-length sequences using the short reads (**Figure 1**).

The correctness of the assembled sequence is evaluated using three strategies as shown in **Figure 1**. For strategy 1, isoCirc was used to determine the full-length circRNA isoforms from long reads. A sequence reconstructed from short reads was considered correct if it was similar to any one of the sequences of isoCirc results. Similarly, for strategy 2, CIRI-long was used to reconstruct full-length circRNA sequences using long reads.

Another evaluation strategy (strategy 3) used long reads to evaluate the correctness of the assembled circRNA sequences directly. Three main steps of strategy three were 1) we moved a 20 bp fragment on the upstream of the full-length sequence to the end of the full-length sequence, which forms a new full-length sequence with back splice sites; 2) long reads were mapped to the new full-length sequences of circRNAs using minimap2 (H. Li, 2018) with default parameters (-a); 3) for each alignment, mapped_ratio (M/L, where M is the number of mapped bases, and L is the number of bases of circRNA full-length sequences) was calculated; and 4) we discarded any alignment record with mapped_ratio >1 or <0.8, or they contained more than two bp mismatch, insertion, or deletion.

Evaluation Metrics

In all evaluation strategies, full-length circRNAs that were verified correct by long reads were defined as true positives, while those not verified by long reads were defined as false positives. Full-length circRNAs were verified correct in other assembly strategies, but those not assembled in the currently evaluated assembly strategy were defined as false negatives. The assembly performance is assessed using precision, sensitivity, and F1 score and defined as follows:

$$precision = \frac{TP}{TP + FP}$$

$$sensitivity = \frac{TP}{TP + FN}$$

$$F1 = \frac{2*precision*sensitivity}{precision + sensitivity}$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives. F1 score weights precision and sensitivity equally and serves as a balanced metric to evaluate whether a tool achieves favorable precision and sensitivity simultaneously.

RESULTS

Identification of circRNAs Based on Short Reads

Several identification algorithms have been developed for circRNA identification based on short reads. In this study, we selected four algorithms to identify circRNA in human and mouse datasets, including CIRI, CIRCexplorer, circRNA_finder, and find_circ. Among the identified circRNAs, 13,027 (31.60%) were observed between all four algorithms (**Figure 2A**), while 11,890 (28.80%) were only found by a single algorithm. A total of 25,634 distinct circRNAs candidates were identified by CIRI, 23,763 (92.70%) of which were generated from exons, and the remaining were generated from introns or intergenic regions. For circRNA_finder and find_circ, 25,925 and 29,828 circRNAs were identified, respectively. Similarly, most of these circRNAs were derived from exons; only less than 10% were derived from introns and

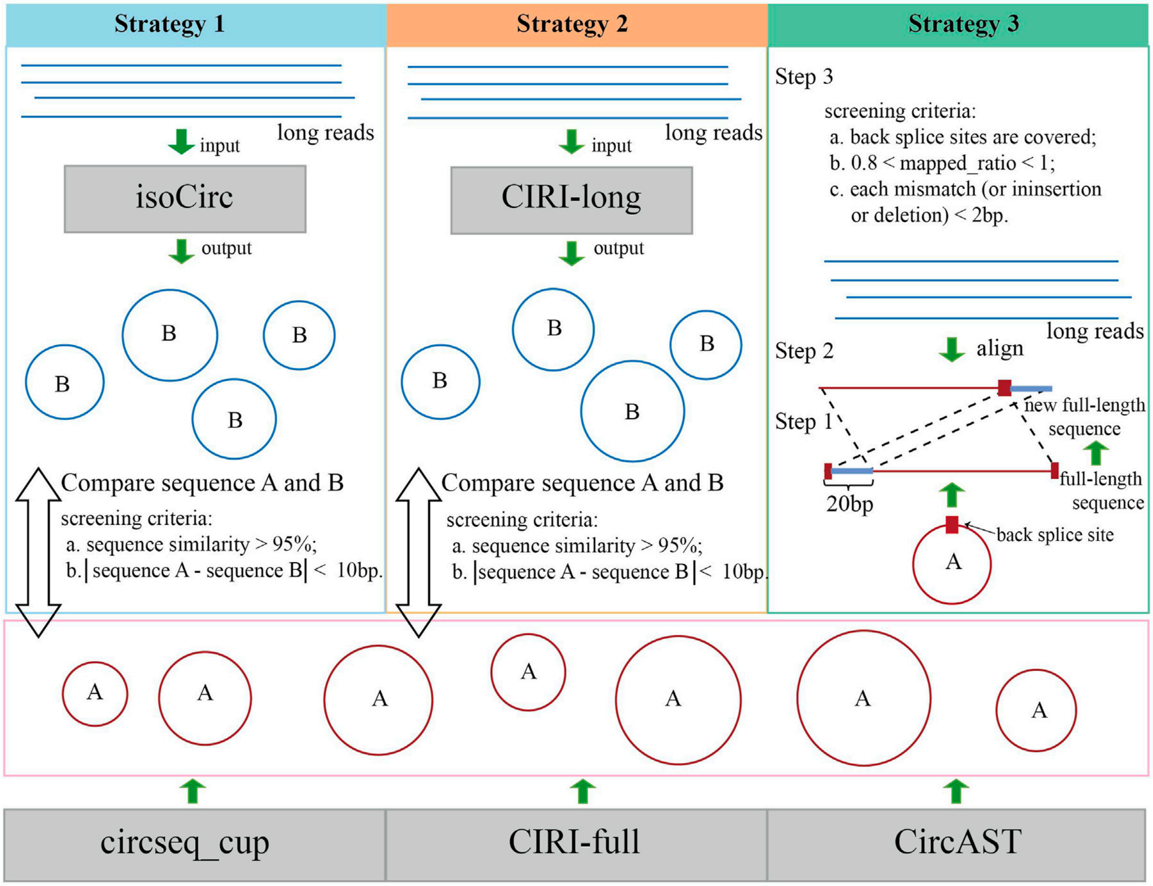

FIGURE 1 | Evaluation of circRNA full-length sequences using long reads. Blue lines and circles **(B)** represent long reads or circRNAs identified using long reads; red lines and circles **(A)** represent assembled full-length sequence and circRNAs identified using short reads.

intergenic regions. However, among the circRNAs identified by CIRCexplorer, 23,304 (99.08%) were exonic, and 217 (0.92%) were intronic, but they were no intergenic circRNAs (**Figure 2B**). The number of circRNA candidates in each sample is shown in Table S2. By counting the number of back splice reads, 71.50% of circRNAs were supported by less than five back splice reads (**Figure 2C**), which agreed with the fact that circRNAs usually showed lower expression than linear transcripts (X. Li et al., 2018). CIRI produced a larger average number of back splice reads per circRNA in human and mouse than other algorithms (**Figure 2D**). In our results, more circRNAs were identified from mouse than human (Table S2), and circRNAs in mouse were supported by more back splice reads than in human (**Figure 2D**); these phenomena can be attributed to longer reads length (human: 101 bp and mouse: 151 bp) and greater sequencing depth of mouse datasets (**Supplementary Table S1**).

Reconstruction of circRNA Full-Length Sequences Using Short Reads

Full-length sequences are important to analyze the function of circRNAs, such as miRNA sponges, RBP sites, and expression. Three popular methods, circseq_cup, CircAST, and CIRI-full, were used in this study for reconstructing full-length sequences of circRNA for short reads datasets.

As shown in **Figrue 3** (A and B), less than 5% of the full-length circRNAs (circRNA that has the assembled full-length sequence) were common among all the three assembly tools for human and mouse datasets, and more than 95% of the reconstructed sequences of these pieces of software/methods were different. Thus, it is difficult for experimental biologists to select the circRNA sequences, and the functional study of circRNAs could be unreliable due to the wrongly selected circRNA sequences.

Among three assembly tools, full-length circRNAs assembled using CIRI-full were more than those assembled using CircAST and circseq_cup. For example, for the circRNA identification result of CIRI on sample SRR10612068, 300 (6.21%) and 1868 (38.69%) full-length circRNAs were assembled using CircAST and CIRI-full, whereas circsesq_cup identified 323 full-length circRNAs for sample SRR10612068 (**Table 1** and **Supplementary Table S2**). In addition, some unique circRNAs that were only identified using a single circRNA identification algorithm were reconstructed successfully

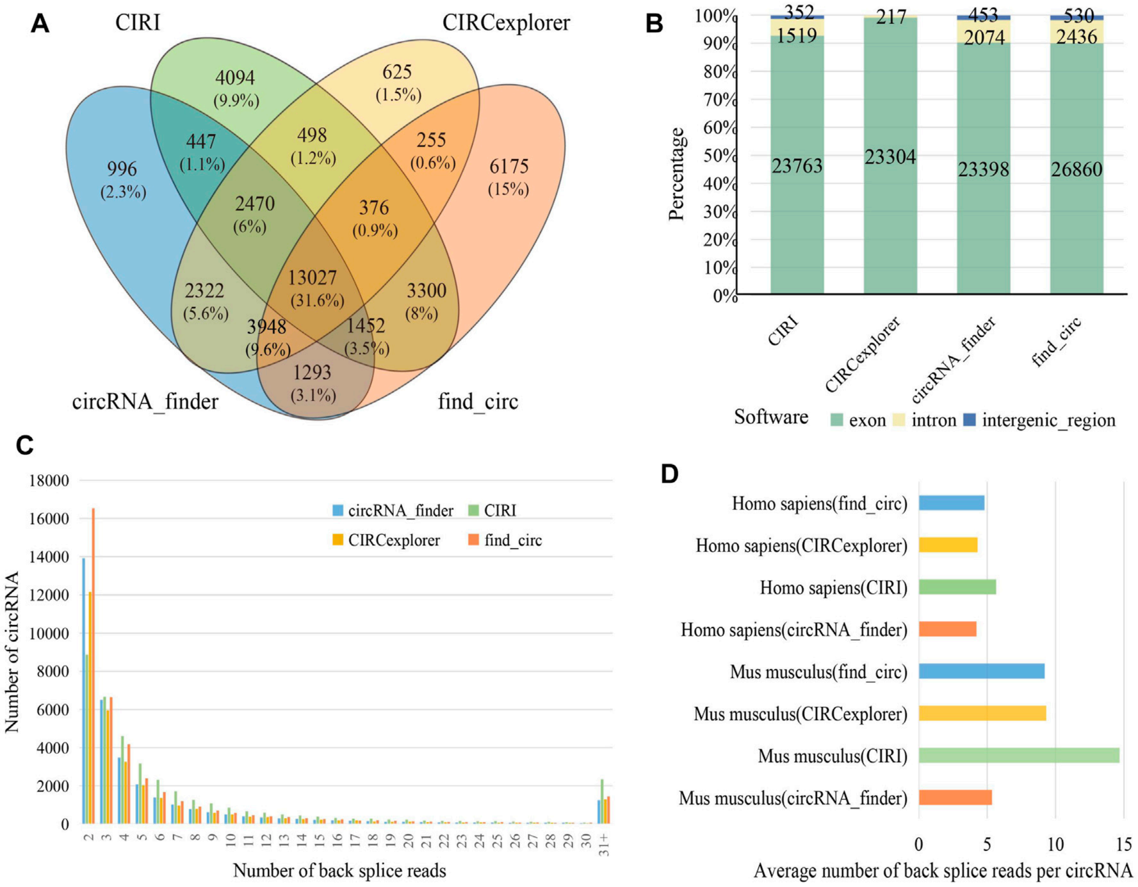

FIGURE 2 | Identification and characterization of circRNAs. **(A)** Venn diagram depicting the overlap between the four different circRNA identification algorithms. **(B)** The percentage of different genomic origins of circRNA. **(C)** The distribution of back splice reads number in four identification algorithms. **(D)** Barplot showing average number of back splice reads per circRNA.

(**Supplementary Figure S1**), indicating that the selection of circRNA identification software had impact on CircAST and CIRI-full. Using CIRI as a circRNA identification tool, CircAST and CIRI-full generated more circRNA full-length sequences than other identification tools (CIRCexplorer, circRNA_finder, and find_circ). For common circRNA candidates in four circRNA identification algorithms, most full-length circRNAs (60%–90%) produced by CircAST and circseq_cup were constructed from the common candidates, while less than half of full-length circRNAs by CIRI-full were involved in common candidates (**Figure 3C**). It was found that the lengths of most full-length circRNAs were shorter than 1,000 bp (**Figure 3D**). CircAST can assemble longer sequences for human and mouse, which is consistent with the advantage of CircAST that it can assemble long circRNAs without using long sequencing reads. However, the performance of CIRI-full was not consistent in PE100 and PE150 (**Figure 3D**). Origin also is an important factor in reconstructing full-length sequences; most full-length circRNAs (94%) were derived from the exon region on the genome in our results (**Figure 3E**), which can be explained by the following: first, more than 90% circRNA candidates belong to exonic circRNAs and second, exonic circRNAs were usually supported by more back splice reads.

Evaluation of Different Sequence Assembly Strategies From Short Reads

There are three assembly tools for assembly of circRNA full-length sequences from short reads, but it is unknown which one has the best performance. Here, we used three evaluation strategies (read alignment, CIRI-long, and isoCirc) to evaluate the performance of nine assembly strategies due to different combinations of circRNA identification software (CIRI, CIRIexplorer, circRNA_finder, and find_circ) and assembly tools (circseq_cup, CIRI_full, and CircAST).

As shown in **Figure 4**, circseq_cup showed different precision (56.57–89.26%) when evaluated using different evaluation strategies in human datasets and lower than 30% sensitivity. In mouse datasets, circseq_cup showed lower precision and sensitivity. For human datasets, CircAST achieved precision higher than 85% and sensitivity lower than 30%, and CIRI-full

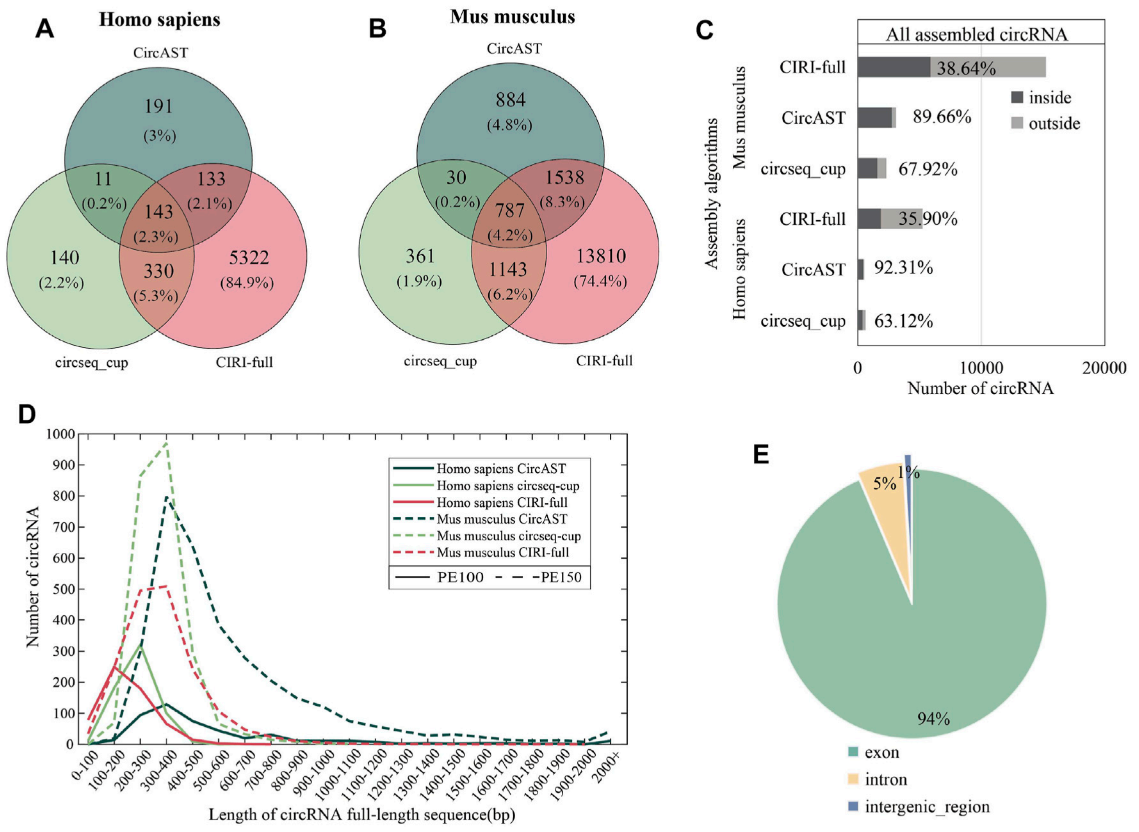

FIGURE 3 | Assembly results of three assembly tools. **(A,B)** Venn diagram depicting the overlap between different assembly algorithms in human and mouse datasets. **(C)** The proportion of full-length circRNAs constructed from the common circRNA candidates. "inside" (dark gray) represents assembled circRNAs belonging to common circRNAs among four identification tools, and 'outside' (light gray) represents assembled circRNAs not belonging to common circRNA among four identification tools. **(D)** Length distribution of circRNA full-length sequences (the result of CIRI-full is scaled by 1/10). **(E)** The percentage of circRNA categories in all assembled circRNA results.

TABLE 1 | Assembly rate and assembly number of circRNA using different assembly tools.

	CircAST^a				CIRI-full^a				circseq_cup^a
	CIRI^b	CIRCexplorer^b	circRNA_finder^b	find_circ^b	CIRI^b	CIRCexplorer^b	circRNA_finder^b	find_circ^b	
SRR10612068	300 (6.21%)	129 (3.98%)	128 (3.80%)	248 (4.86%)	1868 (38.69%)	1,121 (34.61%)	1,131 (33.56%)	1,661 (32.55%)	323
SRR10612069	256 (5.95%)	96 (3.71%)	95 (3.55%)	201 (4.51%)	1723 (40.03%)	948 (36.66%)	967 (36.11%)	1,452 (32.56%)	286
SRR10612070	259 (5.99%	111 (3.98%)	96 (3.37%)	204 (4.37%)	1723 (39.85%	950 (34.10%)	940 (33.01%)	1,508 (32.31%)	285
CRR194214	1958 (16.15%)	1,254 (11.64%)	1,155 (9.92%)	1,292 (10.81%)	7,353 (60.64%)	5,410 (50.23%)	5,658 (48.61%)	5,919 (49.54%)	1,509
CRR194215	2,724 (19.87%)	1852 (13.91%)	1706 (11.55%)	1769 (11.99%)	8,480 (61.86%)	6,526 (49.02%)	6,923 (46.89%	7,095 (48.10%)	1847

The table displays the number of full-length circRNA, and the assembly rate for CircAST, and CIRI-full (The numbers in parenthesis is the assembly rate); and the last column displays the number of full-length circRNA, for circseq_cup. The superscript 'a' indicates that the term is an assembly tool, and superscript 'b' indicates that the term is a identification algorithm. Assembly rate = A/I, where A is number of assembled circRNA, I is number of all identified circRNA.

gained precision lower than 60% and sensitivity higher than 39%. CircAST and CIRI-full showed the same trend in mouse datasets. circseq_cup and CircAST showed high precision and low sensitivity whereas CIRI-full displayed low precision and high sensitivity. It is feasible to improve the precision at the cost of sensitivity for CIRI-full.

In addition, the assembly strategy of CIRI plus CIRI-full showed the highest F1 score (read alignment: 0.6348, CIRI-long:

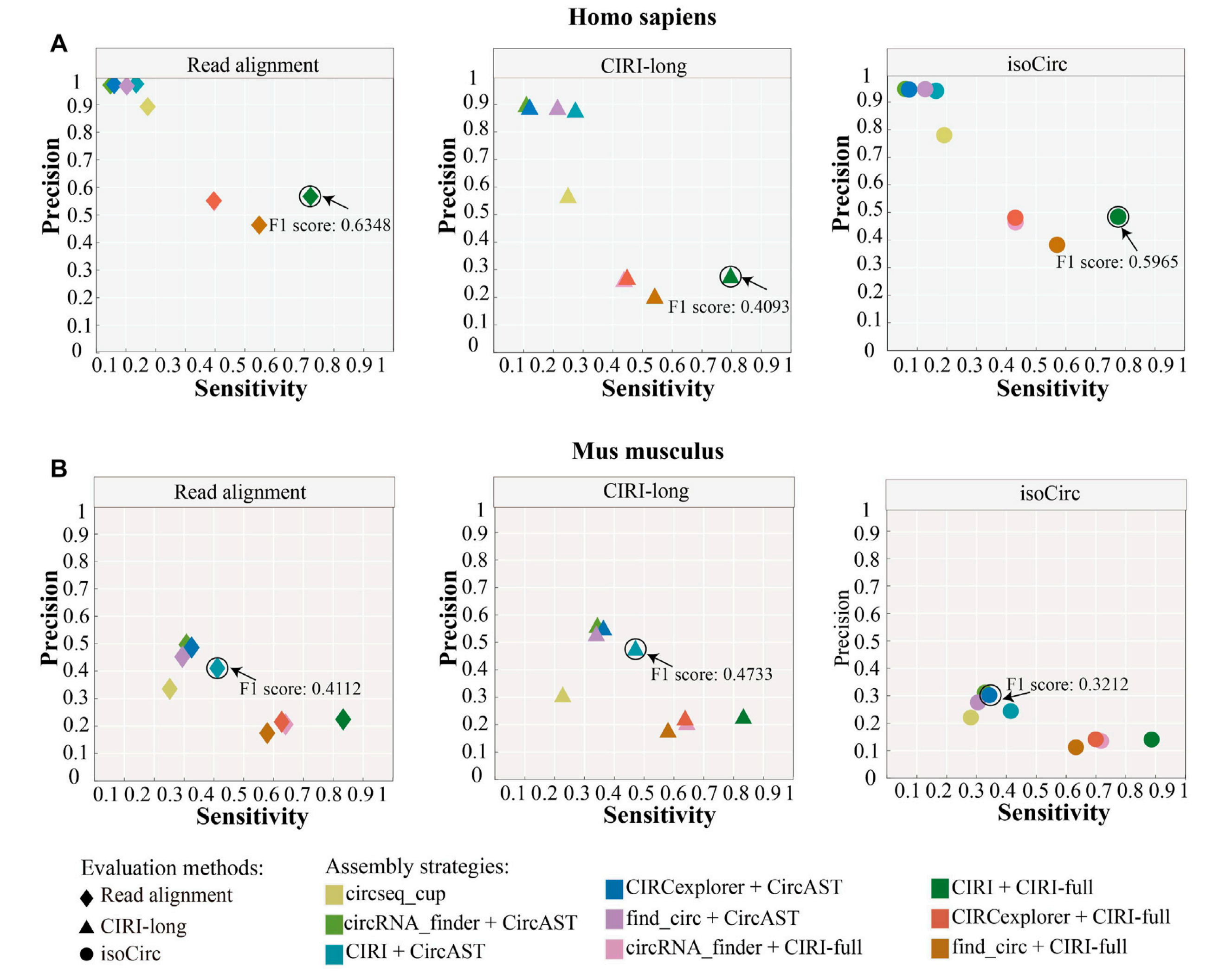

FIGURE 4 | Performance of different assembly strategies in terms of sensitivity and precision. Marked points are the best assembly strategy under different evaluation methods. **(A)** *Homo sapiens.* **(B)** *Mus musculus.*

0.4093, and isoCirc: 0.5965) using all three evaluation strategies in human datasets (**Figure 4**, **Supplementary Table S3**). However, CircAST performed better than CIRI-full in mouse datasets. For mouse datasets, using read alignment and CIRI-long as evaluation strategies, the combination of CIRI and CircAST showed the highest F1 score (read alignment: 0.4112, CIRI-long: 0.4733), and the combination of CIRCexplorer and CircAST produced the highest F1 score (0.3212) when using isoCirc as the evaluation strategy. Overall, CIRI-full showed better performance for human datasets, and CircAST showed better performance for mouse datasets.

Comparison of Evaluation Strategies

As shown in **Figure 1**, three evaluation strategies (see the Method section) were used to evaluate circRNA full-length sequence assembly using long reads.

In **Supplementary Figure S2**, for each evaluation strategy, we combined all positive datasets (full-length circRNAs that were verified correctly) of nine assembly strategies to compare the evaluation strategies. Of all correct full-length circRNAs in human datasets, 1,337 full-length circRNAs (39.1%) were observed between all evaluation strategies, and read alignment confirmed 3,217 full-length circRNA that accounted for about 94% of all verified results (**Supplementary Figure S2A**). Similarly, there were 1,391 (34.9%) verified full-length circRNAs found in the results of all three evaluation strategies in mouse datasets. For mouse datasets, instead of read alignment, CIRI-long generated the largest number of verified circRNA sequences (3,128, 78.5%) (**Supplementary Figure S2B**).

