MACHINE LEARNING TECHNIQUES ON GENE FUNCTION PREDICTION VOLUME II

EDITED BY: Quan Zou, Arun Kumar Sangaiah and Dariusz Mrozek

PUBLISHED IN: Frontiers in Genetics



frontiers Research Topics

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88976-632-1
DOI 10.3389/978-2-88976-632-1

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

MACHINE LEARNING TECHNIQUES ON GENE FUNCTION PREDICTION VOLUME II

Topic Editors:
Quan Zou, University of Electronic Science and Technology of China, China
Arun Kumar Sangaiah, VIT University, India
Dariusz Mrozek, Silesian University of Technology, Poland

Citation: Zou, Q., Sangaiah, A. K., Mrozek, D., eds. (2023). Machine Learning Techniques on Gene Function Prediction Volume II. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-632-1

Table of Contents

05 ***Editorial: Machine Learning Techniques on Gene Function Prediction Volume II***
Ren Qi, Arun Kumar Sangaiah, Dariusz Mrozek and Quan Zou

07 ***Application of Machine Learning for Drug–Target Interaction Prediction***
Lei Xu, Xiaoqing Ru and Rong Song

15 ***ACP-DA: Improving the Prediction of Anticancer Peptides Using Data Augmentation***
Xian-gan Chen, Wen Zhang, Xiaofei Yang, Chenhong Li and Hengling Chen

24 ***A Large-Scale and Serverless Computational Approach for Improving Quality of NGS Data Supporting Big Multi-Omics Data Analyses***
Dariusz Mrozek, Krzysztof Stępień, Piotr Grzesik and Bożena Małysiak-Mrozek

40 ***Gene Targets Network Analysis for the Revealing and Guidance of Molecular Driving Mechanism of Lung Cancer***
Risheng Huang, Xiao Xiang, Kangliang Zhang, Yuanliang Zheng, Chichao Wang and Guanqiong Hu

49 ***Predicting Human Protein Subcellular Locations by Using a Combination of Network and Function Features***
Lei Chen, ZhanDong Li, Tao Zeng, Yu-Hang Zhang, ShiQi Zhang, Tao Huang and Yu-Dong Cai

62 ***Research on Mental Stress Recognition of Depressive Disorders in Patients With Androgenic Alopecia Based on Machine Learning and Fuzzy K-Means Clustering***
Yulong Li, Baojin Wu, Xiujun Li, Qin Zhou, Xin Yang and Yufei Li

75 ***Improving the Classification of Alzheimer's Disease Using Hybrid Gene Selection Pipeline and Deep Learning***
Nivedhitha Mahendran, P. M. Durai Raj Vincent, Kathiravan Srinivasan and Chuan-Yu Chang

88 ***Prediction of Protein–Protein Interaction Sites Based on Stratified Attentional Mechanisms***
Minli Tang, Longxin Wu, Xinyu Yu, Zhaoqi Chu, Shuting Jin and Juan Liu

96 ***Prediction of Hormone-Binding Proteins Based on K-mer Feature Representation and Naive Bayes***
Yuxin Guo, Liping Hou, Wen Zhu and Peng Wang

106 ***KK-DBP: A Multi-Feature Fusion Method for DNA-Binding Protein Identification Based on Random Forest***
Yuran Jia, Shan Huang and Tianjiao Zhang

115 ***Bioinformatics Analysis of Long Non-coding RNA and Related Diseases: An Overview***
Yuxin Gong, Wen Zhu, Meili Sun and Lei Shi

124 ***Testing Gene-Gene Interactions Based on a Neighborhood Perspective in Genome-wide Association Studies***
Yingjie Guo, Honghong Cheng, Zhian Yuan, Zhen Liang, Yang Wang and Debing Du

134 ***Application of Sparse Representation in Bioinformatics***
Shuguang Han, Ning Wang, Yuxin Guo, Furong Tang, Lei Xu, Ying Ju and Lei Shi

146 ***Using Machine Learning Approaches to Predict Target Gene Expression in Rice T-DNA Insertional Mutants***
Ching-Hsuan Chien, Lan-Ying Huang, Shuen-Fang Lo, Liang-Jwu Chen, Chi-Chou Liao, Jia-Jyun Chen and Yen-Wei Chu

158 ***SNAREs-SAP: SNARE Proteins Identification With PSSM Profiles***
Zixiao Zhang, Yue Gong, Bo Gao, Hongfei Li, Wentao Gao, Yuming Zhao and Benzhi Dong

167 ***VTP-Identifier: Vesicular Transport Proteins Identification Based on PSSM Profiles and XGBoost***
Yue Gong, Benzhi Dong, Zixiao Zhang, Yixiao Zhai, Bo Gao, Tianjiao Zhang and Jingyu Zhang

177 ***Screening and Identification of Survival-Associated Splicing Factors in Lung Squamous Cell Carcinoma***
Min Chen, Rui Zhu, Fangzhou Zhang and Liucun Zhu

187 ***Immunoglobulin Classification Based on FC* and GC* Features***
Hao Wan, Jina Zhang, Yijie Ding, Hetian Wang and Geng Tian

195 ***Predicting Ischemic Stroke Outcome Using Deep Learning Approaches***
Gang Fang, Zhennan Huang and Zhongrui Wang

203 ***MIMRDA: A Method Incorporating the miRNA and mRNA Expression Profiles for Predicting miRNA-Disease Associations to Identify Key miRNAs (microRNAs)***
Xianbin Li, Hannan Ai, Bizhou Li, Chaohui Zhang, Fanmei Meng and Yuncan Ai

223 ***A SNARE Protein Identification Method Based on iLearnPlus to Efficiently Solve the Data Imbalance Problem***
Dong Ma, Zhihua Chen, Zhanpeng He and Xueqin Huang

231 ***Analysis of Breast Cancer Based on the Dysregulated Network***
Yanhao Huo, Xianbin Li, Peng Xu, Zhenshen Bao and Wenbin Liu

242 ***Clinical Prediction of Heart Failure in Hemodialysis Patients: Based on the Extreme Gradient Boosting Method***
Yanfeng Wang, Xisha Miao, Gang Xiao, Chun Huang, Junwei Sun, Ying Wang, Panlong Li and Xu You

251 ***Identifying Functions of Proteins in Mice With Functional Embedding Features***
Hao Li, ShiQi Zhang, Lei Chen, Xiaoyong Pan, ZhanDong Li, Tao Huang and Yu-Dong Cai

frontiers | Frontiers in Genetics

EDITORIAL
published: 30 June 2022
doi: 10.3389/fgene.2022.949285

Editorial: Machine Learning Techniques on Gene Function Prediction Volume II

Ren Qi[1,2], *Arun Kumar Sangaiah*[3], *Dariusz Mrozek*[4] *and Quan Zou*[1,5]*

[1]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [2]School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, [3]School of Computing Science and Engineering, VIT University, Vellore, India, [4]Institute of Informatics, Silesian University of Technology, Gliwice, Poland, [5]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China

Keywords: machine learning, genetics, bioinformatics, feature selection, deep learning, single cell sequencing data

Editorial on the Research Topic

Machine Learning Techniques on Gene Function Prediction Volume II

Predicting the function of genes is a critical problem in biology. The current generation rate of new gene sequences is too fast to discover and validate them experimentally, emphasizing the importance of machine learning. Machine learning techniques have advanced our understanding of gene function, which have been widely employed to study amongst other things the interaction among genes and proteins, diseases and differentiation. The power of a combination of machine learning and biological analysis can be found in our first installment, Machine Learning Techniques on Gene Function Prediction Volume I, especially in predicting gene and ncRNA function.

We believe it is not the end, so we planned the second special issue on this subject, Machine Learning Techniques on Gene Function Prediction Volume II. In the first installment, we found most authors paid attention to gene and ncRNA function prediction. This Research topic will further explore the potential for machine learning applied to gene function prediction. Moreover, we would also like to share some works on single-cell sequencing data analysis and related machine learning methods. We are pleased to receive many submissions with the new sight of machine learning techniques combined with gene function prediction. All of these papers were accepted for publication with the assistance of professional referees. Twenty-four papers are finally selected from all submissions after rigorous reviews.

There are seven papers describing protein function prediction or protein identification. Ma et al. identified Soluble N-ethylmaleimide sensitive factor activating protein receptor (SNARE) proteins based on iLearnPlus and solved the problem of data imbalance. Coincidentally, Zhang et al. proposed a machine learning method to recognize SNARE proteins based on SVM and improved the identification accuracy compared with existing methods. In addition, Wan et al. distinguished immunoglobulins and non-immunoglobulins by FC* and GC* features, where immunoglobulins are critical in disease regulation. To identify hormone-binding proteins (HBPs), which are important to organisms' growth, Guo et al. present a prediction model HBP_NB, combining with k-mer feature representation, feature selection, and Naive Bayes. Furthermore, Gong et al. developed a machine learning method to identify vesicular transport proteins. Besides protein recognition, there are still two papers on protein function prediction. Li et al. presented a new multi-label classifier to explore protein function by embedding multi-type features. Meanwhile, Chen et al. provided a computing method that combined knowledge of the protein-protein interaction network and functional characteristics to help predict human protein subcellular localization patterns and their potential biological importance. Chien et al. paid attention to gene expression prediction in T- DNA mutants through machine learning methods on rice functional gene research.

OPEN ACCESS

Edited and Reviewed by:
Richard D. Emes, University of Nottingham, United Kingdom

***Correspondence:**
Quan Zou
zouquan@nclab.net

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

Received: *20 May 2022*
Accepted: *13 June 2022*
Published: *30 June 2022*

Citation:
Qi R, Sangaiah AK, Mrozek D and Zou Q (2022) Editorial: Machine Learning Techniques on Gene Function Prediction Volume II. Front. Genet. 13:949285. doi: 10.3389/fgene.2022.949285

With a wide range of mathematical statistics capabilities, composable machine learning methods can help automate and analyze complex relationships between genes and disease. Three papers focused on this issue in general diseases, which contain heart disease, Alzheimer's disease, and androgenic alopecia. Wang et al. predicted the occurrence of heart failure (HF) events in hemodialysis (HD) patients by the extreme gradient boosting method. To overcome the curse of dimensionality on Alzheimer's disease gene expression datasets, Wan et al. designed a hybrid gene selection pipeline combined with deep learning methods to improve the classification of this disease. Li et al. focused on mental stress recognition of depression disorders in patients with androgenic alopecia (AGA), and they analyzed the effect of psychological interventions in the rehabilitation of AGA patients by machine learning and fuzzy K-means clustering method FAW-FS, which combined with metaheuristic, the Filter and Wrapper algorithms. Li et al. created MIMRDA to classify top-ranked miRNAs. The method incorporated miRNA and mRNA expression profiles to predict associations between miRNA and disease and identify key miRNAs and recommended potential biomarkers as well. Two papers reviewed machine learning and deep learning methods applied in disease-gene related. Gong et al. discussed the research progress of lncRNAs and reviewed disease-related lncRNA methods and the relationships between lncRNAs and diseases. Fang et al. reviewed machine learning and deep learning methods for Ischemic Stroke disease.

Five papers focus on discussing the gene and cancer relationship. Chen et al. provided a computing method based on machine learning to predict anticancer peptides (ACPs), which is very important for the discovery of new cancer treatment drugs. Huo et al. employed dysregulated networks (DNs) to analyze subtypes of breast cancer, they measured the regulation strength between genes based on gene expression values. Huang et al. explored the function of DEG between lung cancer tissues and revealed the molecular driving mechanism of lung cancer. Chen et al. concentrated on Lung squamous cell carcinoma (LUSC) study, they screened key factors that regulate the initiation and progression of this disease by a metric learning analysis method. In addition, Han et al. reviewed sparse representation methods' applications in bioinformatics, such as cancer molecules and gene expression profiles fields.

Another aspect is studying the interactions between gene and protein, gene and gene, protein and protein. DNA-protein interactions, such as gene expression and transcriptional regulation, are tightly linked to DNA-binding proteins (DBP), Jia et al. proposed a feature extraction method that fused multiple PSSM features to predict DBP. To deepen the understanding of the principle of protein interactions, Tang et al. proposed a new hierarchical attention network structure HANPPIS predicting protein-protein interaction sites. For the study of gene-gene interactions (GGIs), most methods are based on assumptions about GGIs forms. To improve statistical performance, Guo et al. tested GGIs based on a maximal neighborhood coefficient perspective in Genome-wide association studies, which outperformed other baseline methods. Besides, Xu et al. reviewed drug-target interaction and specific applications of machine learning technique prediction methods. Notably, our special issue showed a new sight on multi-omics data storage and parallel processing. Mrozek et al. gave a large-scale and serverless computational approach for improving the quality of NGS data supporting big Multi- Omics data analyses.

To conclude, papers in this special issue have demonstrated the power of machine learning techniques in a broad range of gene function studies, especially in inferring the relationships between genes and diseases. We highly expect such studies will get great attention. Especially, more insightful results are desirable for promoting the development and progress of biology. Finally, we thank all efforts of the authors, reviewers, and staff at the Frontiers in Genetics editorial office.

AUTHOR CONTRIBUTIONS

RQ wrote the manuscript draft. DM and AS helped to revise the text. QZ gave some helpful suggestions.

FUNDING

The work was supported by the National Natural Science Foundation of China (No. 62131004, No. 61922020), the Sichuan Provincial Science Fund for Distinguished Young Scholars (2021JDJQ0025), and the Municipal Government of Quzhou under Grant Number 2020D003 and 2021D004.

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.

Copyright © 2022 Qi, Sangaiah, Mrozek and Zou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 21 June 2021
doi: 10.3389/fgene.2021.680117

Application of Machine Learning for Drug–Target Interaction Prediction

*Lei Xu[1], Xiaoqing Ru[2] and Rong Song[1]**

[1] School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China, [2] Department of Computer Science, University of Tsukuba, Tsukuba, Japan

Exploring drug–target interactions by biomedical experiments requires a lot of human, financial, and material resources. To save time and cost to meet the needs of the present generation, machine learning methods have been introduced into the prediction of drug–target interactions. The large amount of available drug and target data in existing databases, the evolving and innovative computer technologies, and the inherent characteristics of various types of machine learning have made machine learning techniques the mainstream method for drug–target interaction prediction research. In this review, details of the specific applications of machine learning in drug–target interaction prediction are summarized, the characteristics of each algorithm are analyzed, and the issues that need to be further addressed and explored for future research are discussed. The aim of this review is to provide a sound basis for the construction of high-performance models.

Keywords: machine learning, drug–target interactions, data, features, task algorithms, drug development

OPEN ACCESS

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

Reviewed by:
Ying Hong Li,
Chongqing University of Posts and Telecommunications, China
Changli Feng,
Taishan University, China

****Correspondence:***
Rong Song
sr1@szpt.edu.cn

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

Received: *13 March 2021*
Accepted: *28 May 2021*
Published: *21 June 2021*

Citation:
Xu L, Ru X and Song R (2021) Application of Machine Learning for Drug–Target Interaction Prediction. Front. Genet. 12:680117. doi: 10.3389/fgene.2021.680117

INTRODUCTION

Tens of thousands of known diseases threatening human health, and new ones are being added every year. They include emerging diseases (e.g., the currently prevalent COVID-19) and diseases that have plagued the public for many years and have no cure so far (e.g., Parkinson's disease and Alzheimer's disease) (Xu et al., 2018a, 2019). Rapidly and accurately discovering drugs that can effectively treat diseases is very important for the development of society. Long cycle and high cost are common phenomena in current drug development, but these fail to guarantee a high success rate. Many steps are required from drug development to final marketing, including drug discovery, preclinical and clinical trials, and marketing approval (Srivastava et al., 2019; Li Z. et al., 2020). The overall success rate of drug discovery and preclinical studies, which are part of the laboratory development phase, is approximately 0.05–0.1%, and less than 1% of the candidate compounds are likely to have the expected effect and proceed to the clinical trial phase. Investigating drug–target interactions is an important step in the drug discovery process and can improve the success rate of new drug discovery (Chen et al., 2019; Huang et al., 2020; Zeng et al., 2020b). These not only signal the need to expend significant resources to find and test candidate compounds one by one during the drug development phase to confirm that they meet expectations, but also demonstrate the importance of drug–target interaction prediction in the overall drug development process. Supplementally, an obvious drawback of biomedical experiment is that it does not allow for rapidly finding and solving problems, which can be detrimental to the treatment of emerging and highly infectious diseases. Therefore, machine learning methods have been introduced into the prediction of drug–target interactions.

Machine learning, a computer technology for data analysis designed to build predictive models using datasets, has become an important means of modern biological research (Xu et al., 2018b; Yang et al., 2018; Liu et al., 2019, 2020; Tang et al., 2020; Zeng et al., 2020a). It has become a mainstream technique for analyzing and solving problems involved in drug–target interaction prediction studies (Cai et al., 2018; Stephenson et al., 2019; Zeng et al., 2019; Fu et al., 2020; Wang J. et al., 2020).

THREE FACTORS

The existing data background, powerful toolkits, and current status and requirements have promoted machine learning to become the mainstream method of drug–target interaction prediction.

(1) Existing databases. With the emergence of sequencing technology, high-throughput technology and computer-aided drug design method, a large number of proteins have been sequenced and many compounds have been synthesized. On the basis of existing related works and accumulated experience, relevant data has been organized and various databases have been constructed. Most of the data in these databases are publicly available and free to download, which provides a good data foundation for solving drug–target interaction prediction problems by machine learning. Researchers can collect datasets from databases that cover different information according to their needs (Zheng et al., 2019, 2020). Some representative databases are briefly described here.

UniProt database[1]: UniProt is supported by many institutions, and is the most informative and comprehensive protein database (Consortium, 2015). It consists of five sub-databases: Swiss-Prot, TrEMBL, UniRef, UniParc, and Proteomes. Each sub-database has its own unique function. For example, Swiss-Prot is a high-quality, manually annotated, non-redundant database, in which protein annotations are derived mainly from the literature or E-value verification calculation analysis results. Proteomes is a database that provides proteomic information for species with fully sequenced genomes.

PubChem database[2]: PubChem is an open chemistry database that collects information including chemical structures, identifiers, physicochemical properties, and biological activities of chemical molecules (Kim et al., 2016, 2021). It is the world's largest database with free access to chemical information, and currently covers 109 million compounds. PubChem has become an important chemical information resource for scientists, students, and the public.

DrugBank database[3]: As a bioinformatics and cheminformatics resource, DrugBank combines detailed drug data (i.e., chemical, pharmacological, and pharmaceutical) with comprehensive target information (i.e., sequence, structure, and pathway) (Wishart et al., 2018). The latest DrugBank release (version 5.1.8.) contains 14,443 drug molecules and 5,244 non-redundant protein sequences associated with these drugs. The database describes not only clinical information on drugs, namely drug side effects and drug–drug interactions, but also contains molecular-level data, such as chemical structures of drugs and proteins targeted by drugs (Wishart et al., 2008). One significant function of DrugBank is that it supports comprehensive and complex searches, so it is used widely by the pharmaceutical industry, medicinal chemists, pharmacists, physicians, students, and the general public.

KEGG database[4]: KEGG was established in 1995 by the Kanehisa Laboratories at the Bioinformatics Center, Kyoto University, Japan, and is now one of the most commonly used international bioinformatics databases (Kanehisa and Goto, 2000). KEGG is a database used to understand the high-level functions and practicability of biological systems from molecular-level information (Li H. et al., 2020; Wang et al., 2021a) (especially large-scale molecular datasets generated by genome sequencing and other high-throughput techniques), of which the data information can be roughly classified into four major categories: system information, genetic information, chemical information, and medical information.

BindingDB database[5]: BindingDB is a publicly available, web-accessible database for measuring binding affinity, focusing on the interactions between proteins considered to be drug targets and drug-like small molecules (Liu et al., 2007). BindingDB currently contains 2,114,159 binding data between 8,202 protein targets and 928,022 small molecules.

(2) Powerful toolkits and web servers. Bioinformatics and cheminformatics are emerging interdisciplinary fields that use computers to solve biological and chemical problems. Many toolkits and web servers have been developed (Zuo et al., 2017; Zou et al., 2019; Lin et al., 2020; Pang and Liu, 2020; Shao et al., 2021), which can help to solve problems in drug–target interaction prediction.

STITCH[6]: STITCH not only includes experimentally validated drug–target interaction data, but also integrates predicted drug–target relationships (Kuhn et al., 2007). This website can clearly depict the protein–protein interactions, protein–compound interactions, and the strength of the interactions.

SwissTargetPrediction[7]: SwissTargetPrediction can estimate the most likely macromolecule to be targeted by a biologically active small molecule and count the percentage of each target type targeted by the small molecule (Gfeller et al., 2014).

RDkit[8]: RDkit is a powerful python toolkit for chemical information, which has functions such as acquiring molecule information from multiple formats, obtaining information about atoms, bonds, and rings in molecules, generating molecular descriptors and molecular fingerprints of compounds, and calculating similarities of compound structures (Landrum, 2013).

OpenChem[9]: OpenChem is a pytorch-based deep learning toolkit for computational chemistry and drug design,

[1]https://www.uniprot.org/

[2]https://pubchem.ncbi.nlm.nih.gov/

[3]https://go.drugbank.com/

[4]https://www.genome.jp/kegg/

[5]https://www.bindingdb.org/bind/index.jsp

[6]http://stitch.embl.de/

[7]http://www.swisstargetprediction.ch/

[8]https://www.rdkit.org/

[9]https://mariewelt.github.io/OpenChem/html/index.html

which contains Feature2Label, Smiles2Label, Graph2Label, SiameseModel, GenerativeRNN, and MolecularRNN (Korshunova et al., 2021). Users can train predictive models for classification, regression, and multi-task problems, and develop generative models for generating novel molecules with optimized properties. Its goal is to make deep learning an easy-to-use tool for researchers in computational chemistry and drug design.

iFeature[10]: iFeature is a python toolkit that can compute various structural and physicochemical property descriptors from protein and peptide sequences. iFeature can compute and extract comprehensive spectra for 18 major sequence coding schemes, including 53 different types of feature descriptors. In addition, iFeature integrates 12 different types of commonly used feature clustering, selection, and dimensionality reduction algorithms (Chen et al., 2018).

Pse-in-one[11]: Pse-in-one is a python toolkit that generates all possible pseudo-components of DNA, RNA, and protein sequences. It covers a total of 28 different patterns, 14 for DNA sequences, 6 for RNA sequences, and 8 for protein sequences (Liu et al., 2015, 2017). This toolkit is widely and increasingly used by researchers to tackle various problems in computational biology, and a more specific and detailed version BioSeq-Analysis (Liu, 2019) has recently been released.

(3) Current status and requirements. With the development of high-throughput technologies, many compounds and proteins have been mined. The human genome contains more than 20,000 genes, and approximately 80% of them can encode one or more proteins. Only a small number of proteins have been identified as pharmacologically active and are targets for currently approved drugs. The pharmacological functions of most proteins remain to be demonstrated. This is also true for most compounds. For example, there are currently 111 million compounds in the PubChem database, but proteins that could interact with many of these compounds are unknown. In addition, it is obvious that the traditional approach of wet experiments is not feasible for some emerging, highly infectious and destructive new pathogens, such as the SARS, H7N9, Ebola, Mers, and COVID-19 viruses (Cheng et al., 2021). Considering the huge amounts of available data and large numbers of diseases that cause serious social health risks, using computational chemistry-related theories and computer simulation methods to computationally predict drug–target interaction can effectively improve efficiency. Machine learning-based methods have become effective ways to compensate for the shortcomings of traditional biochemical experimental methods.

APPLICATIONS

The current drug–target interaction prediction procedures are shown in **Figure 1**. Existing studies on drug–target interaction prediction have shown that using different calculation or optimization methods in the steps of data set acquisition, feature extraction and processing, and task algorithm selection can build models with good performance.

(1) Dataset acquisition. Redundant data, unbalanced categories, and unrepresentative samples can lead to long experimental cycles, as well as inaccurate and biased experimental results. Different data acquisition methods have been used to avoid or reduce the impact of these problems on model construction. For example, Wang et al. (2010) collected negative examples by random selection to solve the data imbalance problem. Wang et al. (2018) also used random selection to extract negative examples, and this operation was performed five times to reduce the impact of the unverified negative samples. Pdti-EssB (Mahmud et al., 2020) used random under-sampling and under-sampling clustering to address the data imbalance problem.

Currently, most target molecules are proteins, of which four protein families [kinases, G protein-coupled receptors (GPCRs), ion channels, and nuclear receptors] account for 44% of the target molecules, and 70% of the currently developed drugs are targeted to these four protein families. Datasets established by Yamanishi et al. (2008), which contain the interactions between these four proteins and drugs,have been widely used (Öztürk et al., 2018; Mahmud et al., 2020). The relevant data can be downloaded from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. Most of the computational approaches based on these datasets have focused on binary classification, that is, they only explore whether a drug can interact with a particular protein. To further accelerate process and reduce cost, drug–target affinity has been explored in some studies. Drug–target affinity is a key property that determines the strength of the interaction between the small molecule drug and the target. The commonly used datasets for predicting drug–target affinity are the Kinase (Davis et al., 2011) and KIBA (Tang et al., 2014) datasets.