Then, we compared precision of nine assembly strategies under three evaluation methods. In human datasets, read alignment showed the highest precision for all nine assembly strategies, while for mouse datasets, CIRI-long showed the

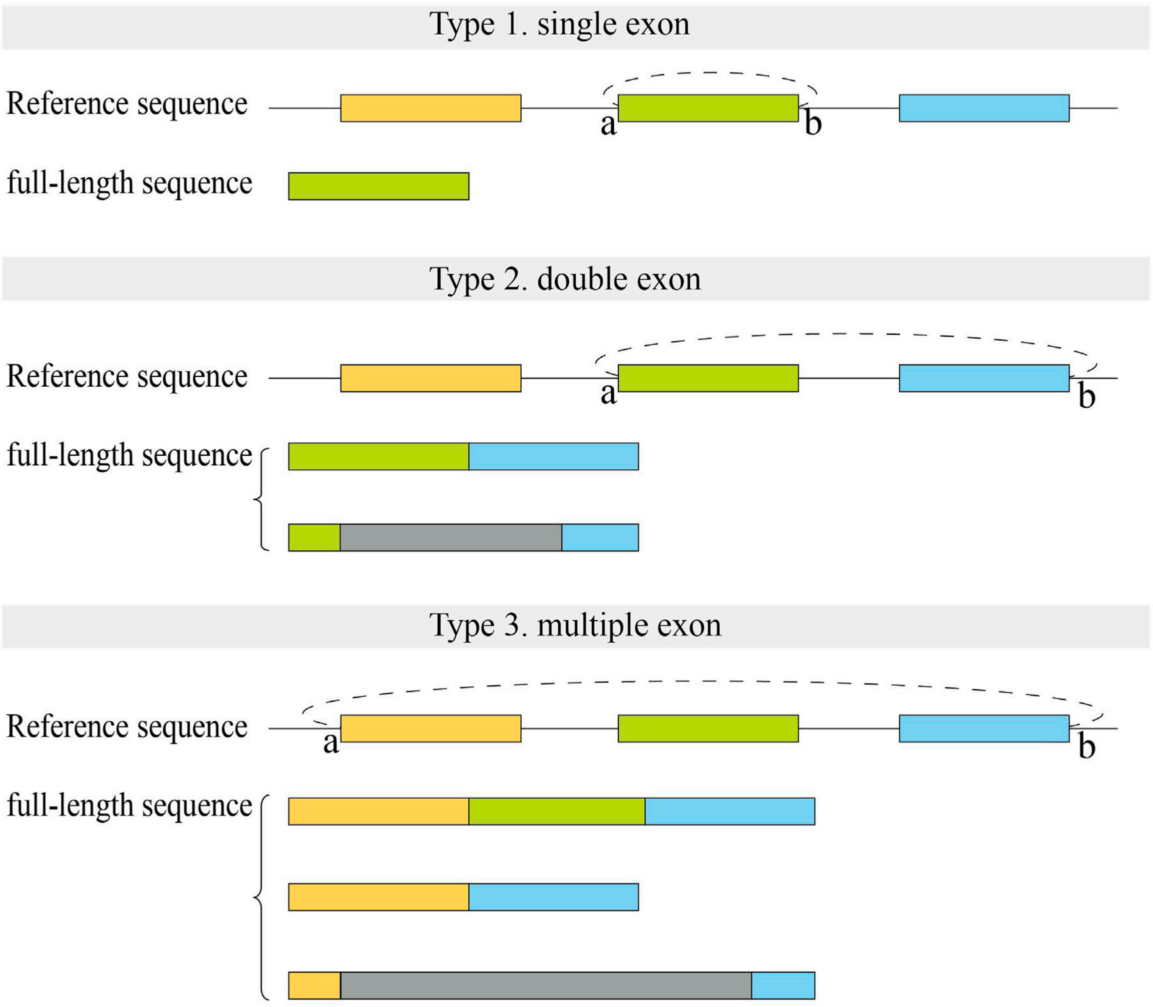

FIGURE 5 | Structure of full-length sequences reconstructed by CIRI-full in human datasets. Small letters **(A,B)** represent two back splice sites. Color rectangles represent exons, and gray rectangles represent the uncertain region which may include exons or introns.

highest precision for eight assembly strategies (**Supplementary Figure S2C,D**). Evaluation strategies showed various performances in human and mouse datasets. The precision of CIRI-long was higher than that of isoCirc for human datasets, while for mouse datasets, the opposite trend was observed.

To analyze the reason for the opposite trend observed between CIRI-long and isoCirc, we generated five subset samples from SRR10612050 according to read length (<1,000 bp, 2000–2,300 bp, 3,500–3,530 bp, 5,000–5,050 bp, and 6,900–7,000 bp) (Table S4). The majority of circRNAs were identified by CIRI-long for read lengths less than 1,000 bp, and isoCirc identified more circRNAs when read length was longer than 1,000 bp. The results showed that CIRI-long and isoCirc tend to behave differently for different read lengths.

From the above analysis, it was found that using circRNA sequences that are verified by all three evaluation methods are more reliable; however, in order to generate enough number of circRNA sequences, we chose to use the circRNA sequences verified by at least two of the three evaluation strategies. In the flowing analysis, we combined all the correct full-length circRNAs verified by at least two evaluation strategies.

Number of Back Splice Reads Affects the Quality of Reconstructed circRNA Sequences

It is found that the circRNA assembly results of circseq_cup and CircAST displayed higher precision than CIRI-full, whereas CIRI-full displayed the highest sensitivity. In this part, we analyzed the impact of the back splice reads on the precision of creditable full-length circRNAs which were verified by at least two evaluation methods.

Supplementary Figure S3 illustrates the change of precision of assembly tools with the increasing number of back splice reads given in human datasets. With the increasing number of back

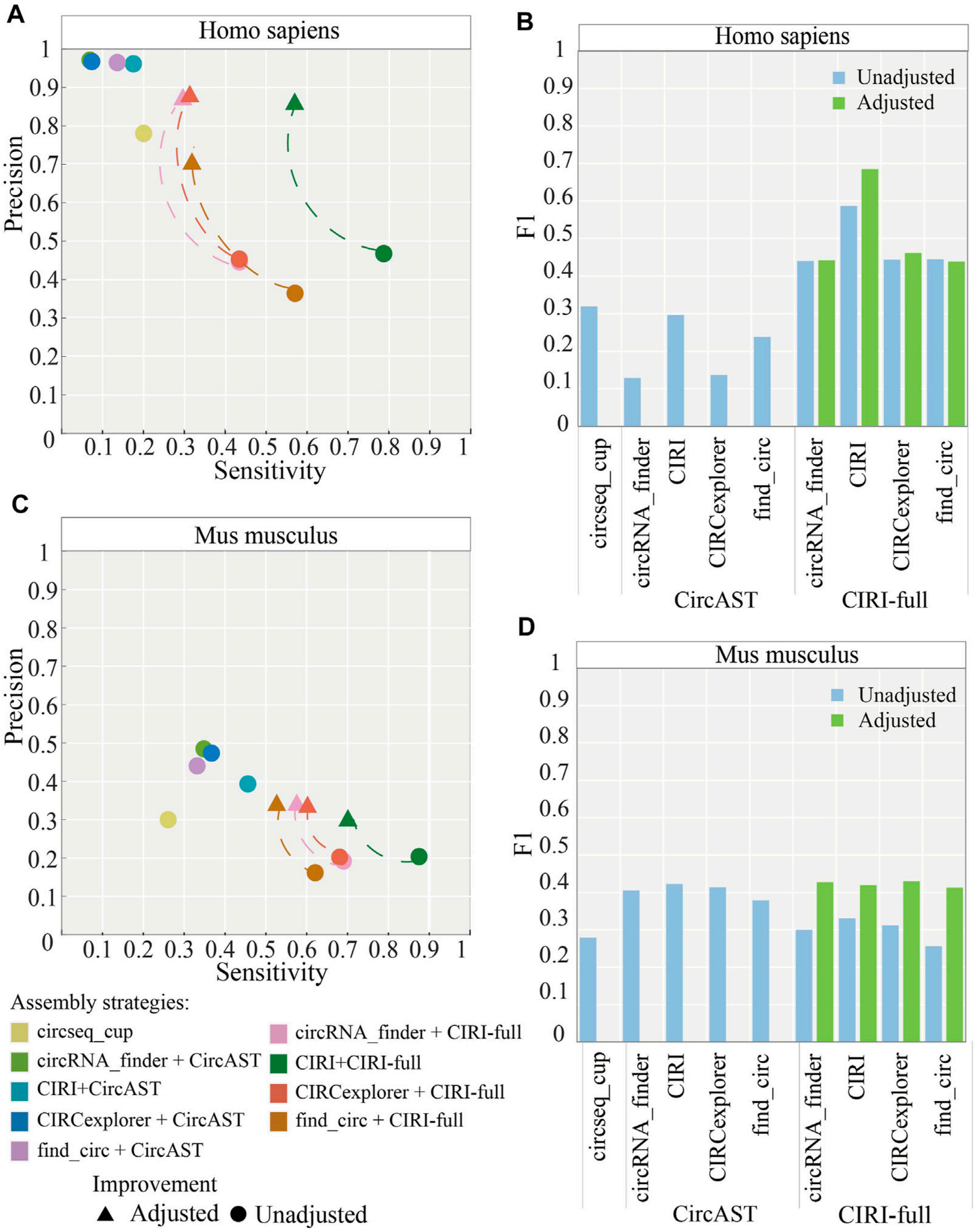

FIGURE 6 | Performance of assembly strategies related to CIRI-full after adjustment (screening). **(A,B)** Performance of assembly strategies in human and mouse datasets. **(C,D)** F1 score of assembly strategies in human and mouse datasets. “Adjusted” represents performance of CIRI-full after screening and “Unadjusted” represents performance of CIRI-full before screening.

splice reads, the precision of circseq_cup and CIRI-full were also increased. However, the precision of CircAST did not show a similar trend (**Supplementary Figure S3**). The curves of CircAST showed larger fluctuations due to its low sensitivity, and a lower number of wrong circRNAs causes a sharp decrease in precision. In mouse datasets, the precision of all assembly strategies increased with the increasing number of back splice reads (**Supplementary Figure S4**). We can assemble more reliable full-length sequences when circRNAs were supported by many back splice reads.

Improving circRNA Sequence Assembly for CIRI-Full

Previous results showed that for human datasets, circseq_cup and CircAST assembled a lower number of circRNA sequences with high precision and low sensitivities, and most of them (~80%) were verified as correct. Meanwhile, CIRI-full generated more full-length sequences of circRNAs, and only less than 57% of circRNA sequences were evaluated as correct. Therefore, one can improve the precision by screening more credible sequences at the cost of sensitivity.

We first analyzed the sequences of exonic full-length circRNAs in CIRI-full for human datasets (**Figure 5**). For full-length circRNAs that were derived from a single exon, more than 90% of circRNA full-length sequences were full exon sequences in assembly results (Type 1). In the reconstructed results of circRNAs derived from two adjacent exons, about 40–50% of sequences contained two complete exons with no intron sequences (Type 2). Fewer (~16%) full-length circRNAs derived from multiple exons consisted of all exon sequences between back splice sites (Type 3).

In addition, we calculated the ratio between full-length circRNAs that consisted of all exon back splice sites from CIRI-full and the correct ones. It was found that more than 80% of full-length sequences consisting of all exons between back splice sites were verified correctly. Thus, to improve the precision of CIRI-full, we screened exonic circRNA that full-length sequences consisted of all exon sequences between back splice sites; these sequences were considered more reliable and were selected as correct sequences. After applying the screening protocol, the average precision of CIRI-full over four circRNA identification algorithms increased from 43.26 to 82.77% in human datasets (**Figure 6A**), and the average F1 score increased from 0.4788 to 0.5069 (**Figure 6C**).

The same screening rule was also applied in the mouse datasets; the average precision of CIRI-full over four circRNA identification algorithms increased from 18.96 to 32.82% (**Figure 6B**), and the average F1 score increased from 0.2995 to 0.4223 (**Figure 6D**). CIRI-full showed higher F1 score than CircAST in mouse datasets after screening.

DISCUSSION

Reconstruction of circRNA full-length sequences is vital for its function identification. Three assembly tools were developed to assemble full-length sequences using short reads, and two of them, CircAST and CIRI-full, require identification information of circRNA to complete assembly.

Here, we calculated the assembly rate of CircAST and CIRI-full in all datasets and the number of full-length circRNAs on circseq_cup (**Table 1**). For the same sample, CIRI-full produced more circRNAs full-length sequences than CircAST and circseq_cup.

As we know, in addition to BSJ, CIRI-full also proposed a new feature, named RO (Y. Zheng et al., 2019). The combination of BSJ and RO could assemble full-length sequences of some circRNAs, these circRNAs lacking support reads on internal sequences when they were assembled only using BSJ. Besides, incomplete full-length sequences were also included in the results. Thus, CIRI-full had the highest sensitivity and lowest precision among the three assembly tools (**Figure 4**). CircAST and circseq_cup chose another way and provided full-length sequences with high precision (Wu et al., 2019). CircAST had a low assembly rate due to filtered out circRNAs that were supported by less than 12 back splice reads. circseq_cup screened reliable back splice reads by several criteria to ensure the correctness of full-length sequences. High precision and sensitivity are our ultimate goal. In this study, we screened some circRNA full-length sequences that consisted of all exons between back splice sites in CIRI-full as final results. This procedure increased the precision and F1 score of CIRI-full (**Figure 6**).

In addition, as shown in **Table 1**, assembly tools displayed higher assembly rate in mouse than human, whereas assembly tools displayed poor performance in mouse datasets when we evaluated the performance using three evaluation strategies based on long reads (**Supplementary Table S3**). High assembly rate in mouse datasets is due to the feature of short reads. Short reads of mouse had bigger sequencing depth and longer sequence reads than human datasets (**Supplementary Table S1**) (X. Li et al., 2020). The number and length of back splice reads affect the assembly rate of assembly tools. Mouse datasets find it easier to assemble more circRNA full-length sequences than human datasets. Evaluation of performance was based on corresponding long reads in this study. For short reads of mouse, long reads datasets and short reads are not matched perfectly. The small long reads datasets lead to only part of full-length sequences that could be verified. Big short reads datasets and small long reads datasets make assembly tools show poor performance and low precision and sensitivity.

As shown in **Figure 6A** and **Figure 6C**, the precision of CIRI-full is improved by about 40% in human datasets and about 10% in mouse datasets. The difference was caused by sequencing datasets. The size of short reads and long reads are similar in human datasets; long reads could be used to verify most candidate circRNAs. By removing part of low-confidence full length circRNAs, the precision of CIRI-full was greatly improved. The short reads data are much bigger than long reads in mouse datasets; thus, only a small part of candidate circRNAs was verified by the long reads, and the precision of CIRI-full for mouse datasets was not improved as much as for human datasets.

This work indicated that the combination of CIRI and CIRI-full is a better assembly strategy for the single assembly algorithm, and several reported assembly tools should be used simultaneously to obtain comprehensive and reliable results. However, we only used two datasets (in human and mouse) to evaluate the performance of assembly tools, and human and mouse are both mammals. Thus, our conclusion is more applicable to mammals, and whether it is applicable to other animals or plants still needs further verification. In addition, developing a new assembly algorithm that has the advantages of lower data requirements and more reliable assembly results is more significant.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://bigd.big.ac.cn/gsa https://www.ncbi.nlm.nih.gov/sra.

AUTHOR CONTRIBUTIONS

YW and JZ planned experiments; JZ and MH analyzed data; YW and JZ wrote the manuscript; WL, YP, and YP modified the manuscript.

FUNDING

This work was partly supported by the National Key Research and Development Program of China under grant No. 2018YFB0204403, the Strategic Priority CAS Project XDB38050100, the National Science Foundation of China under grant no. U1813203, the National Natural Youth Science Foundation of China under grant no. 31601028, the Shenzhen Basic Research Fund under grant nos. JCYJ20190808163801777, KQTD20200820113106007, and RCYX2020071411473419, and the CAS Key Lab under grant no. 2011DP173015.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.816825/full#supplementary-material

REFERENCES

Arnberg, A. C., Van Ommen, G.-J. B., Grivell, L. A., Van Bruggen, E. F. J., and Borst, P. (1980). Some Yeast Mitochondrial RNAs Are Circular. *Cell* 19 (2), 313–319. doi:10.1016/0092-8674(80)90505-x

Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., et al. (2014). circRNA Biogenesis Competes with Pre-mRNA Splicing. *Mol. Cel* 56 (1), 55–66. doi:10.1016/j.molcel.2014.08.019

Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., FitzGerald, G. A., and Grant, G. R. (2017). Simulation-based Comprehensive Benchmarking of RNA-Seq Aligners. *Nat. Methods* 14 (2), 135–139. doi:10.1038/nmeth.4106

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* 29 (1), 15–21. doi:10.1093/bioinformatics/bts635

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an Efficient and Unbiased Algorithm for De Novo Circular RNA Identification. *Genome Biol.* 16 (1), 4. doi:10.1186/s13059-014-0571-3

Gao, Y., Zhang, J., and Zhao, F. (2018). Circular RNA Identification Based on Multiple Seed Matching. *Brief Bioinform* 19 (5), 803–810. doi:10.1093/bib/bbx014

Gao, Z., Li, J., Luo, M., Li, H., Chen, Q., Wang, L., et al. (2019). Characterization and Cloning of Grape Circular RNAs Identified the Cold Resistance-Related Vv-circATS1. *Plant Physiol.* 180 (2), 966–985. doi:10.1104/pp.18.01331

Glažar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a Database for Circular RNAs. *Rna* 20 (11), 1666–1670. doi:10.1261/rna.043687.113

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA Circles Function as Efficient microRNA Sponges. *Nature* 495 (7441), 384–388. doi:10.1038/nature11993

Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., et al. (2014). A Multi-Split Mapping Algorithm for Circular RNA, Splicing, Trans-splicing and Fusion Detection. *Genome Biol.* 15 (2), R34. doi:10.1186/gb-2014-15-2-r34

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs Are Abundant, Conserved, and Associated with ALU Repeats. *Rna* 19 (2), 141–157. doi:10.1261/rna.035667.112

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions. *Genome Biol.* 14 (4), R36. doi:10.1186/gb-2013-14-4-r36

Kim, D., and Salzberg, S. L. (2011). TopHat-Fusion: an Algorithm for Discovery of Novel Fusion Transcripts. *Genome Biol.* 12 (8), R72. doi:10.1186/gb-2011-12-8-r72

Kristensen, L. S., Andersen, M. S., Stagsted, L. V. W., Ebbesen, K. K., Hansen, T. B., and Kjems, J. (2019). The Biogenesis, Biology and Characterization of Circular RNAs. *Nat. Rev. Genet.* 20 (11), 675–691. doi:10.1038/s41576-019-0158-7

Kristensen, L. S., Andersen, M. S., Stagsted, L. V. W., Ebbesen, K. K., and Hansen, T. B. (1986). The Hepatitis delta (delta) Virus Possesses a Circular RNA. *Nature* 323 (6088), 558–560. doi:10.1038/323558a0

Larsen, P. A., Heilman, A. M., and Yoder, A. D. (2014). The Utility of PacBio Circular Consensus Sequencing for Characterizing Complex Gene Families in Non-model Organisms. *BMC Genomics* 15 (1), 720. doi:10.1186/1471-2164-15-720

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H. (2018). Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* 34 (18), 3094–3100. doi:10.1093/bioinformatics/bty191

Li, X., Yang, L., and Chen, L.-L. (2018). The Biogenesis, Functions, and Challenges of Circular RNAs. *Mol. Cel* 71 (3), 428–442. doi:10.1016/j.molcel.2018.06.034

Li, X., Zhang, B., Li, F., Yu, K., and Bai, Y. (2020). The Mechanism and Detection of Alternative Splicing Events in Circular RNAs. *PeerJ* 8, e10032. doi:10.7717/peerj.10032

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs Are a Large Class of Animal RNAs with Regulatory Potency. *Nature* 495 (7441), 333–338. doi:10.1038/nature11928

Piwecka, M., Glažar, P., Hernandez-Miranda, L. R., Memczak, S., Wolf, S. A., Rybak-Wolf, A., et al. (2017). Loss of a Mammalian Circular RNA Locus Causes miRNA Deregulation and Affects Brain Function. *Science* 357 (6357). doi:10.1126/science.aam8526

Qu, S., Yang, X., Li, X., Wang, J., Gao, Y., Shang, R., et al. (2015). Circular RNA: A New star of Noncoding RNAs. *Cancer Lett.* 365 (2), 141–148. doi:10.1016/j.canlet.2015.06.003

Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J., and Kleinschmidt, A. K. (1976). Viroids Are Single-Stranded Covalently Closed Circular RNA Molecules Existing as Highly Base-Paired Rod-like Structures. *Proc. Natl. Acad. Sci.* 73 (11), 3852–3856. doi:10.1073/pnas.73.11.3852

Shi, Y., Jia, X., and Xu, J. (2020). The New Function of circRNA: Translation. *Clin. Transl Oncol.* 22 (12), 2162–2169. doi:10.1007/s12094-020-02371-1

Szabo, L., Morey, R., Palpant, N. J., Wang, P. L., Afari, N., Jiang, C., et al. (2015). Statistically Based Splicing Detection Reveals Neural Enrichment and Tissue-specific Induction of Circular RNA during Human Fetal Development. *Genome Biol.* 16 (1), 126. doi:10.1186/s13059-015-0690-5

Szabo, L., and Salzman, J. (2016). Detecting Circular RNAs: Bioinformatic and Experimental Challenges. *Nat. Rev. Genet.* 17 (11), 679–692. doi:10.1038/nrg.2016.114

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends Genet.* 34 (9), 666–681. doi:10.1016/j.tig.2018.05.008

Westholm, J. O., Miura, P., Olson, S., Shenker, S., Joseph, B., Sanfilippo, P., et al. (2014). Genome-wide Analysis of drosophila Circular RNAs Reveals Their Structural and Sequence Properties and Age-dependent Neural Accumulation. *Cel Rep.* 9 (5), 1966–1980. doi:10.1016/j.celrep.2014.10.062

Wu, J., Li, Y., Wang, C., Cui, Y., Xu, T., Wang, C., et al. (2019). CircAST: Full-Length Assembly and Quantification of Alternatively Spliced Isoforms in Circular RNAs. *Genomics, Proteomics & Bioinformatics* 17 (5), 522–534. doi:10.1016/j.gpb.2019.03.004

Ye, C.-Y., Zhang, X., Chu, Q., Liu, C., Yu, Y., Jiang, W., et al. (2017). Full-length Sequence Assembly Reveals Circular RNAs with Diverse Non-GT/AG Splicing Signals in rice. *RNA Biol.* 14 (8), 1055–1063. doi:10.1080/15476286.2016.1245268

Yin, S., Tian, X., Zhang, J., Sun, P., and Li, G. (2021). PCirc: Random forest-based Plant circRNA Identification Software. *BMC Bioinformatics* 22 (1), 10. doi:10.1186/s12859-020-03944-1

Zhang, J., Hao, Z., Yin, S., and Li, G. (2020a). GreenCircRNA: a Database for Plant circRNAs that Act as miRNA Decoys. *Database* 2020, baaa039. doi:10.1093/database/baaa039

Zhang, J., Liu, R., Zhu, Y., Gong, J., Yin, S., Sun, P., et al. (2020b). Identification and Characterization of circRNAs Responsive to Methyl Jasmonate in *Arabidopsis thaliana*. *Ijms* 21 (3), 792. doi:10.3390/ijms21030792

Zhang, M., Zhao, K., Xu, X., Yang, Y., Yan, S., Wei, P., et al. (2018). A Peptide Encoded by Circular Form of LINC-PINT Suppresses Oncogenic Transcriptional Elongation in Glioblastoma. *Nat. Commun.* 9 (1), 4475. doi:10.1038/s41467-018-06862-2

Zhang, X.-O., Wang, H.-B., Zhang, Y., Lu, X., Chen, L.-L., and Yang, L. (2014). Complementary Sequence-Mediated Exon Circularization. *Cell* 159 (1), 134–147. doi:10.1016/j.cell.2014.09.001

Zhao, J., Lee, E. E., Kim, J., Yang, R., Chamseddin, B., Ni, C., et al. (2019). Transforming Activity of an Oncoprotein-Encoding Circular RNA from Human Papillomavirus. *Nat. Commun.* 10 (1), 2300. doi:10.1038/s41467-019-10246-5

Zheng, S., Zhang, X., Odame, E., Xu, X., Chen, Y., Ye, J., et al. (2021). CircRNA-Protein Interactions in Muscle Development and Diseases. *Ijms* 22 (6), 3262. doi:10.3390/ijms22063262

Zheng, Y., Ji, P., Chen, S., Hou, L., and Zhao, F. (2019). Reconstruction of Full-Length Circular RNAs Enables Isoform-Level Quantification. *Genome Med.* 11 (1), 2. doi:10.1186/s13073-019-0614-1

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 17 March 2022
doi: 10.3389/fgene.2022.862741

A Novel Necroptosis-Related lncRNA Signature Predicts the Prognosis of Lung Adenocarcinoma

*Yinliang Lu, XueHui Luo, Qi Wang, Jie Chen, Xinyue Zhang, YueSen Li, Yuetong Chen, Xinyue Li and Suxia Han**

Department of Radiation Oncology, First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

Background: Necroptosis is closely related to the tumorigenesis and development of cancer. An increasing number of studies have demonstrated that targeting necroptosis could be a novel treatment strategy for cancer. However, the predictive potential of necroptosis-related long noncoding RNAs (lncRNAs) in lung adenocarcinoma (LUAD) still needs to be clarified. This study aimed to construct a prognostic signature based on necroptosis-related lncRNAs to predict the prognosis of LUAD.

Methods: We downloaded RNA sequencing data from The Cancer Genome Atlas database. Co-expression network analysis, univariate Cox regression, and least absolute shrinkage and selection operator were adopted to identify necroptosis-related prognostic lncRNAs. We constructed the predictive signature by multivariate Cox regression. Kaplan–Meier analysis, time-dependent receiver operating characteristics, nomogram, and calibration curves were used to validate and evaluate the signature. Subsequently, we used gene set enrichment analysis (GSEA) and single-sample gene set enrichment analysis (ssGSEA) to explore the relationship between the predictive signature and tumor immune microenvironment of risk groups. Finally, the correlation between the predictive signature and immune checkpoint expression of LUAD patients was also analyzed.

Results: We constructed a signature composed of 7 necroptosis-related lncRNAs (AC026355.2, AC099850.3, AF131215.5, UST-AS2, ARHGAP26-AS1, FAM83A-AS1, and AC010999.2). The signature could serve as an independent predictor for LUAD patients. Compared with clinicopathological variables, the necroptosis-related lncRNA signature has a higher diagnostic efficiency, with the area under the receiver operating characteristic curve being 0.723. Meanwhile, when patients were stratified according to different clinicopathological variables, the overall survival of patients in the high-risk group was shorter than that of those in the low-risk group. GSEA showed that tumor- and immune-related pathways were mainly enriched in the low-risk group. ssGSEA further

Edited by:
Juan Wang,
Inner Mongolia University, China

Reviewed by:
Junwei Luo,
Henan Polytechnic University, China
Teng Zhixia,
Northeast Forestry University, China

****Correspondence:***
Suxia Han
shan87@mail.xjtu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *26 January 2022*
Accepted: *11 February 2022*
Published: *17 March 2022*

Citation:
Lu Y, Luo X, Wang Q, Chen J, Zhang X, Li Y, Chen Y, Li X and Han S (2022) A Novel Necroptosis-Related lncRNA Signature Predicts the Prognosis of Lung Adenocarcinoma. Front. Genet. 13:862741. doi: 10.3389/fgene.2022.862741

Abbreviations: AUC, area under the ROC curve; GSEA, gene set enrichment analysis; GSVA, gene set variation analysis; LASSO, least absolute shrinkage and selection operator; lncRNA, long non-coding RNA; LUAD, lung adenocarcinoma; NRlncRNAs, necroptosis-related lncRNAs; NSCLC, non-small cell lung cancer; OS, overall survival; ROC, receiver operating characteristic curve; ssGSEA, single-sample gene set enrichment analysis; TCGA, the cancer genome atlas; TIME, tumor immune microenvironment.

confirmed that the predictive signature was significantly related to the immune status of LUAD patients. The immune checkpoint analysis displayed that low-risk patients had a higher immune checkpoint expression, such as CTLA-4, HAVCR2, PD-1, and TIGIT. This suggested that immunological function is more active in the low-risk group LUAD patients who might benefit from checkpoint blockade immunotherapies.

Conclusion: The predictive signature can independently predict the prognosis of LUAD, helps elucidate the mechanism of necroptosis-related lncRNAs in LUAD, and provides immunotherapy guidance for patients with LUAD.

Keywords: lung adenocarcinoma, necroptosis gene, long noncoding RNA, tumor immune microenvironment, prognostic signature

INTRODUCTION

Lung cancer is one of the most frequently diagnosed cancers and the leading cause of cancer-related deaths worldwide (Ferlay et al., 2021). Lung cancer is usually divided into non-small cell lung cancer (NSCLC) and small cell lung cancer; 85% of patients are NSCLC, of which lung adenocarcinoma (LUAD) accounts for about 50% (Thai et al., 2021). Recently, substantial improvements, such as chemotherapy, radiotherapy, and immunotherapy, have been made in the treatment of NSCLC patients. However, there is still a proportion of patients with distant metastasis that cannot be effectively treated at an early stage due to the lack of specific biomarkers, resulting in poor 5-year survival rates (Jurisic et al., 2020). Therefore, the identification of a reliable and specific biomarker for diagnosis and prognosis is urgently crucial for NSCLC.

Necroptosis is a form of programmed inflammatory cell death mediated by receptor-interacting protein kinases RIPK1, RIPK3, and mixed lineage kinase domain-like protein (MLKL). Necroptosis is characterized by early loss of plasma membrane integrity, leakage of intracellular contents, and organelle swelling (Krysko et al., 2017; Jiao et al., 2018). Recent studies have indicated that necroptosis has an important role in tumorigenesis, tumor metastasis, and tumoral immune response (Gong et al., 2019). Of note is the fact that necroptosis appears to be antitumorigenic or protumorigenic, depending on the tumor type and conditions during tumorigenesis (Yan et al., 2022). RIPK3 may restrict myeloid leukemogenesis and the differentiation of leukemia-initiating cells by promoting RIPK3–MLKL-mediated necroptosis (Höckendorf et al., 2016). Necroptosis could promote pancreatic cancer cell migration and invasion by the release of CXCL5 (Ando et al., 2020). Necroptosis blockage by MLKL ablation could substantially decrease the lung metastasis of breast cancer cells (Jiao et al., 2018). In addition, necroptosis is expected to develop an inflammatory tumor immune microenvironment *via* releasing damage-associated molecular patterns (DAMPs), cytokines, and/or chemokines in the tumor microenvironment, resulting in tumor-promoting or anti-tumor effects (Sprooten et al., 2020). On one hand, necroptotic tumor cells attract macrophages and DC cells, which are activated by DAMPs and cytokines. The activated DC cells migrate to the lymph nodes and activate naive CD4$^+$ and CD8$^+$ T cells. The naive T cells are activated and differentiated into effector T cells that leave the lymph nodes, re-enter the blood circulation, and infiltrate into tumor tissue to produce anti-tumor effects (Sancho et al., 2009). RIPK1 expression and NF-κB activation during necroptotic cell death are necessary for efficient cross-priming and antitumor immunity (Yatim et al., 2015). Consistently, vaccination with necroptotic cancer cells could also induce efficient antitumor immunity in an experimental mouse model (Aaes et al., 2016). On the other hand, necroptotic tumor cells also attract myeloid suppressor cells and tumor-associated macrophages, resulting in tumor-associated immunosuppression. Necroptosis-induced CXCL1 promoted pancreatic cancer progression *via* tumor-associated macrophage-induced immune suppression (Seifert et al., 2016). What is mentioned above implies the potential of targeting necroptosis as a novel cancer therapy, especially for immunotherapy.

Long non-coding RNAs (lncRNAs) are non-coding RNAs with transcripts of more than 200 nucleotides. Growing evidence has ascertained that lncRNAs are involved in the progression and metastasis of NSCLC and were associated with the immune pathway (Pang et al., 2021). LINC01748 exerted carcinogenic effects in NSCLC cell lines by regulating the microRNA-520a-5p/HMGA1 axis (Tan et al., 2022). lncRNA-SChLAP1 was verified to induce NSCLC progression and immune evasion by regulating the AUF1/PD-L1 pathway (Du et al., 2021). In addition, several studies demonstrated that lncRNA could also regulate necroptosis *via* functioning as competitive RNAs to influence the expression of target genes. lncRNA-107053293 was demonstrated to regulate necroptosis by acting as a competing endogenous RNA of miR-148a-3p (Wang W et al., 2020). The depletion of Linc00176 disrupted the cell cycle and induced necroptosis in hepatocellular carcinoma *via* regulating the expression of miRNAs, such as miR-9 and miR-185 (Tran et al., 2018). Based on the important role of lncRNA on the tumor, the prognostic signatures based on lncRNAs of LUAD patients have been widely introduced (Chen H et al., 2021; Xu et al., 2021). Nevertheless, research on necroptosis-related lncRNAs (NRlncRNAs) in LUAD prognosis and tumor immune microenvironment (TIME) has not been reported.

In this study, we constructed a novel predictive signature based on NRlncRNAs to forecast the prognosis of LUAD. We also validated its clinical value and confirmed that this signature can be used as a predictor of immunotherapy, which may offer a guiding function for clinicians.

MATERIALS AND METHODS

Preparation of Transcriptomic Data and Clinical Information

We downloaded the transcriptome RNA sequencing data of LUAD samples from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/). Meanwhile, we obtained the corresponding clinical parameters of these patients and excluded patients with missing overall survival (OS) or poor OS (less than 60 days) to reduce statistical bias in this analysis.

Identification of Necroptosis-Related lncRNA

A list of 67 necroptosis genes was obtained from previously reported literature (Zhao et al., 2021). The correlations between 67 necroptosis-related genes and lncRNA expression were analyzed *via* Pearson correlation analysis. All NRlncRNAs (2,154) should conform to the standard of correlation coefficients (|Pearson R|) >0.4 and p <0.001. Then, we obtained 1,061 differentially expressed lncRNAs [log2 fold change > 1, false discovery rate (FDR) <0.05] after screening the synthetic data matrix by Strawberry Perl V-5.30.0 (https://www.perl.org/) and R software V-4.1.2 (https://www.r-project.org/) with limma R package.

Establishment and Validation of the Risk Signature According to Necroptosis-Related lncRNAs in LUAD

The entire 481 TCGA set of LUAD was divided into a train risk set and a test risk set randomly by the caret R package. The ratio was 1: 1. The train set was used to construct a necroptosis-related lncRNA signature, and the test set and entire set were applied to validate the signature. Combined with the clinical information of LUAD in TCGA, we screened and obtained 40 NRlncRNAs linked to OS significantly by univariate Cox (uni-Cox) regression analysis (p < 0.05). Subsequently, we performed least absolute shrinkage and selection operator (LASSO) Cox analysis (using the penalty parameter estimated by 10-fold cross-validation) *via* the glmnet R package to screen out optimal lncRNAs associated with LUAD prognosis. This method aims to prevent over-fitting during modeling. Finally, a prognostic risk signature based on the optimal lncRNAs was established with the multivariate Cox (multi-Cox) regression analysis, and the risk score of every patient with LUAD was calculated based on the following formula:

$$\text{risk score} = \sum_{i=1}^{n} \text{Coef}\,(i) \times \text{Expr}\,(i)$$

Coef(i) and Expr(i) represent the regression coefficient of the multi-Cox regression analysis for each lncRNA and each lncRNA expression level, respectively. The patients were stratified into low- and high-risk groups, with the risk score as the cutoff. Kaplan–Meier method and log-rank test were conducted to analyze whether there is a difference in the OS of LUAD patients between the low- and high-risk groups using the survival R package.

We evaluated the prognostic value of the established risk signature between the model and the clinical characteristics *via* chi-square test. Uni-Cox and multi-Cox regression analyses were performed to explore whether the prognostic signature was a potential independent prognostic indicator for patients with LUAD, and the results were visualized with two forest maps. Several receiver operating characteristic (ROC) curves were generated, and the area under the ROC curve (AUC) was calculated by the survival, survminer, and timeROC R packages to validate the predictive value of the prognostic signature.

Nomogram and Calibration

We combined the risk score with the clinical variables of age, gender, N stage, T stage, M stage, and tumor stage to set up a nomogram for the 1-, 3-, and 5-year OS of LUAD patients by the rms R package. Correction curves based on the Hosmer–Lemeshow test were applied to illustrate the uniformity between the actual outcome and the signature prediction outcome.

Enrichment of Functions and Pathways in the Risk Prognosis Signature

We used gene set enrichment analyses (GSEA) software 4.1.2 (http://www.gsea-msigdb.org/gsea/index.jsp) to carry out GSEA and to identify significantly enriched pathways between the low- and high-risk groups. Values of p <0.05 and FDR <0.25 were considered the thresholds for statistical significance. The results were visualized by the gridExtra, grid, and ggplot2 R packages.

Estimation of the Tumor Immune Microenvironment of the Prognostic Signature

To figure out the relationship between this signature and TIME, firstly, we calculated the infiltration values for TCGA-LUAD dataset samples based on 7 algorithms: XCELL (Aran et al., 2017), TIMER (Li T et al., 2017; Li et al., 2020), QUANTISEQ (Finotello et al., 2019), MCPCOUNTER (Dienstmann et al., 2019), EPIC (Racle et al., 2017), CIBERSORT-ABS (Tamminga et al., 2020), and CIBERSORT (Chen et al., 2018). Using Spearman correlation analysis, the relationship of immune cell subpopulations and risk score value was evaluated. Wilcoxon signed-rank test, limma, scales, ggplot2, ggtext, tidyverse, and ggpubr R packages were applied, and the results are displayed in a bubble chart. Then, we explored the abundance of immune cells and stromal cells between different groups. The StromalScore, ImmuneScore, and ESTIMATEScore (StromalScore + ImmuneScore) of each patient were calculated. Their differences were compared using the Wilcoxon signed-rank test, and p <0.05 was considered to be significant. Subsequently, single-sample GSEA (ssGSEA) was

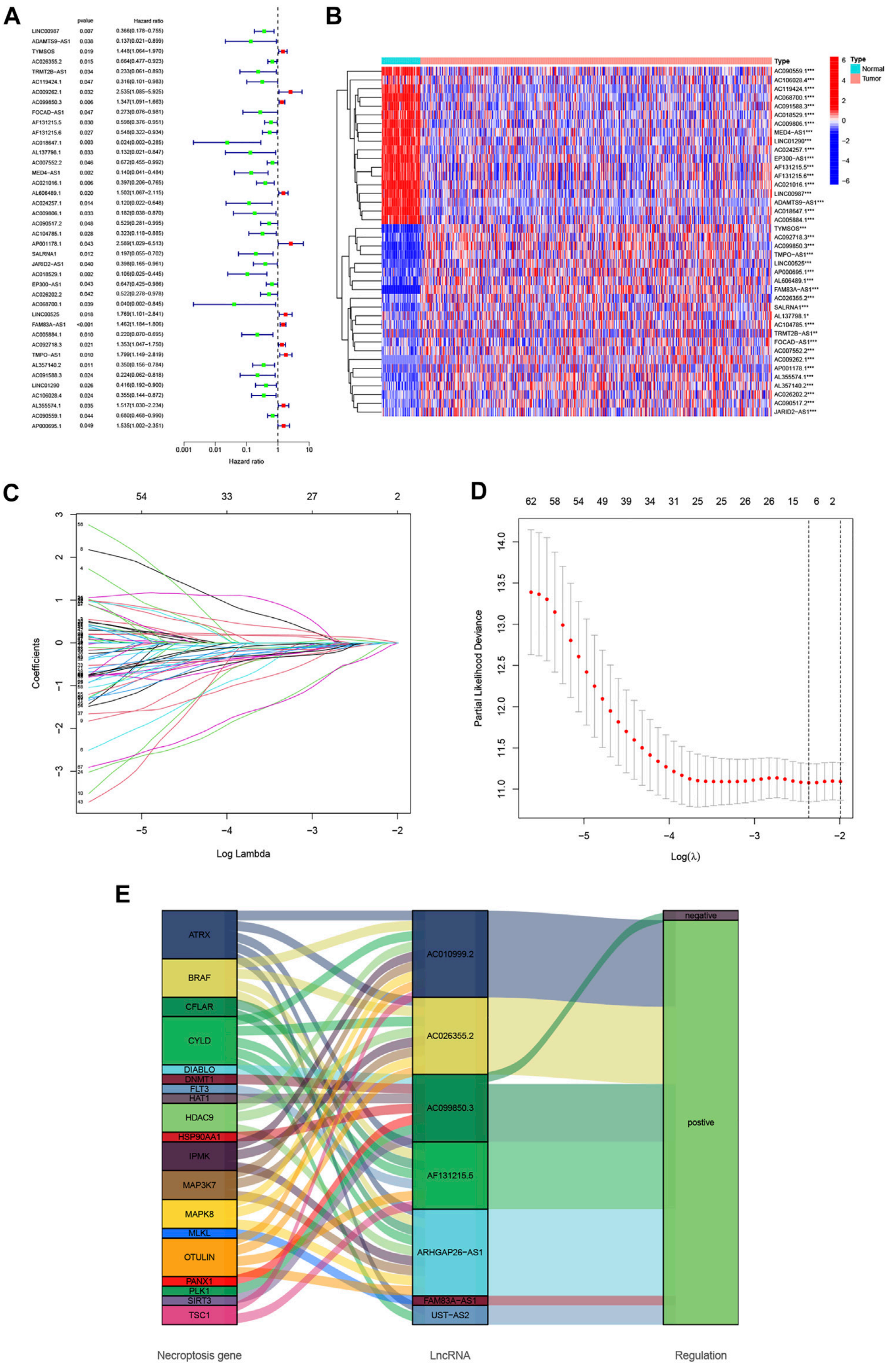

FIGURE 1 | Identification of necroptosis-related lncRNA prognostic signature in lung adenocarcinoma (LUAD). **(A)** Forest plot of 40 necroptosis-related lncRNAs selected by univariate Cox regression analysis. **(B)** The differential expressions of 40 necroptosis-related lncRNAs linked to survival between LUAD and normal samples. **(C)** The 10-fold cross-validation for variable selection in the least absolute shrinkage and selection operator (LASSO) algorithm. **(D)** The LASSO coefficient profile of necroptosis-related lncRNAs. **(E)** The Sankey diagram of the connection between 19 necroptosis genes and 7 necroptosis-related lncRNAs. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

conducted for scoring LUAD-infiltrating immune cells to quantify their relative content *via* the "GSVA" package. The scores of immune cells and pathways in different groups are shown on multi-boxplots, respectively. Finally, we also made comparisons about the immune checkpoint activation between low- and high-risk groups by the ggpubr R package.

RESULTS

Identification of Necroptosis-Related lncRNAs in LUAD Patients

The detailed flow diagram of our study is exhibited in **Supplementary Figure S1**. The transcriptome data of LUAD downloaded from TCGA included 59 normal samples and 539 tumor samples. We distinguished the mRNAs and lncRNAs by GTF files. According to the expression of 67 necroptosis genes and differentially expressed lncRNAs between normal and tumor samples, we finally obtained 1,016 NRlncRNAs (**Supplementary Table S1**), including 97 downregulated lncRNAs and 919 upregulated ones (**Supplementary Figure S2**).

Construction of the Necroptosis-Related lncRNA Predictive Signature

Using uni-Cox regression analysis in the TCGA train set, we obtained 40 NRlncRNAs which were significantly correlated with OS and made a heat map (**Figures 1A,B**). To avoid overfitting and improve the accuracy and explainability of the prognostic signature, we performed the LASSO-penalized Cox analysis on these lncRNAs and extracted 19 lncRNAs related to necroptosis in LUAD when the first-rank value of Log(λ) was the minimum likelihood of deviance (**Figures 1C,D**). Subsequently, we constructed the predictive signature composed of 7 NRlncRNAs (AC026355.2, AC099850.3, AF131215.5, UST-AS2, ARHGAP26-AS1, FAM83A-AS1, and AC010999.2) *via* multi-Cox regression analysis. Of those lncRNAs, 6 lncRNAs were regulated positively by necroptosis genes in the Sankey diagram (**Figure 1E**). Meanwhile, some of those lncRNAs (AC099850.3, AF131215.5, and FAM83A-AS1) were demonstrated to be highly associated with NSCLC previously. Subsequently, the risk score of every LUAD patient was calculated based on correlation coefficients calculated by multivariate Cox regression analysis, and the patients were divided into low- and high-risk groups according to the median value of the risk score. The risk score was calculated as follows: risk score = (−0.3641 × AC026355.2 expression) + (0.1747 × AC099850.3 expression) + (−0.3943 × AF131215.5 expression) + (−0.6257 × UST-AS2 expression) + (−2.8454 × ARHGAP26-AS1 expression) + (0.3281 × FAM83A-AS1 expression) + (−2.1752 × AC010999.2 expression) (**Supplementary Table S2**).

Prognosis Values of the Necroptosis-Related lncRNA Signature

To value the prognostic ability of the risk signature, we compared the distribution of risk score, the pattern of survival time, the survival status, and the relevant expression of 7 NRlncRNAs between the low- and high-risk groups in the train, test, and entire sets (**Figures 2A–L**). These all indicated that the low-risk group had better prognoses. Meanwhile, the LUAD patients were separated into groups according to age, gender, stage, T stage, N stage, and M stage to study the relationship between the risk signature and the prognosis of LUAD patients among universal clinicopathological variables. For different classifications, except T3-4 and M1 stage (**Figures 3H, L**), the OS of the patients in the low-risk group was significantly longer than that of the patients in the high-risk group (**Figures 3A–G**, **Figures 3I–K**). The possible explanation of the T3–T4 and M1 stage might be the limited number of patients due to poor prognoses in advanced NSCLC. These results suggest that the predictive signature can also predict the prognosis of LUAD patients in a different group of age, gender, stage, N stage, T1-2 stage, and M0 stage.

An Independent LUAD Prognostic Indicator of the Necroptosis-Related lncRNA Signature

To determine whether the predictive signature is an independent prognostic factor for LUAD patients, Cox regression analysis was performed in the entire set. The Uni-Cox regression analysis showed that stage, T stage, N stage, and risk score were significantly associated with the OS of LUAD patients (**Figure 4A**). The multi-Cox regression analysis showed that only risk score (hazard ratio = 1.331, confidence interval = 1.175–1.507, $p < 0.001$) was an independent predictor of OS in LUAD patients (**Figure 4B**). Then, we used AUC to validate the sensitivity and the specificity of the signature in the entire set. The AUC of the risk score was 0.723, which was better than that of clinicopathological variables in predicting the prognosis of LUAD patients (**Figure 4C**). The AUCs of 1-, 3-, and 5-year survival were 0.723, 0.679, and 0.715, respectively, which indicated a good predictive value (**Figure 4D**). These results further implied that the signature was a promising biomarker for indicating the prognosis risk of LUAD.

Construction and Evaluation of the Prognostic Nomogram

The nomogram including clinicopathological variables and the risk score were constructed to predict the 1-, 3-, and 5-year prognosis of LUAD patients (**Figure 5A**). The calibration curves indicated a good consistency between the actual OS rates and the predicted survival rates at 1, 3, and 5 years (**Figure 5B**).

Tumor Immune Microenvironment of the Necroptosis-Related lncRNA Signature

Based on the different prognoses of patients in the high- and low-risk groups, we conducted GSEA to explore the underlying differences in biological functions between risk groups. We found that the T/B cell receptor signaling pathway, Fc epsilon RI signaling pathway, cytokine receptor interaction, and JAK-STAT signaling pathway were significantly enriched in the low-

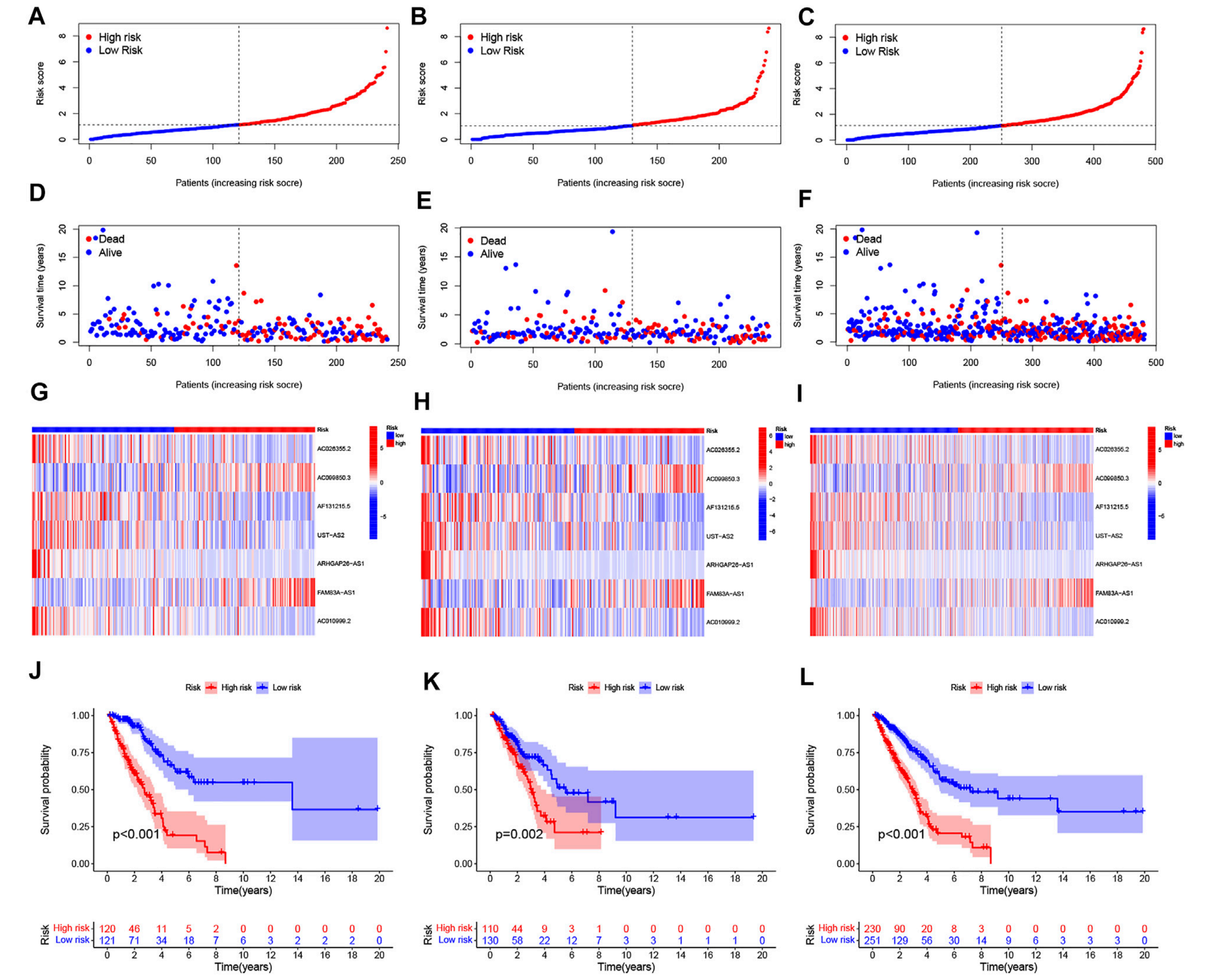

FIGURE 2 | Prognosis values of the 7 necroptosis-related lncRNA signatures in the train, test, and entire sets. The distribution of risk scores **(A–C)**, survival time and survival status **(D–F)**, heat maps of 7 lncRNA expressions **(G–I)**, and Kaplan–Meier survival curves of overall survival of LUAD patients **(J–L)** between low- and high-risk groups in the train, test, and entire sets, respectively.