(2) Feature extraction and processing. Accurate and comprehensive descriptions of the biological or chemical functional information of drugs and targets in numerical form play an important role in the construction of high-performance models. Feature extraction of drugs and targets can be performed from different perspectives (Cheng, 2019; Zhao T. et al., 2020). For example, iGPCR-Drug (Xiao et al., 2013) obtains drug features by discrete Fourier transform of drug molecular fingerprints and extracts GPCR features according to pseudo amino acid compositions. DrugE-Rank (Yuan et al., 2016) represents drug features according to general descriptors and extracts target features according to amino acid composition, transformation, and distribution. TargetGDrug (Hu J. et al., 2016) extracts drug features by applying wavelet transform to drug molecular fingerprints and extracts GPCR features according to evolutionary information. Ru et al. (2020) extracted protein features using the distance-based top-n-gram algorithm and obtained drug features according to general descriptors. Chemical databases store information in a textual representation and the simplified molecular input line entry specification (SMILES) format is a common standard used in many cheminformatics software. Each SMILES string encodes structural information that can be used to predict complex chemical properties, and a large number of machine learning

[10]https://ifeature.erc.monash.edu/

[11]http://bioinformatics.hitsz.edu.cn/Pse-in-One/

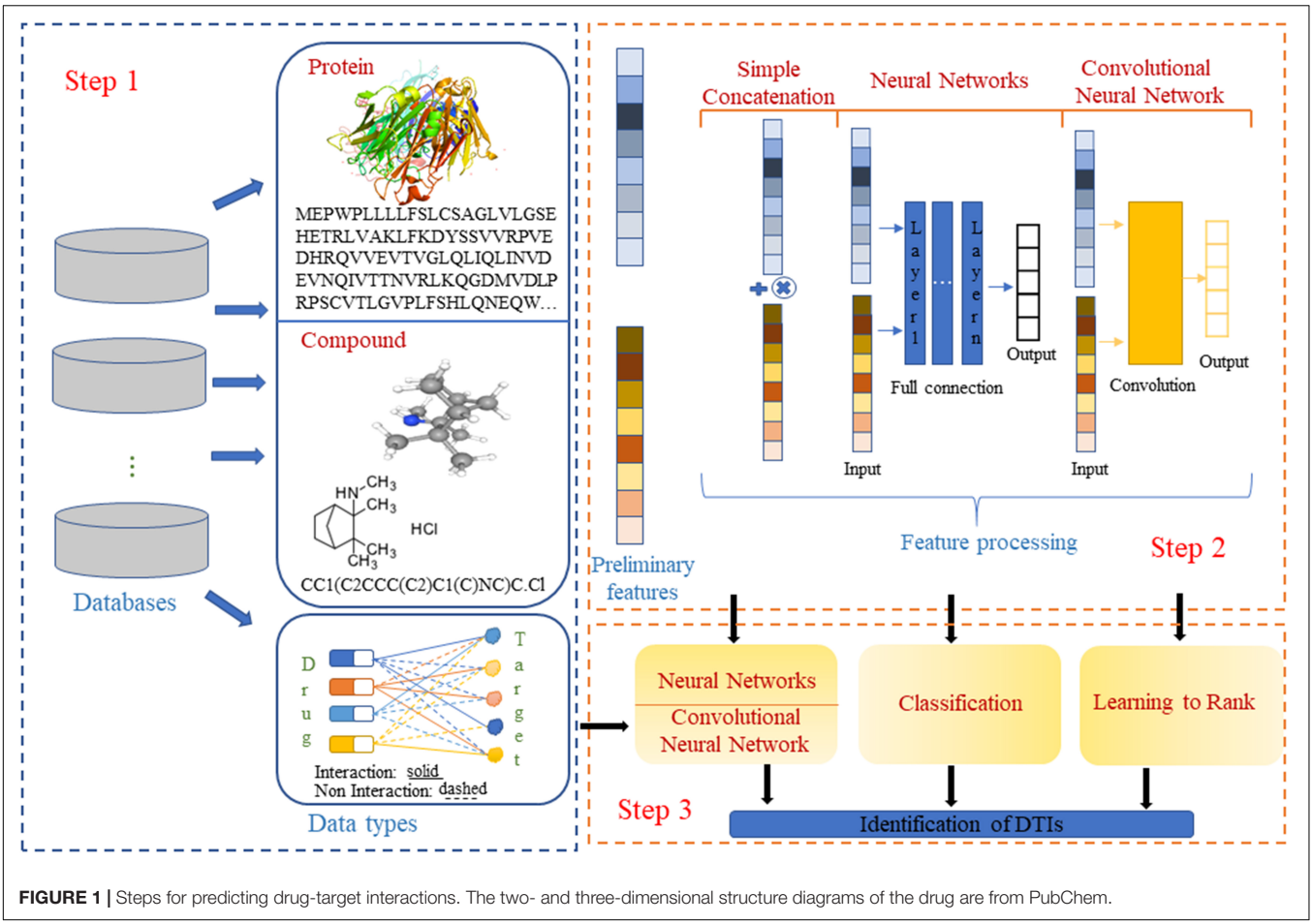


FIGURE 1 | Steps for predicting drug-target interactions. The two- and three-dimensional structure diagrams of the drug are from PubChem.

models can extract molecular features of compounds according to SMILES strings. Recently, convolutional neural networks (CNNs) and recurrent neural networks have been used for molecular feature extraction. Hirohara et al. (2018) transformed SMILES strings into two-dimensional matrices and used CNNs to extract molecular features. Goh et al. (2017) applied natural language processing to SMILES feature extraction and used recurrent neural networks for molecular strings.

The presence of invalid or redundant features not only reduces the accuracy of the experiment result but also lengthens the experimental period. Low-dimensional and comprehensive information feature sets are expected. Therefore, a variety of methods for processing features have been applied to related rearch (Zou et al., 2016a,b; Guo et al., 2020; Zhang G. et al., 2020; Zhao X. et al., 2020). For example, to reduce the noise between features, Li et al. (2017) used principal component analysis (PCA) to reduce the dimensionality of drugs and targets features. Tabei et al. (2012) combined 881 substructures of drugs and 876 Pfam domain structures of targets by tensor product to form feature vectors of drug–target pairs. MFDR (Hu P.-W. et al., 2016) used autoencoders as the building blocks of a deep network to reconstruct drug and protein features into a low-dimensional new representation. DeepConv-DT (Lee et al., 2019) used CNNs on raw protein sequences to capture local amino acid residue information by convolving amino acid subsequences of various lengths.

(3) Selection of task algorithms. Several task algorithms have been used for drug–target interaction prediction, such as classification algorithms, learning to rank algorithms, and deep learning algorithms (Cheng et al., 2019; Lv et al., 2019; Tao et al., 2020; Zhang Y. et al., 2020).

Most of the existing studies treat drug–target interaction prediction as binary tasks, and different classification algorithms have been applied. For example, Bleakley and Yamanishi (2009) proposed a bipartite local model (BLM) based on a support vector machine (SVM) kernel to predict drug–target relationships. LRF-DTI (Shi et al., 2019) is a drug–target interaction prediction method using Lasso for feature extraction and random forest for classification. Yamanishi et al. (2010) used a distance learning algorithm as a classifier. Pred-binding (Shar et al., 2016) extracted molecular structure and protein sequence features, and used support vector machines and random forests to classify whether drugs and targets can be docked.

Drug–target interaction prediction can be regarded as a ranking task. Exploring the strength of drug–target interactions can shorten the drug development process and save expenses. Zhang et al. (2015) applied six learning to rank algorithms (Prank, RankNet, RankBoost, SVMRank, AdaRank, and ListNet)

to virtual screening of drugs,their study showed that learning to rank is an effective computational strategy, especially because of its novel use in cross-target virtual screening and heterogeneous data integration. DrugE-Rank (Yuan et al., 2016) used protein amino acid composition, transformation and distribution information, compound descriptor information, and output information of six classifiers as features to be input into the learning to ranking algorithm to improve the performance of drug-target interaction prediction.

Neural networks have also been used to solve related problems in the prediction of drug–target interactions. Prado-Prado et al. (2011) used the entropy information of drug–protein complexes and neural networks to predict drug–target affinity values. DeepDTA (Öztürk et al., 2018) proposed a deep-learning based model that used only sequence information of both targets and drugs, One novel approach used in this work is the modeling of protein sequences and compound 1D representations with CNNs. GraphDTA (Nguyen et al., 2019) focused on the fact that molecules are by nature formed by chemical bonding of atoms, and used graph convolutional network to learn drug-target binding affinity.

DISCUSSION

Under the background of the existing chemical and biological computing theory, big data and rapid development of computer technology, the use of machine learning for drug-target interaction prediction does have many benefits, but there are still some problems that need to be further explored.

(1) Data heterogeneity. Most of the existing studies are based on publicly available data in databases that collect data with different focuses, and each database has its own criteria for judging the data. Drugs, targets, and related data from different databases often have different terminological descriptions and different organization structures, such inconsistencies make data integration difficult.

(2) Effective representation of biological and chemical features. Feature engineering is a key concern in building machine learning models. There are often technical difficulties in how to effectively extract key features and how to deal with data with high dimensionality. Existing studies have shown that the features of proteins and drugs can be extracted from a variety of angles, and the combination of information from these angles can achieve complementary effects. Most drug–target interaction prediction studies only extract relatively one-sided information, and do not comprehensively consider the information from multiple perspectives. In addition, most studies have focused on extracting drug molecule and protein features separately, ignoring the potentially valid association that may exist between drug and target. Moreover, the direct concatenation of biologically unrelated features may lead to a decrease in prediction accuracy.

(3) Characteristics of task algorithms. The classification, ranking, or deep learning methods used in drug–target interaction prediction all have their own characteristics. Different computational approaches can be used to solve different problems in drug–target interaction prediction, however, these algorithms also have shortcomings. Classification is the simplest and most understandable task. However, there is an obvious and long-standing defect in this task that it is necessary to collect negative samples. Most existing classification studies take experimentally validated drug–target pairs with known interactions as positive samples, and unvalidated or unknown drug–target pairs as negative examples. Among these negative examples, there may be positive samples that have not been accurately validated, the performance of a model that is based on such a dataset will be biased.

On the basis of the existence of one-to-many or many-to-many relationships between queries and documents, learning to rank can be used in multi-target drug discovery. Early drug development followed the "one drug, one target" principle, with the aim of finding high-affinity, high-selective drugs for a specific receptor associated with a particular disease. However, the number of complex diseases is increasing and the proteins associated with these diseases are not limited to one, therefore drug combinations are used to achieve the optimal therapeutic effect. Clinical pharmacology studies have shown that drug combinations greatly increase the incidence of adverse drug reactions, but because of the lack of multi-target drugs, such risks have to be taken. Multi-target drugs are undoubtedly an important area for future research. Therefore, using the characteristics of learning to rank to tackle the multi-target problem of drugs deserves to be explored further. Learning to rank was originally applied for information retrieval. Its output is a relative score of correlation between queries and documents (Cheng, 2020; Ru et al., 2021). This is not sufficient for studies that require accurate prediction of drug–target affinities.

The use of neural networks for predicting accurate drug–target affinity values has shown great potential in this research area. Neural networks can fuse drug and target features, which have changed the current situation of simple concatenation or tensor products of drug and target features. Deep learning contains more neural network structures with multiple implicit layers compared with traditional machine learning, which allows deep learning to handle large datasets and identify complex patterns from the learning process. But for the same reason, neural networks require much more execution time than classification or ranking algorithms. It will lead to overfitting when the drug and target feature dimensions are high.

Although existing machine learning methods have opened a new area in drug–target interaction prediction, they have not achieved satisfactory results so far. Therefore, there is still a need to develop new theoretical and computational methods for drug–target interaction prediction.

CONCLUSION

Drug–target interaction prediction can help to screen out unsuitable compounds and is an important step in the development of new drugs. In this review, we describe the importance of drug–target interaction prediction, analyze in

detail the three main reasons why machine learning has become a mainstream technique, summarize the specific applications of machine learning methods in each step of building machine learning models, analyze the shortcomings of existing research methods, and discuss several aspects that can be further explored (Wei et al., 2014, 2017a,b, 2018, 2019; Ding et al., 2017, 2019, 2020a,b; Jin Q. et al., 2019; Jin S. et al., 2019; Li J. et al., 2020; Su et al., 2020; Wang H. et al., 2020; Zeng et al., 2020c,d; Zhai et al., 2020; Wang et al., 2021b). This review provides meaningful perspectives for future drug–target interaction prediction studies, especially the application of learning to rank to deal with multi-target drug problems.

AUTHOR CONTRIBUTIONS

XR drafted the manuscript. LX and RS initiated the idea, conceived the whole process, and finalized the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the natural science foundation of Guangdong province (grant No. 2018A0303130084) and the Grant of Shenzhen Polytechnic (No. 6021310015K).

REFERENCES

Bleakley, K., and Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 25, 2397–2403. doi: 10.1093/bioinformatics/btp433

Cai, J., Cai, H., Chen, J., and Yang, X. (2018). Identifying "many-to-many" relationships between gene-expression data and drug-response data via sparse binary matching. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 165–176.

Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2019). HOGMMNC: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. *Bioinformatics* 35, 602–610. doi: 10.1093/bioinformatics/bty662

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

Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/1566523219041910221133 07

Cheng, L. (2020). Omics Data and Artificial Intelligence: New Challenges for Gene Therapy. *Curr. Gene Ther.* 20:1. doi: 10.2174/1566523220012006041500 41

Cheng, L., Han, X., Zhu, Z., Qi, C., Wang, P., and Zhang, X. (2021). Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2. *Brief. Bioinform.* 22, 1442–1450. doi: 10.1093/bib/bbab042

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

Consortium, U. (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 29, 1046–1051.

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inform. Sci.* 418, 546–560. doi: 10.1016/j.ins.2017.08.045

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

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

Ding, Y., Tang, J., and Guo, F. (2020b). Identification of drug-target interactions via fuzzy bipartite local model. *Neural Comput. Appli.* 23, 10303–10319. doi: 10.1007/s00521-019-04569-z

Fu, X., Cai, L., Zeng, X., and Zou, Q. J. B. (2020). StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. *Bioinformatics* 36, 3028–3034. doi: 10.1093/bioinformatics/btaa131

Gfeller, D., Grosdidier, A., Wirth, M., Daina, A., Michielin, O., and Zoete, V. (2014). SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res.* 42, W32–W38.

Goh, G. B., Hodas, N. O., Siegel, C., and Vishnu, A. (2017). Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv* 171202034.

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

Hirohara, M., Saito, Y., Koda, Y., Sato, K., and Sakakibara, Y. (2018). Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC bioinformatics* 19:526. doi: 10.1186/s12859-018-2523-5

Hu, J., Li, Y., Yang, J.-Y., Shen, H.-B., and Yu, D.-J. (2016). GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure. *Comput. Biol. Chem.* 60, 59–71. doi: 10.1016/j.compbiolchem.2015.11.007

Hu, P.-W., Chan, K. C., and You, Z.-H. (2016). "Large-scale prediction of drug-target interactions from deep representations," in *2016 International Joint Conference on Neural Networks (IJCNN* (Vancouver: IEEE), 1236–1243.

Huang, J., Chen, J., Zhang, B., Zhu, L., and Cai, H. (2020). Evaluation of gene–drug common module identification methods using pharmacogenomics data. *Brief. Bioinform.* 22:bbaa087.

Jin, Q., Meng, Z., Tuan, D. P., Chen, Q., Wei, L., and Su, R. (2019). DUNet: A deformable network for retinal vessel segmentation. *Knowl. Based Syst.* 178, 149–162. doi: 10.1016/j.knosys.2019.04.025

Jin, S., Zeng, X., Fang, J., Lin, J., Chan, S. Y., and Erzurum, S. C. (2019). Cheng FJNsb, applications: A network-based approach to uncover microRNA-mediated disease comorbidities and potential pathobiological implications. *NPJ Syst. Biol. Appl.* 5, 1–11.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2021). PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* 49, D1388–D1395.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., et al. (2016). PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213.

Korshunova, M., Ginsburg, B., Tropsha, A., and Isayev, O. (2021). OpenChem: A Deep Learning Toolkit for Computational Chemistry and Drug Design. *J. Chem. Inform. Model.* 61, 7–13. doi: 10.1021/acs.jcim.0c00971

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., and Bork, P. (2007). STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.* 36, D684–D688.

Landrum, G. (2013). Rdkit documentation. *Release* 1:4.

Lee, I., Keum, J., and Nam, H. (2019). DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* 15:e1007129. doi: 0.1371/journal.pcbi.1007129

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.* doi: 0.1093/bib/bbaa342 [Epub Online ahead of print].

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.* 22:bbaa159.

Li, Z., Han, P., You, Z.-H., Li, X., Zhang, Y., Yu, H., et al. (2017). In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences. *Sci. Rep.* 7, 1–13.

Li, Z., Zhang, T., Lei, H., Wei, L., Liu, Y., Shi, Y., et al. (2020). Research on Gastric Cancer's Drug-resistant Gene Regulatory Network Model. *Curr. Bioinform.* 15, 225–234. doi: 10.2174/1574893614666190722102557

Lin, X., Quan, Z., Wang, Z., Huang, H., and Zeng, X. (2020). A novel molecular representation with BiGRU neural networks for learning atom. *Brief. Bioinform.* 21, 2099–2111. doi: 10.1093/bib/bbz125

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., 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:e127. doi: 10.1093/nar/gkz740

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71.

Liu, B., Wu, H., and Chou, K.-C. (2017). Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nat. Sci.* 9:67. doi: 10.4236/ns.2017.94007

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* 35, D198–D201.

Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., and Zou, Q. (2020). Zeng XJBib: Computational methods for identifying the critical nodes in biological networks. *Brief. Bioinform.* 21, 486–497. doi: 10.1093/bib/bbz011

Lv, Z. B., Ao, C. Y., and Zou, Q. (2019). Protein Function Prediction: From Traditional Classifier to Deep Learning. *Proteomics* 19:2.

Mahmud, S. H., Chen, W., Meng, H., Jahan, H., Liu, Y., and Hasan, S. M. (2020). Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting. *Anal. Biochem.* 589:13507.

Nguyen, T., Le, H., and Venkatesh, S. (2019). GraphDTA: prediction of drug–target binding affinity using graph convolutional networks. *BioRxiv* doi: 10.1101/684662

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 34, i821–i829.

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. Bioinform.* doi: 10.1109/TCBB.2020.3031888 [Epub Online ahead of print].

Prado-Prado, F., García-Mera, X., Abeijón, P., Alonso, N., Caamaño, O., Yáñez, M., et al. (2011). Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica. *Eur. J. Med. Chem.* 46, 1074–1094. doi: 10.1016/j.ejmech.2011.01.023

Ru, X., Wang, L., Li, L., Ding, H., Ye, X., and Zou, Q. (2020). Exploration of the correlation between GPCRs and drugs based on a learning to rank algorithm. *Comput. Biol. Med.* 119:103660. doi: 10.1016/j.compbiomed.2020.103660

Ru, X., Ye, X., Sakurai, T., and Zou, Q. (2021). Application of learning to rank in bioinformatics tasks. *Brief. Bioinform.* doi: 10.1093/bib/bbaa1394 [Epub Online ahead of print].

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

Shar, P. A., Tao, W., Gao, S., Huang, C., Li, B., Zhang, W., et al. (2016). Pred-binding: large-scale protein–ligand binding affinity prediction. *J. Enzyme Inhib. Med. Chem.* 31, 1443–1450. doi: 10.3109/14756366.2016.1144594

Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., and Yu, B. (2019). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics* 111, 1839–1852. doi: 10.1016/j.ygeno.2018.12.007

Srivastava, N., Mishra, B. N., and Srivastava, P. (2019). In-Silico Identification of Drug Lead Molecule Against Pesticide Exposed-neurodevelopmental Disorders Through Network-based Computational Model Approach. *Curr. Bioinform.* 14, 460–467. doi: 10.2174/1574893613666181112130346

Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., et al. (2019). Survey of Machine Learning Techniques in Drug Discovery. *Curr. Drug Metab.* 20, 185–193. doi: 10.2174/1389200219666180820112457

Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2020). Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief. Bioinform.* 21, 408–420. doi: 10.1093/bib/bby124

Tabei, Y., Pauwels, E., Stoven, V., Takemoto, K., and Yamanishi, Y. (2012). Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics* 28, i487–i494.

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., et al. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inform. Model.* 54, 735–743. doi: 10.1021/ci400709d

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, 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.

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., Liang, P., Zheng, L., Long, C., Li, H., and Zuo, Y. (2021a). eHSCPr discriminating the cell identity involved in endothelial to hematopoietic transition. *Bioinformatics*

Wang, H., Tang, J., Ding, Y., and Guo, F. (2021b). Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment. *Brief. Bioinform.* doi: 10.1093/bib/bbaa409 [Epub Online ahead of print].

Wang, J., Wang, H., Wang, X., and Chang, H. (2020). Predicting drug-target interactions via FM-DNN learning. *Curr. Bioinform.* 15, 68–76. doi: 10.2174/1574893614666190227160538

Wang, L., You, Z.-H., Chen, X., Xia, S.-X., Liu, F., Yan, X., et al. (2018). A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *J. Comput. Biol.* 25, 361–373. doi: 10.1089/cmb.2017.0135

Wang, Y.-C., Yang, Z.-X., Wang, Y., and Deng, N.-Y. (2010). Computationally probing drug-protein interactions via support vector machine. *Lett. Drug Des. Discov.* 7, 370–378. doi: 10.2174/157018010791163433

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

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. Bioinform.* 11, 192–201. doi: 10.1109/tcbb.2013.146

Wei, L., Wan, S., Guo, J., and Wong, K. K. L. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019). Fast Prediction of Protein Methylation Sites Using a Sequence-Based Feature Selection Technique. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 16, 1264–1273. doi: 10.1109/tcbb.2017.2670558

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017b). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906.

Xiao, X., Min, J.-L., Wang, P., and Chou, K.-C. (2013). iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PLoS One* 8:e72234. doi: 10.1371/journal.pone.0072234

Xu, L., Liang, G., Shi, S., and Liao, C. (2018a). SeqSVM: A Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins. *Int. J. Mol. Sci.* 19:1773. doi: 10.3390/ijms19061773

Xu, L., Liang, G., Liao, C., Chen, G. D., and Chang, C. C. (2018b). An Efficient Classifier for Alzheimer's Disease Genes Identification. *Molecules* 23:13.

Xu, L., Liang, G., Liao, C., Chen, G. D., and Chang, C. C. (2019). k-Skip-n-Gram-RF: A Random Forest Based Method for Alzheimer's Disease Protein Identification. *Front. Genet.* 10:7. doi: 10.3389/fgene.2019.00033

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–i240.

Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26, i246–i254.

Yang, X., Han, G., Chen, J., and Cai, H. (2018). Finding correlated patterns via high-order matching for multiple sourced biological data. *IEEE Trans. Biomed. Eng.* 66, 1017–1025. doi: 10.1109/tbme.2018.2866266

Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S. (2016). DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 32, i18–i27.

Zeng, X., Lin, Y., He, Y., Lv, L., and Min, X. (2020a). Deep collaborative filtering for prediction of disease genes. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1639–1647.

Zeng, X., Song, X., Ma, T., Pan, X., Zhou, Y., Hou, Y., et al. (2020b). Cheng FJJopr: Repurpose open data to discover therapeutics for COVID-19 using deep learning. *J. Proteome Res.* 19, 4624–4636. doi: 10.1021/acs.jproteome.0c00316

Zeng, X., Zhu, S., Hou, Y., Zhang, P., Li, L., Li, J., et al. (2020c). Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics* 36, 2805–2812. doi: 10.1093/bioinformatics/btaa010

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 35, 5191–5198. doi: 10.1093/bioinformatics/btz418

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020d). Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 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. Cell Dev. Biol.* 8:591487. doi: 10.3389/fcell.2020.591487

Zhang, G., Yu, P., Wang, J., and Yan, C. (2020). Feature Selection Algorithm for High-dimensional Biomedical Data Using Information Gain and Improved Chemical Reaction Optimization. *Curr. Bioinform.* 15, 912–926. doi: 10.2174/1574893615666200204154358

Zhang, W., Ji, L., Chen, Y., Tang, K., Wang, H., Zhu, R., et al. (2015). When drug discovery meets web search: learning to rank for ligand-based virtual screening. *J. Cheminform.* 7, 1–13.

Zhang, Y., Yan, J., Chen, S., Gong, M., Gao, D., Zhu, M., et al. (2020). Review of the Applications of Deep Learning in Bioinformatics. *Curr. Bioinform.* 15, 898–911. doi: 10.2174/1574893615999200711165743

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. 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* 21:43. doi: 10.1186/s12859-020-3388-y

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

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

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

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

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

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

The handling editor declared a past co-authorship with one of the authors, LX.

Copyright © 2021 Xu, Ru and Song. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 June 2021
doi: 10.3389/fgene.2021.698477

ACP-DA: Improving the Prediction of Anticancer Peptides Using Data Augmentation

Xian-gan Chen[1,2,3], Wen Zhang[4,5], Xiaofei Yang[1,2,3], Chenhong Li[1,2,3] and Hengling Chen[1,2,3]*

[1] School of Biomedical Engineering, South-Central University for Nationalities, Wuhan, China, [2] Hubei Key Laboratory of Medical Information Analysis and Tumor Diagnosis & Treatment, South-Central University for Nationalities, Wuhan, China, [3] Key Laboratory of Cognitive Science (South-Central University for Nationalities), State Ethnic Affairs Commission, Wuhan, China, [4] College of Informatics, Huazhong Agricultural University, Wuhan, China, [5] Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan, China

Anticancer peptides (ACPs) have provided a promising perspective for cancer treatment, and the prediction of ACPs is very important for the discovery of new cancer treatment drugs. It is time consuming and expensive to use experimental methods to identify ACPs, so computational methods for ACP identification are urgently needed. There have been many effective computational methods, especially machine learning-based methods, proposed for such predictions. Most of the current machine learning methods try to find suitable features or design effective feature learning techniques to accurately represent ACPs. However, the performance of these methods can be further improved for cases with insufficient numbers of samples. In this article, we propose an ACP prediction model called ACP-DA (Data Augmentation), which uses data augmentation for insufficient samples to improve the prediction performance. In our method, to better exploit the information of peptide sequences, peptide sequences are represented by integrating binary profile features and AAindex features, and then the samples in the training set are augmented in the feature space. After data augmentation, the samples are used to train the machine learning model, which is used to predict ACPs. The performance of ACP-DA exceeds that of existing methods, and ACP-DA achieves better performance in the prediction of ACPs compared with a method without data augmentation. The proposed method is available at http://github.com/chenxgscuec/ACPDA.

Keywords: anticancer peptide prediction, data augmentation, feature representation, multilayer perception, machine learning

OPEN ACCESS

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

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Hua Tang,
Southwest Medical University, China

***Correspondence:**
Wen Zhang
zhangwen@mail.hzau.edu.cn

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

Received: *21 April 2021*
Accepted: *07 June 2021*
Published: *30 June 2021*

Citation:
Chen X-g, Zhang W, Yang X, Li C and Chen H (2021) ACP-DA: Improving the Prediction of Anticancer Peptides Using Data Augmentation. Front. Genet. 12:698477. doi: 10.3389/fgene.2021.698477

INTRODUCTION

With the increase in population age, cancer has become one of the most threatening diseases for humans (Bray et al., 2018; Zhang et al., 2020). The complexity and heterogeneity of cancer make it difficult to treat. Traditional clinical methods such as surgery, radiotherapy, and chemotherapy can be used to treat cancer, but the side effects of these methods are very obvious and can cause great discomfort for patients (Doja et al., 2020). Although traditional anticancer drugs are effective,

their shortcomings, such as gastrointestinal damage (Mitchell, 2006), are also notable and can easily cause multidrug tumor resistance (Holohan et al., 2013; Wijdeven et al., 2016). In view of these problems, it is urgent to find and design novel cancer treatments and anticancer agents to fight cancer. In recent years, due to their high specificity, low production cost, and low toxicity profile, peptides have emerged as alternative anticancer agents (Otvos, 2008).

Anticancer peptides (ACPs), a class of naturally occurring important defense substances, provide a new direction for research involving novel anticancer drugs. ACPs are usually short peptides with a length of 10–50 amino acids. Since ACPs only interact with the anionic cell membrane components of cancer cells, they exhibit extensive cytotoxicity against a variety of cancer cells but not normal cells (Barras and Widmann, 2011; Boohaker et al., 2012). There are currently many peptide-based therapies being evaluated for their efficacy in treating tumors. However, only a few peptides can be used for clinical treatment. Therefore, the discovery of new ACPs is of great significance to the successful clinical application of these peptides. An increasing number of ACPs from protein sequences have been identified and verified by experiments (Tyagi et al., 2013), but it is time consuming and expensive to use experimental methods to identify ACPs. Therefore, computational methods for ACP identification are urgently needed.

There are many computational methods in the field of bioinformatics, that are used to solve different kinds of bioinformatics problems (Zou et al., 2018; Zou, 2019; Deng et al., 2020; Huang et al., 2020). There are many computational methods, especially machine learning-based methods, for the identification of ACPs. Anti-CP was the first computational tool based on a support vector machine (SVM), which used sequence-based features and binary profiles (Tyagi et al., 2013). Hajisharifi et al. (2014) considered Chou's pseudo amino acid composition (PseAAC) and local alignment kernel for the prediction of ACPs (Amanat et al., 2020; Hasan et al., 2020; Naseer et al., 2020). Chen et al. (2016) developed a method based on the optimization of g-gap dipeptide components. Li and Wang (2016) selected the amino acid composition, average chemical shifts, and reduced amino acid composition to represent ACPs. Wei et al. (2018) developed a feature representation learning model with a two-step feature selection technique to improve the prediction of ACPs. Xu et al. (2018) proposed using 400-dimensional features with g-gap dipeptide features for ACPs. Boopathi et al. (2019) applied a two-step method to obtain optimal feature vectors, which were used as inputs for a SVM. Ge et al. (2019) proposed a generalized chaos game representation (CGR) for ACP identification. Ge et al. (2020) used different features and multiple classifiers and the classifier outputs were used as inputs for a SVM, which was used to identify ACPs. Yu et al. (2020) explored three different deep-learning architectures and found that recurrent neural networks are superior to other architectures. Zhao et al. (2020) used a deep belief network to encode the sequences and chemical features of ACPs and applied random relevance vector machines to identify ACPs. Yi et al., 2019 proposed a deep learning long short-term memory (LSTM) neural network model called ACP-DL to predict novel ACPs. Agrawal et al., 2020 used various features and different machine learning classifiers on two datasets for the prediction of ACPs.

However, the number of ACPs involved in the above methods did not exceed 1000 cases, which is not a large number. The performance of these methods could potentially be further improved if additional ACPs are considered. In this article, we use data augmentation to increase the number of samples in the training set and further improve the performance of ACP prediction methods based on machine learning. Specifically, we propose an ACP prediction model with Data Augmentation, named ACP-DA. In our method, binary profile features (BPFs) and the features that describe the physicochemical properties of amino acids are concatenated to represent peptides, and the samples in the training set are augmented in the feature space. The samples after data augmentation are used to train a machine learning model, which is used for the prediction of ACPs.

The flowchart of ACP-DA is shown in **Figure 1**. There are four major steps in our method. First, given peptide sequences as the input, each sequence is preprocessed to equal length. Second, the peptide sequences are represented by concatenating BPFs and AAindex features selected based on minimum redundancy maximum relevance (mRMR). Third, data augmentation is performed in the feature space of samples in the training set. Ultimately, the data-augmented samples are used to train a multilayer perception (MLP) model, and the trained MLP model assigns labels to the samples in the testing set. To evaluate the performance of our method, we used five-fold cross-validation to evaluate ACP-DA based on two benchmark datasets: ACP740 and ACP240. We discuss the performance of this method with different parameters and evaluate the effect of data augmentation based on different classifiers. The experimental results show that data augmentation can help improve the prediction of ACPs under the condition of using suitable classifiers, and our method is suitable for ACP prediction.

MATERIALS AND METHODS

Datasets

A good dataset is very important for establishing a reliable ACP prediction model. In recent years, several excellent datasets have been established (Wei et al., 2018; Yi et al., 2019). We selected two benchmark datasets, ACP740 and ACP240, which have more samples (Yi et al., 2019) than others. The similarities between the two datasets were as follow: ACPs verified in the experiment were regarded as positive samples, and anti-microbial peptides (AMPs) without anticancer function were regarded as negative samples. CD-HIT was used to remove the peptide sequences with a similarity of more than 90%. The difference was that ACP740 was from Chen et al.'s and Wei et al.'s studies, while ACP240 was from Yi et al.'s studies. There were 376 positive samples and 364 negative samples in ACP740, and there were 129 positive samples and 111 negative samples in ACP240. There were no overlapping data between ACP740 and ACP240, and both are non-redundant datasets. These two datasets are available at https://github.com/haichengyi/ACP-DL.

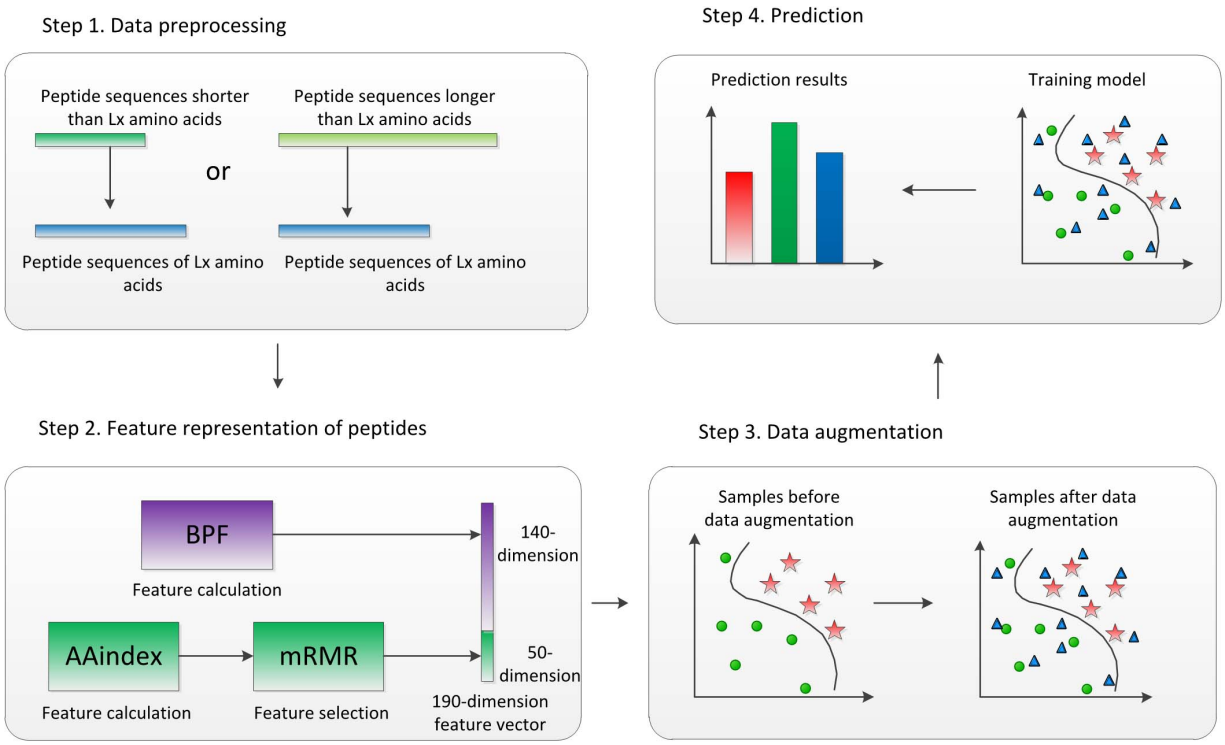


FIGURE 1 | Flowchart of ACP-DA. Binary profile features (BPFs) and AAindex features after feature selection were concatenated to represent peptides, and the samples in the training set were augmented in the feature space. The samples after data augmentation were used to train the multilayer perception (MLP) model, which was used for the prediction of anticancer peptides (ACPs).

Prediction Framework

To identify potential ACPs, we propose an ACP prediction model called ACP-DA. **Figure 1** illustrates the framework of the proposed method. First, we preprocess the peptide sequences to equal length, and the length is selected to be L_X amino acids so that the next feature calculation can be performed. Second, the AAindex in the iFeature Python package (Chen et al., 2018) is used to calculate the physicochemical properties of the amino acids in each sequence, and mRMR (Peng et al., 2005) is then used for feature selection. BPFs and AAindex features after feature selection for each peptide sequence are concatenated to represent a peptide. Third, data augmentation is performed in the feature space of samples in the training set for subsequent processing. Finally, the data-augmented samples are used to train the MLP model; the trained MLP model assigns labels to the samples in the testing set. The following sections describe the steps in our framework in detail.

Preprocessing

Since the AAindex in the iFeature Python package can only encode peptides of the same length, we need to preprocess the original peptide sequences to obtain peptides of the same lengths. To obtain the best sequence length, we need to know the length distribution of all samples. We performed statistical analyses of the length of the peptides in the ACP740 and ACP240 datasets. As shown in **Figure 2**, most of the peptides were less than 60 amino acids in length. To obtain peptides of the same length, we processed each peptide as follows. For sequences less than L_X amino acids, each peptide was padded with "X" until L_X amino acids were reached. For sequences greater than L_X amino acids, the extra amino acids after L_X were removed, and only the first L_X amino acids were reserved. L_X can be selected as 40, 50, or 60. We think the best length can be derived from the three numbers.

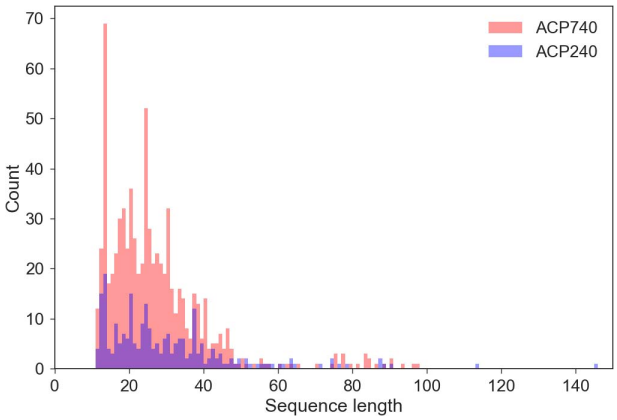


FIGURE 2 | Sequence length statistics for all peptides in the ACP740 and ACP240 dataset.

Representation of Peptides

The main objective of feature representation is to convert peptides of different lengths into fixed-length feature vectors (Zhang and Liu, 2019). The unprocessed peptide sequence P can be represented as:

$$P = p_1 p_2 ... p_L$$

where p_1 is the first residue of P and p_L is the last residue of P. L is the length of P. p_i $(1 \leq i \leq L)$ is an element of the standard amino acid alphabet {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. After preprocessing, the peptide sequence can be expressed as:

$$P = p_1 p_2 ... p_{L_X}$$

Different feature representation methods describe different properties of peptide sequences. If two features have complementary characteristics, combining the two features will help to improve the performance of the predictive model. However, more features don't necessarily lead to better performance. Too many features may cause redundancy, resulting in performance degradation. So, we tested three feature representation methods and their concatenations: BPFs, AAindex, and K-mer sparse matrix.

Binary Profile Features

There are 20 different amino acids in the standard amino acid alphabet. In BPFs, each amino acid is encoded by a feature vector composed of zeroes and ones. The first amino acid type A in the above amino acid alphabet is encoded as f(A) = (1, 0, ..., 0), the second amino acid type C is encoded as f(C) = (0, 1, ..., 0), and so on. For each given peptide, the N-terminus of k amino acids is encoded as the following feature vector:

$$F_{BPF} = [f(p_1), f(p_2), ..., f(p_k)]$$

Experiments derived from ACP-DL (Yi et al., 2019) show that the result is best when k is 7, which means that only the first 7 amino acids in each peptide sequence are encoded; therefore, the dimension of F_{BPF} is $20 \times 7 = 140$.

AAindex

The physicochemical properties of amino acids represent the characteristics of biochemical reactions and have been widely used in bioinformatics research. The AAindex is a database of amino acid indices representing the physicochemical and biochemical properties of amino acids (Kawashima et al., 2008). We used the AAindex in the iFeature Python package to represent peptides. The AAindex descriptor can only be used to encode peptides of equal length (Tung and Ho, 2008). The preprocessing steps in the previous section changed peptides of different lengths to peptides of equal length for encoding based on the AAindex. If L_X in the above section is set to 40, the AAindex descriptor calculated for a peptide of 40 amino acids will result in a 21,240-dimensional feature vector. The dimension of the features is too large, which may cause dimension disaster. mRMR is an excellent dimension reduction technology, and it has good application properties in many scenarios. Therefore, we selected the 50 most informative candidate features by using mRMR for more efficiency.

K-mer Sparse Matrix

K-mer of peptides is composed of K amino acids. Suppose the length of a peptide sequence is L, there will be 7^K different possible K-mer and an $L - K + 1$ step appearing in the sequence. One peptide sequence is transformed to a $7^K \times (L - K + 1)$ K-mer sparse matrix M, which is a low-rank matrix. Singular value decomposition (SVD) is used to convert this matrix into a 343-dimension feature vector.

A new feature vector is formed to represent peptides by concatenating the above feature representations of each peptide. If BPFs and AAindex are selected, the dimension of the new feature vector is 140 + 50 = 190.

Data Augmentation

When using machine learning technology to solve scientific problems, insufficient data (Han et al., 2020) or data imbalance (Fu et al., 2020; Gao et al., 2020; Mahmud et al., 2021) issues are common. Collecting more data can certainly solve these problems, but sometimes it may be difficult to obtain more data due to cost restrictions or other reasons. In such cases, data augmentation can potentially be efficient. Data augmentation has mainly historically been used in the field of computer vision (Chaitanya et al., 2021; Wang et al., 2021), and novel samples can be obtained by flipping, rotating, scaling, and cropping the original samples in the methods of data augmentation. In the field of bioinformatics, there will sometimes be data imbalances (Zou et al., 2016; Wan et al., 2017; Meng et al., 2019). Data augmentation can be used to solve data imbalance problems (Chen et al., 2020). Here, we are facing the insufficient sample problem, which can be solved by data augmentation. Four oversampling techniques are used to generate new samples in feature space to improve the performance of the RNA coding potential prediction model (Chen et al., 2020). Noise adding oversampling (NAO) is the best. We also use this technique to generate new samples.

To improve the performance of the ACP prediction model, we augmented the positive and negative samples in the datasets, respectively. Data augmentation is achieved by adding perturbation values to the original samples in the feature space to generate pseudosamples. The features of a peptide include two parts: BPFs and the AAindex. BPFs are binary codes composed of 0 and 1, which are not suitable for adding perturbations. So, we only add perturbations to the AAindex, and the BPFs remain unchanged. A new sample F_{new} is generated by the following mathematical formula:

$$\mathrm{F_{new}} = \mathrm{F_i} * \mathrm{V} * \mathrm{a} + \mathrm{F_i}$$

where F_i is a random sample from the training samples of peptide sequences and $i = 1, ..., N$. N is the total number of positive (negative) samples. V is a 190-dimensional vector for generating perturbations that corresponds to F_i. Because perturbations are not added to BPFs, V is composed of two parts: one is a 140-dimensional vector of zeros corresponding to BPFs, and the other is a 50-dimensional random vector with values between 0 and 1 corresponding to AAindex. Thus, perturbations are added to AAindex features, and BPFs are kept unchanged in the

pseudosample set F_{new}. a is the coefficient of the perturbation and is set to 0.005 for the ACP740 dataset. We repeat the sampling process N times to obtain N new samples.

Classifier

The MLP classifier is an artificial neural network composed of an input layer, a hidden layer, and an output layer. The hidden layer can be a single layer or multiple layers, and the layers are fully connected. A back propagation (BP) algorithm is used to train the MLP classifier (Rumelhart et al., 1986). Due to its excellent classification performance, this classifier has been used in many fields of bioinformatics, as noted in Auer et al. (2020). For implementation, we used the scikit-learn Python package; the hidden layer was composed of 6 sublayers, each with 100 neurons. The L2 penalty (regularization term) parameter was 0.01, and the other parameters were set to default values. We employed the MLP classifier to train our predictive model. In this work, we also tested other classifiers, including SVM (Fan et al., 2008), random forest (RF) (Breiman, 2001), MLP, decision tree (DT), and extremely randomized trees (ExtraTrees) (Geurts et al., 2006) classifiers, to build prediction models based on the augmented data in the training set. Among these classifiers, the MLP classifier works best according to the experiments section.

Performance Evaluation

We used five-fold cross-validation to evaluate the performance of ACP-DA. In the evaluation, five metrics were used in the binary classification tasks. The five metrics were accuracy (ACC), precision (PRE), sensitivity (SN), specificity (SP), and the Matthews correlation coefficient (MCC), which were widely used in bioinformatics (Zhang et al., 2019). These metrics are defined as follows:

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

$$PRE = \frac{TP}{TP + FP}$$

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP*TN - FP*TN}{\sqrt{(TP + FN)*(TP + FP)*(TN + FP)*(TN + FN)}}$$

where TP stands for true positives and FN, TN, and FP stand for false negatives, true negatives, and false positives, respectively. MCC is a comprehensive performance evaluation metric.

RESULTS

In this section, we first discuss the effects of two important parameters on the performance of our method and then compare the performance of the model for different features. We also analyze the effect of data augmentation in the case of using different classifiers. Finally, we compare our method with existing methods.

Parameter Discussion

Two parameters affect the performance of the model. One is L_X in the preprocessing step, which is the length of the peptide sequence after preprocessing. L_X can be set to 40, 50, or 60. The other parameter is N, which is related to the number of new positive (negative) samples in the data augmentation step. Here, we use the training samples after data augmentation to build the prediction model with 100, 200, or 300% of the original positive (negative) sample number as new samples. Thus, N may be set to 100, 200, or 300% of the original positive (negative) sample number.

For the ACP740 and ACP240 datasets, the performance of the prediction models established based on different parameters is shown in **Tables 1**, **2**. MCC is a comprehensive performance evaluation metric, and larger MCC values mean better performance. Therefore, we choose the best parameters L_X 40 and N 100% for ACP740 and L_X 40 and N 300% for ACP240 according to the maximum MCC value. The N value of ACP240 is larger than that of ACP740, which means that more pseudosamples are needed for ACP240 than ACP740 because ACP240 has fewer samples than ACP740.

Comparisons With Different Features

Binary profile feature and k-mer sparse matrix have been proved to be effective in ACP-DL (Yi et al., 2019), and AAindex has also been mentioned in physicochemical property based therapeutic

TABLE 1 | Performance of ACP-DA with different parameters based on ACP740 (The best metrics are in bold).

L_X	N	ACC (%)	PRE (%)	SE (%)	SP (%)	MCC (%)
40	100%	81.89	**84.14**	80.59	83.23	**64.71**
40	200%	82.02	83.46	80.89	**83.26**	64.56
40	300%	81.49	82.89	80.88	82.15	63.40
50	100%	80.41	83.35	79.02	81.88	62.59
50	200%	**82.03**	81.51	84.57	79.36	64.68
50	300%	80.27	77.23	**86.98**	73.35	61.17
60	100%	79.19	80.18	79.54	78.85	58.89
60	200%	78.37	77.72	81.67	75.01	57.21
60	300%	79.73	79.14	81.93	77.47	59.61

TABLE 2 | Performance of ACP-DA with different parameters based on ACP240 (The best metrics are in bold).

L_X	N	ACC (%)	PRE (%)	SE (%)	SP (%)	MCC (%)
40	100%	85.42	83.43	92.28	77.59	71.57
40	200%	87.92	87.17	91.48	83.91	76.03
40	300%	**88.33**	**90.11**	88.37	**88.30**	**76.68**
50	100%	85.00	84.71	88.43	81.11	70.10
50	200%	83.75	84.80	86.12	81.10	68.10
50	300%	85.42	86.48	86.86	83.83	71.03
60	100%	86.25	84.35	92.28	79.37	72.97
60	200%	87.08	86.89	90.74	83.04	74.64
60	300%	87.92	85.70	**93.78**	81.11	76.26

peptide predictor (PPTPP) (Zhang and Zou, 2020). BPF and AAindex were introduced in the previous subsection. The k-mer sparse matrix was proposed to represent protein sequences (You et al., 2016), and later used in the representation of peptide sequences (Yi et al., 2019). To obtain more effective features or feature combinations, we use the MLP classifier to build ACP prediction models and test the performance of each model based on three features and their pairwise concatenations without data augmentation.

The three features are BPFs, the AAindex, and the k-mer sparse matrix (k-mer). The concatenations of the three features are BPF + AAindex, BPF + k-mer, AAindex + k-mer and BPF + AAindex + k-mer. The performance of the models for different features and feature concatenations is shown in **Figure 3**. When the three features are used alone, BPF and AAindex yield the best performance. Among the four feature concatenations, BPF + AAindex yields the best performance for ACP240 and BPF + AAindex + k-mer yields the best performance for ACP740. The performance of BPF + AAindex + k-mer on ACP240 is even worse than that of BPF alone. On the basis of comprehensive consideration of various factors, we chose the concatenation of BPF + AAindex to represent the peptide sequence.