risk group (**Figure 6A**), indicating that low-risk patients are closely related to tumor- and immune-related pathways. The GSEA results also revealed that the Notch signaling pathway, Wnt signaling pathway, and p53 signaling pathway, pathways in cancer and small cell lung cancer, were significantly enriched in the high-risk group. The Notch pathway plays a vital role in lung tumorigenesis and progression. Researchers found that cigarette smoke could promote LUAD progression *via* activating the Notch-1 pathway (Chiappara et al., 2022). Additionally, Notch-1 signaling synergized with Hif-1α could upregulate the expression of survivin in LUAD cell line A549 (Chen et al., 2011). The overexpression of Wnt pathway-activating genes and the down-expression of negative regulators of the pathway are closely correlated with NSCLC tumorigenesis, prognosis, and resistance to therapy (Stewart, 2014; Zeybek et al., 2022). The Wnt responder cells showed an increased tumor propagation ability, suggesting that they have features of normal tissue stem cells (Tammela et al., 2017). These mechanisms may explain why the high-risk group has a worse prognosis. Then, we studied the correlation between risk scores and tumor-infiltrating immune cells (**Figure 6B**). More immune cells are closely related to the low-risk group on different platforms. Consistently, we also found that StromalScore, ImmuneScore, and ESTIMATEScore in low-risk patients were significantly higher than those of high-risk patients (**Figures 6C–E**). To further explore the correlation between risk scores and immune cells and functions, we quantified the enrichment scores of ssGSEA for different immune cell subgroups, related functions, or pathways. The results exhibited that activated dendritic cells (aDCs), B cells, DCs, immature dendritic cells (iDCs), mast cells, neutrophils, T helper cells, T follicular helper (Tfh) cells, tumor-infiltrating lymphocyte (TIL), and T regulatory

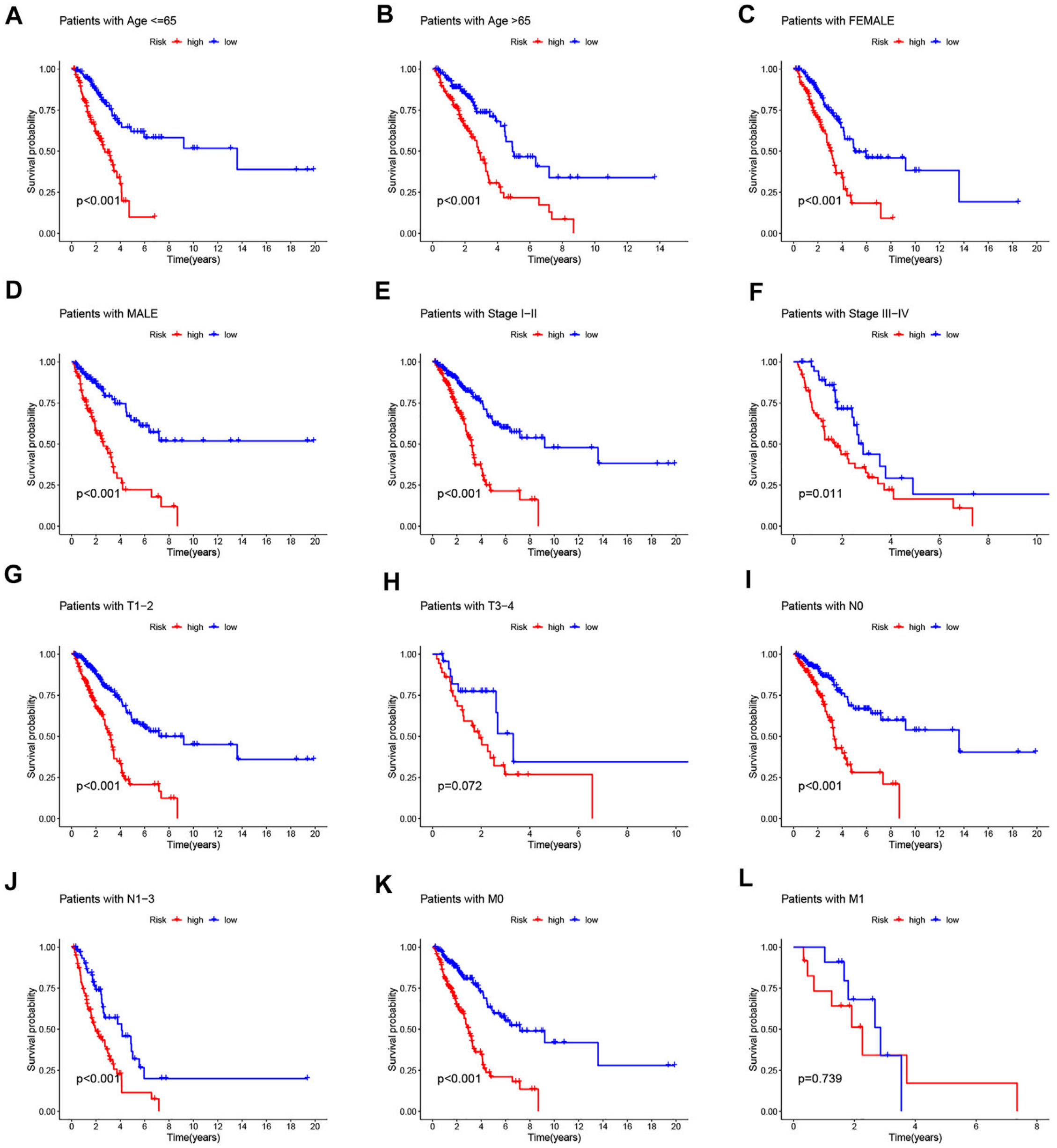

FIGURE 3 | Kaplan–Meier survival curves of low- and high-risk groups sorted by different clinicopathological variables. **(A,B)** Age, **(C,D)** sex, **(E,F)** stage, **(G,H)** T stage, **(I,J)** N stage, and **(K,L)** M stage.

cells (Tregs) were significantly negatively correlated with the risk score (**Figure 6F**). Compared with the high-risk group, several immune pathways, *e.g.*, checkpoint, cytolytic activity, human leukocyte antigen (HLA), T cell co-inhibition, T cell co-stimulation, and type II IFN response were higher in the low-risk group (**Figure 6G**). Furthermore, by comparing immune checkpoint activation between different risk groups, we found that almost all the immune checkpoints expressed more activity in the low-risk group, such as CTLA-4, HAVCR2 (TIM3), PDCD1 (PD-1), TIGIT, and CD70 (**Figure 6H**). These findings suggested that, in the low-risk group, the immunological function is more active and might be more sensitive to immunotherapy.

DISCUSSION

As the most common subtype of lung cancer, LUAD still poses a huge threat to human health worldwide, with mounting morbidity and mortality. The identification of a specific and reliable prognostic signature for LUAD patients is extremely vital to improve the prognosis. Although there are a lot of other signatures using lncRNAs to predict the survival outcomes of LUAD, a necroptosis-related lncRNA predictive signature has not been reported. Herein we constructed a necroptosis-related lncRNA signature to explore the prognosis and TIME of LUAD patients.

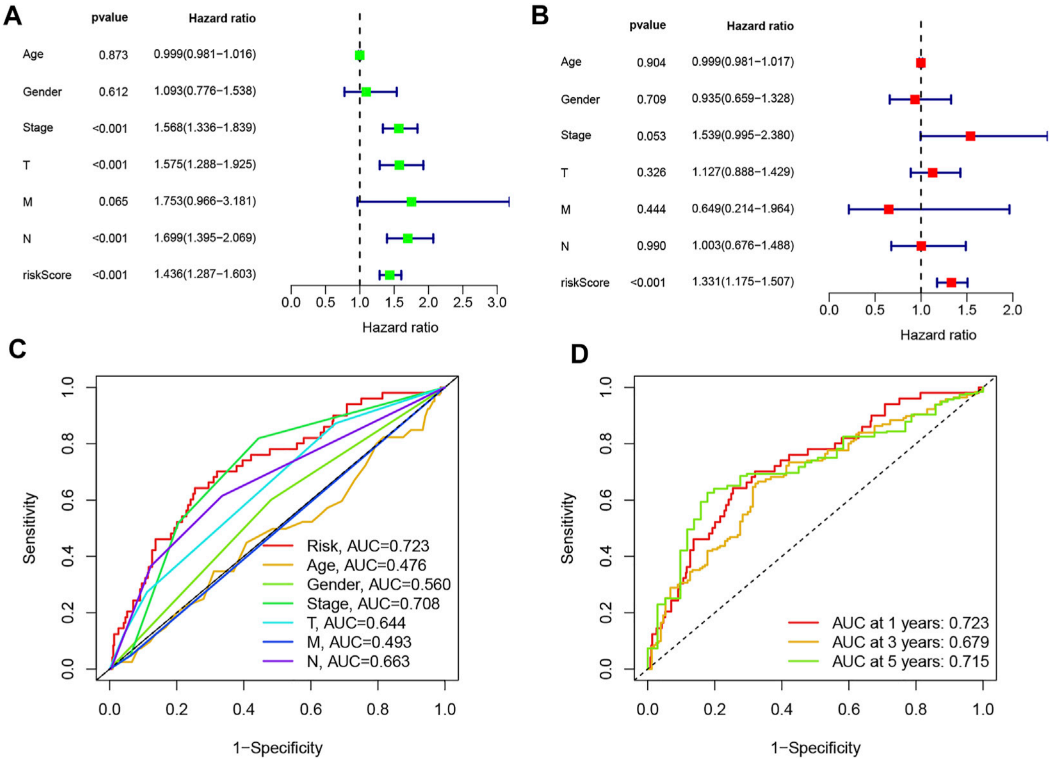

FIGURE 4 | Validation of the predictive value of the prognostic signature. **(A)** Uni-Cox and **(B)** multi-Cox regression analyses of clinical characteristics and risk score with overall survival. **(C)** Comparison of the prediction accuracy of the risk model with clinicopathological features, such as age, gender, stage, T stage, M stage, and N stage. **(D)** Accuracy of the risk signature in predicting 1-, 3-, and 5-year receiver operating characteristic curves based on the entire sets.

In this study, 1,016 differentially expressed NRlncRNAs were acquired to explore the prognostic function. We conducted univariate, LASSO, and multivariate Cox regression analyses and identified seven NRlncRNAs (AC026355.2, AC099850.3, AF131215.5, UST-AS2, ARHGAP26-AS1, FAM83A-AS1, and AC010999.2) significantly linked to the OS of LUAD patients to construct the necroptosis-related lncRNA signature. Among those lncRNAs, AC099850.3 has been reported to be highly expressed in tumors and closely related to the development and procession of NSCLC (Zhou et al., 2021); AC099850.3 is demonstrated to promote proliferation and migration in hepatocellular carcinoma and is also an important member of the prognosis model in hepatocellular carcinoma and colorectal cancer (Wu et al., 2021; Zhang et al., 2021). AF131215.5 also represented the independent prognostic significance of OS in patients with LUAD (Hou and Yao, 2021). FAM83A-AS1 could increase FAM83A expression by enhancing FAM83A pre-mRNA stability and promote the tumorigenesis of LUAD (Wang et al., 2021). FAM83A-AS1 was also verified to contribute to LUAD proliferation and stemness *via* the HIF-1α/glycolysis axis (Chen et al., 2022). Other lncRNAs (AC026355.2, UST-AS2, ARHGAP26-AS1, and AC010999.2) were revealed for the first time. It is noteworthy that knowledge on those newly distinguished NRlncRNAs could develop a better mechanistic understanding of LUAD, which might be new targets for cancer treatment. Then, the LUAD patients were divided into high- and low-risk groups based on the median value of the risk score. The results all indicated that the low-risk group had a better prognosis than the high-risk group, and risk score was an independent predictor of OS in LUAD patients. The ROC analysis showed that the signature was superior to conventional clinical characteristics in the survival prediction of LUAD. Similarly, the predictive nomogram established also showed a perfect consistency between the observed and predicted rates for the 1-, 3-, and 5-year OS. Collectively, these studies mentioned above indicate that our necroptosis-related lncRNA signature could predict the prognosis of LUAD patients accurately.

Researchers have demonstrated that necroptosis is strongly associated with tumorigenesis, tumor immune response, and poor prognosis (Gong et al., 2019), especially in solid tumors, but the specific role of necroptosis in those processes is still largely unknown. Therefore, we continued to explore the underlying mechanism of necroptosis-related lncRNA signature among different risk groups.

GSEA showed that the T/B cell receptor signaling pathway, Fc epsilon RI signaling pathway, cytokine receptor interaction, and JAK/STAT signaling pathway were significantly enriched in the low-risk group. Researchers found that the aberrant activation of the JAK/STAT signaling pathway was closely related to the occurrence, development, metastasis, and drug resistance of lung cancer (Li S. D. et al., 2017). The overexpression of JAK2 induced the proliferation, migration, and invasion abilities of lung adenocarcinoma A549 cells; conversely, the downregulation of JAK2 could suppress the protumorigenic effect (Xu et al.,

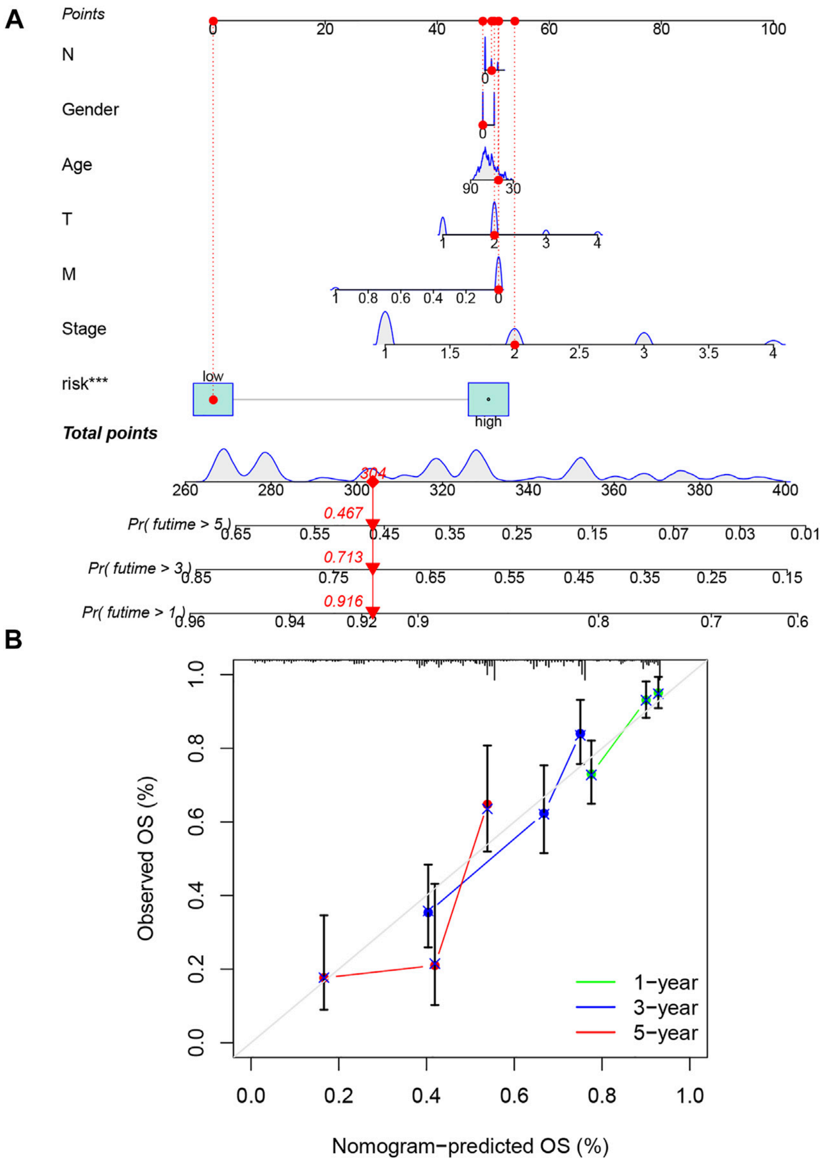

FIGURE 5 | Construction and verification of the nomogram. **(A)** A nomogram combining clinicopathological variables and risk score predicts the 1-, 3-, and 5- year overall survival of lung adenocarcinoma patients. **(B)** The calibration curves test the consistency between the actual outcome and the predicted outcome at 1, 3, and 5 years.

2017). EGFR tyrosine kinase inhibitors (TKIs), such as afatinib and dacomitinib, could activate STAT3 *via* autocrine interleukin-6 (IL-6) production, and that blockade of the IL-6R/JAK1/STAT3 signaling pathway potentiated sensitivity to those EGFR TKIs in NSCLC cells (Kim et al., 2012). In addition, the researcher found that zVAD (a pan-caspase inhibitor) induced necroptotic death in TLR3- and TLR4-activated macrophages *via* the JAK/STAT1/ROS pathway (Chen Y. S. et al., 2021). IFN-activated JAK/STAT signaling induced the robust expression of ZBP1, which complexed with RIPK3 to trigger MLKL-driven necroptosis (Ingram et al., 2019). Similarly, TNF-α synergized with IFN-γ could induce epithelial cell necroptosis through the CASP8-JAK1/2-STAT1 module (Woznicki et al., 2021). Taken together, we speculated that necroptosis probably contributed to the occurrence and development of LUAD through the JAK/STAT signaling pathway.

According to the role of necroptosis in regulating tumor immunity and the enrichment of immune-related pathways in low-risk groups, we performed ssGSEA to explore the immune status in different groups. The immune cells (aDCs, B cells, DCs, iDCs, mast cells, neutrophils, T helper cells, Tfh cells, TIL, and Tregs) and immune functions (checkpoint, cytolytic activity, HLA, T cell co-inhibition, T cell co-stimulation, and type II IFN response) were mainly active among the low-risk groups, some of which were closely linked to necroptosis. Necroptotic cells can provide both tumor-specific antigens and inflammatory cytokines to DCs for antigen cross-priming which activates cytotoxic $CD8^+$ T

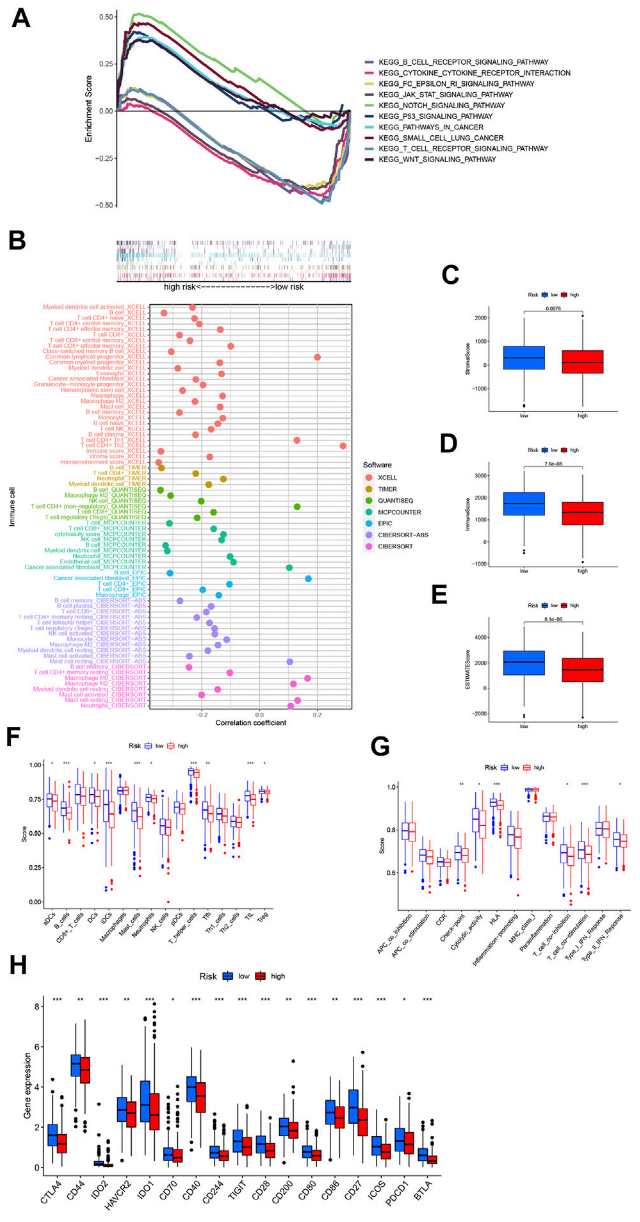

FIGURE 6 | The differences of tumor immune microenvironment between the low- and high-risk groups. **(A)** Gene set enrichment analysis of the top 10 pathways significantly enriched in the risk groups. **(B)** The immune cell bubble of risk groups. **(C–E)** The box plots of the comparison of StromalScore, ImmuneScore, and ESTIMATEScore, respectively, between low- and high-risk groups. **(F, G)** The ssGSEA scores of immune cells and immune functions in the risk groups. **(H)** The difference of common immune checkpoint expression in the risk groups. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

lymphocytes. RIPK3 was necessary for the regulation of cytokine expression in DCs, which could participate in innate and adaptive immune systems (Park et al., 2021). Wang X *et al.* found that a serine protease was involved in the RIPK3–MLKL-mediated necroptotic death pathway in neutrophils (Wang X et al., 2020). These results further illustrated that necroptosis might be involved in the progression of LUAD by regulating tumor immunity. Subsequently, we analyzed the correlation between common immune checkpoint expression and necroptosis-related lncRNA signature. Some researchers have indicated that the expression levels of immune checkpoint genes are highly associated with the efficacy of immunotherapy (Ahluwalia et al., 2021; Hu et al., 2021). Our findings demonstrated that most of the immune checkpoints' expression was elevated in low-risk LUAD patients compared to the high-risk group. Among those, PD-1 and CTLA-4 inhibitors have been validated to benefit patients with advanced NSCLC in clinical trials (Paz-Ares et al., 2021). In addition, TIM3, TIGIT, and CD70 have been under investigation, and drugs blocking these immune checkpoints are in clinical or preclinical developments (Bewersdorf et al., 2021; Hansen et al., 2021). Therefore, this signature implied that it would be more advantageous for LUAD patients at a lower risk to receive immunotherapy.

However, our research has several limitations and shortcomings. Firstly, it was better to include more clinical databases for external validation. Secondly, the underlying molecular mechanisms of the NRlncRNAs in LUAD should be further validated by experiments. Thus, we will recollect and expand clinical samples and attempt to validate the accuracy of this model *via* more external experiments in our following work.

In conclusion, the necroptosis-related lncRNA predictive signature can independently predict the prognosis of LUAD patients, helps elucidate the process and mechanism of NRlncRNAs in LUAD, and provides immunotherapy guidance for patients with LUAD, but it still needs further experimental verification in the future.

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

YLu designed the study and wrote the manuscript. XLu, QW, JC, and XZ performed the analysis. YLi, YC, and XLi collected the dataset. SH reviewed and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Innovation Capacity Support Plan of Shaanxi Province (under grant no. 2018TD-002).

ACKNOWLEDGMENTS

All authors would like to appreciate the TCGA database for sharing data.