Classifier Discussion

After determining that the concatenation of BPF + AAindex should be used to represent peptides, we need to consider which classifier is the best in our method. We analyzed the performance of the prediction model with data augmentation on several different classifiers. We considered five different classifiers, namely, SVM, RF, MLP, ExtraTrees, and DT classifiers, to build the prediction models. Since MCC is a comprehensive metric, we used it to evaluate the performance of the prediction models. The performance of the models on ACP740 and ACP240 is shown in **Figure 4**.

As shown in **Figure 4A**, based on the ACP740 dataset, for the prediction models built using the MLP, RF, SVM, and ExtraTrees classifiers, data augmentation can improve the prediction performance according to the MCC value. However, data augmentation will cause the performance to decrease for the prediction model established based on the DT. As shown in **Figure 4B**, for the ACP240 dataset, data augmentation can improve the performance of the prediction models established based on the MLP, SVM, and ExtraTrees classifiers and cause the prediction performance of the models based on the RF and DT classifiers to decrease. Therefore, when using the MLP, SVM, and ExtraTrees classifiers, data augmentation is helpful for improving the performance of the ACP prediction model. These results show that the effectiveness of data augmentation is related to the choice of classifier. RF and DT classifiers are not suitable for our prediction model.

As deep learning technology has the advantages of strong learning ability and good portability, it has outstanding performance in various fields in recent years. Combined with the MCC value of the two datasets, we chose the MLP classifier to build the final predictive model.

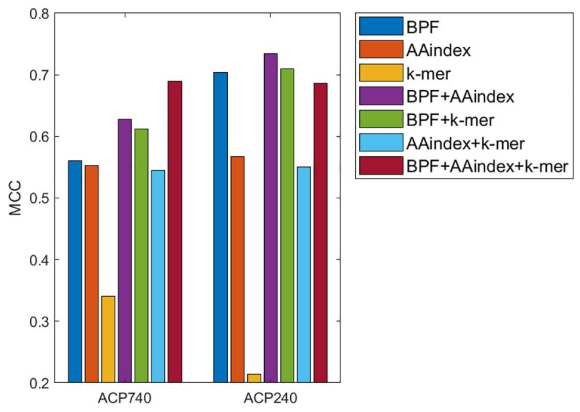


FIGURE 3 | Comparison of prediction models using BPFs, the AAindex, the k-mer sparse matrix (k-mer), and their concatenations based on ACP740 and ACP240.

Comparison With Existing Methods

To verify the effectiveness of our proposed method, we compared our method ACP-DA with ACP-DL (Yi et al., 2019), AntiCP 2.0 (Agrawal et al., 2020), and DeepACP (Yu et al., 2020). The results on ACP740 and ACP240 are shown in **Figure 5**.

Compared with ACP-DL, the advantage of our method lies in the use of data augmentation. In addition, our method used the AAindex feature instead of the k-mer sparse matrix in ACP-DL. Our method with data augmentation outperforms ACP-DL in most metrics, especially on the two metrics of most importance MCC and ACC.

As shown in **Figure 5A**, the performance of our method on ACP740 was better than that of ACP-DL and AntiCP2.0 and worse than that of DeepACP according to the MCC value and ACC value. **Figure 5B** shows that our method performed better than other methods on ACP240. The number of samples on ACP240 was less than that on ACP740. Our method performed better on ACP240, which indicated that our method was more suitable for the case of insufficient samples.

DISCUSSION

As a complex disease, cancer involves complex biological processes. The complex mechanisms of cancer make it difficult to trace the cause. Despite the emergence of various cancer treatment strategies, most of the strategies have been unsatisfactory. Due to its high specificity, high tissue penetration, low production cost and other advantages, treatment based on ACPs has become a potential cancer treatment method. Most ACPs come from protein sequences. The development of high-throughput sequencing technology has brought an increase in the number of available proteins, and it is expected that the number of ACPs will also increase. It is time-consuming and expensive to use experimental methods to discover ACPs from protein sequence data. Therefore, it is urgent to develop computational methods to speed up the identification of ACPs.

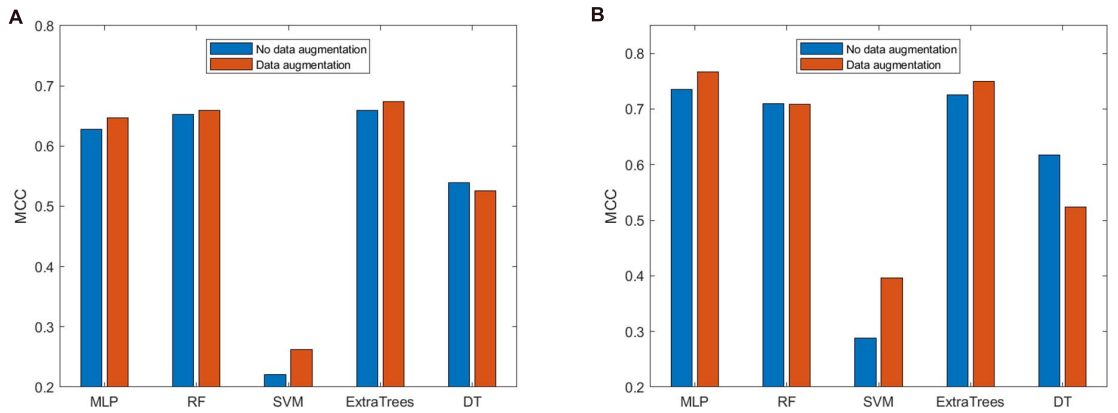


FIGURE 4 | Comparison of the prediction models with and without data augmentation based on **(A)** ACP740 and **(B)** ACP240.

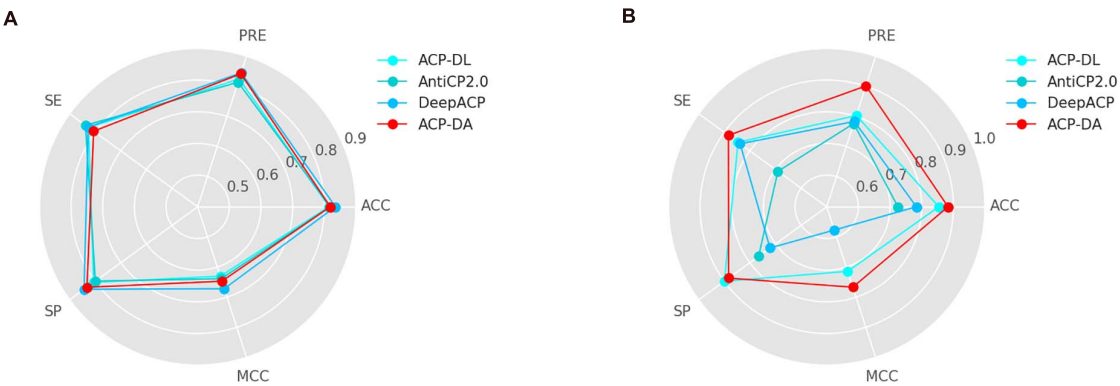


FIGURE 5 | Comparison of ACP-DA with existing methods on **(A)** ACP740 and **(B)** ACP240.

In this paper, an ACP prediction method called ACP-DA is proposed. According to the results on the two datasets of ACP740 and ACP240, our model has good overall performance. Compared with existing methods, our method has a better effect in identifying whether the peptide sequence is ACP, and its accuracy may be attributed to the following reasons.

First, how to use effective feature representation methods to characterize peptide sequences is a major challenge in current prediction methods. To find an effective feature or feature combination, we tested 3 feature representation methods and their feature combinations: BPF, AAindex, k-mer, BPF + AAindex, BPF + k-mer, AAindex + k-mer, BPF + AAindex + k-mer. Experiments on the ACP740 and ACP240 datasets show that BPF + AAindex obtains the best performance, so we use BPF + AAindex to represent the peptide sequences.

Second, we used data augmentation to increase the samples in the training set for the insufficient samples. Data augmentation is achieved by generating pseudosamples based on the original samples. The specific method of generating pseudosamples is to add disturbances in the feature space of the original sample. The feature space of the sample is formed by the concatenation of BPF and AAindex. BPF is a code composed of 0 and 1, which is not suitable for adding disturbance, so, we only add disturbance on AAindex to generate pseudosamples. The model is trained with the augmented data to further improve the performance of the prediction model.

Finally, various classifiers show good performance in many classification tasks of bioinformatics. However, it is still unknown whether our data augmentation method can improve the performance of prediction models under various types of classifiers. Therefore, we tested the effect of this method in the case of using five different classifiers. The results show that data augmentation is effective when using MLP, SVM, and ExtraTrees, and data augmentation may not be effective when using RF or DT. Therefore, we choose the MLP with the best overall performance as the final classifier.

The main innovation of this article lies in the use of data augmentation methods. From the experimental results, the method is of great significance. When using MLP, SVM, and ExtraTrees as classifiers, the use of data augmentation can significantly improve the performance of the prediction model. Moreover, a comparative analysis with other methods shows that ACP-DA is better than other methods in most cases.

In short, we provide a new idea for the identification of ACPs, and hope that ACP-DA will play an important role in the development of new anticancer drugs.

CONCLUSION

In this work, we proposed a novel ACP prediction model called ACP-DA. To establish an effective prediction model, we concatenated BPFs and the AAindex to represent peptide sequences. We performed data augmentation in the feature space and used the augmented data to train the prediction model. The experimental results show that the proposed method can effectively distinguish ACPs and non-ACPs. Compared with the method without data augmentation, ACP-DA achieves better performance. ACP-DA will be a useful tool for the discovery of novel potential ACPs.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/haichengyi/ACP-DL.

AUTHOR CONTRIBUTIONS

X-GC conceived the algorithm, performed the experiments, analyzed the data, and drafted the manuscript. WZ designed the experiments and revised the manuscript. XY, CL, and HC provided suggestions for the study design and the writing of the manuscript. All authors approved the final manuscript.

FUNDING

This work was funded by the National Natural Science Foundation of China (62072206, 61772381, 61976227, and 31870771), the National Science Foundation of Hubei (2020CFA025), the "Fundamental Research Funds for the Central Universities", South-Central University for Nationalities (Grant Number CZY20039), and Huazhong Agricultural University Scientific & Technological Self-innovation Foundation.

REFERENCES

Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N., and Raghava, G. P. S. (2020). AntiCP 2.0: an updated model for predicting anticancer peptides. *Brief Bioinform.* 22:bbaa153. doi: 10.1093/bib/bbaa153

Amanat, S., Ashraf, A., Hussain, W., Rasool, N., and Khan, Y. D. (2020). Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PseAAC. *Curr. Bioinform.* 15, 396–407. doi: 10.2174/1574893614666190723114923

Auer, A., Strauss, M. T., Strauss, S., and Jungmann, R. (2020). nanoTRON: a Picasso module for MLP-based classification of super-resolution data. *Bioinformatics* 36, 3620–3622. doi: 10.1093/bioinformatics/btaa154

Barras, D., and Widmann, C. (2011). Promises of apoptosis-inducing peptides in cancer therapeutics. *Curr. Pharm. Biotechnol.* 12, 1153–1165. doi: 10.2174/138920111796117337

Boohaker, R. J., Lee, M. W., Vishnubhotla, P., Perez, J. M., and Khaled, A. R. (2012). The use of therapeutic peptides to target and to kill cancer cells. *Curr. Med. Chem.* 19, 3794–3804. doi: 10.2174/092986712801661004

Boopathi, V., Subramaniyam, S., Malik, A., Lee, G., Manavalan, B., and Yang, D. C. (2019). mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.* 20:1964. doi: 10.3390/ijms20081964

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

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

Chaitanya, K., Karani, N., Baumgartner, C. F., Erdil, E., Becker, A., Donati, O., et al. (2021). Semi-supervised task-driven data augmentation for medical image segmentation. *Med. Image Anal.* 68:101934. doi: 10.1016/j.media.2020.101934

Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K. C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 7, 16895–16909. doi: 10.18632/oncotarget.7815

Chen, X. G., Liu, S., and Zhang, W. (2020). Predicting coding potential of RNA sequences by solving local data imbalance. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2020.3021800 [Epub ahead of print].

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

Deng, Y., Xu, X., Qiu, Y., Xia, J., Zhang, W., and Liu, S. (2020). A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics* 36, 4316–4322. doi: 10.1093/bioinformatics/btaa501

Doja, M. N., Kaur, I., and Ahmad, T. (2020). Current state of the art for survival prediction in cancer using data mining techniques. *Curr. Bioinform.* 15, 174–186. doi: 10.2174/1574893614666190902152142

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., and Lin, C. J. (2008). LIBLINEAR: a library for large linear classification. *JMLR* 9, 1871–1874.

Fu, G. H., Wu, Y. J., Zong, M. J., and Pan, J. (2020). Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data. *BMC Bioinform.* 21:121. doi: 10.1186/s12859-020-3411-3

Gao, Y., Chuai, G., Yu, W., Qu, S., and Liu, Q. (2020). Data imbalance in CRISPR off-target prediction. *Brief Bioinform.* 21, 1448–1454. doi: 10.1093/bib/bbz069

Ge, L., Liu, J., Zhang, Y., and Dehmer, M. (2019). Identifying anticancer peptides by using a generalized chaos game representation. *J. Math. Biol.* 78, 441–463. doi: 10.1007/s00285-018-1279-x

Ge, R., Feng, G., Jing, X., Zhang, R., Wang, P., and Wu, Q. (2020). EnACP: an ensemble learning model for identification of anticancer peptides. *Front. Genet.* 11:760. doi: 10.3389/fgene.2020.00760

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learn.* 63, 3–42. doi: 10.1007/s10994-006-6226-1

Hajisharifi, Z., Piryaiee, M., Mohammad Beigi, M., Behbahani, M., and Mohabatkar, H. (2014). Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* 341, 34–40. doi: 10.1016/j.jtbi.2013.08.037

Han, C., Xie, Q., and Lin, S. (2020). Are dropout imputation methods for scRNA-seq effective for scHi-C data? *Brief Bioinform.* doi: 10.1093/bib/bbaa289 [Epub ahead of print].

Hasan, M. A. M., Ben Islam, M. K., Rahman, J., and Ahmad, S. (2020). Citrullination site prediction by incorporating sequence coupled effects into PseAAC and resolving data imbalance issue. *Curr. Bioinform.* 15, 235–245. doi: 10.2174/1574893614666191202152328

Holohan, C., Van Schaeybroeck, S., Longley, D. B., and Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer* 13, 714–726. doi: 10.1038/nrc3599

Huang, F., Yue, X., Xiong, Z., Yu, Z., Liu, S., and Zhang, W. (2020). Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Brief Bioinform.* 22:bbaa140. doi: 10.1093/bib/bbaa140

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205. doi: 10.1093/nar/gkm998

Li, F. M., and Wang, X. Q. (2016). Identifying anticancer peptides by using improved hybrid compositions. *Sci. Rep.* 6:33910. doi: 10.1038/srep33910

Mahmud, S. M. H., Chen, W., Liu, Y., Awal, M. A., Ahmed, K., Rahman, M. H., et al. (2021). PreDTIs: prediction of drug-target interactions based on multiple feature information using gradient boosting framework with data balancing and

feature selection techniques. *Brief Bioinform.* doi: 10.1093/bib/bbab046 [Epub ahead of print].

Meng, C., Wei, L., and Zou, Q. (2019). SecProMTB: support vector machine—based classifier for secretory proteins using imbalanced data sets applied to *Mycobacterium tuberculosis*. *Proteomics* 19:e1900007. doi: 10.1002/pmic.201900007

Mitchell, E. P. (2006). Gastrointestinal toxicity of chemotherapeutic agents. *Semin. Oncol.* 33, 106–120. doi: 10.1053/j.seminoncol.2005.12.001

Naseer, S., Hussain, W., Khan, Y. D., and Rasool, N. (2020). Sequence-based identification of arginine amidation sites in proteins using deep representations of proteins and PseAAC. *Curr. Bioinform.* 15, 937–948. doi: 10.2174/1574893615666200129110450

Otvos, L. Jr. (2008). Peptide-based drug design: here and now. *Methods Mol. Biol.* 494, 1–8. doi: 10.1007/978-1-59745-419-3_1

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

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Tung, C. W., and Ho, S. Y. (2008). Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinform.* 9:310. doi: 10.1186/1471-2105-9-310

Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., and Raghava, G. P. (2013). In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.* 3:2984. doi: 10.1038/srep02984

Wan, S., Duan, Y., and Zou, Q. (2017). HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* 17:1700262. doi: 10.1002/pmic.201700262

Wang, Y., Huang, G., Song, S., Pan, X., Xia, Y., and Wu, C. (2021). Regularizing deep networks with semantic data augmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* doi: 10.1109/TPAMI.2021.3052951 [Epub ahead of print].

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Wijdeven, R. H., Pang, B., Assaraf, Y. G., and Neefjes, J. (2016). Old drugs, novel ways out: drug resistance toward cytotoxic chemotherapeutics. *Drug Resist Updat.* 28, 65–81. doi: 10.1016/j.drup.2016.07.001

Xu, L., Liang, G., Wang, L., and Liao, C. (2018). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes (Basel)* 9:158. doi: 10.3390/genes9030158

Yi, H. C., You, Z. H., Zhou, X., Cheng, L., Li, X., Jiang, T. H., et al. (2019). ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Mol. Ther. Nucleic Acids* 17, 1–9. doi: 10.1016/j.omtn.2019.04.025

You, Z.-H., Zhou, M. C., Luo, X., and Li, S. (2016). Highly efficient framework for predicting interactions between proteins. *IEEE Trans. Cybernet.* 47, 731–743. doi: 10.1109/tcyb.2016.2524994

Yu, L., Jing, R., Liu, F., Luo, J., and Li, Y. (2020). DeepACP: a novel computational approach for accurate identification of anticancer peptides by deep learning algorithm. *Mol. Ther. Nucleic Acids* 22, 862–870. doi: 10.1016/j.omtn.2020.10.005

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

Zhang, L., He, Y., Song, H., Wang, X., Lu, N., Sun, L., et al. (2020). Elastic net regularized softmax regression methods for multi-subtype classification in cancer. *Curr. Bioinform.* 15, 212–224. doi: 10.2174/1574893613666181112141724

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019). A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 405–415. doi: 10.1109/TCBB.2019.2931546

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, 3982–3987. doi: 10.1093/bioinformatics/btaa275

Zhao, T., Hu, Y., and Zang, T. (2020). DRACP: a novel method for identification of anticancer peptides. *BMC Bioinform.* 21(Suppl. 16):559. doi: 10.1186/s12859-020-03812-y

Zou, Q. (2019). Latest machine learning techniques for biomedicine and bioinformatics. *Curr. Bioinform.* 14, 176–177.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* 9:515.

Zou, Q., Xie, S., Lin, Z., Wu, M., and Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. *Big Data Res.* 5, 2–8. doi: 10.1016/j.bdr.2015.12.001

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.

Copyright © 2021 Chen, Zhang, Yang, Li and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 July 2021
doi: 10.3389/fgene.2021.699280

A Large-Scale and Serverless Computational Approach for Improving Quality of NGS Data Supporting Big Multi-Omics Data Analyses

Dariusz Mrozek[1]*, *Krzysztof Stępień*[1], *Piotr Grzesik*[1] and *Bożena Małysiak-Mrozek*[2]

[1] Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland, [2] Department of Graphics, Computer Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland

Various types of analyses performed over multi-omics data are driven today by next-generation sequencing (NGS) techniques that produce large volumes of DNA/RNA sequences. Although many tools allow for parallel processing of NGS data in a Big Data distributed environment, they do not facilitate the improvement of the quality of NGS data for a large scale in a simple declarative manner. Meanwhile, large sequencing projects and routine DNA/RNA sequencing associated with molecular profiling of diseases for personalized treatment require both good quality data and appropriate infrastructure for efficient storing and processing of the data. To solve the problems, we adapt the concept of Data Lake for storing and processing big NGS data. We also propose a dedicated library that allows cleaning the DNA/RNA sequences obtained with single-read and paired-end sequencing techniques. To accommodate the growth of NGS data, our solution is largely scalable on the Cloud and may rapidly and flexibly adjust to the amount of data that should be processed. Moreover, to simplify the utilization of the data cleaning methods and implementation of other phases of data analysis workflows, our library extends the declarative U-SQL query language providing a set of capabilities for data extraction, processing, and storing. The results of our experiments prove that the whole solution supports requirements for ample storage and highly parallel, scalable processing that accompanies NGS-based multi-omics data analyses.

Keywords: next-generation sequencing, data quality, cloud computing, big data, data lake, OMICS data, serverless, querying

OPEN ACCESS

Edited by:
Leyi Wei,
Shandong University, China

Reviewed by:
Renhai Chen,
Tianjin University, China
Che-Lun Hung,
National Yang Ming Chiao Tung University, Taiwan

***Correspondence:**
Dariusz Mrozek
dariusz.mrozek@polsl.pl

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

Received: *23 April 2021*
Accepted: *28 May 2021*
Published: *13 July 2021*

Citation:
Mrozek D, Stępień K, Grzesik P and Małysiak-Mrozek B (2021) A Large-Scale and Serverless Computational Approach for Improving Quality of NGS Data Supporting Big Multi-Omics Data Analyses. Front. Genet. 12:699280. doi: 10.3389/fgene.2021.699280

1. INTRODUCTION

Several commercially available sequencing platforms on the market today allow thousands or even millions of DNA/mRNA sequence fragments (sequence reads) to be obtained simultaneously. Raw data obtained once the sequencing is complete include a set of many short genome sequence reads that usually undergo several phases of data analysis. The NGS data pre-processing scheme preceding a secondary data analysis should include sequence quality control and data processing phase, covering the removal of low-quality sequences and bases, demultiplexing, removal of adapters, primers, and contamination, error correction, and detection of enrichment biases. Each

nucleotide in the DNA/mRNA read is accompanied by information about the probability of its misidentification. This probability directly determines the *phred* quality score, which is given for the DNA sequence reads in FASTQ files. The quality score Q for a base-call is a logarithmic measure depending on the probability P of incorrect nucleotide identification (Ewing and Green, 1998):

$$Q = -10 \times \log_{10} P. \quad (1)$$

High values of the quality score correspond to low probabilities of misidentification errors, and conversely. Low-quality bases are often located at the very beginning of the sequence. The probability of misidentifying nucleotides also increases with the position in the read. In addition, raw data may be contaminated with fragments of the adapter sequences that do not belong to the sequenced material. Therefore, the quality improvement of NGS sequence reads is vital for further analysis of genomic data analyses since the presence of poor quality or technical sequences may degrade the results of the analyses.

At the same time, as a high-throughput technology, NGS sequencing generates vast amounts of biomedical data. This raises challenges of Big Data (Mrozek, 2014, 2018) not only due to the volume of data generated, but also due to the *velocity* (i.e., speed) in which the data is produced in various projects, and the *variety* of formats in which the data is delivered. These three V characteristics (i.e., volume, velocity, variety), typical for Big Data problems, largely influence the *value* that can be retrieved from the data. The implementation of even small projects that require data from the NGS sequencing of multiple genomes can pose many problems related to the infrastructure necessary to perform the task. The infrastructure must provide the needed storage space and computing power to process large amounts of information efficiently. Therefore, highly distributed and scalable environments are recently used to solve the challenges of NGS Big Data processing and NGS-based analyses performed at various steps of analysis workflows, from primary to tertiary.

These environments rely on a broad Hadoop ecosystem and its tools. For example, SeqPig (Schumacher et al., 2013) as a dedicated library for distributed analysis and processing of large NGS sequencing data on Hadoop clusters extends the processing capabilities of Apache Pig and the Pig Latin scripting language. Apart from processing files in FASTA and FASTQ formats, the library enables the assessment of the quality of sequences. Several Hadoop-based solutions were proposed for the secondary NGS data analysis steps, including initial alignment of short reads to a reference genome with BigBWA tool (implementing the Burrows-Wheeler Aligner (Abuín et al., 2015), tag SNPs selection (Hung et al., 2015), and construction of phylogenetic trees based on ultra-large DNA sequences (Zou, 2016; Zou et al., 2016). Within the tertiary analysis of NGS data, the GenoMetric Query Language (GMQL) (Masseroli et al., 2015, 2018) simplifies the variant analysis in genomic data stored in Hadoop Distributed File System (HDFS) with a declarative query language, distributed processing, and integration of heterogeneous biomedical data sources (Masseroli et al., 2016). Furthermore, Wiewiórka et al. (2019) proposed a library for scalable depth of coverage calculations over genomic data on Apache Spark. These solutions prove that distributed processing can solve the problems of voluminous and quickly produced data.

On the other hand, the *variety* of data, which next to the *volume* is one of the challenges affecting the NGS data obtained in several formats after particular phases of data production, processing, and analysis, causes the need for efficient and scalable data storage. Big Data lakes that allow storing the data before and after data analyses in the native formats facilitate gathering all the data in one place. However, processing the data must be accompanied by specific steps of data extraction. We first introduced the Extract, Process, and Store (EPS) process in Małysiak-Mrozek et al. (2018) for processing biomedical data with the use of fuzzy techniques. It clearly exposed the extraction and storing phases that can also be parallelized while processing big data in a distributed manner.

We adopt this idea in the NGS data processing performed in this paper to improve processing performance for large amounts of NGS data while at the same time reducing the operational overhead by taking advantage of the serverless nature of the Data Lake Analytics service. However, we also show limitations of the used data lake platform and the EPS while processing NGS data.

1.1. Related Works

The growing body of research shows that the quality of NGS data is important for future NGS-based multi-omics data analyses. There are many approaches and tools dedicated to processing and cleaning the DNA/RNA sequences obtained with single and paired-end sequencing techniques in the literature. First of them, Trimmomatic, introduced by Bolger et al. (2014), is a tool dedicated to trimming and filtering next-generation sequencing reads, supporting both single and paired-end reads. For trimming, it offers two algorithms, one called "simple," which tries to find an approximate match between provided adapter sequence and read, and the second, called "palindrome mode," which is dedicated to detecting contaminants at the end of the reads. It also offers to filter sequences based on Illumina quality score. According to performance experiments presented in the paper, it is faster than comparable tools such as AdapterRemoval, Reaper, or Cutadapt. Schubert et al. (2016) propose AdapterRemoval v2, an improvement to an AdapterRemoval introduced previously in Lindgreen (2012). It is a tool that allows for the trimming of adapter sequences from both single-end and paired-end FASTQ reads. It takes advantage of a modified Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). Additionally, it also allows for the merging of overlapping paired-ended reads into consensus sequences. According to the performance experiments presented in the paper, it offers performance comparable to Trimmomatic. Another tool that takes advantage of the Needleman-Wunsch algorithm has been introduced by Roehr et al. (2017). The authors present FLEXBAR 3.0, an improvement to previously introduced FLEXBAR (Dodt et al., 2012), which is a sequence trimming software dedicated to processing NGS reads and trimming barcode and adapter sequences. It supports five trimming modes, LEFT, LEFT-TAIL, RIGHT, RIGHT-TAIL, and ANY. In version 3.0, it introduced multi-threading and SIMD

vectorization to improve performance over previous versions. According to benchmarks presented by the authors, it offers better trimming quality than Trimmomatic but takes two times longer to process the same number of reads. Criscuolo and Brisse (2013) introduced AlienTrimmer, a tool dedicated to the removal of alien sequences such as primer, adapters, or barcode sequences from raw next-generation sequencing data. The tool supports the removal of such sequences from both 5′ and 3′ ends. It uses an algorithm based on the k-mer decomposition of specified alien sequences and then tries to find occurrences of such k-mers in the sequence. The authors highlight that k-mer decomposition-based algorithms, such as the one used in AlienTrimmer, are prone to decreased accuracy in case of sequencing errors and when handling short fragments of alien sequences. Another tool dedicated to trimming adapters and low-quality bases in next-generation sequencing data, Btrim, has been introduced by Kong in Kong (2011). When performing adapter trimming, the tool uses an algorithm that is based on a modified version of Myers' bit-vector dynamic programming algorithm. When performing trimming of bases with low quality, it switches to a moving window algorithm that trims bases if the average quality score is lower than the predefined threshold. It supports data in FASTQ for both Sanger and Illumina reads. Smeds and Künstner (2011) proposed ConDeTri, a content-dependent trimming solution dedicated to processing and trimming Illumina reads. It supports the removal of sequencing errors from the 3' end as well as the removal of reads with low-quality bases. The algorithm allows keeping low-quality bases (below the threshold) if they are surrounded by high-quality bases. Martin, in his work (Martin, 2011), introduces Cutadapt, which is another tool dedicated to removing adapter sequences from Illumina reads. Cutadapt trims at most one adapter sequence in a single run and does not offer other trimming capabilities. According to the benchmarks presented in Schubert et al. (2016), it offers slower performance than AdapterRemoval and Trimmomatic. Unlike Cutadapt, PEAT (Paired-End Adapter Trimmer) (Li et al., 2015) does not require providing adapter sequence but instead detects adapter sequence by finding mutually reverse-complement region between paired reads. It is also not capable of processing barcode sequences on 5' ends, does not take the read quality scores into account, but for benchmarked datasets, it offered much better performance in terms of speed than tools such as AdapterRemoval and Trimmomatic. For trimming paired-end NGS reads, Skewer (Jiang et al., 2014) adapter trimmer offers better memory efficiency while being slower than solutions like Trimmomatic and Btrim.

In addition to tools that are dedicated mostly to trimming adapter sequences, there are also toolkits, like Kraken (Davis et al., 2013), FASTX-Toolkit (Gordon, 2008), or ERNE (Del Fabbro et al., 2013), that allow building advanced pipelines for analyzing NGS data, where filtering and adapter trimming is only one of the steps. In terms of the declarative nature of the adapter trimming, Fuzzysplit, a flexible fuzzy search library (Liu, 2019) provides a pattern language that can be used to define adapter patterns that should be detected in target sequences. However, it does not support any other matching algorithms and does not consider quality scores from FASTQ formats. It offers great flexibility at the cost of a steep learning curve and the requirement to write custom templates for each supported format.

In terms of addressing Big Data challenges, Expósito et al. (2020) proposed SeQual for large-scale processing of NGS reads on Apache Spark. It implements filtering, trimming, and formatting procedures, operates on FASTQ and FASTA data formats, and offers a user-friendly graphical user interface. However, it requires access to the Spark computational cluster.

While there are also other local tools dedicated to trimming NGS data, such as ea-utils (Aronesty, 2011, 2013), PRINSEQ (Schmieder and Edwards, 2011), SeqPurge (Sturm et al., 2016), PE-Trimmer (Liao et al., 2020), StreamingTrim (Bacci et al., 2014), AfterQC (Chen et al., 2017), ClinQC (Pandey et al., 2016), UrQT (Modolo and Lerat, 2015), pTrimmer (Zhang et al., 2019), Fastq_clean (Zhang et al., 2014), they are often designed as separate programs instead of libraries and only one of them, Fuzzysplit, offers declarative interface, but has limited functionality. They are often also not designed for Big Data processing that takes advantage of Cloud Computing technologies, except SeQual, which is built on top of the Apache Spark framework. The downside of SeQual is that the underlying Apache Spark cluster has to be provided and managed, which adds operational complexity and requires knowledge about managing the computing cluster itself.

1.2. Scope of the Work

It is worth noting that most of the works mentioned in previous sections do not focus on improving the quality of NGS data at a large scale. Moreover, only one of them provides declarative querying capabilities for this purpose but with limited NGS data quality improvement capabilities. Our solution hybridizes different technological approaches, which finally leads to possessing three fundamental properties—it is declarative, addresses challenges of Big NGS Data, and is scalable on the Cloud. Moreover, unlike SeQual, it does not require complex management of the computational cluster.

To solve the problems of Big NGS Data, in this paper, we present the scalable solution that utilizes the Data Lake ecosystem and serverless computing on the Microsoft Azure platform, enabling NGS data cleaning in the Cloud. Furthermore, we show how we can use the Data Lake ecosystem to build an environment for distributed storing and analyzing NGS data. This will be demonstrated by implementing solutions designed to control and improve the quality of reads from raw data. The results of our experiments show that the storage method and the degree of parallelism have the most significant impact on the time necessary to pre-process the sequence in terms of their quality improvement and thus on the costs of using the Cloud platform incurred.

2. MATERIALS AND METHODS

The approach we propose for big NGS data cleaning assumes storing the genomic data in NGS data lake in the Azure Data Lake Store in Microsoft Azure cloud and performing serverless but highly scalable processing of the data by formulating processing

queries in the declarative U-SQL language. The data lake is the place where data can be stored in its original format, including structured, semi-structured, and unstructured data. This allows applying the *schema on read* approach while processing the same data for various purposes. In contrast to the *schema on write* approach widely used in transactional systems, the *schema on read* approach schematizes the data when it is needed. Furthermore, the U-SQL language combines the declarative nature of the SQL language with imperative capabilities of C# programming language to process data in a scalable manner, which fits the scenario of big NGS data processing. Finally, serverless computing allows skipping the management of the servers responsible for data processing and frees the user from keeping the running servers all the time (which usually increases costs). By applying the serverless approach, we rely on the computing resources that are allocated by the cloud provider only when we need to execute the processing jobs.

In our approach, we process big NGS data stored in NGS Data Lake in three phases—Extract, Process, and Store (EPS)—as it is shown in **Figure 1**. Particular phases of the EPS allow for the following:

1. Extract—uses various *extractors* to extract appropriate data stored in the data lake, read it, and load the data for further processing,
2. Process—applies developed *processors* for NGS data to perform a set of transformations on the extracted NGS data set; these transformations cover the process of improving the quality of data,
3. Store—uses various *outputters* to store the processed data back in the NGS data lake.

2.1. NGS Data Extraction

Data extraction allows reading data from the specified files in the data lake. General workflow for data extraction from a single NGS data file is presented in Algorithm 1. Standard files (e.g., in FASTQ format) are extracted as a whole (by a single computational unit, called Allocation Unit or AU)—lines 8–12. Large files in the row-oriented format (see later in this section) are additionally split into smaller chunks and extracted in parallel. In both cases, for each DNA sequence read r_j in the file or chunk, the extractor E extracts the data appropriately (depending on the format) and represents it in the row-oriented format. The sequence read in a row-oriented format r_j^T is added to the data chunk c^* (a resultant rowset, line 5 and 10).

The symbol T (line 10) denotes transposition, and we use it when the NGS data is extracted from FASTQ files, where each DNA sequence read r_j is represented by a quadruple:

$$r_j = \begin{bmatrix} d_1 \\ s \\ d_2 \\ q \end{bmatrix}, \tag{2}$$

where d_1 contains sequence identifier and an optional description, s is a raw sequence, d_2 is a separator line beginning with a plus (+) sign with an optional description, q contains encoded quality scores for base calls in the sequence s.

Algorithm 1: NGS data extraction for a single file located in big NGS data lake.

Input :
f : a large file of NGS data to extract;
E : an extractor;
Output: c^*: a rowset with extracted data in a row-oriented format

```
1  Extract(f, E)
2  if filetype = row-o then
3  |  C ← SplitIntoChunks(f);
4  |  parallel foreach c_i ⊂ C do
5  |  |  c* ← c* ∪ E(r_j);
6  |  end
7  end
8  else
9  |  foreach r_j ∈ f do
10 |  |  c* ← c* ∪ E(r_j)^T;
11 |  end
12 end
13 return c*;
```

Files in the FASTQ format contain many of such reads and can be represented as:

$$f_{FASTQ} = \begin{bmatrix} r_1 \\ r_2 \\ \dots \\ r_{|f|} \end{bmatrix}, \tag{3}$$

where $|f|$ denotes the number of sequence reads in a file. The extraction process is a function that temporary changes the format of the data to the row-oriented one:

$$\begin{bmatrix} r_1 \\ r_2 \\ \dots \\ r_{|f|} \end{bmatrix} \xrightarrow{E} \begin{bmatrix} r_1^T \\ r_2^T \\ \dots \\ r_{|f|}^T \end{bmatrix}. \tag{4}$$

For the paired-end sequencing, we operate on two files with forward (left) and reverse (right) sequence reads

$$f^f_{FASTQ} = \begin{bmatrix} r^f_1 \\ r^f_2 \\ \dots \\ r^f_{|f|} \end{bmatrix} \text{ and } f^r_{FASTQ} = \begin{bmatrix} r^r_1 \\ r^r_2 \\ \dots \\ r^r_{|f|} \end{bmatrix}, \tag{5}$$

where r_j^f and r_j^r are corresponding forward (left) and reverse (right) sequence reads. Therefore, the extraction process provides an appropriate row-oriented representation for them that looks as follows:

$$f^{paired}_{ROW-O} = \begin{bmatrix} r^f_1 & r^r_1 \\ r^f_2 & r^r_2 \\ \dots & \dots \\ r^f_{|f|} & r^r_{|f|} \end{bmatrix}. \tag{6}$$

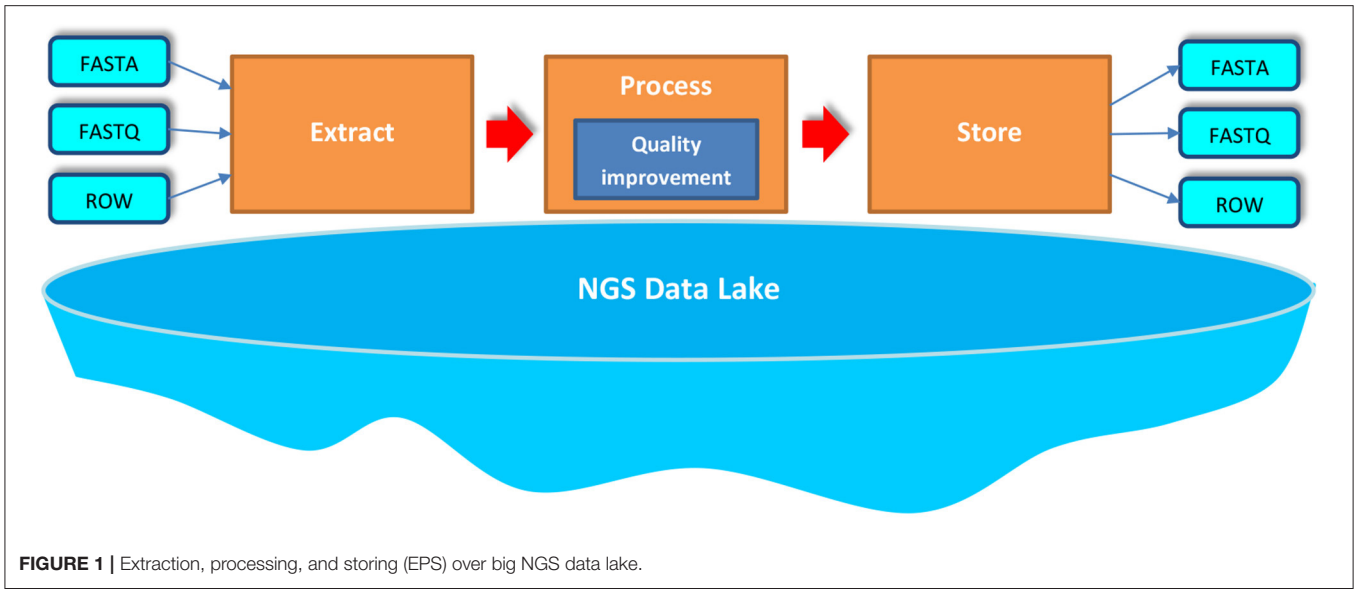


FIGURE 1 | Extraction, processing, and storing (EPS) over big NGS data lake.

If there are many files with independent genomic data to be extracted or in case a large genomic data file is divided into smaller files (e.g., intentionally), the extraction process can be further parallelized on many Allocation Units (Algorithm 2, line 1). Each file f_l in a collection of files F undergoes the same extraction steps (line 2) as in Algorithm 1. This produces many rowsets c_l in the row-oriented format that are either independent partitions of data for multiple genomic data files or are merged in a single rowset, when operating on many smaller files for one sequencing experiment (line 3). The collection of rowsets or a merged rowset C is then returned for further processing in the Process phase of the EPS (line 5).

Algorithm 2: Parallelization of NGS data extraction from files located in big NGS data lake.

Input : F: a set of files to extract;
E: an extractor;
Output: C: a collection of rowsets (if processing many independent files) or a merged rowset (if processing many smaller files for one genomic experiment) with extracted data in a row-oriented format.

1 **parallel foreach** $f_l^{in} \in F$ **do**
2 $\quad c_l \leftarrow \text{Extract}(f_l^{in}, E)$;
3 $\quad C \leftarrow C \cup c_l$;
4 **end**
5 **return** C;

Reading data from files located in the NGS data lake is implemented in the EXTRACT expression of the U-SQL language. The EXTRACT expression consists of a list of attributes extracted, a FROM clause followed by a file path, and a USING clause followed by an instance of the extractor that defines how the files should be read (like in Listing 1). The library that we developed allows extraction from three file formats used to store raw NGS data. With the library, we can read data from FASTQ file format, dedicated to storing NGS raw data. Additionally, we designed a dedicated row-oriented format for processing NGS data on the Azure Data Lake platform, which improves the performance of the processing. The new data format assumes that all data related to one sequence is kept in a single row, in sections separated by a delimiter, which is a vertical bar "|." This format was specially designed during the implementation of this work to make the best use of the possibilities of the Data Lake services. The layout of a single row that stores information describing the corresponding reads (paired-end) in the row-oriented file is shown below and implements the representation from formula 6.

```
<Description of read 1>|<Sequence 1>|<Optional descr.>|
<Quality values for read 1>|
<Description of read 2>|<Sequence 2>|<Optional descr.>|
<Quality values for read 2>|
```

Consequently, for the new row-oriented format, we also implemented appropriate extractors that enable reading NGS data stored in it. We also provided the ability to read data from FASTA format files. However, files in this format do not store information on the sequence reads quality. Therefore, no mechanism for cleaning data stored in this format has been implemented in the Process phase. Simple operations on FASTA files can be performed using U-SQL expressions (shortening the sequence to a specific length, removing short sequences, etc.). In summary, the following extractors were prepared for reading NGS data:

- `FastaExtractor`—for reading data from files in the FASTA format.
- `FastqExtractor`—for reading data from files in the FASTQ format. As an argument, the extractor takes a *Boolean* value that indicates whether the identifier taken from the first description line of a read should be written to a separate column. By default this value is set to *true*.

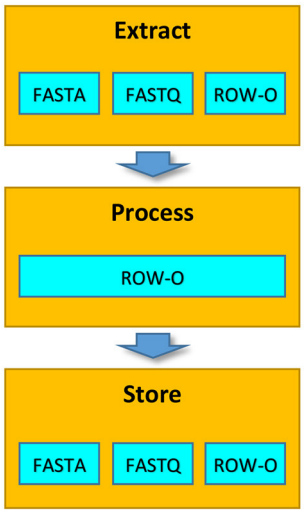


FIGURE 2 | Formats of data representation in particular phases of the EPS process.

- `FormattedFastaExtractor`—for reading from files in a row-oriented version of the FASTA format.
- `FormattedFastqExtractor`—for reading from files in a row-oriented version of the FASTQ format.
- `FormattedPairedEndExtractor`—for reading data from files in the row-oriented version of the FASTQ format, in which data from paired-end sequencing related to a single read are stored in one row of the file.

It is worth noting that although before extraction the data can be stored in various formats, in the Process phase, the extracted data is always represented in the row-oriented format (**Figure 2**). This representation allows processing the data more efficiently (see section 2.4 for details).

Examples of data reading with the use of the implemented extractors are presented in Listing 1. U-SQL enables reading data in parallel from multiple files located in a given location. Information on sequence reads resulting from paired-end sequencing is usually stored in two separate FASTQ files (like in Listing 1, lines 3 and 8). In order to be able to process such data in the successive steps, it is required to link the corresponding reads from these two files (lines 11–16).

```
1  @SRR_1 = EXTRACT id int, name string, sequence
       string,
2              optional string, quality string
3    FROM @forwardFilePath
4    USING new NGSQualityControl.Domain.Extractors.
       FastqExtractor();
5
6  @SRR_2 = EXTRACT id int, name string, sequence
       string,
7              optional string, quality string
8    FROM @reverseFilePath
9    USING new NGSQualityControl.Domain.Extractors.
       FastqExtractor();
10
11 @SRR_1_2 = SELECT r1.name AS name_r1,  r2.name
       AS name_r2,
12       r1.sequence AS sequence_r1, r2.sequence AS
        sequence_r2,
13       r1.optional AS optName_r1, r2.optional AS
       optName_r2,
14       r1.quality AS qualScore_r1, r2.quality AS
       qualScore_r2
15     FROM @SRR_1 AS r1 JOIN @SRR_2 AS r2
16          ON r1.id == r2.id;
```

Listing 1 | Reading data from two files and linking reads related to the same sequence.

The Extraction process can be quite complex, and the invocation of extractors according to the U-SQL syntax may cause troubles for those users and NGS analysts who are not familiar with programming. Therefore, to facilitate using the above-mentioned solutions, we added wrapping functions that enable the same functionality of reading NGS data. Examples of these functions are presented in Listing 2.

```
1  // extracting from two FASTQ files, for the
       paired-end sequencing
2  @SRR988072 = ExtractPairedEndSequences(
3     @"/SRR988072_Compressed/SRR988072_1.gz",
4     @"/SRR988072_Compressed/SRR988072_2.gz"
5  );
6
7  // extracting from a FASTQ file, for the single-
       read sequencing
8  @SRR988072 =  ExtractSingleEndSequences(
9     @"/SRR988075_FULL/SRR988075_2.fastq"
10 );
```

Listing 2 | Invocation of wrapping functions for extraction of data from FASTQ files with data obtained with the paired-end and single-read sequencing techniques.

2.2. NGS Data Processing: Improving NGS Data Quality

NGS data processing covers applying a set of transformations for the rowset generated in the Extract phase. The phase is parallelized for large rowsets c provided at the input (Algorithm 3). First, the rowset c is divided into many data chunks (line 2). Then, each data chunk c_i is processed in parallel on allocation units by applying cleaning transformations $t_k \in T_R$ for each row (sequence read in a row-oriented format) r_j of the data chunk c_i. The cleaning covers single reads in the single-read mode (lines 5–11) or forward (left) and reverse (right) reads in the paired-end sequencing mode (lines 12–18). Results are stored in the new rowset c_i^* (lines 10 and 17). At the end, all new data chunks are merged together into new rowset c^* with cleaned data (line 21, $|C|$ is the number of data chunks the input rowset c was divided into).

Improving NGS data quality is implemented in the U-SQL and performed through a set of transformations implemented in the Process phase of the EPS process. The set of transformations is modeled based on the capabilities of the Trimmomatic tool (Bolger et al., 2014). Trimmomatic works in two modes: single-read and paired-end. We have implemented the following commands for improving data quality in our tool:

Algorithm 3: Processing a row-oriented data partition (a rowset) from a single NGS data file (or a pair of files for the paired-end sequencing).

Input :
c : a rowset with extracted sequence reads;
T_R : a set of transformations;
Output: c^* : a rowset of cleaned data;

```
Process(c, T_R)
C ← SplitIntoChunks(c);
parallel foreach c_i ⊂ C do
    foreach r_j ∈ c_i do
        if mode = single-read then
            r_j^* ← r_j;
            foreach t_k ∈ T_R do
                r_j^* ← t_k(r_j^*);
            end
            c_i^* ← c_i^* ∪ r_j^*;
        end
        else
            (r_j^{f,*}, r_j^{r,*}) ← (r_j^f, r_j^r);
            foreach t_k ∈ T_R do
                (r_j^{f,*}, r_j^{r,*}) ← t_k(r_j^{f,*}, r_j^{r,*});
            end
            c_i^* ← c_i^* ∪ (r_j^{f,*}, r_j^{r,*});
        end
    end
end
c^* ← ⋃_{i=1}^{|C|} c_i^*;
return c^*;
```

- ILLUMINACLIP—removes Illumina adapters from sequence reads,
- SLIDINGWINDOW—removes nucleotides using the sliding window method; starts scanning at the 5' end and cuts off the read when the average quality in the window falls below the threshold value,
- MAXINFO—removes nucleotides with an adaptive method, by balancing the read length and error level to maximize the quality of each read,
- LEADING—removes nucleotides from the beginning of the sequence as long as their quality is lower than the specified threshold,
- TRAILING—removes nucleotides from the end of the sequence as long as their quality is below the specified threshold,
- CROP—reduces reads to a specified length,
- HEADCROP—deletes the specified number of nucleotides from the beginning of the read,
- TAILCROP—deletes the specified number of nucleotides from the end of the sequence,
- MINLEN—deletes the read if its length is shorter than the specified value,
- AVGQUAL—deletes the sequence if the average quality of its nucleotides is lower than the specified threshold.

NGS data transformations are performed by invoking dedicated *processors* for the U-SQL queries that are used for parallel processing in the Data Lake environment. We developed two data processors that allow cleaning the NGS reads:

1. `FastqPairedEndTrimmerProcessor` (wrapped by the `ProcessPairedEnd` processing function)—allows processing sequence reads obtained as a result of the paired-end sequencing.
2. `FastqSingleEndTrimmerProcessor` (wrapped by the `ProcessSingleEnd` processing function)—allows processing sequence reads obtained as a result of the single-read sequencing.