SUPPLEMENTARY MATERIAL

The supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.862741/full#supplementary-material

REFERENCES

Aaes, T. L., Kaczmarek, A., Delvaeye, T., De Craene, B., De Koker, S., Heyndrickx, L., et al. (2016). Vaccination with Necroptotic Cancer Cells Induces Efficient Anti-tumor Immunity. *Cel Rep.* 15 (2), 274–287. doi:10.1016/j.celrep.2016.03.037

Ahluwalia, P., Ahluwalia, M., Mondal, A. K., Sahajpal, N., Kota, V., Rojiani, M. V., et al. (2021). Immunogenomic Gene Signature of Cell-Death Associated Genes with Prognostic Implications in Lung Cancer. *Cancers* 13 (1), 155. doi:10.3390/cancers13010155

Ando, Y., Ohuchida, K., Otsubo, Y., Kibe, S., Takesue, S., Abe, T., et al. (2020). Necroptosis in Pancreatic Cancer Promotes Cancer Cell Migration and Invasion by Release of CXCL5. *PLoS One* 15 (1), e0228015. doi:10.1371/journal.pone.0228015

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biol.* 18 (1), 220. doi:10.1186/s13059-017-1349-1

Bewersdorf, J. P., Shallis, R. M., and Zeidan, A. M. (2021). Immune Checkpoint Inhibition in Myeloid Malignancies: Moving beyond the PD-1/pd-L1 and CTLA-4 Pathways. *Blood Rev.* 45, 100709. doi:10.1016/j.blre.2020.100709

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., and Alizadeh, A. A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* 1711, 243–259. doi:10.1007/978-1-4939-7493-1_12

Chen, H., Hu, Z., Sang, M., Ni, S., Lin, Y., Wu, C., et al. (2021). Identification of an Autophagy-Related lncRNA Prognostic Signature and Related Tumor Immunity Research in Lung Adenocarcinoma. *Front. Genet.* 12, 767694. doi:10.3389/fgene.2021.767694

Chen, Y.-S., Chuang, W.-C., Kung, H.-N., Cheng, C.-Y., Huang, D.-Y., Sekar, P., et al. (2021). Pan-Caspase Inhibitor zVAD Induces Necroptotic and Autophagic Cell Death in TLR3/4-Stimulated Macrophages. *Mol. Cell* [Epub ahead of print]. doi:10.14348/molcells.2021.0193

Chen, Y., Li, D., Liu, H., Xu, H., Zheng, H., Qian, F., et al. (2011). Notch-1 Signaling Facilitates Survivin Expression in Human Non-small Cell Lung Cancer Cells. *Cancer Biol. Ther.* 11 (1), 14–21. doi:10.4161/cbt.11.1.13730

Chen, Z., Hu, Z., Sui, Q., Huang, Y., Zhao, M., Li, M., et al. (2022). LncRNA FAM83A-AS1 Facilitates Tumor Proliferation and the Migration via the HIF-1α/Glycolysis axis in Lung Adenocarcinoma. *Int. J. Biol. Sci.* 18 (2), 522–535. doi:10.7150/ijbs.67556

Chiappara, G., Di Vincenzo, S., Sangiorgi, C., Di Sano, C., D'Anna, C., Zito, G., et al. (2022). Cigarette Smoke Upregulates Notch-1 Signaling Pathway and Promotes Lung Adenocarcinoma Progression. *Toxicol. Lett.* 355, 31–40. doi:10.1016/j.toxlet.2021.11.002

Dienstmann, R., Villacampa, G., Sveen, A., Mason, M. J., Niedzwiecki, D., Nesbakken, A., et al. (2019). Relative Contribution of Clinicopathological Variables, Genomic Markers, Transcriptomic Subtyping and Microenvironment Features for Outcome Prediction in Stage II/III Colorectal Cancer. *Ann. Oncol.* 30 (10), 1622–1629. doi:10.1093/annonc/mdz287

Du, Z., Niu, S., Wang, J., Wu, J., Li, S., and Yi, X. (2021). SChLAP1 Contributes to Non-small Cell Lung Cancer Cell Progression and Immune Evasion through Regulating the AUF1/PD-L1 axis. *Autoimmunity* 54 (4), 1–9. doi:10.1080/08916934.2021.1913582

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., et al. (2021). Cancer Statistics for the Year 2020: An Overview. *Int. J. Cancer* [Epub ahead of print]. doi:10.1002/ijc.33588

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., et al. (2019). Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-Seq Data. *Genome Med.* 11 (1), 34. doi:10.1186/s13073-019-0638-6

Gong, Y., Fan, Z., Luo, G., Yang, C., Huang, Q., Fan, K., et al. (2019). The Role of Necroptosis in Cancer Biology and Therapy. *Mol. Cancer* 18 (1), 100. doi:10.1186/s12943-019-1029-8

Hansen, K., Kumar, S., Logronio, K., Whelan, S., Qurashi, S., Cheng, H.-Y., et al. (2021). COM902, a Novel Therapeutic Antibody Targeting TIGIT Augments Anti-tumor T Cell Function in Combination with PVRIG or PD-1 Pathway Blockade. *Cancer Immunol. Immunother.* 70 (12), 3525–3540. doi:10.1007/s00262-021-02921-8

Höckendorf, U., Yabal, M., Herold, T., Munkhbaatar, E., Rott, S., Jilg, S., et al. (2016). RIPK3 Restricts Myeloid Leukemogenesis by Promoting Cell Death and Differentiation of Leukemia Initiating Cells. *Cancer Cell* 30 (1), 75–91. doi:10.1016/j.ccell.2016.06.002

Hou, J., and Yao, C. (2021). Potential Prognostic Biomarkers of Lung Adenocarcinoma Based on Bioinformatic Analysis. *Biomed. Res. Int.* 2021, 1–14. doi:10.1155/2021/8859996

Hu, F.-F., Liu, C.-J., Liu, L.-L., Zhang, Q., and Guo, A.-Y. (2021). Expression Profile of Immune Checkpoint Genes and Their Roles in Predicting Immunotherapy Response. *Brief Bioinform* 22 (3), bbaa176. doi:10.1093/bib/bbaa176

Ingram, J. P., Thapa, R. J., Fisher, A., Tummers, B., Zhang, T., Yin, C., et al. (2019). ZBP1/DAI Drives RIPK3-Mediated Cell Death Induced by IFNs in the Absence of RIPK1. *J.I.* 203 (5), 1348–1355. doi:10.4049/jimmunol.1900216

Jiao, D., Cai, Z., Choksi, S., Ma, D., Choe, M., Kwon, H.-J., et al. (2018). Necroptosis of Tumor Cells Leads to Tumor Necrosis and Promotes Tumor Metastasis. *Cell Res* 28 (8), 868–870. doi:10.1038/s41422-018-0058-y

Jurisic, V., Vukovic, V., Obradovic, J., Gulyaeva, L. F., Kushlinskii, N. E., and Djordjević, N. (2020). EGFRPolymorphism and Survival of NSCLC Patients Treated with TKIs: A Systematic Review and Meta-Analysis. *J. Oncol.* 2020, 1–14. doi:10.1155/2020/1973241

Kim, S. M., Kwon, O.-J., Hong, Y. K., Kim, J. H., Solca, F., Ha, S.-J., et al. (2012). Activation of IL-6R/JAK1/STAT3 Signaling Induces De Novo Resistance to Irreversible EGFR Inhibitors in Non-small Cell Lung Cancer with T790M Resistance Mutation. *Mol. Cancer Ther.* 11 (10), 2254–2264. doi:10.1158/1535-7163.Mct-12-0311

Krysko, O., Aaes, T. L., Kagan, V. E., D'Herde, K., Bachert, C., Leybaert, L., et al. (2017). Necroptotic Cell Death in Anti-cancer Therapy. *Immunol. Rev.* 280 (1), 207–219. doi:10.1111/imr.12583

Li, S. D., Ma, M., Li, H., Waluszko, A., Sidorenko, T., Schadt, E. E., et al. (2017). Cancer Gene Profiling in Non-small Cell Lung Cancers Reveals Activating Mutations in JAK2 and JAK3 with Therapeutic Implications. *Genome Med.* 9 (1), 89. doi:10.1186/s13073-017-0478-1

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res.* 77 (21), e108–e110. doi:10.1158/0008-5472.Can-17-0307

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells. *Nucleic Acids Res.* 48 (W1), W509–w514. doi:10.1093/nar/gkaa407

Pang, Z., Chen, X., Wang, Y., Wang, Y., Yan, T., Wan, J., et al. (2021). Long Non-coding RNA C5orf64 Is a Potential Indicator for Tumor Microenvironment and Mutation Pattern Remodeling in Lung Adenocarcinoma. *Genomics* 113 (1 Pt 1), 291–304. doi:10.1016/j.ygeno.2020.12.010

Park, H.-H., Kim, H.-R., Park, S.-Y., Hwang, S.-M., Hong, S. M., Park, S., et al. (2021). RIPK3 Activation Induces TRIM28 Derepression in Cancer Cells and Enhances the Anti-tumor Microenvironment. *Mol. Cancer* 20 (1), 107. doi:10.1186/s12943-021-01399-3

Paz-Ares, L., Ciuleanu, T.-E., Cobo, M., Schenker, M., Zurawski, B., Menezes, J., et al. (2021). First-line Nivolumab Plus Ipilimumab Combined with Two Cycles of Chemotherapy in Patients with Non-small-cell Lung Cancer (CheckMate 9LA): an International, Randomised, Open-Label, Phase 3 Trial. *Lancet Oncol.* 22 (2), 198–211. doi:10.1016/s1470-2045(20)30641-0

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E., and Gfeller, D. (2017). Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data. *Elife* 6, e26476. doi:10.7554/eLife.26476

Sancho, D., Joffre, O. P., Keller, A. M., Rogers, N. C., Martínez, D., Hernanz-Falcón, P., et al. (2009). Identification of a Dendritic Cell Receptor that Couples Sensing of Necrosis to Immunity. *Nature* 458 (7240), 899–903. doi:10.1038/nature07750

Seifert, L., Werba, G., Tiwari, S., Giao Ly, N. N., Alothman, S., Alqunaibit, D., et al. (2016). The Necrosome Promotes Pancreatic Oncogenesis via CXCL1 and Mincle-Induced Immune Suppression. *Nature* 532 (7598), 245–249. doi:10.1038/nature17403

Sprooten, J., De Wijngaert, P., Vanmeerbeek, I., Martin, S., Vangheluwe, P., Schlenner, S., et al. (2020). Necroptosis in Immuno-Oncology and Cancer Immunotherapy. *Cells* 9 (8), 1823. doi:10.3390/cells9081823

Stewart, D. J. (2014). Wnt Signaling Pathway in Non-small Cell Lung Cancer. *JNCI J. Natl. Cancer Inst.* 106 (1), djt356. doi:10.1093/jnci/djt356

Tammela, T., Sanchez-Rivera, F. J., Cetinbas, N. M., Wu, K., Joshi, N. S., Helenius, K., et al. (2017). A Wnt-Producing Niche Drives Proliferative Potential and Progression in Lung Adenocarcinoma. *Nature* 545 (7654), 355–359. doi:10.1038/nature22334

Tamminga, M., Hiltermann, T. J. N., Schuuring, E., Timens, W., Fehrmann, R. S., and Groen, H. J. (2020). Immune Microenvironment Composition in Non-small Cell Lung Cancer and its Association with Survival. *Clin. Transl Immunol.* 9 (6), e1142. doi:10.1002/cti2.1142

Tan, Y., Xu, F., Xu, L., and Cui, J. (2022). Long Non-coding RNA LINC01748 Exerts Carcinogenic Effects in Nonsmall Cell Lung Cancer Cell Lines by Regulating the microRNA-520a-5p/HMGA1 axis. *Int. J. Mol. Med.* 49 (2), 22. doi:10.3892/ijmm.2021.5077

Thai, A. A., Solomon, B. J., Sequist, L. V., Gainor, J. F., and Heist, R. S. (2021). Lung Cancer. *The Lancet* 398 (10299), 535–554. doi:10.1016/s0140-6736(21)00312-3

Tran, D. D. H., Kessler, C., Niehus, S. E., Mahnkopf, M., Koch, A., and Tamura, T. (2018). Myc Target Gene, Long Intergenic Noncoding RNA, Linc00176 in Hepatocellular Carcinoma Regulates Cell Cycle and Cell Survival by Titrating Tumor Suppressor microRNAs. *Oncogene* 37 (1), 75–85. doi:10.1038/onc.2017.312

Wang, W., Shi, Q., Wang, S., Zhang, H., and Xu, S. (2020). Ammonia Regulates Chicken Tracheal Cell Necroptosis via the LncRNA-107053293/MiR-148a-3p/FAF1 axis. *J. Hazard. Mater.* 386, 121626. doi:10.1016/j.jhazmat.2019.121626

Wang, W., Zhao, Z., Xu, C., Li, C., Ding, C., Chen, J., et al. (2021). LncRNA FAM83A-AS1 Promotes Lung Adenocarcinoma Progression by Enhancing the Pre-mRNA Stability of FAM83A. *Thorac. Cancer* 12 (10), 1495–1502. doi:10.1111/1759-7714.13928

Wang, X., Avsec, D., Obreza, A., Yousefi, S., Mlinaricõ-Rasõcõan, I., and Simon, H.-U. (2020). A Putative Serine Protease Is Required to Initiate the RIPK3-MLKL-Mediated Necroptotic Death Pathway in Neutrophils. *Front. Pharmacol.* 11, 614928. doi:10.3389/fphar.2020.614928

Woznicki, J. A., Saini, N., Flood, P., Rajaram, S., Lee, C. M., Stamou, P., et al. (2021). TNF-α Synergises with IFN-γ to Induce Caspase-8-jak1/2-STAT1-dependent Death of Intestinal Epithelial Cells. *Cell Death Dis* 12 (10), 864. doi:10.1038/s41419-021-04151-3

Wu, F., Wei, H., Liu, G., and Zhang, Y. (2021). Bioinformatics Profiling of Five Immune-Related lncRNAs for a Prognostic Model of Hepatocellular Carcinoma. *Front. Oncol.* 11, 667904. doi:10.3389/fonc.2021.667904

Xu, F., Huang, X., Li, Y., Chen, Y., and Lin, L. (2021). m6A-related lncRNAs Are Potential Biomarkers for Predicting Prognoses and Immune Responses in Patients with LUAD. *Mol. Ther. - Nucleic Acids* 24, 780–791. doi:10.1016/j.omtn.2021.04.003

Xu, Y., Jin, J., Xu, J., Shao, Y. W., and Fan, Y. (2017). JAK2 Variations and Functions in Lung Adenocarcinoma. *Tumour Biol.* 39 (6), 101042831771114. doi:10.1177/1010428317711140

Yan, J., Wan, P., Choksi, S., and Liu, Z.-G. (2022). Necroptosis and Tumor Progression. *Trends Cancer* 8 (1), 21–27. doi:10.1016/j.trecan.2021.09.003

Yatim, N., Jusforgues-Saklani, H., Orozco, S., Schulz, O., Barreira da Silva, R., Reis e Sousa, C., et al. (2015). RIPK1 and NF-Kb Signaling in Dying Cells Determines Cross-Priming of CD8 + T Cells. *Science* 350 (6258), 328–334. doi:10.1126/science.aad0395

Zeybek, A., Öz, N., Kalemci, S., Tosun, K., Edgünlü, T. G., Kızıltuğ, M. T., et al. (2022). The Role of Wnt Pathway Antagonists in Early-Stage Lung Adenocarcinoma. *Mol. Biol. Rep.* 49 (1), 9–17. doi:10.1007/s11033-021-06759-2

Zhang, W., Fang, D., Li, S., Bao, X., Jiang, L., and Sun, X. (2021). Construction and Validation of a Novel Ferroptosis-Related lncRNA Signature to Predict Prognosis in Colorectal Cancer Patients. *Front. Genet.* 12, 709329. doi:10.3389/fgene.2021.709329

Zhao, Z., Liu, H., Zhou, X., Fang, D., Ou, X., Ye, J., et al. (2021). Necroptosis-Related lncRNAs: Predicting Prognosis and the Distinction between the Cold and Hot Tumors in Gastric Cancer. *J. Oncol.* 2021, 1–16. doi:10.1155/2021/6718443

Zhou, J., Zhang, M., Dong, H., Wang, M., Cheng, Y., Wang, S., et al. (2021). Comprehensive Analysis of Acetylation-Related lncRNAs and Identified AC099850.3 as Prognostic Biomarker in Non-small Cell Lung Cancer. *J. Oncol.* 2021, 1–19. doi:10.1155/2021/4405697

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 March 2022
doi: 10.3389/fgene.2022.855629

MultiGATAE: A Novel Cancer Subtype Identification Method Based on Multi-Omics and Attention Mechanism

Ge Zhang[1], Zhen Peng[1], Chaokun Yan[1], Jianlin Wang[1], Junwei Luo[2] and Huimin Luo[1]*

[1]School of Computer and Information Engineering, Henan University, Kaifeng, China, [2]College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

Cancer is one of the leading causes of death worldwide, which brings an urgent need for its effective treatment. However, cancer is highly heterogeneous, meaning that one cancer can be divided into several subtypes with distinct pathogenesis and outcomes. This is considered as the main problem which limits the precision treatment of cancer. Thus, cancer subtypes identification is of great importance for cancer diagnosis and treatment. In this work, we propose a deep learning method which is based on multi-omics and attention mechanism to effectively identify cancer subtypes. We first used similarity network fusion to integrate multi-omics data to construct a similarity graph. Then, the similarity graph and the feature matrix of the patient are input into a graph autoencoder composed of a graph attention network and omics-level attention mechanism to learn embedding representation. The K-means clustering method is applied to the embedding representation to identify cancer subtypes. The experiment on eight TCGA datasets confirmed that our proposed method performs better for cancer subtypes identification when compared with the other state-of-the-art methods. The source codes of our method are available at https://github.com/kataomoi7/multiGATAE.

Keywords: cancer subtype identification, multi-omics, graph attention network, omics-level attention mechanism, cluster

Edited by:
Juan Wang,
Inner Mongolia University, China

Reviewed by:
Junwei Han,
Harbin Medical University, China
Guosheng Han,
Xiangtan University, China
Yanjuan Li,
Quzhou University, China

Correspondence:
Huimin Luo
luohuimin@henu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *15 January 2022*
Accepted: *14 February 2022*
Published: *21 March 2022*

Citation:
Zhang G, Peng Z, Yan C, Wang J, Luo J and Luo H (2022) MultiGATAE: A Novel Cancer Subtype Identification Method Based on Multi-Omics and Attention Mechanism. Front. Genet. 13:855629. doi: 10.3389/fgene.2022.855629

1 INTRODUCTION

Cancer is one of the leading causes of death worldwide and is a serious threat to human health (Sung et al., 2021). Cancer is extremely heterogeneous, and distinct molecular subtypes have different clinical outcomes (Zhao and Yan, 2019). The goal of cancer subtype identification is to discover patient groups with different clinical outcomes, thus facilitating personalized treatment (Liang et al., 2021). For instance, four potential molecular subtypes of gastric cancer, i.e., EBV, MSI, GS, and CIN, were uncovered by The Cancer Genome Atlas (TCGA) project (Bass et al., 2014), and each of these four molecular subtypes has specific clinical significance signatures (Sohn et al., 2017). Therefore, cancer subtype identification is of great importance.

The rapid development of high throughput sequencing technology has made a massive amount of omics data from the different levels available. This provides an opportunity to investigate the heterogeneity of cancer and to identify cancer subtypes (Zhao et al., 2019). Since omics data lack labels associated with cancer subtypes, cancer subtype identification is usually addressed using clustering (Xu et al., 2019). Earlier studies usually used only single-omics data; however, single-omics data provide only a very limited view on cancer subtype identification (Gomez-Cabrero et al., 2014; Le Van et al., 2016). Thus, many researchers integrate multi-omics data to identify cancer subtypes.

Yang et al. (2021a) proposed a computational method called Deep Subspace Mutual Learning (DSML). DSML constructed branching models for each type of omics data and then constructed a main stem model to optimize the feature representation learned from single-omics data. Finally, spectral clustering was applied to the learned representation to identify cancer subtypes. Chaudhary et al. (2018) applied an autoencoder to process multi-omics data to gain low-dimensional features, then the features were further filtered using Cox-PH analysis. Finally, K-means was applied to the resulting features to cluster cancer subtypes. While using multi-omics data provides a comprehensive view, it also introduces additional computational costs.

Apart from the differences in the used data, some studies have typically focused on analyzing the features of omics data and the distribution of each data type to identify cancer subtypes. Shen et al. (2009) proposed an integrative clustering method named iCluster. iCluster models the subtypes of cancer as latent variables which can be simultaneously estimated from the omics data. Yang et al. (2021) introduced a deep-learning method named Subtype-GAN for cancer subtyping. Subtype-GAN consists of three modules: encoder, decoder, and discriminator. The encoder takes multi-omics data as input and encodes them into low-dimensional representation. The decoder reconstructs the original input using the low-dimensional representation. The discriminator is used to force the representation encoded by the encoder to follow the prior Gaussian distribution. Finally, Consensus GMM clustering is applied to the low-dimensional representation to determine the most appropriate clustering number and to predict the subtype results. However, these methods are limited by strong assumptions on the distribution of the omics data (Song et al., 2021). Noise in the omics data may affect the results of cancer subtyping. Similarity-based approaches for multi-omics data can avoid this problem (Song et al., 2021). Wang et al. (2014) proposed a method named Similarity Network Fusion (SNF) for integrating multi-omics data. SNF first generates a sample similarity network for each type of data and then iteratively fuses these similarity networks. Zhao and Yan (2019) proposed a cancer subtyping method named Molecular and Clinical Networks Fusion (MCNF), which integrates multi-omics and clinical data. MCNF first applies unsupervised random forest to multi-omics and clinical data to generate a patient affinity network and then uses random walk to fuse the patient affinity networks. After obtaining the fused network, PAM clustering is used to identify the cancer subtypes. Yang et al. (2021b) introduced a clustering method, Deep Subspace Fusion Clustering (DSFC), for cancer subtype prediction. DSFC calculates data self-expressiveness to generate a patient similarity network, and then fuses these patient similarity networks to gain a combined network. Finally, spectral clustering is performed on the combined similarity network to find cancer subtypes. Similarity-based approaches usually just use the omics data to generate a similarity network, and completely disregard the feature information of the omics data in subsequent calculations. This may lead to incomplete subtype results.

To make full use of the feature information of the omics data and the similarity graph, a graph-based neural network was used because it takes both the feature information as well as the similarity graph into consideration (Wu et al., 2021). In this work, we proposed a deep-learning method named multiGATAE for cancer subtype identification. multiGATAE first applies multi-omics data to construct a similarity graph and then establish a graph autoencoder network which is composed of a graph attention network and an omics-level attention mechanism to obtain the embedding representation. Finally, the K-means clustering method is applied to the embedding representation to identify cancer subtypes. multiGATAE was compared with serval state-of-the-art methods on eight public cancer datasets, and the results demonstrated that our proposed method performs better.

The remainder of this article is organized as follows. In **section 2**, we present the proposed method. The datasets we used and the experiment results are shown in **section 3**. In **section 4**, we conclude this article and discuss the future work.

2 MATERIALS AND METHODS

In this section, the details of our proposed-method multiGATAE are described. Our proposed method consists of three parts. Firstly, a similarity graph is constructed by integrating multi-omics data. Then, the similarity graph and omics data are input to a graph autoencoder composed of a graph attention network and omics-level attention mechanism to learn the embedding representation. Finally, the K-means method is applied to the embedding representation to identify the cancer subtypes. The workflow of multiGATAE is shown in **Figure 1**.

2.1 Construction of Similarity Graph

A network fusion method named SNF (Wang et al., 2014) was used to construct the similarity graph. SNF first generated specific similarity graphs for each omics, and then iteratively integrated them to construct the combined similarity graph. Suppose that there are n patients and m views (such as mRNA, miRNA, and DNA methylation). The similarity graph is defined as a graph G = (V, E), where V is the set of patients \{x_1, x_2, x_3 . . . , x_n\} and the edges E correspond to the similarity between vertices v ∈ V. The edge weights are represented by an n × n similarity matrix W, and W is computed by **Eq. 1**.

$$W_{i,j} = \exp\left(-\frac{\phi^2\left(x_i, x_j\right)}{\alpha\gamma_{i,j}}\right) \tag{1}$$

where α is a hyperparameter, ϕ (x_i, x_j) is the Euclidean distance between patients $x_{i,}$ and x_j, and $\gamma_{i,j}$ is used to eliminate the scaling problem. In order to compute the fused matrix from multiple types of data, the similarity matrix is normalized as **Eq. 2**.

$$P_{i,j} = \begin{cases} \dfrac{W_{i,j}}{2\sum_{k\neq i} W_{i,k}} & j \neq i \\ \dfrac{1}{2} & j = i \end{cases} \tag{2}$$

assuming N_i is a set of x_i 's neighbors. Then, the local affinity matrix S is calculated by **Eq. 3**.

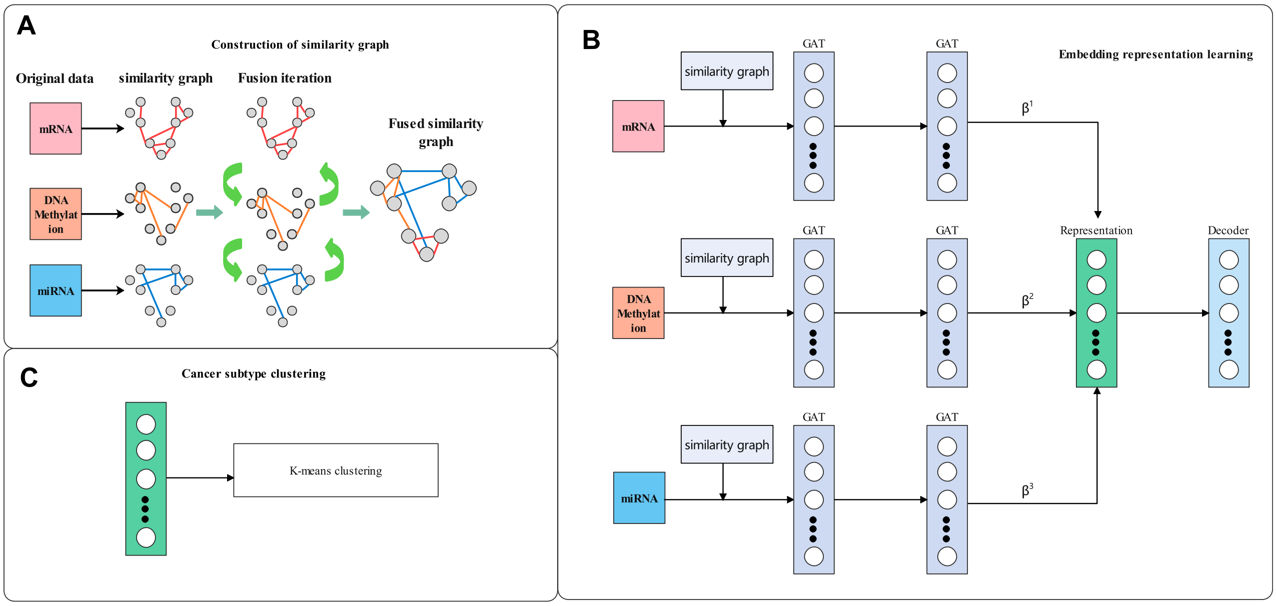

FIGURE 1 | Workflow of multiGATAE. **(A)** Construction of similarity graph. **(B)** Embedding representation learning. **(C)** Cancer subtype clustering.

$$S_{i,j} = \begin{cases} \dfrac{W_{i,j}}{\sum_{k\in N_i} W_{j,k}} & j \in N_i \\ 0 & otherwise \end{cases} \tag{3}$$

Let $P_t^{(h)}$ represent the normalized similarity matrix of h-th type data ($1 \leq h \leq m$) in the t-th iteration; $P_t^{(h)}$ is updated according to **Eq. 4**.

$$P_{t+1}^{(h)} = S^{(h)} \left(\frac{\sum_{k\neq h} P_t^{(k)}}{m-1} \right) \left(S^{(h)}\right)^T \tag{4}$$

where $S^{(h)}$ represents the local affinity matrix of the h-th type data. Through this process of continuous iterative fusion, the combined similarity graph, which contains complementary information from three omics datasets, is finally obtained and then taken as the input of multiGATAE to learn the embedding representation.