An example of how to process the extracted rowset with developed processors is shown in Listing 3. The processing statement consumes the processed data set with extracted sequence reads and quality information in the `PROCESS` clause (lines 2 and 10) and generates a new rowset with cleaned NGS sequence reads (lines 1 and 9). The rowset consists of information specified in the `PRODUCE` clause (lines 3 and 11). Processing is performed with the use of one of the two data processors invoked in the `USING` clause. These processors accept several arguments. The first one is the `String` value with a list of cleaning commands (`@command`, lines 6 and 16). Commands are issued in the order in which they are given. Command arguments are given after the colon symbol ":." It is recommended that the removal of adapters from NGS reads be performed first. The `@illuminaAdaptors` argument (defaults to *null*) is a `String` value that takes the location to the file with adapters to be removed from the input sequences. The last argument takes the quality score coding (`PHRED33`—set by default, or `PHRED64`). The `@keepUnpaired` argument of the `Boolean` type (for the paired-end data processor) is used to set the flag (*false* by default), forcing the storage of reads that, as a result of cleaning, were deprived of the associated read stored in the second (paired) file.

```
1  @SRRSingleEnd_result =
2      PROCESS @SRRSingleEnd //processed rowset
3      PRODUCE name string, sequence string,
4              optionalName string, qualityScore
       string
5      USING new NGSQualityControl.Domain.
       Processors.
6  FastqSingleEndTrimmerProcessor(@command,
       @illuminaAdaptors,
7   (NGSQualityControl.Helper.Infrastructure.
       QualityEncodingType)@qualityType);
8
9  @SRRPairedEnd_result =
10     PROCESS @SRRPairedEnd_1_2 //processed rowset
11     PRODUCE name_r1 string, name_r2 string,
12             sequence_r1 string, sequence_r2
       string,
13             optionalName_r1 string,
       optionalName_r2 string,
```

```
14              qualityScore_r1 string,
        qualityScore_r2 string
15      USING new NGSQualityControl.Domain.
        Processors.
16  FastqPairedEndTrimmerProcessor(@command,
        @keepUnpaired, @illuminaAdaptors, (
        NGSQualityControl.Helper.Infrastructure.
        QualityEncodingType)@qualityType);
```

Listing 3 | Cleaning NGS data with the developed processors in U-SQL.

As in the case of the extractors used for reading the NGS data, also for the processing phase, we developed the wrapping functions that simplify the use of prepared solutions. These functions and an example of how to use them are presented in Listing 4. They accept the rowset with extracted NGS data as a first argument (lines 2 and 10) and a set of cleaning transformations (commands) as a second argument (lines 3 and 11). The third argument for the paired-end sequencing data processor tells it what to do with the reads left unpaired after the cleaning (the DEFAULT value means not to keep them, line 4).

```
1  @SRR988074_result = ProcessPairedEnd(
2     @SRR988074,
3     @"ILLUMINACLIP:2:30:10 TAILCROP:10 LEADING:20
        TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:30",
4     DEFAULT,
5     @Res_Lookup,
6     DEFAULT
7  );
8
9  @SRR988075_result = ProcessSingleEnd(
10    @SRR988075,
11    @"ILLUMINACLIP:2:30:10 TAILCROP:10 LEADING:20
        TRAILING:20 MINLEN:30 SLIDINGWINDOW:4:20",
12    @Res_Lookup,
13    DEFAULT
14 );
```

Listing 4 | Wrapping functions for cleaning NGS data with the developed processors in the Data Lake environment.

The last two arguments correspond to the location of the dictionary of adapters to be removed (@Res_Lookup, lines 5 and 12) and the quality score encoding (lines 6 and 13, DEFAULT means PHRED33). Both functions return cleaned rowsets of NGS data.

2.3. NGS Data Storing

Storing data completes the EPS process for the NGS data. It is performed according to Algorithm 4. The procedure accepts the rowset c with the processed data, a dedicated outputter O, and the name of the output file (or files, depending on the mode). The rowset is split into several data chunks (line 2) that are written into several parts of the file(s). The offset is determined by the data chunk c_i (lines 6 and 9–10). The degree of parallelism depends on the size of data written, the used outputter, and the maximum number of AUs specified by the user while executing the job. Custom outputters (storage processors) may, however, serialize this part of the EPS (see **Table 1**). Each read r_j in the rowset is stored appropriately depending on the destination format specified (e.g., it is transposed again to be represented in the FASTQ format, unless we use the row-oriented format to store the data in the output files).

Algorithm 4: Storing a processed rowset of NGS data to output files with a dedicated outputter O.

Input :
c : a rowset with processed sequence reads;
O : a dedicated outputter; $f_{out}|f_{out1},f_{out2}$: output files(s);
Output: $f_{out}|f_{out1},f_{out2}$: output files(s) with cleaned data;

```
1  Store(c, O, f_out|f_out1,f_out2 )
2  C ← SplitIntoChunks(c);
3  parallel foreach c_i ⊂ C do
4      foreach r_j ∈ c_i do
5          if mode = single-read then
6              f_out[c_i + j] ← O(r_j);
7          end
8          else
9              f_out1[c_i + j] ← O(r_j^f);
10             f_out2[c_i + j] ← O(r_j^r);
11         end
12     end
13 end
```

The Store phase implemented in U-SQL covers saving the output of processing scripts to a file in the Data Lake or a database. The data is written to the file using one of the dedicated outputters that we have developed for various formats that NGS data can be stored in. Five different output interfaces have been prepared for this purpose:

1. FastaOutputter—saves data to a file in the FASTA format,
2. FastqOutputter (with the SavePairedEndRowsetDecompressed and the SaveSingleEnd-RowsetDecompressed functions)—saves data to a file in the FASTQ format,
3. FastqGzipOutputter (with the SavePairedEndRowsetCompressed and the SaveSingleEnd RowsetCompressed functions)—saves data to a compressed FASTQ file.
4. FormattedFastqOutputter (with the SaveFormattedPairedEndRowsetDecompressed and the SaveFormattedSingleEndRowsetDecompressed functions)—saves data to a file in the row-oriented version of the FASTQ format; as an argument, it uses a Boolean value that specifies whether the rowset being stored contains reads resulting from paired-end sequencing or only reads from single-read sequencing,
5. FormattedGzipFastqOutputter (with the SaveFormattedPairedEndRowsetCompressed and the SaveFormattedSingleEndRowsetCompressed functions)—stores data to the compressed, row-oriented version of the FASTQ format.

An example of invoking the proposed outputters to store NGS data after the quality improvement to a file in the Data Lake is presented in Listing 5. The `OUTPUT` clause accepts the rowset with the cleaned NGS data that will be stored (lines 1 and 5). The NGS data will again be stored in the Data Lake in the destination file specified in the `TO` clause either directly or by a `string` variable (lines 2 ad 6). Finally, the data are persisted in the storage space in a specific format by invoking a particular outputter (lines 3 and 7).

```
1 OUTPUT @SRR // NGS rowset after processing to be
      stored
2 TO @destination1 // path for the destination
     file, where data should be stored
3 USING NGSQualityControl.Domain.Factories.
     OutputtersFactory.GetFastqOutputter();
4
5 OUTPUT @SRR
6 TO @destination2
7 USING NGSQualityControl.Domain.Factories.
     OutputtersFactory.GetGzipFastqOutputter();
```

Listing 5 | Sample invocations of the Store phase for two developed outputters saving cleaned NGS data in the uncompressed and compressed FASTQ formats.

To unify the coding related to saving the processed NGS data to a file in the Data Lake, for each of the outputters, we also created wrappers that facilitate the use of implemented mechanisms. Sample functions for storing paired-end NGS data as decompressed and compressed files are presented in Listing 6. As arguments, the wrapping functions take paths to the files to which the NGS data from the NGS resultant rowset should be saved (lines 3–4 and 10–11). The rowset with cleaned data is given as the last argument (lines 6 ad 12).

```
1  SavePairedEndRowsetDecompressed(
2     //paths for the destination files, where data
        should be stored
3     @forward_destination,
4     @reverse_destination,
5     //NGS rowset after processing to be stored
6     @SRR_Paired_End_Result
7  );
8
9  SavePairedEndRowsetCompressed(
10    @forward_destination,
11    @reverse_destination,
12    @SRR_Paired_End_Result
13 );
```

Listing 6 | Sample wrapping functions for storing paired-end NGS data after processing and improving its quality.

2.4. Granularity of Parallelism

Parallel computations can be performed according to various levels of granularity, including fine-grained, medium-grained, and coarse-grained. The granularity defines the amount of computational work performed within a single task. While performing the quality control and NGS data cleaning in the proposed Data Lake environment, we can apply two types of parallelism:

- Coarse-grained parallelism—this granularity applies when multiple, whole FASTA and FASTQ files are processed in compressed and decompressed form.
- Medium-grain parallelism—this granularity applies when multiple large NGS data from FASTA and FASTQ files are divided into many (d) smaller files (e.g., 250 or 750 MB), or when NGS data are stored as a whole in the row-oriented format (then, the splitting is done automatically).

Both levels of granularity are presented symbolically in **Figure 3**. In our solution, the granularity of parallelism depends on the format and sizes of processed data files. In the most typical scenario, when whole NGS data files are uploaded to the data lake in the FASTQ format, coarse-grained parallelism will occur. The coarse-grained parallelism relies here on processing individual NGS data files by Allocation Units (AUs) responsible for data processing-related computations. **Figure 3** shows only three AUs in action, but there can be many more. This level of granularity is applied due to the large sizes and non-standard format of the NGS data files from the viewpoint of processing data in Big Data environments. Each sequence read entity is composed of four successive rows. This requires dedicated extractors and outputters to extract the data before and store the data after processing. Unlike those used for standard row-oriented data, these are not standard extractors and outputters, where each row constitutes an independent entity. The efficiency of such an approach is lower due to longer idle times resulting from uneven sizes of processed data files.

The average idle time for the coarse-grained parallelism T_C^{idle} can be calculated as follows:

$$\bar{T}_C^{idle} = \frac{1}{n-1}\sum_{i=1}^{n-1}(T_{max} - T_{AU_i}), \tag{7}$$

where T_{max} is the longest processing time registered (usually the execution time noted when processing the largest NGS data file), which is equivalent to the execution time of the whole parallel processing, T_{AU_i} is the processing time of the i-th NGS data file (another than the largest one) by another AU (other than the one that processes the longest), n is the number of AUs in use.

Splitting FASTQ data into multiple data chunks d causes changing the granularity of parallelism from coarse-grained to medium-grained and usually increases overall efficiency. This should be visible, especially if the sizes of processed data files differ much. This increase in efficiency is possible due to shorter idle times for AUs that have nothing to process in the final iteration of data processing (assuming that $n < d$, we have to perform several processing iterations with the same AUs for different data chunks).

The idle time for the medium-grained parallelism T^{idle} is equal to the idle time of any AU ($T_{AU_i}^{idle}$) processing a data chunk:

$$T^{idle} = T_{AU_i}^{idle}. \tag{8}$$

The best performance can be achieved when the number of allocated AUs is equal to the number of data chunks ($n = d$). Theoretically, in such a case, the idle time $T^{idle} = 0$. The number

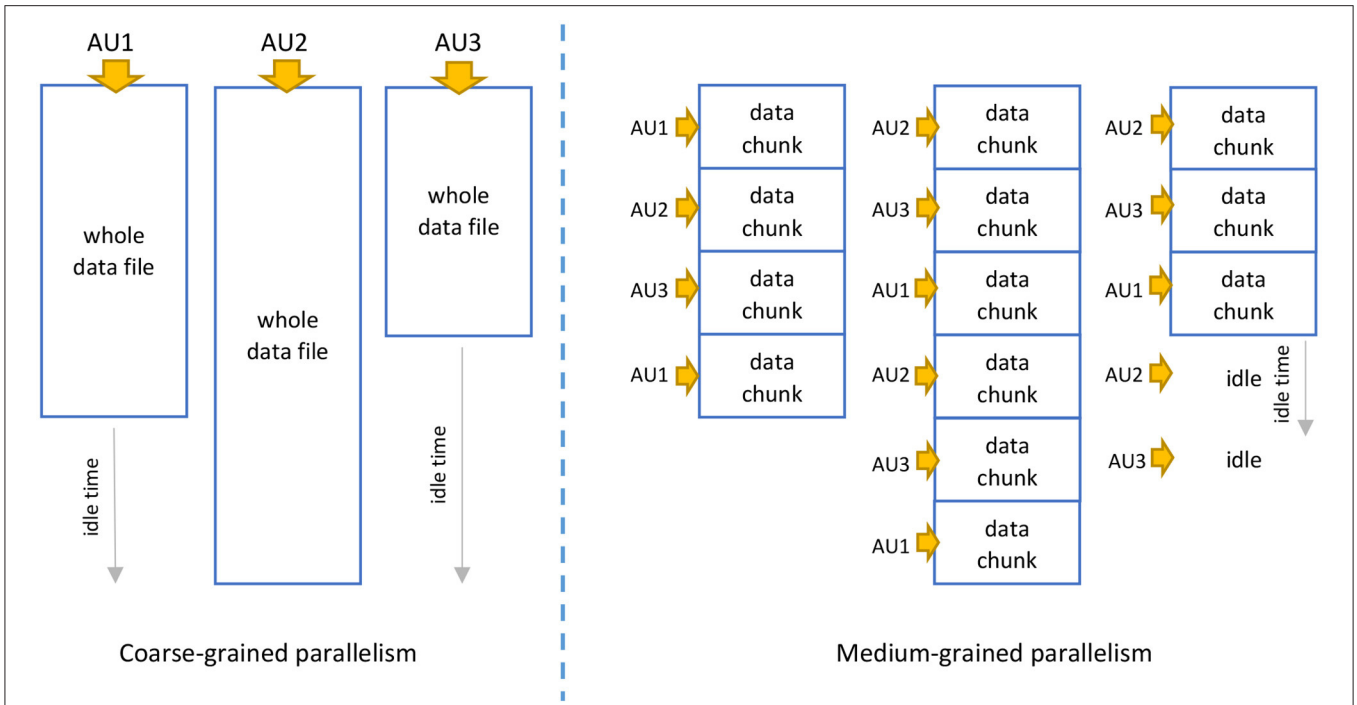


FIGURE 3 | Granularity of parallelism applied in various storage scenarios: **(left)** NGS data processed as a whole by many Allocation Units (AUs), **(right)** processing NGS data divided into chunks.

of allocated AUs can be even greater than the number of data chunks ($n > d$), but it would unnecessarily increase the cost of using the NGS data lake platform as some AUs would have nothing to process (overallocation).

It is worth noting that medium-grained parallelism is automatically applied when the NGS data is stored in a row-oriented format. This is a non-standard format to keep the NGS data in but, at the same time, a standard format for processing data on Big Data platforms. This fact causes that, unlike in previous cases, the medium-grained parallelism occurs in all phases of the EPS process, including extraction and storing. And this is the reason why we proposed a new format to store the NGS sequence data.

Table 1 summarizes the granularity levels used for particular storage formats in particular phases of the EPS process. When processing row-oriented files, we operate on the medium-grained level of parallelism in all phases of the EPS. For native formats (FASTA and FASTQ), we usually operate on the coarse-grained level of parallelism while extracting and storing the data. This is because we use non-standard extractors. Medium-grained parallelism is achievable in the Extract phase when we pre-process the files and physically divide them into many smaller files. This should speed up the Extract phase but requires an additional preparation step.

3. EXPERIMENTAL RESULTS

The presented Data Lake-based approach was tested to verify the quality of results and performance of the NGS data cleaning. We performed tests in the highly parallel Azure Data Lake environment and on local workstations. For the Data Lake environment, we executed the EPS process on the varying number of Allocation Units (AUs).

The purpose of the experiments was to find a way to store NGS data in the Data Lake Store to make the most efficient use of the Data Lake Analytics performing the EPS process, thereby reducing the analysis time and, indirectly, the associated costs of using the scalable platform. In the following sections, we will also briefly present a comparison of the duration of data processing in the cloud and on desktop computers. On the other hand, we also checked the correctness of the obtained results. We checked whether the NGS data processed and cleaned with the use of the developed library are identical to those obtained as a result of analogical processing performed on local workstations with the Trimmomatic program.

During our tests on improving the quality of NGS data, we executed the U-SQL script that looked like the one presented in Listing 7 (executed scripts differed only with the paths to data files extracted and stored as we worked with different data sets). The presented U-SQL script extracts NGS data obtained by means of the paired-end sequencing technique, stored in two files. Then, it processes the files according to the cleaning commands given. Finally, it saves the processed reads to two uncompressed FASTQ files.

```
1 REFERENCE ASSEMBLY [NGSQualityControl.Helper];
2 REFERENCE ASSEMBLY [NGSQualityControl.Domain];
3
```

Table 1 | Granularity levels of parallel computations used for particular storage formats in particular phases of the EPS process.

	FASTA	FASTQ	Row-O
Extract	Coarse- or Medium-grained	Coarse- or Medium-grained	Medium-grained
Process	Medium-grained	Medium-grained	Medium-grained
Store	Coarse-grained	Coarse-grained	Medium-grained

```
DECLARE @Res_Lookup string = @"/Adapters/TruSeq3-PE.fa";
DEPLOY RESOURCE @Res_Lookup;

@SRR988074 = ExtractPairedEndSequences(
   @"/SRR988074_FULL/SRR988074_1.fastq",
   @"/SRR988074_FULL/SRR988074_2.fastq"
);

@SRR988074_result = ProcessPairedEnd(
   @SRR988074,
   @"ILLUMINACLIP:2:30:10 TAILCROP:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:30",
   DEFAULT,
   @Res_Lookup,
   DEFAULT
);

SavePairedEndRowsetDecompressed(
   @"/Result/SRR988074_1.fastq",
   @"/Result/SRR988074_2.fastq",
   @SRR988074_result
);
```

Listing 7 | Sample U-SQL script used for parallel cleaning of NGS data in performed experiments.

During our tests, we used the raw NGS data obtained from the NGS sequencing of the *Drosophila melanogaster* with the paired-end method. We tested our library on four NGS data sets (each providing two FASTQ files containing the corresponding sequence reads from the 3' to 5' end of the sequenced DNA fragment). The data were collected from the Sequence Read Archive database (Leinonen et al., 2010). The following NGS data sets were used in our experiments:

- SRR988072 (two files, 4.95 GB each),
- SRR988073 (two files, 3.61 GB each),
- SRR988074 (two files, 5.2 GB each),
- and SRR988075 (two files, 11.7 GB each).

The total amount of data was about 50.1 GB for uncompressed data. For the compressed data (gzip-based compression), the total amount of data was ∼14.8 GB (SRR988072—1.48 and 1.30 GB, SRR988073—1.08 and 951 MB, SRR988074—1.62 and 1.45 GB, and SRR988075—3.63 and 3.33 GB). These files contained the NGS data characterized by low quality and contamination with Illumina adapters. For this reason, they were selected for our tests related to NGS sequence cleaning.

3.1. Processing Multiple Genomes

One of the advantages of the Data Lake ecosystem is the possibility of processing the NGS data of many genomes simultaneously. In this section, we present the results of performance experiments carried out for parallel processing of all data sets (SRR988072, SRR988073, SRR988074, and SRR988075) for various storage scenarios, file formats, and compression used. Experiments were performed with 8 AUs. For this experiment, the NGS sequence reads were stored in their native FASTQ files and the row-oriented files. Additionally, we also divided the NGS data into 250 MB FASTQ files to increase the level of parallelism (manually apply the medium-grained parallelism) and verify whether it affects the performance of the EPS process. The 250 MB size of the files fits exactly one block of data, called an extent, assigned to a single unit of parallelism—a vertex in the execution graph.

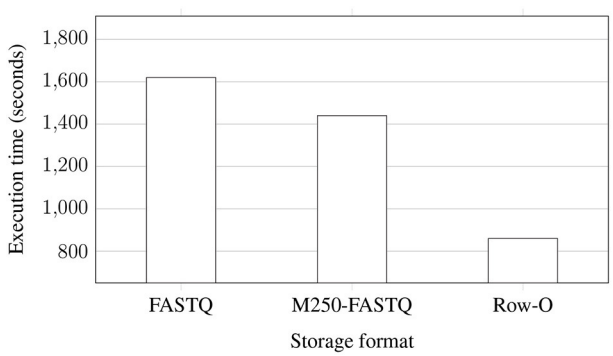


FIGURE 4 | Processing times of NGS data extracted from uncompressed files and saved to eight uncompressed FASTQ files (two files for each of the processed genomes) with 8 AUs for various storage formats: regular FASTQ, multiple 250 MB FASTQ (M250-FASTQ), and row-oriented (Row-O).

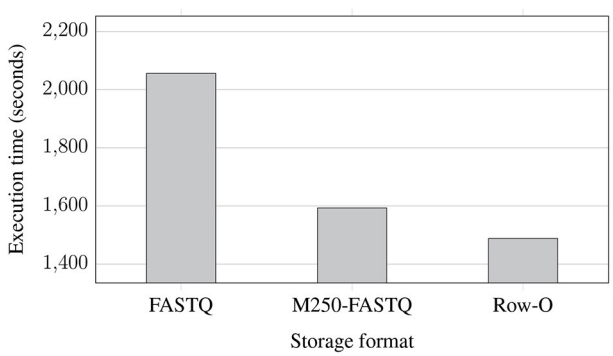


FIGURE 5 | Processing times of NGS data extracted from compressed files and saved to eight compressed FASTQ files (two files for each of the processed genomes) with 8 AUs for various storage formats: regular FASTQ, multiple 250 MB FASTQ (M250-FASTQ), and row-oriented (Row-O).

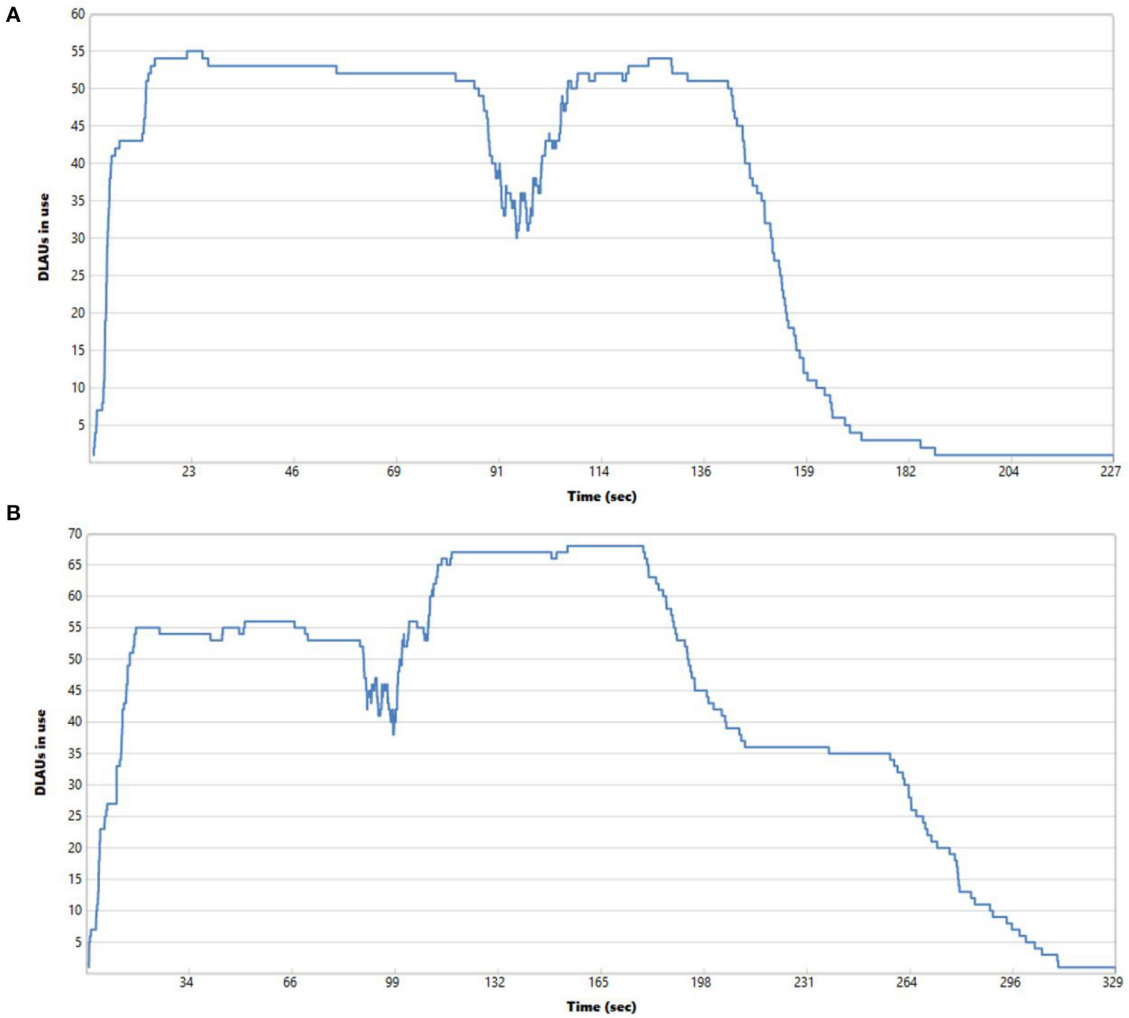


FIGURE 6 | Utilization of Allocation Units (computational units) over time while processing data from four uncompressed **(A)** and compressed **(B)** row-oriented files with forward and reverse reads and saving to eight uncompressed **(A)** and compressed **(B)** FASTQ files (two files for each of the processed genomes) with 55 AUs **(A)** and 68 AUs **(B)**.