2.2 Embedding Representation Learning

Cancer subtype identification is a typical clustering problem because of the lack of labels associated with the cancer subtypes (Xu et al., 2019). A key problem of clustering is how to capture the feature information of the nodes and the relationship between the nodes (Wang et al., 2019). A graph-based neural network may be able to solve this problem because it considers both the feature information of the nodes as well as the similarity relationships (Wu et al., 2021). In this work, we constructed a graph autoencoder composed of a graph attention network and omics-level attention mechanism to learn the embedding representation. We first introduce the Graph Convolutional Network (GCN) (Kipf and Welling, 2016a). The aim of the GCN is to learn a latent representation Z based on the node feature matrix X, which describes every node in the graph, and a similarity matrix A, which encodes the similarities between the nodes. The layer-wise propagation rule of GCN can be formulated as **Eq. 5**.

$$Z^L = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} Z^{L-1} W^{L-1}\right) \tag{5}$$

where $\tilde{A} = A + E$, which is a similarity matrix adding self-connections. $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$. $\sigma(\cdot)$ is a nonlinear activation function. Z^L is the output of the L layer. However, a limitation of GCN is that it does not assign different weights to different nodes in the neighborhood (Veličković et al., 2017). In a practical situation, different neighbor nodes may play different roles for the current node. Therefore, we chose to use GAT (Veličković et al., 2017) which aggregates the neighbor nodes through the self-attention mechanism (Vaswani et al., 2017) and enables the adaptive assignment of weights to different neighbors. GAT first computes the attention coefficients by **Eq. 6**

$$e_{ij} = \alpha\left(Wx_i, Wx_j\right) \tag{6}$$

where $\alpha(\cdot)$ is a shared attentional mechanism, and x_i and x_j represent the features of node i and node j, respectively. The attention coefficients indicate the importance of node j's features to node i. To make the attention coefficients comparable across different nodes, the softmax function is used to normalize them:

$$\alpha_{ij} = softmax\left(e_{ij}\right) \tag{7}$$

The normalized attention coefficients are then used to compute the final output Z as **Eq. 8**

$$Z^L = \sigma\left(\alpha_{ij} \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} Z^{L-1} W^{L-1}\right) \tag{8}$$

In order to make the output Z more approximate to the similarity graph A, we propose an omics-level attention mechanism to aggregate the output of multi-omics. The attention score is defined as **Eq. 9**

$$w^i = v^T tanh\left(W_z \cdot Z^i + W_a \cdot A\right) \tag{9}$$

where w^i and Z^i represent the attention score and the output of omics i. v, W_z, and W_a are trainable vectors. As mentioned above, we normalize the omics-level attention scores using the softmax function as **Eq. 10**

$$\beta^i = softmax\left(w^i\right) \tag{10}$$

We then obtain the final representation Z^{final} by aggregating the output of multi-omics as **Eq. 11**.

$$Z^{final} = \sum\left(\beta^i Z^i\right) \tag{11}$$

The final representation Z^{final} is input into the decoder to reconstruct the original similarity graph. The decoder is defined as **Eq. 12** (Kipf and Welling, 2016b).

$$\hat{A} = \tau\left(Z^{final} Z^{final^T}\right) \tag{12}$$

After the neural network optimization is completed, a standard clustering method named K-means (Ding and He, 2004) is applied to the final representation Z^{final} to identify cancer subtypes.

3 EXPERIMENTS AND RESULTS

To evaluate the performance of our proposed-method multiGATAE, we compared it with eight state-of-the-art clustering methods, namely, DLSF (Zhang et al., 2022), subtype-WESLR (Song et al., 2021), SNF (Wang et al., 2014), NEMO (Rappoport and Shamir, 2019), iClusterBayes (Mo et al., 2018), moCluster (Meng et al., 2016), LRAcluster (Wu et al., 2015), and PFA (Shi et al., 2017) on eight public cancer multi-omics datasets. Here, we first introduce the details of these eight state-of-the-art methods, then we introduce the datasets used in this section and show the experiment results on these eight datasets.

- NEMO is a multi-omics clustering method based on the neighborhood. NEMO first constructs inter-patient similarity network for each omics and then integrates these networks into one network. Finally, the network is used for clustering.
- iClusterBayes adopts latent variables to capture the inherent structure of multi-omics datasets. The latent variable space is then used to identify cancer subtypes.
- moCluster investigates the joint patterns among multi-omics datasets. It uses multi-block multivariate analysis to define a set of latent variables and passes it to the clustering method to identify the cancer subtypes.
- LRAcluster discovers shared latent subspaces of the multi-omics data based on the integrative probabilistic model. The shared latent subspaces can be applied to identify subtypes.
- SNF is a network fusion method. It generates similarity networks for single-omics data and fuses these independent similarity networks into a combined network. This combined network can be used for cancer clustering.
- PFA is a pattern fusion analysis framework. It can capture intrinsic structure from multi-omics data for cancer clustering.
- subtype-WESLR uses a weighted ensemble strategy to fuse base clustering obtained by distinct methods as prior knowledge and maps each omics data into a common latent subspace. The common latent subspace is optimized iteratively to identify cancer subtypes.
- DLSF is a novel cancer clustering method based on deep neural network. It uses a cycle autoencoder which has a shared self-expressive layer to merge latent representation at each omics level into a fused representation at the multi-omics level. The fused representation can be used to identify cancer subtypes.

3.1 Data Set and Data Preprocessing

Eight TCGA cancer public datasets including kidney renal clear cell carcinoma (KIRC), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), skin cutaneous melanoma (SKCM), lung squamous cell carcinoma (LUSC), glioblastoma multiforme (GBM), liver hepatocellular carcinoma (LIHC), and ovarian serous cystadenocarcinoma (OV) were used in this work. They were downloaded from TCGA (Cancer Genome Atlas Research Network, 2008), and each of them contains four types of data: miRNA expression, mRNA expression, DNA methylation, and clinical profiles. These three datasets are preprocessed by the following steps. Outlier removal is the first step. The features with missing values in more than 20% samples were deleted. Similarly, samples which have more than 20% features were removed. Finally, 206 samples in KIRC, 623 in BRCA, 214 in COAD, 439 in SKCM, 271 in GBM, 337 in LUSC, 404 in LIHC, and 290 in OV remained in this step. The next step is missing-data imputation. K nearest neighbor (Troyanskaya et al., 2001) imputation had been applied to impute the missing values. Finally, all of these datasets were normalized as **Eq. 13**:

$$\tilde{f} = \frac{f - E(f)}{\sqrt{Var(f)}} \tag{13}$$

where $E(f)$ is the mean of f, and Var(f) is the variance of f.

3.2 Optimal Number of Clusters

Since the K-means clustering method cannot automatically determine the optimal number of clusters, a silhouette width (Rand, 1971) was adopted to find the optimal clustering number. The parameters of our proposed method were also adjusted according to the silhouette width. We determined the optimal hidden layers, learning rate (Lr), and the dropout according to the grid search method. The optimal hidden layers were 2, Lr was

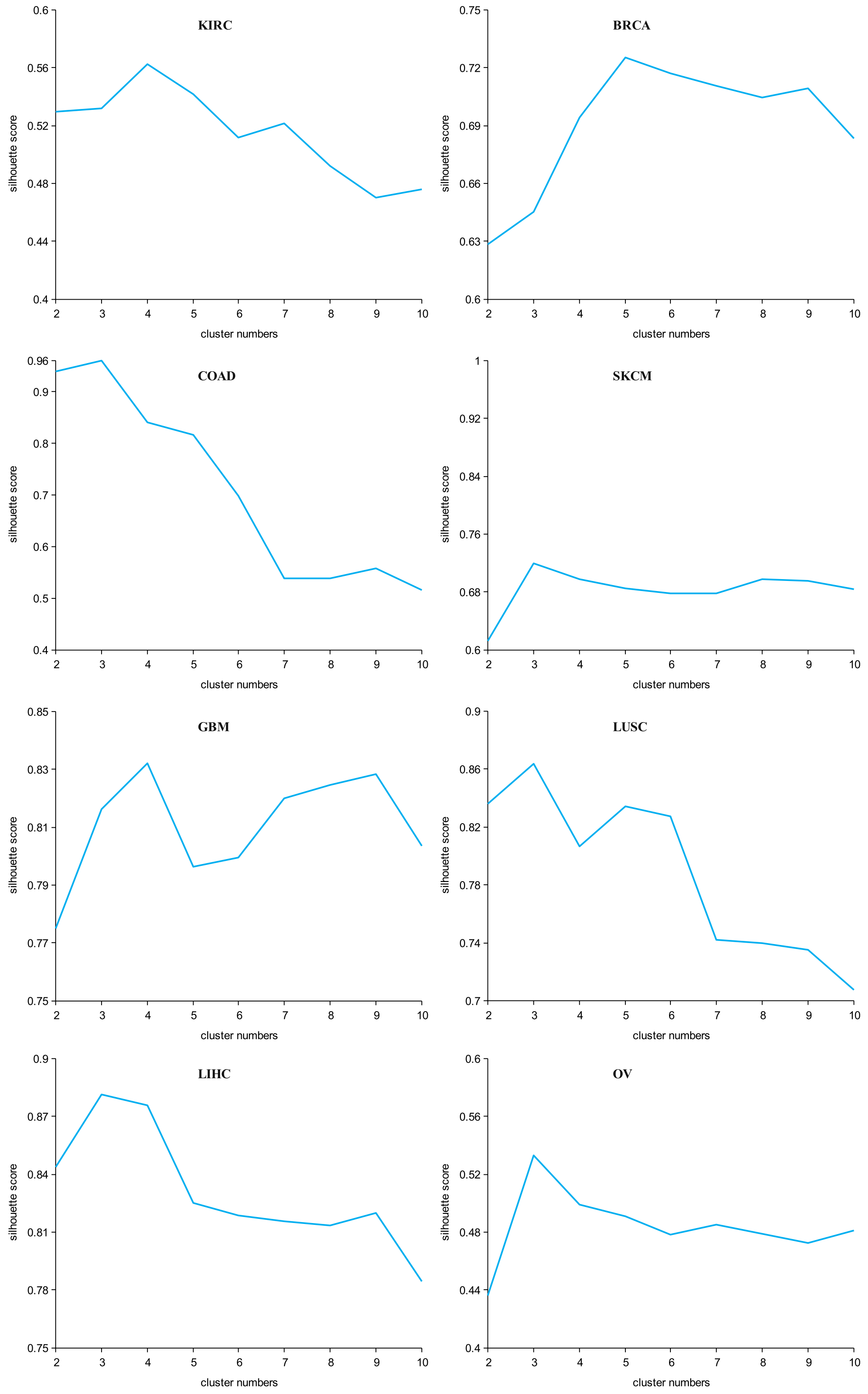

FIGURE 2 | Silhouette width multiGATAE achieved on the eight datasets.

0.01, and dropout was 0.5, which achieved the best silhouette width and were finally applied in this work. In addition, for the compared methods, the parameters as given in their original articles were slightly modified to make them more suitable for our dataset. The silhouette width that our proposed method achieved on the eight datasets is shown in **Figure 2**.

TABLE 1 | Results of comparison methods and the proposed method, the first value is cluster number and the second is the negative log10 *p*-value.

Metric	Algorithm	KIRC	BRCA	COAD	SKCM	GBM	LUSC	LIHC	OV
p-value	NEMO	3/4.48	4/0.31	4/0.96	4/2.74	3/2.96	3/2.15	3/1.60	3/0.05
	iClusterBayes	4/2.51	5/1.06	4/0.09	4/1.85	3/0.22	3/1.24	3/1.11	3/1.48
	moCluster	3/2.82	5/3.31	3/1.04	4/2.98	3/1.96	3/2.31	2/1.02	3/1.60
	LRAcluster	3/2.07	5/2.23	4/1.17	3/3.25	3/2.00	3/2.35	3/0.39	3/2.96
	SNF	3/3.40	4/2.82	3/1.07	4/2.31	3/2.92	3/2.03	3/1.54	3/1.15
	PFA	2/2.08	5/2.89	3/1.00	4/2.64	2/2.23	3/1.04	2/2.64	3/0.05
	subtype-WESLR	4/4.76	**5/5.24**	4/2.43	5/5.00	3/3.84	5/2.30	**4/5.21**	3/3.44
	DLSF	4/2.76	3/1.89	4/0.05	5/3.85	**5/4.53**	3/0.11	3/3.15	4/0.03
	multiGATAE	**4/5.30**	5/1.68	**3/3.12**	**3/5.52**	4/4.0	**3/2.60**	3/3.51	**3/5.40**
C-index	NEMO	0.654	0.526	0.557	0.56	0.533	0.565	0.535	0.514
	iClusterBayes	0.617	0.535	0.552	0.542	0.515	0.516	0.557	0.536
	moCluster	0.626	0.588	0.543	0.566	0.538	0.576	0.553	0.56
	LRAcluster	0.597	0.539	0.579	0.562	0.551	0.572	0.541	0.584 2
	SNF	0.638	0.587	0.568	0.565	0.544	0.566	0.538	0.543
	PFA	0.581	0.544	0.57	0.564	0.538	0.52	0.555	0.567
	subtype-WESLR	**0.66**	0.595	0.632	0.58	0.559	0.587	0.594	0.581
	DLSF	0.623	**0.627**	0.539	0.578	0.582	0.527	0.575	0.563
	multiGATAE	0.618	0.574	**0.644**	**0.594**	**0.587**	**0.614**	**0.599**	**0.61**

Bold values indicates the best values.

Since the sample size of the cancer omics data is not very large, an excessive number of clusters may introduce bias. Thus, the number of clusters adopted in this work ranged from two to 10. The range of the silhouette width was from −1 to 1, and the closer it was to 1 meant the better the clustering performance was. We can see from **Figure 2** that within a certain range, the silhouette width exhibited an increasing tendency. After reaching the optimal cluster number, the silhouette width started to gradually decrease. Specifically, for the KIRC datasets, the silhouette width achieved was the best when the cluster number was set to 4. This meant that the best clustering results were obtained when KIRC was clustered into four subtypes. Similarly, the BRCA was finally clustered into five subtypes, the COAD into three subtypes, the SKCM into three subtypes, the GBM into four subtypes, the LUSC into three subtypes, the LIHC into three subtypes, and the OV dataset into three subtypes. We can see that all the optimal numbers are within five, and this may indicate that the amount of available data was not sufficient to identify numerous cancer subtypes.

3.3 Comparison With Other Methods

To validate the performance of our proposed-method multiGATAE, we compared it with eight state-of-the-art methods on eight cancer datasets. Due to the lack of labels for the omics data, the negative log10 *p*-value and C-index of log-rank test were used as the metric. The log-rank test of the Cox regression (Hosmer and Lemeshow, 1999) is a statistical model and is used to assess the difference in survival profiles between subtypes. The *p*-value represents whether the observed differences are significant. If the *p*-value is less than 0.05, the observed subtypes are considered significantly different. To facilitate comparison, the negative and log operations were performed. The C-index was used to assess the predictive performance of the survival model. The results are shown in **Table 1**.

It can be seen from **Table 1** that our proposed-method multiGATAE achieved the best performance on most datasets. Specifically, on the KIRC dataset, the negative log10 *p*-value that multiGATAE achieved was 5.30, which is 0.54 higher than the best remaining method subtype-WESLR. As for COAD, SKCM, LUSC, and OV datasets, the multiGATAE achieved 0.69, 0.52, 0.3, and 1.96 improvements compared with the best remaining method. As for the C-index, except for KIRC and BRCA, multiGATAE outperformed the compared methods on the other datasets. This demonstrates that the subtypes identified by our proposed method are indeed survival distinct. To illustrate the difference between the subtypes identified by our proposed method clearly, the survival curves for the eight cancer datasets are shown in **Figure 3**. As can be seen in **Figure 3**, except for BRCA, the cancer subtypes identified by our method on the other seven datasets all exhibit significantly different survival curves. The survival curve was significantly different between the subtypes, and this difference became progressively greater with time, indicating that the probability of survival varies between subtypes. For example, in the case of KRIC, subtype 3 showed a very low survival probability compared to the other subtypes when the time was above 1,000. This suggests that our method could identify groups of patients with different prognoses and help with precision treatment.

3.4 Analysis of Identified Subtypes on Lung Squamous Cell Carcinoma

In order to further validate our proposed method, we selected LUSC for a relevant biological analysis of identified subtypes. There were three subtypes identified by our proposed method, and in order to discover the differences at the molecular level between these three subtypes, we performed differential

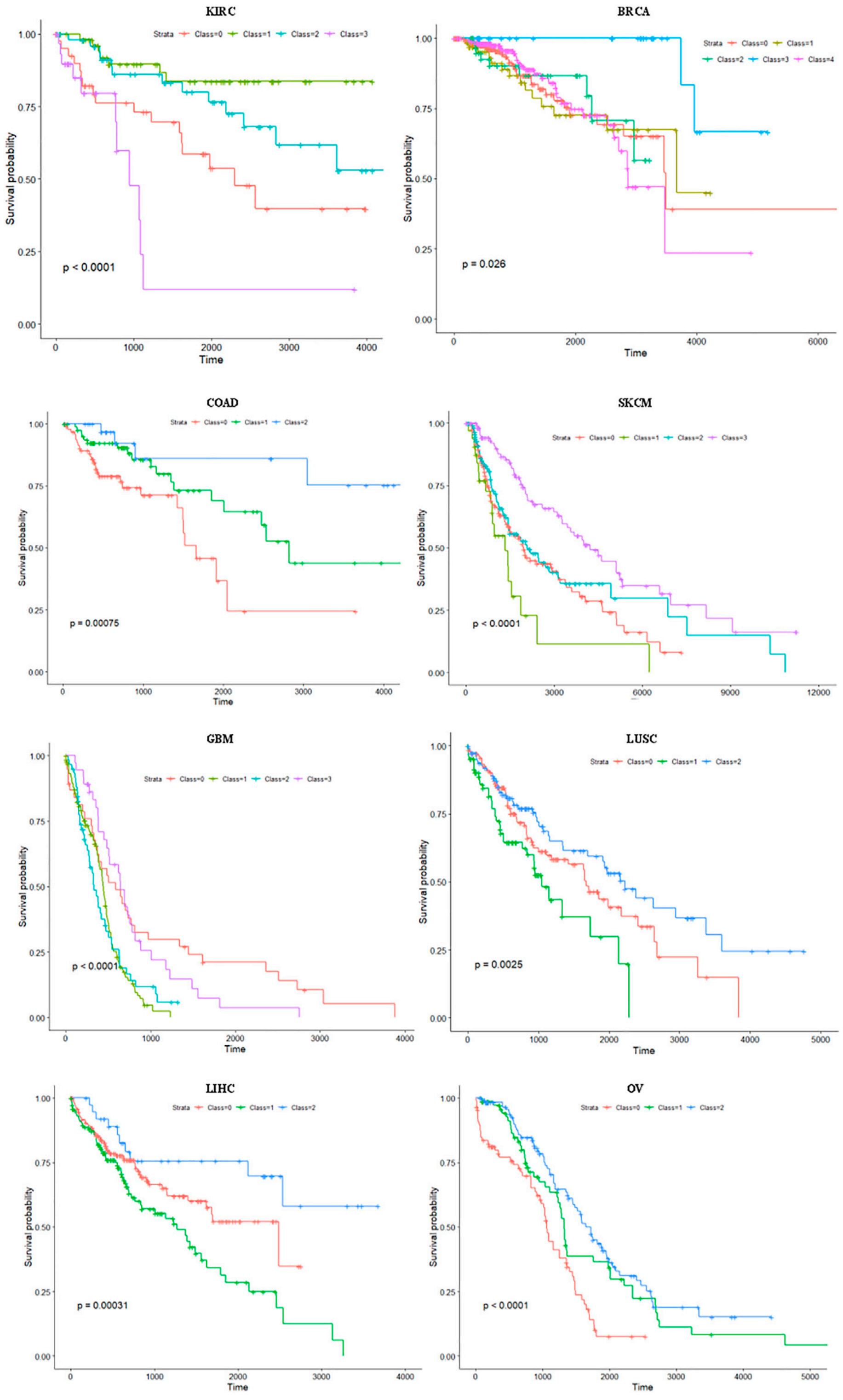

FIGURE 3 | Survival curves for eight cancer datasets.

mRNA expressions by R package limma (Smyth, 2005). The differentially expressed mRNAs are shown by the heat map in **Figure 4**. As we can see from **Figure 4**, there are mRNAs which are significantly differentially expressed. This demonstrates that the subtypes identified by our proposed method have molecular-level differences.

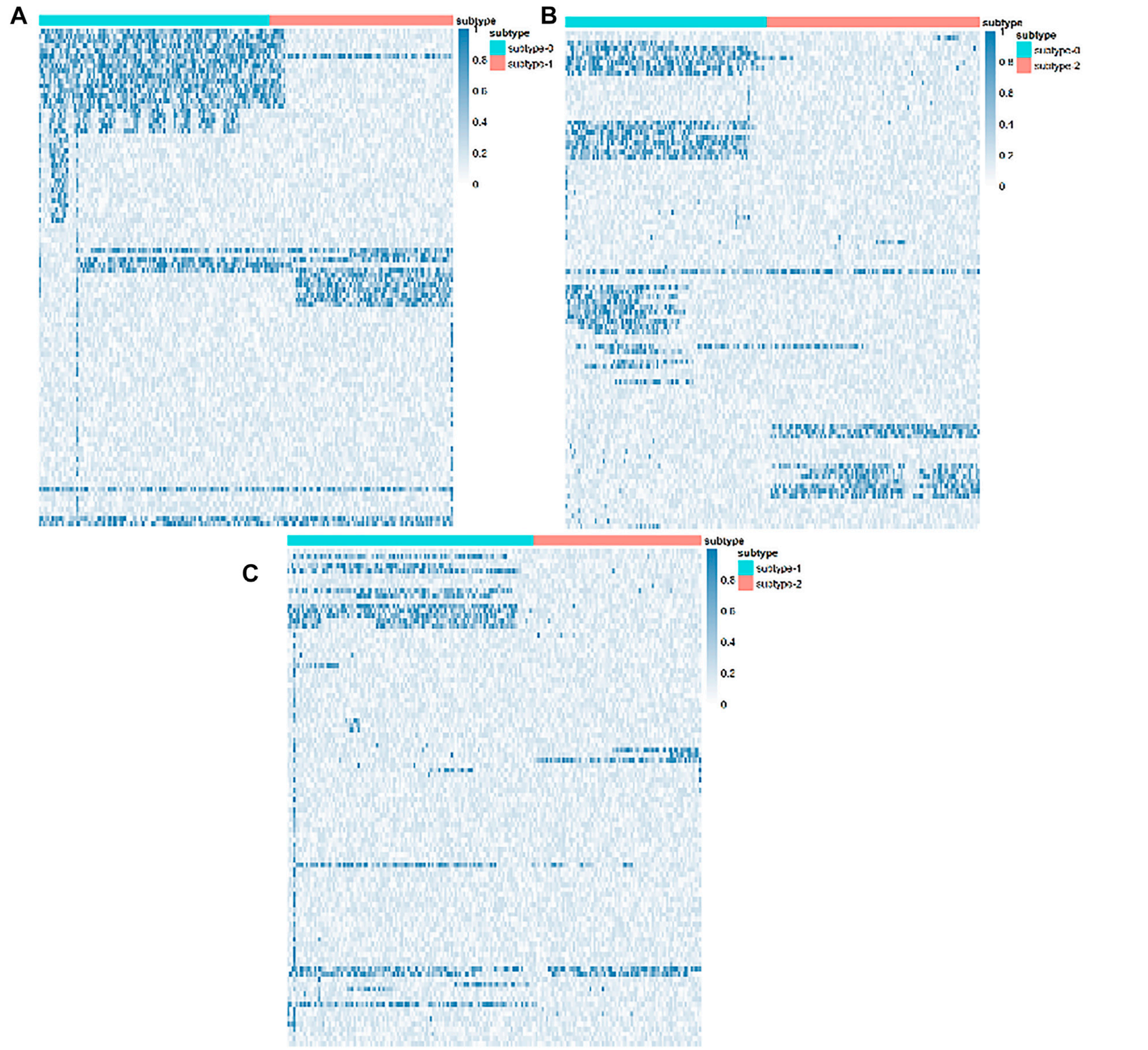

FIGURE 4 | Differentially expressed mRNAs of the LUCS subtypes. **(A)** Subtype 0 and subtype 1. **(B)** Subtype 0 and subtype 2. **(C)** Subtype 1 and subtype 2.

TABLE 2 | Results of multi-omics and single-omics, the first value is cluster number and the second is the negative log10 *p*-value.

	KIRC	BRCA	COAD	SKCM	GBM	LUSC	LIHC	OV
mRNA	4/1.31	3/0.20	3/0.24	3/1.52	4/1.27	3/0.38	3/0.8	3/0.97
DNA methylation	3/1.75	3/0.71	3/0.73	3/1.69	4/1.71	3/0.03	3/0.87	3/2.85
miRNA	4/1.57	4/0.39	3/0.98	3/1.98	4/1.24	4/0.53	3/0.667	3/1.35
Multi-omics	4/5.30	5/1.68	3/3.12	3/5.52	4/4.0	3/2.60	3/3.51	3/5.40

3.5 Effectiveness of Multi-Omics Data

In this work, we used multi-omics data in order to obtain a comprehensive view on cancer subtype identification. To investigate the difference in results between single-omics and multi-omics data, we carried out experiments with single-omics data. The results are shown in **Table 2**. It can be seen from **Table 2** that multiGATAE with multi-omics data performed better than using single-omics data. This suggests that integrating multi-omics data helps to capture a better embedded expression and thus identify more stable cancer subtypes. Besides, the DNA methylation data showed relatively better results compared with the other omics data. This may indicate that the DNA methylation data contains more information that facilitates cancer subtype identification.

4 CONCLUSION

Cancer is a highly heterogeneous disease that causes a large number of deaths every year. Cancer subtype identification aims to identify groups of patients with different clinical outcomes for precise treatment. In this work, we proposed a novel cancer subtype identification method named multiGATAE. multiGATAE first constructed a similarity graph by integrating multi-omics data, and then input the similarity graph and the omics data into a graph autoencoder network which is composed of a graph attention network and an omics-level attention mechanism to obtain the embedding representation. Once gaining the embedding representation, the K-means clustering method was applied to it to identify subtypes. multiGATAE was compared with eight state-of-the-art methods on eight public cancer datasets. The results demonstrate that our proposed method can identify distinct subtypes with different survival outcomes. In the future, we consider integrating more data to develop our method. In addition, when learning embedding representation, taking clustering losses into consideration is also a way to improve our method.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/.

AUTHOR CONTRIBUTIONS

GZ and ZP conceived and designed the approach. ZP performed the experiments. HL and JL analyzed the data. GZ and ZP wrote the manuscript. CY and JW supervised the whole study process and revised the manuscript. All authors have read and approved the final version of manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos 61802113, 61802114, 61972134); Science and Technology Development Plan Project of Henan Province (Nos 202102210173, 212102210091); China Postdoctoral Science Foundation (No. 2020M672212); and Henan Province Postdoctoral Research Project Founding.