Figure 4 shows the execution time of the whole EPS process performed for improving the quality of NGS data stored in the three formats. The data were extracted from uncompressed data files and stored again in uncompressed files after improving the quality. As can be observed, processing the whole FASTQ files takes the longest, while row-oriented files are processed almost two times faster. The distribution of data to multiple FASTQ files brings some improvement, but it is not huge.

Figure 5 shows the execution time of the whole EPS process performed for improving the quality of NGS data stored in the same three formats. However, in contrast to the previous experiment, the data were extracted from compressed data files and stored in compressed files after improving the quality. In terms of storage format, conclusions are similar to those from the previous experiment. However, we can observe that for the compressed data, the use of the row-oriented format does not bring such a huge improvement in the execution time. Comparing the results of both experiments, we can also notice that processing the compressed data takes more time, which is caused by additional decompression and compression steps while extracting and storing the data in the EPS process.

It is worth noting that in both cases of processing compressed and uncompressed files, the row-oriented format turned out to be highly scalable. When scaling out to many AUs for the same collection of data, we could decrease the execution time to 227 s when processing uncompressed row-oriented files and to 329 s when processing compressed row-oriented files for all data sets and storing the cleaned NGS data to eight uncompressed and compressed FASTQ files in both scenarios. **Figure 6** shows the utilization of Allocation Units over time while processing the data. It can be observed that AUs are not evenly utilized during the whole execution time. Especially in the final phases of the job

execution, their utilization is low due to storing in FASTQ files, for which we cannot rely on the medium-grained parallelism.

3.2. Performance Gain Over Local Processing

In the next series of experiments, we compared the execution time of the whole EPS process performed in the Data Lake environment and on local workstations. The data processing time on the local computers was checked using two machines with different configurations. The first workstation had an Intel Core 2 Duo 3.6 GHz CPU, 3 GB DDR memory, and 320/7200/16 hard disk drive. The second workstation had much better compute capabilities. It was equipped with an Intel Core i7-4790K 4 GHz processor, 16 GB DDR3 memory, and the same type of hard disk drive. For improving the quality of data, we used the original Trimmomatic program. NGS data were processed to achieve the best possible quality scores. The following commands were used during the data processing phase for particular NGS data sets (SRR988072, SRR988073, SRR988074, and SRR988075), analogous to those used to perform data cleaning in the Data Lake ecosystem:

```
>ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:20 TRAILING:20
 SLIDINGWINDOW:4:20 MINLEN:30
>ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:20 TRAILING:20
SLIDINGWINDOW:4:20 MINLEN:30
>ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 TAILCROP:10 LEADING:20
TRAILING:20 SLIDINGWINDOW :4:20 MINLEN:30
>ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 TAILCROP:10 LEADING:20
TRAILING:20 SLIDINGWINDOW :4:20 MINLEN:30
```

Figure 7 shows processing times for NGS data extracted from regular uncompressed FASTQ files and saved to eight uncompressed FASTQ files (two files for each of the processed genomes) for various implementations: on Data Lake with 8 AUs (DLA-8), workstation 1 (WS1), and workstation 2 (WS2). This experiment shows a pessimistic case since whole FASTQ files are processed at the coarse-grained level of parallelism. As we can observe, the processing time was reduced more than three times in the Data Lake environment. It is not linearly proportional to the number of AUs in use (8 AUs), but differences in sizes of processed FASTQ files and granularity of parallelism do not allow for better performance gain (some AUs finish their processing earlier and stay idle for some time, as it was presented in **Figure 3**).

3.3. Quality Control

Our library was created to allow scalable processing and improving the quality of NGS raw data stored in the Big NGS Data Lake. At the same time, the library implements the set of functionalities of the Trimmomatic application, an open-source desktop program intended for this purpose. As a part of our experiments, we also validated the effectiveness of our library in terms of the quality of results. Tests were performed for all genome sequences in our NGS Data Lake. In this section, we show the validation results on the example of NGS data marked with the SRR988074 identifier. To validate the effects of the cleaning process performed on FASTQ files with our library for data lake, we used the FastQC tool (Wingett and Andrews, 2018). **Figure 8** shows the comparison of results generated by the FastQC program presenting the assessment of the quality of

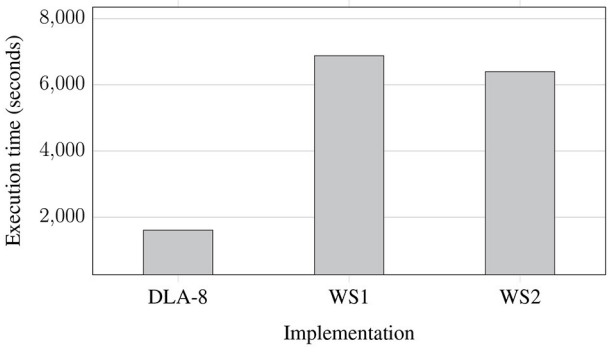


FIGURE 7 | Processing times of NGS data extracted from regular uncompressed FASTQ files and saved to eight uncompressed FASTQ files (two files for each of the processed genomes) for various implementations: on Data Lake with 8 AUs (DLA-8), workstation 1 (WS1), and workstation 2 (WS2).

nucleotides in DNA sequence reads from two NGS files storing data obtained with the paired-end sequencing technique. We can observe that quality scores of the sequence reads before the cleaning process drop below 20 (for the file with forward reads SRR988074_1, **Figure 8A**). After the cleaning process, the quality scores stay above 25 (**Figure 8C**), and even 30 for the file containing reverse reads (SRR988074_2, **Figure 8D**). In the case of both files, definite improvement is visible.

Since we implemented the same set of functionalities as in the Trimmomatic, in terms of the quality improvement, the results are the same as for the original desktop application.

4. DISCUSSION AND FUTURE DIRECTIONS

Improving the quality of NGS data is one of the first steps preceding the secondary analysis of DNA genome sequences and further NGS-based analyses. Our work confirms that the steps of the pre-processing can be performed on a large scale by (1) collecting the massive volumes of NGS data in the NGS data lake, (2) processing them in parallel within the EPS process, and (3) scaling the computations in the Cloud. Our library becomes then a handy element of the early stages of the secondary analysis of NGS data.

Although, as we could see, processing some storage formats (like the whole compressed FASTQ files) do not provide linearly proportional performance gain and do not allow utilizing both types of parallelism, we also found a way to parallelize computations for other storage formats efficiently (e.g., whole native, uncompressed FASTQ files) and take advantages of the platform capabilities and the techniques we propose. Our experiments showed that we could benefit from the coarse-grained parallel processing when we process multiple genomes. Medium-grained parallelism is advantageous mainly for row-oriented files. However, our solution has limitations for handling the whole FASTQ files, for which the medium-grained parallelism cannot be applied in all phases of the EPS process.

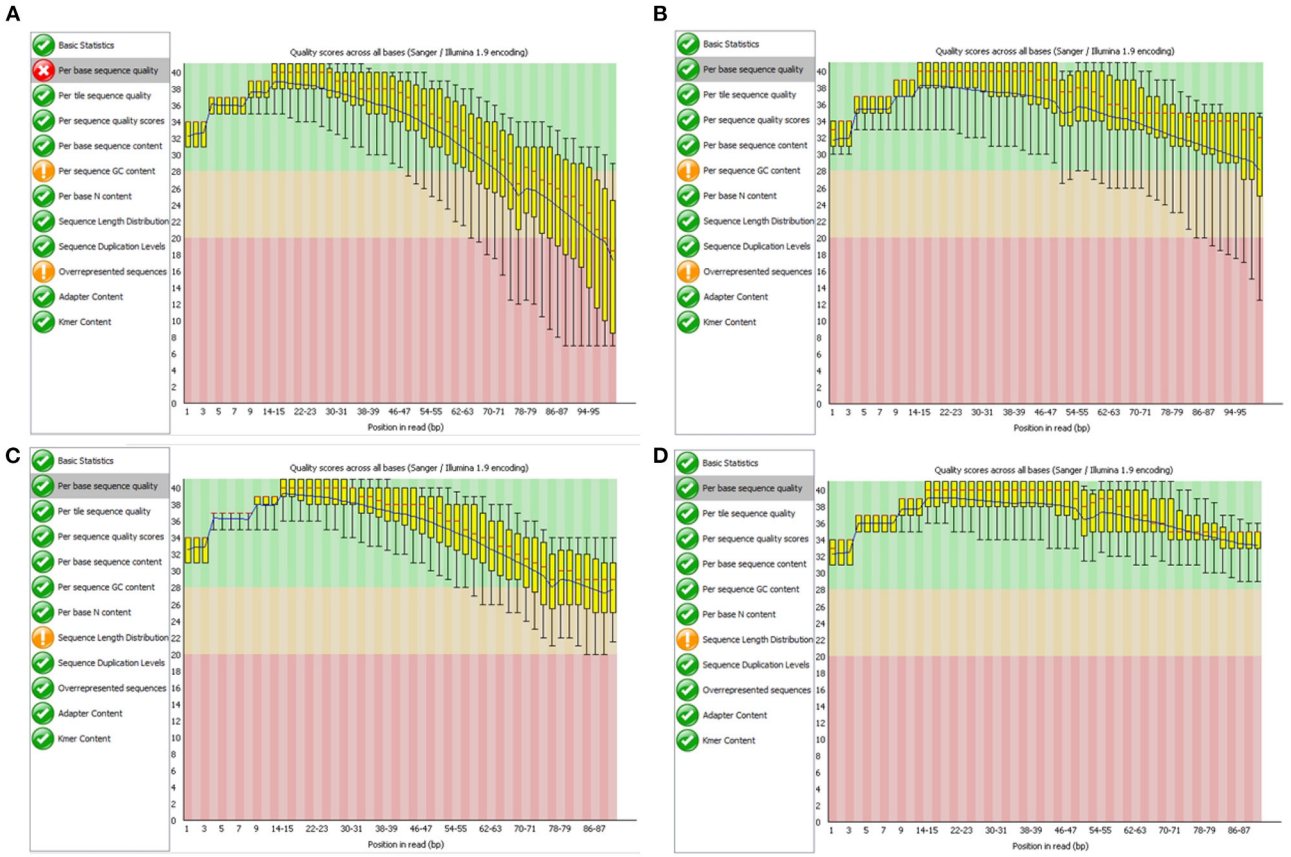


FIGURE 8 | Evaluation of nucleotide sequence quality in the FASTQ files SRR999074_1 **(A,C)** and SRR999074_2 **(B,D)** before **(A,B)** and after **(C,D)** improving the quality with the developed software.

Our experiments also showed that to take full advantage of such data lake platforms, it is advisable to work with data split into many smaller files or work with non-standard formats for storing the NGS data, like the row-oriented format presented in the paper. However, this would require either additional data pre-processing (to split the data) or changing the formats in which data are provided for analysis. The row-oriented format is the best-fitting one for all Big Data platforms and would give the best performance.

Our library complements other solutions presented in section 1 by providing a set of functionalities that are dedicated to cleaning NGS data on a large scale in a highly scalable environment, which was not available so far. In such a way, it extends the data cleaning capabilities of the Trimmomatic package toward large data sets. Like SeqPig and GMQL, the functionality of our library is exposed through a declarative language, but for a different purpose. Also, in our project, we used the U-SQL language that combines capabilities of the SQL language used for querying relational databases with the C# programming language, which is highly extendable. However, the limitation of the adopted Data Lake platform is that it is tightly linked with the Azure cloud. Therefore, unlike the SeqPig or SeQual, it is not portable to other cloud platforms. On the other hand, our Data Lake library allows improving the quality of NGS data obtained with the use of single-read and paired-end sequencing techniques and stored in various formats, which are also significant unique features of our solution.

5. CONCLUDING REMARKS

Secondary and tertiary analyses performed by scientists working in genomics and computational biology require high-quality data and an infrastructure that provides appropriate space to store large amounts of data generated as a result of next-generation sequencing techniques at various stages of the analysis. Data obtained from single genome sequencing can reach up to several hundred gigabytes and may be of various quality. The infrastructure used to analyze the NGS data should also provide computing power to allow rapid and scalable processing of gathered data. The hybridization of tools that enable handling computations over big biomedical data with extensive scaling capabilities of the Cloud proved to be a reasonable solution. This work shows the successful implementation of the early stage pre-processing techniques used

in NGS data analysis pipelines in a distributed environment of the Data Lake ecosystem hosted in the Microsoft Azure cloud computing platform. Data Lake Store allows hosting data in any format and without restrictions on the amount of data stored. These characteristics make Data Lake a perfect place for storing large amounts of data for further analysis. In such a way, our solution addresses the *volume* and *variety* challenges of processing Big NGS Data. The Data Lake Analytics allows then for parallel processing of many genomes simultaneously in a distributed environment, addressing the *velocity* challenge. This would be difficult to achieve on, for example, desktop computers due to the limited capabilities of the processors or hard disk drives. Additionally, the use of the Data Lake Analytics and serverless computing paradigm reduces the maintenance overhead and removes the need to maintain and scale underlying computing clusters manually. Finally, procedures and functions for improving NGS data quality embedded in declarative U-SQL queries simplify the cleaning process that ultimately leads to the increase in the *value* of obtained results.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Sequence Read Archive (https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR988072).

AUTHOR CONTRIBUTIONS

DM proposed the idea and KS extended it. DM conceived and designed the experiments and prepared the experimental environment. KS and DM performed the experiments, verified results, and designed and implemented the tools. DM and BM-M analyzed the data. PG reviewed the related literature. BM-M, KS, PG, and DM wrote the paper and made revisions to address comments of the referees. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by a pro-quality grant for highly scored publications or issued patents (grant No 02/100/RGJ21/0009), the professorship grant (02/020/RGP19/0184) of the Rector of the Silesian University of Technology, Gliwice, Poland, and partially, by Statutory Research funds of Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland (grant 02/100/BK_21/0008, BK-221/RAU7/2021) and Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland (Rau6, 2021), and partially, by the Polish Ministry of Science and Higher Education as a part of the CyPhiS program at the Silesian University of Technology, Gliwice, Poland (Contract No. POWR.03.02.00-00-I007/17-00).

REFERENCES

Abuín, J. M., Pichel, J. C., Pena, T. F., and Amigo, J. (2015). BigBWA: approaching the Burrows?Wheeler aligner to Big Data technologies. *Bioinformatics* 31, 4003–4005. doi: 10.1093/bioinformatics/btv506

Aronesty, E. (2011). *ea-utils: Command-Line Tools for Processing Biological Sequencing Data*. Available online at: https://github.com/ExpressionAnalysis/ea-utils (accessed April 10, 2021).

Aronesty, E. (2013). Comparison of sequencing utility programs. *Open Bioinform. J.* 7, 1–8. doi: 10.2174/1875036201307010001

Bacci, G., Bazzicalupo, M., Benedetti, A., and Mengoni, A. (2014). StreamingTrim 1.0: a Java software for dynamic trimming of 16s rRNA sequence data from metagenetic studies. *Mol. Ecol. Resour.* 14, 426–434. doi: 10.1111/1755-0998.12187

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M., and Gu, J. (2017). AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* 18:80. doi: 10.1186/s12859-017-1469-3

Criscuolo, A., and Brisse, S. (2013). Alientrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* 102, 500–506. doi: 10.1016/j.ygeno.2013.07.011

Davis, M. P. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A. J. (2013). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* 63, 41–49. doi: 10.1016/j.ymeth.2013.06.027

Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS ONE* 8:e85024. doi: 10.1371/journal.pone.0085024

Dodt, M., Roehr, J. T., Ahmed, R., and Dieterich, C. (2012). Flexbar-flexible barcode and adapter processing for next-generation sequencing platforms. *Biology* 1, 895–905. doi: 10.3390/biology1030895

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194. doi: 10.1101/gr.8.3.186

Expósito, R. R., Galego-Torreiro, R., and González-Domínguez, J. (2020). Sequal: big data tool to perform quality control and data preprocessing of large NGS datasets. *IEEE Access* 8, 146075–146084. doi: 10.1109/ACCESS.2020.3015016

Gordon, A. (2008). *FASTX-Toolkit: FASTQ/a Short-Reads Pre-Processing Tools*. Available online at: http://hannonlab.cshl.edu/fastx_toolkit/ (accessed April 10, 2021).

Hung, C.-L., Chen, W.-P., Hua, G.-J., Zheng, H., Tsai, S.-J. J., and Lin, Y.-L. (2015). Cloud computing-based TagSNP selection algorithm for human genome data. *Int. J. Mol. Sci.* 16, 1096–1110. doi: 10.3390/ijms16011096

Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 15:182. doi: 10.1186/1471-2105-15-182

Kong, Y. (2011). Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 98, 152–153. doi: 10.1016/j.ygeno.2011.05.009

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database Collaboration (2010). The sequence read archive. *Nucleic Acids Res.* 39(Suppl. 1), D19–D21. doi: 10.1093/nar/gkq1019

Li, Y.-L., Weng, J.-C., Hsiao, C.-C., Chou, M.-T., Tseng, C.-W., and Hung, J.-H. (2015). Peat: an intelligent and efficient paired-end sequencing adapter trimming algorithm. *BMC Bioinformatics* 16:S2. doi: 10.1186/1471-2105-16-S1-S2

Liao, X., Li, M., Zou, Y., Wu, F., Pan, Y., and Wang, J. (2020). An efficient trimming algorithm based on multi-feature fusion scoring model for NGS data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 17, 728–738. doi: 10.1109/TCBB.2019.2897558

Lindgreen, S. (2012). Adapterremoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* 5:337. doi: 10.1186/1756-0500-5-337

Liu, D. (2019). Fuzzysplit: demultiplexing and trimming sequenced DNA with a declarative language. *PeerJ* 7:e7170. doi: 10.7717/peerj.7170

Małysiak-Mrozek, B., Stabla, M., and Mrozek, D. (2018). Soft and declarative fishing of information in Big Data lake. *IEEE Trans. Fuzzy Syst.* 26, 2732–2747. doi: 10.1109/TFUZZ.2018.2812157

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Masseroli, M., Canakoglu, A., Pinoli, P., Kaitoua, A., Gulino, A., Horlova, O., et al. (2018). Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. *Bioinformatics* 35, 729–736. doi: 10.1093/bioinformatics/bty688

Masseroli, M., Kaitoua, A., Pinoli, P., and Ceri, S. (2016). Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. *Methods* 111, 3–11. doi: 10.1016/j.ymeth.2016.09.002

Masseroli, M., Pinoli, P., Venco, F., Kaitoua, A., Jalili, V., Palluzzi, F., et al. (2015). GenoMetric Query Language: a novel approach to large-scale genomic data management. *Bioinformatics* 31, 1881–1888. doi: 10.1093/bioinformatics/btv048

Modolo, L., and Lerat, E. (2015). UrQt: an efficient software for the unsupervised quality trimming of NGS data. *BMC Bioinformatics* 16:137. doi: 10.1186/s12859-015-0546-8

Mrozek, D. (2014). *High-Performance Computational Solutions in Protein Bioinformatics*. SpringerBriefs in Computer Science. Cham: Springer International Publishing. doi: 10.1007/978-3-319-06971-5

Mrozek, D. (2018). *Scalable Big Data Analytics for Protein Bioinformatics, Vol. 28 of Computational Biology*. Cham: Springer. doi: 10.1007/978-3-319-98839-9

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453. doi: 10.1016/0022-2836(70)90057-4

Pandey, R. V., Pabinger, S., Kriegner, A., and Weinhäusel, A. (2016). ClinQC: a tool for quality control and cleaning of Sanger and NGS data in clinical research. *BMC Bioinformatics* 17:56. doi: 10.1186/s12859-016-0915-y

Roehr, J. T., Dieterich, C., and Reinert, K. (2017). Flexbar 3.0 - SIMD and multicore parallelization. *Bioinformatics* 33, 2941–2942. doi: 10.1093/bioinformatics/btx330

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Schubert, M., Lindgreen, S., and Orlando, L. (2016). Adapterremoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88. doi: 10.1186/s13104-016-1900-2

Schumacher, A., Pireddu, L., Niemenmaa, M., Kallio, A., Korpelainen, E., Zanetti, G., et al. (2013). SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop. *Bioinformatics* 30, 119–120. doi: 10.1093/bioinformatics/btt601

Smeds, L., and Künstner, A. (2011). Condetri - a content dependent read trimmer for illumina data. *PLoS ONE* 6:e26314. doi: 10.1371/journal.pone.0026314

Sturm, M., Schroeder, C., and Bauer, P. (2016). Seqpurge: highly-sensitive adapter trimming for paired-end NGS data. *BMC Bioinformatics* 17:208. doi: 10.1186/s12859-016-1069-7

Wiewiórka, M., Szmurło, A., Kuśmirek, W., and Gambin, T. (2019). SeQuiLa-cov: a fast and scalable library for depth of coverage calculations. *GigaScience* 8:giz094. doi: 10.1093/gigascience/giz094

Wingett, S., and Andrews, S. (2018). FastQ Screen: a tool for multi-genome mapping and quality control. *F1000Research* 7, 1–13. doi: 10.12688/f1000research.15931.1

Zhang, M., Sun, H., Fei, Z., Zhan, F., Gong, X., and Gao, S. (2014). "Fastq_clean: an optimized pipeline to clean the illumina sequencing data with quality control," in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Belfast), 44–48. doi: 10.1109/BIBM.2014.6999309

Zhang, X., Shao, Y., Tian, J., Liao, Y., Li, P., Zhang, Y., et al. (2019). ptrimmer: An efficient tool to trim primers of multiplex deep sequencing data. *BMC Bioinformatics* 20:236. doi: 10.1186/s12859-019-2854-x

Zou, Q. (2016). "Multiple sequence alignment and reconstructing phylogenetic trees with Hadoop," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Shenzhen), 1438–1438. doi: 10.1109/BIBM.2016.7822492

Zou, Q., Shixiang Wan, and Xiangxiang Zeng (2016). "HPTree: reconstructing phylogenetic trees for ultra-large unaligned DNA sequences via NJ model and Hadoop," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Shenzhen), 53–58.

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.

Copyright © 2021 Mrozek, Stępień, Grzesik and Małysiak-Mrozek. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 September 2021
doi: 10.3389/fgene.2021.727201

Gene Targets Network Analysis for the Revealing and Guidance of Molecular Driving Mechanism of Lung Cancer

*Risheng Huang[1], Xiao Xiang[2], Kangliang Zhang[3], Yuanliang Zheng[1], Chichao Wang[1] and Guanqiong Hu[4]**

[1] Department of Thoracic Surgery, Dingli Clinical College of Wenzhou Medical University (Wenzhou Central Hospital), Wenzhou, China, [2] Department of Breast Surgery, First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China, [3] Central Laboratory, Dingli Clinical College of Wenzhou Medical University (Wenzhou Central Hospital), Wenzhou, China, [4] Department of Pediatric, First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China

The objective was to explore the function of gene differential expressions between lung cancer tissues and the interaction between the relevant encoded proteins, thereby analyzing the important genes closely related to lung cancer. A total of 120 samples from the GEO database (including two groups, i.e., 60 lung cancer *in situ* specimens and 60 normal specimens) were taken as the research objects, which were submitted to the analysis of signaling pathway, biological function enrichment, and protein interactions to reveal the molecular driving mechanism of lung cancer. Results: A total of 875 differentially expressed genes were obtained, including 291 up-regulated genes and 584 down-regulated genes. The up-regulated genes were mainly involved in biological processes such as protein metabolism, protein hydrolysis, mitosis, and cell division. Down-regulated genes were mainly involved in neutrophil chemotaxis, inflammatory response, immune response, and angiogenesis. The protein expression of high expression genes and low expression genes in patients were higher than those in the control group. The protein corresponding to the high expression gene was highly expressed in the patient group. Meanwhile, the proteins corresponding to the low expression genes were also expressed in the patient group, which showed that although the proteins corresponding to the low expression genes were low in the patients, they were still the target genes related to lung cancer. In conclusion, the molecular driving mechanism in lung cancer was mainly related to protein metabolism, proteolysis, mitosis, and cell division. It was found that TOP2A, CCNB1, CCNA2, CDK1, and TTK might be the critical target genes of lung cancer.