REFERENCES

Bass, A. J., Thorsson, V., Shmulevich, I., Reynolds, S. M., Miller, M., Bernard, B., et al. (2014). Comprehensive Molecular Characterization of Gastric Adenocarcinoma. *Nature* 513, 202–209. doi:10.1038/nature13480

Cancer Genome Atlas Research Network (2008). Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways. *Nature* 455, 1061. doi:10.1038/nature07385

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res.* 24, 1248–1259. doi:10.1158/1078-0432.CCR-17-0853

Ding, C., and He, X. (2004). "K-means Clustering via Principal Component Analysis." in Proceedings of the 21 st International Conference on Machine Learning, Banff, Canada, July 2004. doi:10.1145/1015330.1015408

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data Integration in the Era of Omics: Current and Future Challenges. *BMC Syst. Biol.* 8, 1–10. doi:10.1186/1752-0509-8-S2-I1

Hosmer, D. W., and Lemeshow, S. (1999). *Applied Survival Analysis: Time-To-Event*, Vol. 317. Hoboken, New Jersey, United States: Wiley-Interscience.

Kipf, T. N., and Welling, M. (2016a). Semi-supervised Classification with Graph Convolutional Networks. *arXiv*. arXiv preprint arXiv:1609.02907.

Kipf, T. N., and Welling, M. (2016b). Variational Graph Auto-Encoders. *arXiv*. arXiv preprint arXiv:1611.07308.

Le Van, T., Van Leeuwen, M., Carolina Fierro, A., De Maeyer, D., Van den Eynden, J., Verbeke, L., et al. (2016). Simultaneous Discovery of Cancer Subtypes and Subtype Features by Molecular Data Integration. *Bioinformatics* 32, i445–i454. doi:10.1093/bioinformatics/btw434

Liang, C., Shang, M., and Luo, J. (2021). Cancer Subtype Identification by Consensus Guided Graph Autoencoders. *Bioinformatics* 37, 4779–4786. doi:10.1093/bioinformatics/btab535

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2016). Mocluster: Identifying Joint Patterns across Multiple Omics Data Sets. *J. proteome Res.* 15, 755–765. doi:10.1021/acs.jproteome.5b00824

Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K. S., and Hilsenbeck, S. G. (2018). A Fully Bayesian Latent Variable Model for Integrative Clustering Analysis of Multi-type Omics Data. *Biostatistics* 19, 71–86. doi:10.1093/biostatistics/kxx017

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* 66, 846–850. doi:10.1080/01621459.1971.10482356

Rappoport, N., and Shamir, R. (2019). Nemo: Cancer Subtyping by Integration of Partial Multi-Omic Data. *Bioinformatics* 35, 3348–3356. doi:10.1093/bioinformatics/btz058

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis. *Bioinformatics* 25, 2906–2912. doi:10.1093/bioinformatics/btp543

Shi, Q., Zhang, C., Peng, M., Yu, X., Zeng, T., Liu, J., et al. (2017). Pattern Fusion Analysis by Adaptive Alignment of Multiple Heterogeneous Omics Data. *Bioinformatics* 33, 2706–2714. doi:10.1093/bioinformatics/btx176

Smyth, G. K. (2005). *Limma: Linear Models for Microarray Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Berlin/Heidelberg, Germany: Springer, 397–420. doi:10.1007/0-387-29362-0_23

Sohn, B. H., Hwang, J.-E., Jang, H.-J., Lee, H.-S., Oh, S. C., Shim, J.-J., et al. (2017). Clinical Significance of Four Molecular Subtypes of Gastric Cancer Identified by the Cancer Genome Atlas Project. *Clin. Cancer Res.* 23, 4441–4449. doi:10.1158/1078-0432.CCR-16-2211

Song, W., Wang, W., and Dai, D.-Q. (2021). Subtype-WESLR: Identifying Cancer Subtype with Weighted Ensemble Sparse Latent Representation of Multi-View Data. *Brief. Bioinform.* 23, bbab398. doi:10.1093/bib/bbab398

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a Cancer J. clinicians* 71, 209–249. doi:10.3322/caac.21660

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing Value Estimation Methods for Dna Microarrays. *Bioinformatics* 17, 520–525. doi:10.1093/bioinformatics/17.6.520

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, Vancouver, December 2004. Editors L. K. Saul, Y. Weiss, and L. Bottou, 5998–6008.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph Attention Networks. *arXiv*. arXiv preprint arXiv:1710.10903.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat. Methods* 11, 333. doi:10.1038/nmeth.2810

Wang, C., Pan, S., Hu, R., Long, G., Jiang, J., and Zhang, C. (2019). "Attributed Graph Clustering: A Deep Attentional Embedding Approach," in Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao China, August 2019. doi:10.24963/ijcai.2019/509

Wu, D., Wang, D., Zhang, M. Q., and Gu, J. (2015). Fast Dimension Reduction and Integrative Clustering of Multi-Omics Data Using Low-Rank Approximation: Application to Cancer Molecular Classification. *BMC genomics* 16, 1022. doi:10.1186/s12864-015-2223-8

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Networks Learn. Syst.* 32, 4–24. doi:10.1109/TNNLS.2020.2978386

Xu, A., Chen, J., Peng, H., Han, G., and Cai, H. (2019). Simultaneous Interrogation of Cancer Omics to Identify Subtypes with Significant Clinical Differences. *Front. Genet.* 10, 236. doi:10.3389/fgene.2019.00236

Yang, B., Xin, T.-T., Pang, S.-M., Wang, M., and Wang, Y.-J. (2021a). Deep Subspace Mutual Learning for Cancer Subtypes Prediction. *Bioinformatics* 37, 3715–3722. doi:10.1093/bioinformatics/btab625

Yang, B., Zhang, Y., Pang, S., Shang, X., Zhao, X., and Han, M. (2021b). Integrating Multi-Omic Data with Deep Subspace Fusion Clustering for Cancer Subtype Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 216–226. doi:10.1109/TCBB.2019.2951413

Yang, H., Chen, R., Li, D., and Wang, Z. (2021c). Subtype-GAN: a Deep Learning Approach for Integrative Cancer Subtyping of Multi-Omics Data. *Bioinformatics* 37, 2231–2237. doi:10.1093/bioinformatics/btab109

Zhang, C., Chen, Y., Zeng, T., Zhang, C., and Chen, L. (2022). Deep Latent Space Fusion for Adaptive Representation of Heterogeneous Multi-Omics Data. *Brief. Bioinform.*, Bbab600. doi:10.1093/bib/bbab600

Zhao, L., Lee, V. H., Ng, M. K., Yan, H., and Bijlsma, M. F. (2019). Molecular Subtyping of Cancer: Current Status and Moving toward Clinical Applications. *Brief. Bioinformatics* 20, 572–584. doi:10.1093/bib/bby026

Zhao, L., and Yan, H. (2019). Mcnf: A Novel Method for Cancer Subtyping by Integrating Multi-Omics and Clinical Data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 17, 1682–1690. doi:10.1109/TCBB.2019.2910515

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fgene.2022.875112

PredMHC: An Effective Predictor of Major Histocompatibility Complex Using Mixed Features

*Dong Chen and Yanjuan Li**

College of Electrical and Information Engineering, Quzhou University, Quzhou, China

The major histocompatibility complex (MHC) is a large locus on vertebrate DNA that contains a tightly linked set of polymorphic genes encoding cell surface proteins essential for the adaptive immune system. The groups of proteins encoded in the MHC play an important role in the adaptive immune system. Therefore, the accurate identification of the MHC is necessary to understand its role in the adaptive immune system. An effective predictor called PredMHC is established in this study to identify the MHC from protein sequences. Firstly, PredMHC encoded a protein sequence with mixed features including 188D, APAAC, KSCTriad, CKSAAGP, and PAAC. Secondly, three classifiers including SGD, SMO, and random forest were trained on the mixed features of the protein sequence. Finally, the prediction result was obtained by the voting of the three classifiers. The experimental results of the 10-fold cross-validation test in the training dataset showed that PredMHC can obtain 91.69% accuracy. Experimental results on comparison with other features, classifiers, and existing methods showed the effectiveness of PredMHC in predicting the MHC.

Keywords: protein classification, major histocompatibility complex, machine learning, feature extraction, identification

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Chunyu Wang,
Harbin Institute of Technology, China
Haiying Zhang,
Xiamen University, China

****Correspondence:***
Yanjuan Li
lyjuan5@163.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *13 February 2022*
Accepted: *07 March 2022*
Published: *25 April 2022*

Citation:
Chen D and Li Y (2022) PredMHC: An Effective Predictor of Major Histocompatibility Complex Using Mixed Features.
Front. Genet. 13:875112.
doi: 10.3389/fgene.2022.875112

INTRODUCTION

As a large locus on vertebrate DNA, the major histocompatibility complex (MHC) contains a tightly linked set of polymorphic genes encoding cell surface proteins that are essential for immune surveillance. These cell surface proteins are called MHC molecules (Kubiniok et al., 2022). MHC molecules are classified into MHC class I, MHC class II, and MHC class III according to variation in molecular structure, function, and distribution (Marcoux et al., 2021). MHC class I molecules are expressed in all nucleated cells and platelets—essentially all cells except red blood cells, which display antigens to signal cytotoxic T lymphocytes, including clusters of differentiation (CD8$^+$) (McShan et al., 2021). MHC class II molecules are expressed in antigen-presenting cells, such as B cells, dendritic cells, and macrophages, where they normally bind to CD4$^+$ receptors on helper T cells to clear foreign antigens. MHC class III genes are interleaved with class I and class II genes on the short arm of chromosome 6, but their proteins play different physiological roles.

MHC molecules are cell surface glycoproteins with a three-dimensional structure and are of vital importance to infection, autoimmunity, transplantation, and tumor immunotherapy. MHC-binding prediction plays an important role in identifying potential novel therapeutic strategies. Mahoney et al. (2021) pointed out that MHC phosphopeptides can be considered potential immunotherapeutic targets for cancer and other chronic diseases. Therefore, many scholars carried out a lot of research work on MHC-binding prediction. The first computational method

(Altuvia et al., 1995) to uncover the MHC-binding peptide was developed by Altuvia et al., which is based on protein structure and is further improved to distinguish candidate peptides that bind to hydrophobic binding pockets of the MHC molecules (Altuvia et al., 1997). The SVRMHC (Liu et al., 2006) is an MHC-binding peptide model which encoded peptides with physicochemical properties and trained support vector machines to construct a prediction model on mice. NetMHC-3.0 (Lundegaard et al., 2008) is a web server with high performance for predicting peptide binders based on artificial neural networks. Boehm et al. proposed a method named ForestMHC (Boehm et al., 2019) to identify immunogenic peptides. ForestMHC encoded a peptide sequence with physicochemical properties and trained a random forest classifier to construct an identification model. Saxena et al. (2020) predicted the binding potential of peptides to the MHC, which is critical for designing peptide-based therapeutics, using a deep learning model named OnionMHC. In consideration of the importance of structural information, the OnionMHC represents peptides with its sequence and structure-based features for peptide-HLA-A*02:01 binding predictions. (Lv et al., 2020) Jiang et al. (2021) gave a comprehensive review of the state-of-the-art literature on MHC-binding peptide prediction and an in-depth evaluation of feature representation methods, prediction models, and model training strategies on benchmark datasets. Based on the limitation of only handling peptide sequences with fixed length, Jiang et al. proposed a novel variable-length MHC-binding prediction model named BVLSTM-MHC. Experimental results on an independent validation dataset showed that BVLSTM-MHC has better performance than the ten mainstream prediction tools.

Scientists are devoted to discover MHC molecules in various vertebrate genomes. Hopkins et al. (1986) described a rat monoclonal antibody which can recognize MHC class II antigens in sheep and seems to recognize determinants which are nonpolymorphic. Moreover, based on the antibody, the distribution of sheep class II molecules is investigated, and the class II- expression variations by cells in efferent lymph and peripheral is also investigated. Westbrook et al. (2015) combined the SMRT sequencing technology and CCS and introduced and validated the technology of SMRT-CCS on identifying class I transcripts in Mauritian-origin cynomolgus macaques. Furthermore, SMRT-CCS was applied to characterize 60 new full-length class I transcriptional sequences expressed in the Chinese cynomolgus monkey population. By using pyrosequencing with high-resolution and Sanger sequencing technology, Shiina et al. (2015) genotyped 127 unrelated animals and identified 112 different alleles. Moreover, the International Society for Animal Genetics (ISAG) standardized the nomenclature and established the IPD-MHC database which is used to scientifically manage the MHC allele sequences and genes from nonhuman organisms (Giuseppe et al., 2017; Maccari et al., 2018; Ali et al., 2021; Burton et al., 2021; Karcioglu and Bulut, 2021; Roy et al., 2021; Safaei et al., 2021; Wang et al., 2021).

At early stages, the research studies related to the MHC are developed based on mice experiments. With the availability of a large amount of data and development of machine learning, developing a machine learning–based model to research the MHC was feasible. Li et al. (2019) proposed an identification method of the MHC based on an extreme learning machine algorithm. Although high accuracy has been achieved, there are still many aspects worthy of further investigation (Lv et al., 2019; Lv et al., 2021a; Lv et al., 2021b). In this study, we aim to propose a new MHC predictor, PredMHC, to further improve prediction performance.

MATERIALS AND METHODS

Framework of PredMHC

In this study, we introduced a novel MHC predictor named PredMHC, the framework of which is shown in **Figure 1**. First, PredMHC encoded a protein sequence with mixed features including 188D, APAAC, KSCTriad, CKSAAGP, and PAAC. Second, three classifiers including SGD, SMO, and random forest were trained on the mixed features of protein sequence. Finally, the prediction result was obtained by the voting of the three classifiers. We will introduce the datasets, feature extraction, and classifiers in detail in the following section.

Dataset

The dataset constructed by Li et al. (2019) is used in this study. A web server called ELM-MHC was developed by Li et al., from which the dataset can be downloaded. The reason that we used the same dataset as ELM-MHC is as follows. First, the dataset is constructed by searching for MHC sequences on the Uniprot database, and it is reliable. Second, the dataset is used cd-hit to de-duplication processing. The protein sequences are clustered based on the parameter setting, and the sequence with the maximum length in every cluster is used as a representative sequence. The redundant and homology-biased sequences are removed in this dataset. Finally, the most important inference was that we can fairly compare with the existing method by using the same dataset. The final dataset contained 13,488 protein sequences, which consists of 6,712 MHC protein sequences (positive examples) and 6,776 nonMHC protein sequences (negative examples). All protein sequences were divided into two groups: 10,790 sequences as a set of 10-fold cross-validation and 2,698 sequences as a set of independent validation. The training dataset (Train-10790) comprised 5,370 MHC protein sequences and 5,420 nonMHC protein sequences, all randomly selected from the set of positive and negative examples, respectively. They were then further randomly divided into five sets for the input of 10-fold cross-validation. The independent testing dataset (Test-2698) contained 1,342 positive and 1,356 negative examples.

Feature Extraction

To classify a protein sequence into different categories using the machine learning method, the first step is to encode the protein sequence with features. A feature that can effectively discriminate positive examples from negative examples can greatly improve the prediction performance of the model. In this study, we try to encode protein sequences with mixed features including 188D,

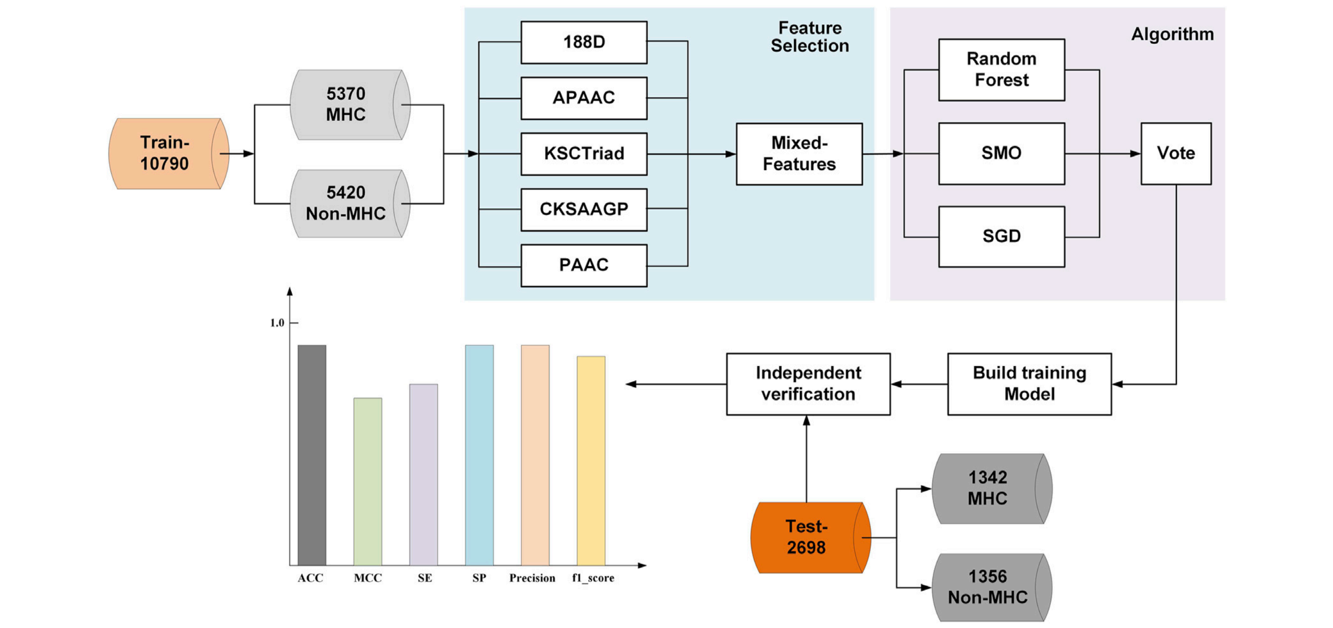

FIGURE 1 | Framework of PredMHC.

APAAC, KSCTriad, CKSAAGP, and PAAC. The mixed features can represent a protein sequence from different prospectives; thus, it can better distinguish different protein sequences.

SVMProt-188D

SVMProt-188D is a feature extraction method based on the amino acid composition and physicochemical properties (Dubchak et al., 1995; Saxena et al., 2021). It encodes each protein sequence as a 188-dimensional feature vector. The first 20 features are the frequencies of the 20 amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y in alphabetical order) occurring in the sequence. The formula is defined as

$$(V_1,\ V_2,\ \ldots,\ V_{20}) = \frac{N_i}{L},$$

where N_i denotes the number of the ith amino acid in the protein sequence and L denotes the length of a sequence. Obviously, $\sum V_i = 1$.

The latter dimensions are correlated with eight physicochemical properties, namely, hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure, and solvent accessibility. Each physicochemical property consists of 21 numbers. In detail, each property consists of three descriptors, composition (C), transition (T), and distribution (D). C indicates the proportion of amino acids with specific physicochemical properties to all amino acids, and the dimension of C is 3; T represents the percentage frequency of amino acids with a specific property behind amino acids with another property, and its dimension is 3; and D represents the proportions of the chain length of 0, 25, 50, 75, and 100% amino acids with a specific property, and its dimension is 8. Therefore, after analyzing the composition and eight physicochemical properties of amino acids, we can obtain a total of 20+(3 + 5+8)×8 = 188 features.

Amphiphilic Pseudo Amino Acid Composition

The concept of amphiphilic pseudo amino acid composition (APAAC), originally proposed by Chou (Chou, 2005; Lv et al., 2021a; Awais et al., 2021; Naseer et al., 2021; Yan et al., 2021), is an effective protein descriptor and has been applied for diverse protein sequence analysis. APAAC is different from traditional AAC. It can incorporate a partial sequence-order effect by using the hydrophobicity and hydrophilicity of the constituent amino acids in a protein. For the convenience of the readers, we will briefly introduce the concept of APAAC. Let $R_1R_2R_3...R_L$ be a protein sequence with length L, where R_1 denotes the residue at position 1, R_2 denotes the residue at positon 2, and so forth. According to the definition of APAAC, a protein can be denoted as a vector P with dimension (20+2λ). Vector P is defined as follows.

$$P = [P_1, \ldots, P_{20}, P_{20+1}, \ldots, P_{20+\lambda}, \ldots, P_{20+2\lambda}], \quad (1)$$

where $P_1, P_2, \ldots, P_{20}$ in **Eq. 1** represent the classic AAC and the next 2λ discrete numbers describe the sequence correlation factor.

K-Spaced Conjoint Triad

The k-spaced conjoint triad (KSCTriad) (Chao et al., 2018; Zhen et al., 2020) is an effective protein descriptor and has been comprehensively applied for diverse biological sequence analyses. Different from the conjoint triad descriptor, KSCTriad not only calculates the number of three continuous amino acid units but also incorporates the continuous amino acid units that are separated by any k-residues.

Composition of K-Spaced Amino Acid Group Pairs

The composition of k-spaced amino acid pairs (CKSAAP) (Chen et al., 2010; Ahmad et al., 2021; Akbar et al., 2021; Al-Qazzaz et al., 2021; Alar and Fernandez, 2021; Alim et al., 2021; Buriro et al., 2021) method describes the order-related information of the protein sequence, which takes the occurrence frequency of two amino acids separated by k-residues in the sequence as a feature element. The protein contains 20 amino acids; thus, a 400-dimensional feature vector can be obtained for each interval. The composition of k-spaced amino acid group pairs (CKSAAGP) is a variation of the CKSAAP method. The 20 amino acids can be classified into five groups based on the chemical properties of their side chains: the aliphatic group, aromatic group, positive charged group, negative charged group, and uncharged group. The CKSAAGP method is based on the frequency of the two groups separated by a k-spaced amino acid.

Pseudo-Amino Acid Composition

The conventional amino acid composition is defined in a 20-D space, and each dimension represents the frequency of the occurrence of one of the 20 native amino acids. Different from the conventional amino acid protein composition, the pseudo-amino acid composition (Chou, 2001; Awais et al., 2021), which is a vector with 20+λ discrete components, will contain much more sequence-order and sequence-length information. According to the concept of pseudo-amino acid composition, the feature is given by

$$P = \begin{bmatrix} p_1 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \end{bmatrix},$$

where the first 20 components are the occurrence frequencies of the 20 amino acids in the protein which is the same as in the conventional amino acid composition, while the additional components $p_{20+1} \cdots p_{20+\lambda}$ are the sequence-order correlation factors of the different ranks.

Classifier

To obtain better classification results, we adopted the voting of three base classifiers as the final classification result. The three classifiers were, respectively, random forest, SMO, and SGD. The three classifiers are popular and have been successfully used in bioinformatics many times.

Random forest is an ensemble classifier based on the decision tree algorithm proposed by Breiman in 2001 (Breiman, 2001). To solve regression or classification tasks, random forests construct many decision trees by extracting subsets from all the samples through the bootstrap technique and obtain the prediction result by voting on these decision trees. Random forests are widely used in bioinformatics because of their low computational overhead and ability of handling unbalanced data.

The support vector machine (SVM) (Hearst et al., 1998) is a well-known machine learning algorithm that completes various classification tasks by constructing a separating hyperplane in the high-dimensional space. However, the training speed of support vector machines is heavily influenced by data size. To solve this problem, the sequential minimum optimization (SMO) (Platt, 1999) algorithm was proposed, which decomposes large quadratic programming problems (OPs) of an original SVM into a series of the smallest possible QP problems. Moreover, the solution process of SMO needs no additional matrix storage, thus saving both time and space costs.

The goal of the stochastic gradient descent (SGD) algorithm is to find a path that leads to optimal result. When using this algorithm, the parameter values are first initialized, and then these values are continuously changed until the target function converges. The SGD algorithm is widely used to process large-scale sparse data, such as text classification tasks.

Measurement

To evaluate the performance of the proposed method, we introduced four indicators commonly used in bioinformatics: sensitivity (SE), specificity (SP), accuracy (ACC), and Matthew's correlation coefficient (MCC). The formulae of these indicators are as follows (Zhang et al., 2021a; Lv et al., 2021b; Zhang et al., 2021b; Zhang et al., 2021c; Zhang et al., 2021d; Zhang et al., 2021e; Zhao et al., 2021; Zhu et al., 2021; Zou et al., 2021; Zhao et al., 2022).

$$SE = \frac{TP}{TP + FN},$$
$$SP = \frac{TN}{TN + FP},$$
$$ACC = \frac{TN + TP}{TN + FP + TP + FN},$$
$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}},$$

where TP is an abbreviation for true positives, representing the number of MHC proteins predicted in positive examples; FP is an abbreviation for false positives, representing the number of MHC proteins predicted in negative examples; TN is an abbreviation for true negatives, representing nonMHC proteins predicted in negative examples; and FN is an abbreviation for false negatives and indicates the number of predicted nonMHC proteins in positive examples. SE and SP represent the predictive accuracy of the model in positive and negative samples, respectively. Both ACC and MCC represent the overall performance of the model. For all the aforementioned metrics , the higher the score they get the better the performance of the model.

RESULT AND DISCUSSION

Cross-Validation Results of Train-10790

In many experiments, we tried a variety of methods to extract highly recognizable features from protein sequences in the training set and used several algorithms to train the model to

TABLE1 | Result of different features on Train-10790.

Feaures	ACC	MCC	SE	SP
(1)-188D	0.8953	0.7927	0.8596	0.9310
(2)-APAAC	0.8329	0.6824	0.9494	0.7108
(3)-KSCTriad	0.8764	0.7580	0.8177	0.9350
(4)-CKSAAGP	0.8682	0.7469	0.7826	0.9529
(5)-PAAC	0.8283	0.6739	0.9485	0.7018
188D + APAAC	0.9003	0.8019	0.8735	0.9276
APAAC + KSCTriad	0.8872	0.7782	0.8386	0.9360
KSCTriad + CKSAAGP	0.8993	0.8039	0.8404	0.9576
CKSAAGP + PAAC	0.8848	0.7728	0.8376	0.9316
188D + APAAC + KSCTriad	0.9121	0.8268	0.8734	0.9511
APAAC + KSCTriad + CKSAAGP	0.9054	0.8155	0.8518	0.9589
KSCTriad + CKSAAGP + PAAC	0.9041	0.8127	0.8516	0.9565
188D + APAAC + KSCTriad + CKSAAGP	0.9157	0.8351	0.8701	0.9618
APAAC + KSCTriad + CKSAAGP + PAAC	0.9065	0.8178	0.8522	0.9608
Our mixed feature	0.9169	0.8370	0.8761	0.9587

TABLE 2 | Result of different classifiers on Train-10790.

Classifiers	ACC	MCC	SE	SP
SGD	0.8794	0.7600	0.8504	0.9081
SMO	0.9038	0.8106	0.8594	0.9478
Random forest	0.8850	0.7699	0.8830	0.8869
Our classification model	0.9169	0.8370	0.8761	0.9587

achieve optimal accuracy. The experimental comparison results of different features are explained in *Performance of Different Features on Cross-Validation*, and the experimental comparison results of different classifiers are explained in *Performance of Different Classifiers on Cross-Validation*.

Performance of Different Features on Cross-Validation

Using the voting of random forest, SMO, and SGD as the classification model, we first tried 188D, APAAC, KSCTriad, CKSAAGP, PAAC, and their combinations. **Table 1** shows the performance of the five single features and several combinations of features with good performance in the 10-fold cross-validation. As shown in **Table 1**, according to the indexes MCC and ACC, the mixed features proposed in this study have the highest score; thus, our method has better overall performance. According to the indicator of SE, the feature of APAAC has the highest score, whereas its value of ACC, MCC, and SP is lower; it verifies that the feature of APAAC was bias to classify a protein into the MHC protein. Similar to APAAC, PAAC also has higher value on the indicator SE and lower value on other indicators. Therefore, from the overall perspective, our method obviously performs better than all other methods.

Performance of Different Classifiers on Cross-Validation

To verify the performance of our used classifier, we compared the classifier used in this study with other classifiers. **Table 2** shows the experimental results. As shown in **Table 2**, the voting of SGD, SMO, and random forest used in our identification system has better performance than other single classifiers. As shown in **Table 2**, our classification model has 0.9169% accuracy and 0.8370 MCC, which are higher than those of other classifiers. It verified that our classification model has better overall performance. According to the number of winning incidences, our classification wins on three indicators and has the highest number of wins. It is shown in **Table 2** that the SE of our classification model was slightly lower than that of random forest. However, the values of ACC, MCC, and SP of our classification model are obviously higher than those of random forest. Therefore, from the overall perspective, our classification model obviously performs better than all other classifiers.

Independent-Validation Results of Test-2698

To evaluate the generalization performance of the proposed model, we tested its performance on the Test-2698 dataset. In detail, we trained the model proposed in this study on the Train-10790 dataset and then computed its performance on the test-2698 dataset. The experimental results are shown in **Tables 3**, **4**. As shown in **Tables 3**, **4**, the feature extraction method and classifier used in this study have better performance than the other feature extraction methods and classifiers, respectively.

Comparison With Other Predictors

To evaluate the performance of the classifier PredMHC, we compared it with ELM-MHC on the same dataset including Train-10790 and Test-2698. The comparison results on the 10-fold cross-validation are shown in **Table 5**. As we can see from **Table 5**, PredMHC has higher score than ELM-MHC on the indicators ACC, MCC, and SP. According to the number of winning incidence, PredMHC has better performance than ELM-MHC. According to ACC and MCC, PredMHC has better overall performance than ELM-MHC. Therefore, PredMHC is superior to the existing methods in the prediction of MHC protein.

TABLE 3 | Result of different features on Test-2698.

Features	ACC	MCC	SE	SP
188D	0.8926	0.7869	0.8593	0.9259
APAAC	0.8357	0.6892	0.9533	0.7139
KSCTriad	0.8741	0.7504	0.8355	0.9127
CKSAAGP	0.8774	0.7614	0.8098	0.9442
PAAC	0.8326	0.6826	0.9527	0.7056
188D + APAAC	0.9010	0.8061	0.8482	0.9530
APAAC + KSCTriad	0.8940	0.7888	0.8697	0.9182
KSCTriad + CKSAAGP	0.9055	0.8155	0.8540	0.9573
CKSAAGP + PAAC	0.8901	0.7818	0.8571	0.9230
188D + APAAC + KSCTriad	0.9172	0.8355	0.8938	0.9412
APAAC + KSCTriad + CKSAAGP	0.9130	0.8287	0.8729	0.9532
KSCTriad + CKSAAGP + PAAC	0.9155	0.8337	0.8769	0.9544
188D + APAAC + KSCTriad + CKSAAGP	0.9198	0.8416	0.8841	0.9550
APAAC + KSCTriad + CKSAAGP + PAAC	0.9134	0.8300	0.8693	0.9574
Our mixed feature	0.9246	0.8502	0.9034	0.9466

TABLE 4 | Result of different classifiers on Test-2698.

Classifier	ACC	MCC	SE	SP
SGD	0.8959	0.7918	0.8935	0.8982
SMO	0.9063	0.8147	0.8682	0.9440
Random forest	0.8948	0.7896	0.8913	0.8982
Our classification model	0.9246	0.8502	0.9034	0.9466

TABLE 5 | Comparison of 10-fold cross-validation with the existing method on all data.

Method	ACC	MCC	SE	SP
ELM-MHC	0.9166	0.822	0.893	0.908
Our method	0.9185	0.8403	0.8741	0.9627

CONCLUSION

In this study, we proposed an efficient, reliable, and simple experimental model for predicting the MHC protein based on mixed features. After a large number of comparative experiments, we selected the mixed features of 188D, APAAC, KSCTriad, CKSAAGP, and PAAC, which showed global performance on the 10-fold cross-validation training dataset and independent test dataset. We then used the voting of SGD, SMO, and random forest to build a prediction model which also achieved the best performance on both training and test datasets. In terms of important indicators, our model obtained an MCC of 0.8370 and ACC of 0.9169 in the 10-fold cross-validation based on the Train-10790 dataset and MCC of 0.8502 and ACC of 0.9246 in the independent validation based on the Test-2698 dataset. In conclusion, we believe that our novel model provides an efficient and reliable method to screen MHCs from a large number of protein sequences. In the future, we will pay more attention to deep learning classifiers and evolution strategies (Tahoces et al., 2021; Tandel et al., 2021; Tavolara et al., 2021; Togacar, 2021; Tsiknakis et al., 2021; Turki and Taguchi, 2021; Usman et al., 2021; Vafaeezadeh et al., 2021; Wang et al., 2021; Watanabe et al., 2021; Yap et al., 2021; Yildirim et al., 2021).

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

Conceptualization, YL; data curation, DC; formal analysis, DC; project administration, DC; writing—original draft, YL; and writing—review and editing, DC.

FUNDING

This work was supported by the Research Start-up Funding Project of Quzhou University (BSYJ202112 and BSYJ202109), the National Natural Science Foundation of China (61901103 and 61671189), and the Natural Science Foundation of Heilongjiang Province (LH 2019F002).

REFERENCES

Ahmad, F., Farooq, A., and Khan, M. U. G. (2021). Deep Learning Model for Pathogen Classification Using Feature Fusion and Data Augmentation. *Cbio* 16 (3), 466–483. doi:10.2174/1574893615999200707143535

Akbar, S., Ahmad, A., Hayat, M., Rehman, A. U., Khan, S., Ali, F., et al. (2021). iAtbP-Hyb-EnC: Prediction of Antitubercular Peptides via Heterogeneous Feature Representation and Genetic Algorithm Based Ensemble Learning Model. *Comput. Biol. Med.* 137, 104778. doi:10.1016/j.compbiomed.2021.104778

Al-Qazzaz, N. K., Alyasseri, Z. A. A., Abdulkareem, K. H., Ali, N. S., Al-Mhiqani, M. N., and Guger, C. (2021). EEG Feature Fusion for Motor Imagery: A New

Robust Framework towards Stroke Patients Rehabilitation. *Comput. Biol. Med.* 137, 104799. doi:10.1016/j.compbiomed.2021.104799

Alar, H. S., and Fernandez, P. L. (2021). Accurate and Efficient Mosquito Genus Classification Algorithm Using Candidate-Elimination and Nearest Centroid on Extracted Features of Wingbeat Acoustic Properties. *Comput. Biol. Med.* 139, 104973. doi:10.1016/j.compbiomed.2021.104973

Ali, F., Akbar, S., Ghulam, A., Maher, Z. A., Unar, A., Talpur, D. B., et al. (2021). AFP-CMBPred: Computational Identification of Antifreeze Proteins by Extending Consensus Sequences into Multi-Blocks Evolutionary Information. *Comput. Biol. Med.* 139, 105006. doi:10.1016/j.compbiomed.2021.105006

Alim, A., Rafay, A., and Naseem, I. (2021). PoGB-pred: Prediction of Antifreeze Proteins Sequences Using Amino Acid Composition with Feature Selection Followed by a Sequential-Based Ensemble Approach. *Cbio* 16 (3), 446–456. doi:10.2174/1574893615999200707141926

Altuvia, Y., Schueler, O., and Margalit, H. (1995). Ranking Potential Binding Peptides to MHC Molecules by a Computational Threading Approach. *J. Mol. Biol.* 249 (2), 244–250. doi:10.1006/jmbi.1995.0293

Altuvia, Y., Sette, A., Sidney, J., Southwood, S., and Margalit, H. (1997). A Structure-Based Algorithm to Predict Potential Binding Peptides to MHC Molecules with Hydrophobic Binding Pockets. *Hum. Immunol.* 58 (1), 1–11. doi:10.1016/s0198-8859(97)00210-3

Awais, M., Hussain, W., Rasool, N., and Khan, Y. D. (2021). iTSP-PseAAC: Identifying Tumor Suppressor Proteins by Using Fully Connected Neural Network and PseAAC. *Cbio* 16 (5), 700–709. doi:10.2174/1574893615666210108094431

Boehm, K. M., Bhinder, B., Raja, V. J., Dephoure, N., and Elemento, O. (2019). Predicting Peptide Presentation by Major Histocompatibility Complex Class I: an Improved Machine Learning Approach to the Immunopeptidome. *BMC Bioinformatics* 20 (1), 7. doi:10.1186/s12859-018-2561-z

Breiman, L. (2001). Random Forests. *Mach Learn.* 45 (1), 5–32. doi:10.1023/a:1010933404324

Buriro, A. B., Ahmed, B., Baloch, G., Ahmed, J., Shoorangiz, R., Weddell, S. J., et al. (2021). Classification of Alcoholic EEG Signals Using Wavelet Scattering Transform-Based Features. *Comput. Biol. Med.* 139, 104969. doi:10.1016/j.compbiomed.2021.104969

Burton, W. S., Myers, C. A., Jensen, A., Hamilton, L., Shelburne, K. B., Banks, S. A., et al. (2021). Automatic Tracking of Healthy Joint Kinematics from Stereo-Radiography Sequences. *Comput. Biol. Med.* 139, 104945. doi:10.1016/j.compbiomed.2021.104945

Chao, Z., Wang, C., Liu, H., Zhou, Q., Qian, L., Guo, Y., et al. (2018). Identification and Analysis of Adenine N6-Methylation Sites in the rice Genome. *Nat. Plants* 4 (8), 554–563. doi:10.1038/s41477-018-0214-x

Chen, K., Jiang, Y., Du, L., and Kurgan, L. (2010). Prediction of Integral Membrane Protein Type by Collocated Hydrophobic Amino Acid Pairs. *J. Comput. Chem.* 30 (1), 163–172. doi:10.1002/jcc.21053

Chou, K-C. (2001). Prediction of Protein Cellular Attributes Using Pseudo-amino Acid Composition. *Proteins Struct. Funct. Bioinformatics* 43 (3), 246–255. doi:10.1002/prot.1035

Chou, K.-C. (2005). Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics* 21 (1), 10–19. doi:10.1093/bioinformatics/bth466

Dubchak, I., Muchnik, I., Holbrook, S. R., and Kim, S. H. (1995). Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence. *Proc. Natl. Acad. Sci.* 92 (19), 8700–8704. doi:10.1073/pnas.92.19.8700

Giuseppe, M., James, R., Keith, B., Guethlein, L. A., Unni, G., Jim, K., et al. (2017). IPD-MHC 2.0: an Improved Inter-species Database for the Study of the Major Histocompatibility Complex. *Nucleic Acids Res.* 45 (D1), D860. doi:10.1093/nar/gkw1050

Hearst, M. A., Dumais, S. T., and Osuna, E. (1998). Support Vector Machines: Training and Applications. *IEEE Intel. Syst. App.* 13 (4), 18–28.

Hopkins, J., Dutia, B. M., and Mcconnell, I. (1986). Monoclonal Antibodies to Sheep Lymphocytes. I. Identification of MHC Class II Molecules on Lymphoid Tissue and Changes in the Level of Class II Expression on Lymph-Borne Cells Following Antigen Stimulation *In Vivo*. *Immunology* 59 (3), 433

Jiang, L., Yu, H., Li, J., Tang, J., Guo, Y., and Guo, F. (2021). Predicting MHC Class I Binder: Existing Approaches and a Novel Recurrent Neural Network Solution. *Brief. Bioinform.* 22 (6), bbab216. doi:10.1093/bib/bbab216

Karcioglu, A. A., and Bulut, H. (2021). The WM-Q Multiple Exact String Matching Algorithm for DNA Sequences. *Comput. Biol. Med.* 136, 104656. doi:10.1016/j.compbiomed.2021.104656

Kubiniok, P., Marcu, A., Bichmann, L., Kuchenbecker, L., Schuster, H., Hamelin, D. J., et al. (2022). Understanding the Constitutive Presentation of MHC Class I Immunopeptidomes in Primary Tissues. *Iscience* 25 (2), 103768. doi:10.1016/j.isci.2022.103768

Li, Y., Niu, M., and Zou, Q. (2019). An Improved MHC Identification Method with Extreme Learning Machine Algorithm. *J. proteome Res.* 18 (3), 1392–1401. doi:10.1021/acs.jproteome.9b00012

Liu, W., Meng, X., Xu, Q., Flower, D. R., and Li, T. (2006). Quantitative Prediction of Mouse Class I MHC Peptide Binding Affinity Using Support Vector Machine Regression (SVR) Models. *BMC Bioinformatics* 7 (1), 182. doi:10.1186/1471-2105-7-182

Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O., and Nielsen, M. (2008). NetMHC-3.0: Accurate Web Accessible Predictions of Human, Mouse and Monkey MHC Class I Affinities for Peptides of Length 8-11. *Nucleic Acids Res.* 36, W509–W512. doi:10.1093/nar/gkn202

Lv, Z., Ao, C., and Zou, Q. (2019). Protein Function Prediction: From Traditional Classifier to Deep Learning. *Proteomics* 19 (14), e1900119. doi:10.1002/pmic.201900119

Lv, Z., Cui, F., Zou, Q., Zhang, L., and Xu, L. (2021). Anticancer Peptides Prediction with Deep Representation Learning Features. *Brief Bioinform* 22 (5), bbab008. doi:10.1093/bib/bbab008

Lv, Z., Ding, H., Wang, L., and Zou, Q. (2021). A Convolutional Neural Network Using Dinucleotide One-Hot Encoder for Identifying DNA N6-Methyladenine Sites in the Rice Genome. *Neurocomputing* 422, 214–221. doi:10.1016/j.neucom.2020.09.056

Lv, Z., Wang, P., Zou, Q., and Jiang, Q. (2020). Identification of Sub-golgi Protein Localization by Use of Deep Representation Learning Features. *Bioinformatics* 36 (24), 5600–5609. doi:10.1093/bioinformatics/btaa1074

Maccari, G., Robinson, J., Bontrop, R. E., Otting, N., de Groot, N. G., Ho, C. S., et al. (2018). IPD-MHC: Nomenclature Requirements for the Non-human Major Histocompatibility Complex in the Next-Generation Sequencing Era. *Immunogenetics* 70 (10), 619–623. doi:10.1007/s00251-018-1072-4

Mahoney, K. E., Shabanowitz, J., and Hunt, D. F. (2021). MHC Phosphopeptides: Promising Targets for Immunotherapy of Cancer and Other Chronic Diseases. *Mol. Cell Proteomics* 20 (640), 100112. doi:10.1016/j.mcpro.2021.100112

Marcoux, G., Laroche, A., Hasse, S., Bellio, M., Mbarik, M., Tamagne, M., et al. (2021). Platelet EVs Contain an Active Proteasome Involved in Protein Processing for Antigen Presentation via MHC-I Molecules. *Blood J. Am. Soc. Hematol.* 138 (25), 2607–2620. doi:10.1182/blood.2020009957

McShan, A. C., Devlin, C. A., Morozov, G. I., Overall, S. A., Moschidi, D., Akella, N., et al. (2021). TAPBPR Promotes Antigen Loading on MHC-I Molecules Using a Peptide Trap. *Nat. Commun.* 12 (1), 3174–3218. doi:10.1038/s41467-021-23225-6

Naseer, S., Hussain, W., Khan, Y. D., and Rasool, N. (2021). NPalmitoylDeep-Pseaac: A Predictor of N-Palmitoylation Sites in Proteins Using Deep Representations of Proteins and PseAAC via Modified 5-Steps Rule. *Cbio* 16 (2), 294–305. doi:10.2174/1574893615999200605142828

Platt, J. C. (1999).*Fast Training of Support Vector Machines Using Sequential Minimal Optimization, Advances in Kernel Methods*. Support Vector Learning

Roy, S., Sharma, B., Mazid, M. I., Akhand, R. N., Das, M., Marufatuzzahan, M., et al. (2021). Identification and Host Response Interaction Study of SARS-CoV-2 Encoded miRNA-like Sequences: an In Silico Approach. *Comput. Biol. Med.* 134, 104451. doi:10.1016/j.compbiomed.2021.104451

Safaei, M., Sundararajan, E. A., Driss, M., Boulila, W., and Shapi'i, A. (2021). A Systematic Literature Review on Obesity: Understanding the Causes & Consequences of Obesity and Reviewing Various Machine Learning Approaches Used to Predict Obesity. *Comput. Biol. Med.* 136, 104754. doi:10.1016/j.compbiomed.2021.104754

Saxena, D., Sharma, A., Siddiqui, M. H., and Kumar, R. (2021). Development of Machine Learning Based Blood-Brain Barrier Permeability Prediction Models Using Physicochemical Properties, MACCS and Substructure Fingerprints. *Cbio* 16 (6), 855–864. doi:10.2174/1574893616666210203104013

Saxena, S., Animesh, S., Fullwood, M., and Mu, Y. (2020). OnionMHC: A Deep Learning Model for Peptide - HLA-A*02:01 Binding Predictions Using Both Structure and Sequence Feature Sets *J. Micromech. Mol. Phys.* 5 (03), 2050009.

Shiina, T., Yamada, Y., Aarnink, A., Suzuki, S., Masuya, A., Ito, S., et al. (2015). Discovery of Novel MHC-Class I Alleles and Haplotypes in Filipino Cynomolgus Macaques (Macaca fascicularis) by Pyrosequencing and Sanger Sequencing. *Immunogenetics* 67 (10), 563–578. doi:10.1007/s00251-015-0867-9

Tahoces, P. G., Varela, R., and Carreira, J. M. (2021). Deep Learning Method for Aortic Root Detection. *Comput. Biol. Med.* 135, 104533. doi:10.1016/j.compbiomed.2021.104533

Tandel, G. S., Tiwari, A., and Kakde, O. G. (2021). Performance Optimisation of Deep Learning Models Using Majority Voting Algorithm for Brain Tumour Classification. *Comput. Biol. Med.* 135, 104564. doi:10.1016/j.compbiomed.2021.104564

Tavolara, T. E., Gurcan, M. N., Segal, S., and Niazi, M. K. K. (2021). Identification of Difficult to Intubate Patients from Frontal Face Images Using an Ensemble of Deep Learning Models. *Comput. Biol. Med.* 136, 104737. doi:10.1016/j.compbiomed.2021.104737

Togacar, M. (2021). Detection of Segmented Uterine Cancer Images by Hotspot Detection Method Using Deep Learning Models, Pigeon-Inspired Optimization, Types-Based Dominant Activation Selection Approaches. *Comput. Biol. Med.* 136, 104659. doi:10.1016/j.compbiomed.2021.104659

Tsiknakis, N., Theodoropoulos, D., Manikis, G., Ktistakis, E., Boutsora, O., Berto, A., et al. (2021). Deep Learning for Diabetic Retinopathy Detection and Classification Based on Fundus Images: A Review. *Comput. Biol. Med.* 135, 104599. doi:10.1016/j.compbiomed.2021.104599

Turki, T., and Taguchi, Y. h. (2021). Discriminating the Single-Cell Gene Regulatory Networks of Human Pancreatic Islets: A Novel Deep Learning Application. *Comput. Biol. Med.* 132, 132. doi:10.1016/j.compbiomed.2021.104257

Usman, S. M., Khalid, S., and Bashir, S. (2021). A Deep Learning Based Ensemble Learning Method for Epileptic Seizure Prediction. *Comput. Biol. Med.* 136. doi:10.1016/j.compbiomed.2021.104710

Vafaeezadeh, M., Behnam, H., Hosseinsabet, A., and Gifani, P. (2021). A Deep Learning Approach for the Automatic Recognition of Prosthetic Mitral Valve in Echocardiographic Images. *Comput. Biol. Med.* 133, 104388. doi:10.1016/j.compbiomed.2021.104388

Wang, X., Wang, S., Fu, H., Ruan, X., Tang, X., and DeepFusion-Rbp (2021). DeepFusion-RBP: Using Deep Learning to Fuse Multiple Features to Identify RNA-Binding Protein Sequences. *Cbio* 16 (8), 1089–1100. doi:10.2174/1574893616666210618145121

Watanabe, S., Sakaguchi, K., Murata, D., and Ishii, K. (2021). Deep Learning-Based Hounsfield Unit Value Measurement Method for Bolus Tracking Images in Cerebral Computed Tomography Angiography. *Comput. Biol. Med.* 137, 104824. doi:10.1016/j.compbiomed.2021.104824

Westbrook, C. J., Karl, J. A., Wiseman, R. W., Mate, S., Koroleva, G., Garcia, K., et al. (2015). No Assembly Required: Full-Length MHC Class I Allele Discovery by PacBio Circular Consensus Sequencing. *Hum. Immunol.* 76 (12), 891–896. doi:10.1016/j.humimm.2015.03.022

Yan, N., Lv, Z., Hong, W., and Xu, X. (2021). Editorial: Feature Representation and Learning Methods with Applications in Protein Secondary Structure. *Front. Bioeng. Biotechnol.* 20219 (822). doi:10.3389/fbioe.2021.748722

Yap, M. H., Hachiuma, R., Alavi, A., Brüngel, R., Cassidy, B., Goyal, M., et al. (2021). Deep Learning in Diabetic Foot Ulcers Detection: A Comprehensive Evaluation. *Comput. Biol. Med.* 135, 104596. doi:10.1016/j.compbiomed.2021.104596

Yildirim, K., Bozdag, P. G., Talo, M., Yildirim, O., Karabatak, M., and Acharya, U. R. (2021). Deep Learning Model for Automated Kidney Stone Detection Using Coronal CT Images. *Comput. Biol. Med.* 135, 104569. doi:10.1016/j.compbiomed.2021.104569

Zhang, J., Sun, Q., and Liang, C. (2021). Prediction of lncRNA-Disease Associations Based on Robust Multi-Label Learning. *Cbio* 16 (9), 1179–1189. doi:10.2174/1574893616666210712091221

Zhang, Q., Zhou, J., and Zhang, B. (2021). Computational Traditional Chinese Medicine Diagnosis: A Literature Survey. *Comput. Biol. Med.* 133, 104358. doi:10.1016/j.compbiomed.2021.104358

Zhang, S., Yuan, Z., Wang, Y., Bai, Y., Chen, B., and Wang, H. (2021). REUR: A Unified Deep Framework for Signet Ring Cell Detection in Low-Resolution Pathological Images. *Comput. Biol. Med.* 136, 104711. doi:10.1016/j.compbiomed.2021.104711

Zhang, Y., Duan, G., Yan, C., Yi, H., Wu, F.-X., and Wang, J. (2021). MDAPlatform: A Component-Based Platform for Constructing and Assessing miRNA-Disease Association Prediction Methods. *Cbio* 16 (5), 710–721. doi:10.2174/1574893616999210120181506

Zhang, Z., Yu, S., Qin, W., Liang, X., Xie, Y., and Cao, G. (2021). Self-supervised CT Super-resolution with Hybrid Model. *Comput. Biol. Med.* 138, 104775. doi:10.1016/j.compbiomed.2021.104775

Zhao, S., Ju, Y., Ye, X., Zhang, J., and Han, S. (2021). Bioluminescent Proteins Prediction with Voting Strategy. *Cbio* 16 (2), 240–251. doi:10.2174/1574893615999200601122328

Zhao, X., Du, Y., and Zhang, R. (2022). A CNN-Based Multi-Target Fast Classification Method for AR-SSVEP. *Comput. Biol. Med.* 141, 105042. doi:10.1016/j.compbiomed.2021.105042

Zhen, C., Pei, Z., Fuyi, L., Marquez-Lago, T. T., André, L., Jerico, R., et al. (2020). 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 (3), 1047–1057. doi:10.1093/bib/bbz041

Zhu, Q., Fan, Y., and Pan, X. (2021). Fusing Multiple Biological Networks to Effectively Predict miRNA-Disease Associations. *Cbio* 16 (3), 371–384. doi:10.2174/1574893615999200715165335

Zou, Y., Wu, H., Guo, X., Peng, L., Ding, Y., Tang, J., et al. (2021). MK-FSVM-SVDD: A Multiple Kernel-Based Fuzzy SVM Model for Predicting DNA-Binding Proteins via Support Vector Data Description. *Cbio* 16 (2), 274–283. doi:10.2174/1574893615999200607173829

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.