Keywords: gene target network analysis, lung cancer, molecular driving mechanism, nursing guidance, protein interaction

OPEN ACCESS

Edited by:
Arun Kumar Sangaiah, VIT University, India

Reviewed by:
Wenbo Zhang, Qiqihar University, China
Zhongyuan Li, First Affiliated Hospital of Qiqihar Medical University, China
Henry Mildredl, Federal Institute for Drugs and Medical Devices, Germany

***Correspondence:**
Guanqiong Hu
huguangqiongvr4d@163.com

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

Received: *18 June 2021*
Accepted: *13 August 2021*
Published: *20 September 2021*

Citation:
Huang R, Xiang X, Zhang K, Zheng Y, Wang C and Hu G (2021) Gene Targets Network Analysis for the Revealing and Guidance of Molecular Driving Mechanism of Lung Cancer. Front. Genet. 12:727201. doi: 10.3389/fgene.2021.727201

INTRODUCTION

The incidence rate of lung cancer is one of the fastest growing malignant tumors (Masters et al., 2017). Lung cancer is mainly divided into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), of which NSCLC accounts for 80% of all lung cancer cases (Shi et al., 2017). Currently, molecular-targeted drug therapy takes the molecules that block the high expression of

cancer cell membrane or cells as the therapeutic target, reduces the fragmentation effect on normal cells by blocking the growth, infiltration, metastasis, and inducing apoptosis of normal cells, and reduces the incidence of adverse drug reactions in patients (Shen et al., 2018; Xu et al., 2018). In the absence of biopsy, the blood samples of patients with lung cancer are the only source of information for analyzing clinically relevant genetic changes, including epidermal growth factor receptor (EGFR), Kirsten rat sarcoma viral oncogene (KRAS), v-raf murine sarcoma viral oncogene homolog B1 (BRAF), c-ros oncogene 1 (ROS1), and anaplastic lymphoma kinase (ALK) (Allan-Blitz et al., 2018). As new treatment options emerge, predictive detection of lung cancer has become a research hot spot in medical field (Horimasu et al., 2017). The diagnosis of lung cancer diseases mainly includes the identification and classification of malignant tumors, molecular tests, and immunohistochemical analysis. Complex diagnostic analysis algorithms have evolved, requiring specific drugs tailored to individual patients and considering the way to make investigations and diagnostic strategies based on individual tumors (Zhang et al., 2018). Some studies have reported that KRAS mutations may be the targets for preventing and treating KRAS mutant lung cancer and other tumor diseases (Krasnov et al., 2017). Studies have shown that the molecular driving mechanisms of lung cancer in different tumor stages are also different, and NKTR may be the target of prevention and treatment of lung cancer diseases (Zhou et al., 2017). Some studies have used the CIBERSORT method to identify and quantify the number of different cells in a tumor sample by reference genes combined with machine learning. Such an approach solves one of the major problems in determining cell types to some extent by using the reference genes (Zins et al., 2018). CIBERSORT is used to estimate the abundance of member cell types in mixed cell population by using gene expression data. It is a tool of bioinformatics analysis method and has important application value in the field of molecular biology.

Bioinformatics uses computers to mine and analyze great information in biological databases, focuses on gene and proteomic analysis, and is widely used in the fields of molecular genetics and genomics. In the field of tumor research, bioinformatics combines suspicious tumor genes with known biological data through the biological network analysis of tumor-related pathways and biological processes, identifies tumor-related functional categories, and excavates tumor networks. It also predicts potential pathogenic proteins and plays an important role in tumor pathogenesis, diagnosis, and treatment. As the gene chip technology continuously develops, it has become a hot topic how to process and analyze tremendous data and find more effective information. At present, gene chip technology is mainly used in the research of tumor-related gene information, such as screening tumor-related genes, measuring tumor mutation genes, studying tumor gene expression profiles, and diagnosing tumor diseases. In this way, it can explore the extent of influences of genetic, environmental, and pharmaceutical factors for tumors on the expression of related genes during the occurrence and development of tumors.

The rapid development of high-throughput technologies, such as MeDip-seq, methylated microarrays, and RNA-seq, has provided technical support for the identification of biomarkers for a variety of diseases such as lung cancer, as well as opportunities for the availability of publicly available data sets. By selecting the gene expression dataset of lung cancer, this study innovatively explores the network of lung cancer target genes through gene expression analysis of different databases, thus exploring the molecular driving mechanism of lung cancer and providing reference for clinical molecular drug treatment and nursing guidance of lung cancer.

MATERIALS AND METHODS

Data Resource and Processing

A total of 120 samples of lung cancer mRNA sample GSE19408 (including two groups: 60 lung cancer *in situ* specimens and 60 normal specimens) were selected from the GEO (Gene Expression Omnibus) database, using open-source software R3.4.2. for preprocessing the differential analysis of sample data.

First, download the sample, import the CEL (cool edit loop) format file into the R program, use the limma package in the R language to count the difference between the lung cancer gene and the normal gene, and then follow the FDR (false discovery rate) and FC (fold change, gene expression fold ratio) from which differentially expressed genes were selected, and the comparison between the two groups of genes must satisfy the requirements of FDR < 0.01 and $|\log2\ FC| \geq 1$.

Signal Pathway Analysis and Biological Function Enrichment

Signaling pathway analysis and biological function enrichment of the screened NSCLC differentially expressed genes were performed using the Functional annotation chart tool under the DAVID platform. First, the differentially expressed genes were introduced into the DAVID list in the form of gene symbol, and the humans were submitted to the task in the species type, and the GO (Gene Ontology) analysis and the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway were performed on both the up-regulated and down-regulated genes (Wang et al., 2020). After the results were obtained, the differentially expressed genes with statistical significance ($P \leq 0.01$) were selected.

Protein Interaction Analysis

Gene data can be applied to gene regulatory network analysis to analyze the differential expression of genes for studying the differential expression of their target genes and the processes that constitute various organisms, such as organ formation, embryo development, and disease pathogenesis. The network of relationships is compared between cell types or states and analyzed further, and specific molecular features and functional blocks can be identified, which are the basis for state transitions. In order to identify key target genes related to lung cancer, this study established a protein interaction network model to explore the regulatory relationship of differential genes at the protein level. The differentially expressed genes obtained by the DAVID platform were subjected to ID (Identity

Document) conversion and input into the STRING 9.1 (the Search Tool for the Retrieval of Interacting Genes) database to establish a differentially expressed gene encoding protein-protein between Interaction network diagram. Proteins at the center of the protein-protein interaction network often play a relatively important role in the development of the disease. The selection criteria for PPI (Protein-protein interaction network) analysis was combination score >0.4 (medium confidence). Enter the PPI value into the visualization tool, that is, the Cytoscape software, and use the analysis plug-in to calculate the edge of the nodes in the network to get the number of protein interactions (Degree). The analysis steps of Cytoscape software are as follows: first, import the node attribute file, file- >import- >table- >file(node.txt) (here is table instead of network), and then set the format of simple network diagram in style. Finally, export the file. The data can be network file, table file, or picture file. The picture file includes a variety of picture formats and PDF format, which can be selected in the toolbar.

Western Blotting Detection

(1) Total protein extraction: Cells were taken out; the culture medium was discarded, and the cells were washed with PBS. Then, 70 μL of cell lysate was added to each well. After 5 min, the cell suspension was transferred to an Eppendorf (EP) tube (TIANGEN Biochemical Technology (Beijing) Co., Ltd., China) and shaken once every 5 min for a total of 6 times. The cell suspension was put into a 4°C centrifuge, centrifuged at 1,000 rpm/min for 15 min. The supernatant was taken for bicinchoninic acid (BCA) protein quantitative determination, and the standard curve was drawn.

(2) Preparation of stacking gel and separation gel: The reagents (purchased from TIANGEN Biochemical Technology (Beijing) Co., Ltd., China) were summarized in **Table 1** below:

(3) Electrophoresis and image development: The glass plate was cleaned thoroughly with distilled water and ethanol. The glass plate was aligned and put in the clamp vertically on the glue rack. The distilled water was added to the glass plate to a suitable position. Then, the device was stood for 8 min to test whether the glass plate was leaking. A 10% separation gel was prepared according to the formula in **Table 1**. After mixing, 6 mL was added to the gap in the middle of the glass plate with a pipette; then, 3 mL of isopropanol was added slowly. Under 37°C condition, once a refraction line appeared between isopropanol and the separation gel, the separation gel solidified. Afterward, the isopropanol was poured out, and the device was washed with distilled water three times for later use. After the stacking gel was configured, 3 mL was added to the glass plate, which should slowly enter the comb to prevent bubbles.

After the concentrated gel was solidified, the glass plate and the plastic replacement plate were sandwiched in the rack with electrodes; then, the device was put into the electrophoresis tank, and the comb was pulled out. Next, 30 μL of the expressed protein supernatant was taken out, added with 10 μL of 5 × loading buffer, mixed evenly, and boiled for 10 min at 100°C.

TABLE 1 | Configuration of stacking gel and separation gel.

Ingredients	Stacking gel	10% Separation gel
Double distilled water	2.6×10^3	3×10^3
30% polyacrylamide liquid	0.64×10^3	3×10^3
1.0 mol/L tromethamine (Tris, pH8.8)	2.3×10^3	2.3×10^3
10% sodium dodecyl sulfate (SDS)	0.03×10^3	0.1×10^3
10% ammonium persulfate	0.04×10^3	0.1×10^3
Tetramethylethylenediamine (TEMED)	0.004×10^3	0.004×10^3

Eventually, 40 μL of the sample was loaded on each well of the electrophoresis gel. Under 80V voltage, the bromophenol blue formed a straight line in the gel, and then the voltage was changed to 120V. When the bromophenol blue ran to the lower edge, the power supply was disconnected, and the membrane was transferred. The membrane transfer process is as follows: soak the glue in the transfer buffer for 10 min, cut six pieces of membrane and filter paper according to the size of the glue, put the transfer buffer for 10 min, place each layer in the order of sponge/3 layers of filter paper/glue/membrane/3 layers of filter paper/sponge, and drive away the bubbles with a test tube. Then put the transfer tank into the ice bath, put the above interlayer, add transfer buffer, and insert the electrode, 100V for 1 h.

After the membrane transfer was completed, the gel image processing system (Unverbindlicher Verkaufspreis, Germany) was used to analyze the target band's molecular weight and net optical density. The relative expression of target protein = target band gray value OD/internal reference gray value OD.

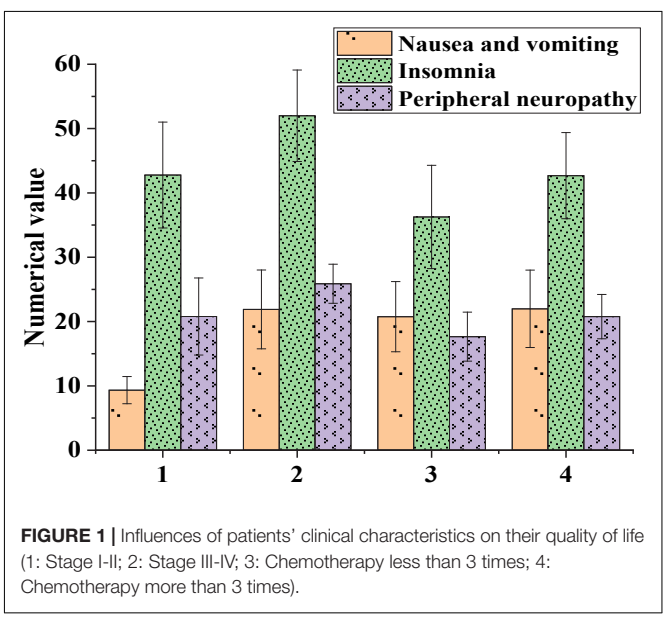


FIGURE 1 | Influences of patients' clinical characteristics on their quality of life (1: Stage I-II; 2: Stage III-IV; 3: Chemotherapy less than 3 times; 4: Chemotherapy more than 3 times).

TABLE 2 | The log_2FC values of first 10 up-regulated genes and first 10 down-regulated genes with relatively large differences.

Oncogenes	Names	log_2FC	*P* value
Up-regulated genes	COL10A1	3.9864	9.69e-32
	COL11A1	3.7236	5.95e-22
	CST1	2.9661	2.90e-23
	CTHRC1	3.2530	9.06e-26
	GREM1	2.8852	2.75e-15
	HS6ST2	3.4452	7.12e-22
	MMP1	3.0086	5.06e-13
	MMP12	3.1327	1.54e-17
	SPINK1	3.3138	1.06e-14
	TOX3	2.8138	2.69e-20
Down-regulated genes	AGER	−3.8451	3.62e-35
	CLDN18	−3.2612	9.98e-19
	FCN3	−3.4334	1.01e-22
	GKN2	−3.2586	3.43e-20
	GPM6A	−3.6053	2.97e-31
	TMEM100	−3.5239	6.07e-21
	SCGB1A1	−3.3605	4.04e-10
	SFTPC	−3.3127	3.23e-15
	SOSTDC1	−3.3652	5.19e-19
	WIF1	−3.7317	2.09e-17

RESULTS AND DISCUSSION

Influence of Patients' Clinical Characteristics on Their Quality of Life

Figure 1 presented the basic clinical characteristics of 60 patients.

Figure 1 suggested that patients in stage III-IV had more severe symptoms, including nausea, vomiting, insomnia, and peripheral neuropathy, than patients in stage I-II. Patients who received more than three chemotherapies had more severe nausea, vomiting, insomnia, and peripheral neuropathy than those who received less than three chemotherapies. This indicated that the more times the chemotherapy patients had, the greater the side effects of the body were. How to make cancer patients achieve the best therapeutic effect within the minimum number of chemotherapy is not only a difficult problem of anti-cancer treatment, but also a key research direction.

Lung Cancer Differential Expression Gene Analysis Results

A total of 875 differentially expressed genes, including 291 up-regulated genes and 584 down-regulated genes, were obtained with FDR $\leq$ 0.05 and log 2 FC $\geq$ 1 criteria. Among these genes, the first 10 up-regulated genes and the first 10 down-regulated genes were shown in **Table 2** and **Figure 2**. The first 10 up-regulated genes were, respectively COL10A1 (log_2FC = 3.9864, $P \leq 0.01$), COL11A1 (log_2FC = 3.7236, $P \leq 0.01$), CST1 (log_2FC = 2.9661, $P \leq 0.01$), CTHRC1 (log_2FC = 3.2530, $P \leq 0.01$), GREM1 (log_2FC = 2.8852, $P \leq 0.01$), HS6ST2 (log_2FC = 3.4452, $P \leq 0.01$), MMP1 (log_2FC = 3.0086, $P \leq 0.01$), MMP12 (log_2FC = 3.1327, $P \leq 0.01$), SPINK1 (log_2FC = 3.3138, $P \leq 0.01$), and TOX3 (log_2FC = 2.8138, $P \leq 0.01$), as shown in **Figure 2A**. The first 10 down-regulated genes were, respectively, AGER (log_2FC = −3.8451, $P \leq 0.01$), CLDN18 (log_2FC = −3.2612, $P \leq 0.01$), FCN3 (log_2FC = −3.4334, $P \leq 0.01$), GKN2 (log_2FC = −3.2586, $P \leq 0.01$), GPM6A (log_2FC = −3.6053, $P \leq 0.01$), TMEM100 (log_2FC = −3.5239, $P \leq 0.01$), SCGB1A1 (log_2FC = −3.3605, $P \leq 0.01$), SFTPC (log_2FC = −3.3127, $P \leq 0.01$), SOSTDC1 (log_2FC = −3.3652, $P \leq 0.01$), and WIF1 (log_2FC = −3.7317, $P \leq 0.01$), as shown in **Figure 2B**.

Up-Regulated Gene Signal Analysis Network

The up-regulated gene COL11A1 was taken as an example; the types of its signal transduction molecules were counted (**Figure 2**).

According to **Figure 3**, the inhibitory conduction signals in normal human tissues were lower than those in lung cancer tissues. In comparison, the activating conduction signals in lung cancer tissues were generally higher than those in normal tissues. This suggested that COL11A1 was involved in the molecular driving mechanism of lung cancer. Next, the types of transferred molecules of COL11A1 were analyzed (**Figure 4**).

As shown in **Figure 4**, the types of metastatic molecules of inhibitory COL11A1 in normal human tissues were lower than

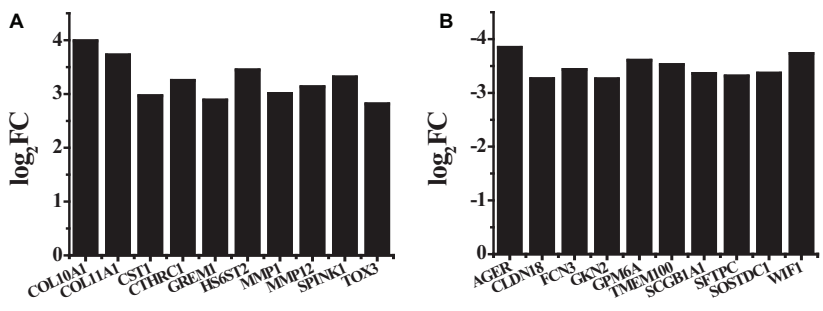


FIGURE 2 | The log_2FC values of first 10 up-regulated genes and first 10 down-regulated genes that had relatively large differences **(A)** the first 10 up-regulated genes; **(B)** the first 10 down-regulated genes.

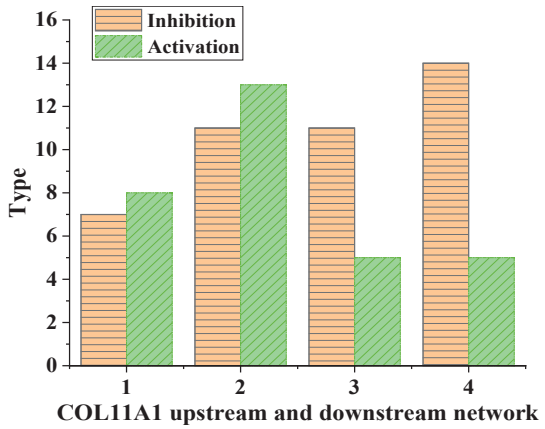


FIGURE 3 | COL11A1 upstream and downstream network signal transduction in different tissues (1 and 3: normal human tissue; 2 and 4: lung cancer tissue).

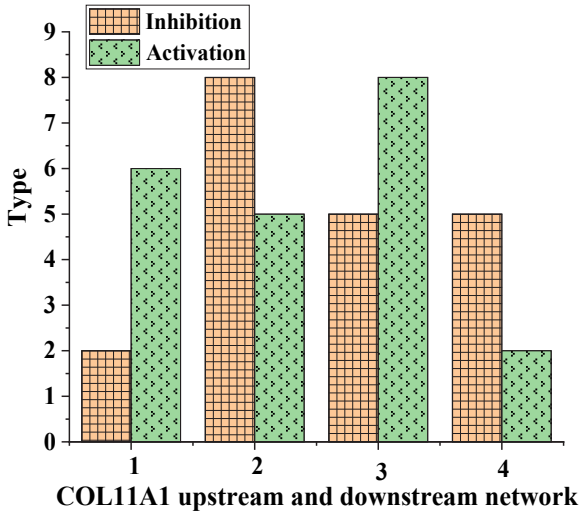


FIGURE 4 | Metastatic molecular types of COL11A1 in different tissues (1 and 3: normal human tissue; 2 and 4: lung cancer tissue).

those in lung cancer tissues. In contrast, the types of metastatic molecules of activating COL11A1 in lung cancer tissues were more than those in normal tissues. This showed that COL11A1 was very metastatic in lung cancer tissues.

Signal Pathway Analysis and Biological Function Enrichment Results

The obtained 291 up-regulated genes and 584 down-regulated genes were input into the DAVID platform for signal pathway analysis and biological function enrichment. The results showed that the differentially expressed genes were enriched in 435, with statistically significant differences ($P \leq 0.01$). The expressed genes were enriched in 166, and the results of GO analysis of the top ten up-regulated genes and the top ten down-regulated genes were shown in **Table 3** and **Figure 5**. Among the up-regulated genes (see **Figure 5A**), BP contains 53 genes with pathway IDs of GO.0030574, GO.0006508, GO.0030199, and GO.0000281; CC contains pathway IDs of GO.0005615, GO.0005576, GO. 203 genes of 0070062 and GO.0005581; MF contains 23 genes with pathway IDs of GO.0004252 and GO.0004556. Among the down-regulated genes (see **Figure 5B**), BP contains 134 genes with pathway IDs of GO.0030593, GO.0006954, GO.0006955, GO.0001525, and GO.0050729; CC contains pathway IDs of GO.0005615, GO. 425 genes of 0005576, GO.0005578 and GO.0005886; MF contains 23 genes with pathway ID GO.0008201. For BP, the up-regulated genes mainly occurred in the process of protein metabolism, proteolysis, mitosis, and cell division. The down-regulated genes were mainly reflected in neutrophilic granulocyte chemoattractant, inflammatory reaction, immune response, and angiogenesis. In the process. For CC, up-regulated genes were mainly enriched in extracellular and collagen trimers and down-regulated genes were mainly enriched outside the cell. For MF, the up-regulated genes were mainly expressed in active serine endonuclease and α-amylase, while the down-regulated genes did not show significant enrichment.

The KEGG pathway analyzed the biological functions of genes from the system level through abundant pathway information, including many complex biological functions such as genetic information transmission, metabolic pathways, and cellular processes. From the annotation analysis of a single gene to the annotation analysis of a gene set, it is judged whether a group of genes appears on a functional node. The KEGG pathway analysis identifies biological processes most relevant to biological phenomena and greatly enhances the reliability of the survey. The results of the KEGG pathway analysis of differentially expressed up-regulated genes and differentially expressed down-regulated genes were shown in **Table 4**. As shown in **Figure 6**, the up-regulated genes include six genes (see **Figure 6A**) with pathway ID 00500: AMY1A, AMY1B, AMY1C, AMY2A, AMY2B, and PGM2L1; 11 genes of 04110: BUB1B, CCNB1, CDC20, CDK1, CDKN2A, MAD2L1, MCM2, ORC6, PTTG1, SFN, and TTK; 7 genes with pathway ID 04115: CCNB1, CDK1, CDKN2A, IGFBP3, RRM2, SFN, and STEAP3 10 genes with pathway ID: 04512: COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, COL11A1, COMP, HMMR, SPP1, and THBS2; 10 genes with pathway ID 04974: ACE2, COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, COL11A1, DPP4, KCNN4, and KCNK5. Down-regulated genes include 16 genes (see **Figure 6B**) with pathway ID 04062: ARRB1, CCL2, CCL4, CCL14, CCL21, CCL23, CXCL3, CXCL12, CXCR2, CX3CL1, ELMO1, FGR, GNG11, PLCB4, PPBP, and PREX1; pathway ID is 04514 Genes: CADM1, CDH5, CD274, CLDN5, CLDN18, CLDN22, ESAM, ICAM1, ICAM2, PECAM1, PTPRM, and SELP; 13 genes with pathway ID 04668: CCL2, CXCL3, CX3CL1, EDN1, FOS, ICAM1, IL1B, IL18R1, IL6, JUNB, MAP3K8, PTGS2, and TNFAIP3; 7 genes with pathway ID 05143: ICAM1, IL1B, IL6, HBA1, HBA2, HBB, and PLCB4; 14 genes with pathway ID 05144: CCL2, CD36, CSF3, GYPC, HBA1, HBA2, HBB, ICAM1, IL1B, IL6, KLRB1, PECAM1, SELE, and SELP. The up-regulated genes were mainly enriched in cell cycle, extracellular matrix receptor interaction, protein digestion and absorption,

TABLE 3 | The GO analysis results of the first 10 up-regulated genes and first 10 down-regulated genes.

Oncogenes	Gene GO classification	Pathway ID	Pathway description	The quantity of genes	*P* value
Up-regulated gene	BP	GO.0030574	Collagen catabolism	14	1.06e-10
	BP	GO.0006508	Proteolysis	26	8.45e-6
	BP	GO.0030199	Collagen fiber tissue	7	1.62e-6
	BP	GO.0000281	Mitosis and cell division	6	4.63e-5
	CC	GO.0005615	Extracellular	53	9.56e-11
	CC	GO.0005576	Extracellular	55	7.45e-9
	CC	GO.0070062	Extracellular	83	1.56e-7
	CC	GO.0005581	Collagen trimer	12	9.63e-7
	MF	GO.0004252	Active serine endonuclease activity	19	1.45e-8
	MF	GO.0004556	α-amylase	4	2.74e-7
Down-regulated gene	BP	GO.0030593	Neutrophilic granulocyte chemoattractant	19	1.07e-13
	BP	GO.0006954	Inflammatory response	38	3.14e-11
	BP	GO.0006955	Immunological reaction	37	1.01e-11
	BP	GO.0001525	Angiogenesis	26	9.45e-8
	BP	GO.0050729	Inflammatory reaction	14	4.06e-8
	CC	GO.0005615	Extracellular	106	9.88e-17
	CC	GO.0005576	Extracellular	108	1.04e-18
	CC	GO.0005578	Extracellular matrix	33	1.16e-13
	CC	GO.0005886	Cytoplasm membrane	178	6.05e-8
	MF	GO.0008201	Heparin binding	23	4.33e-8

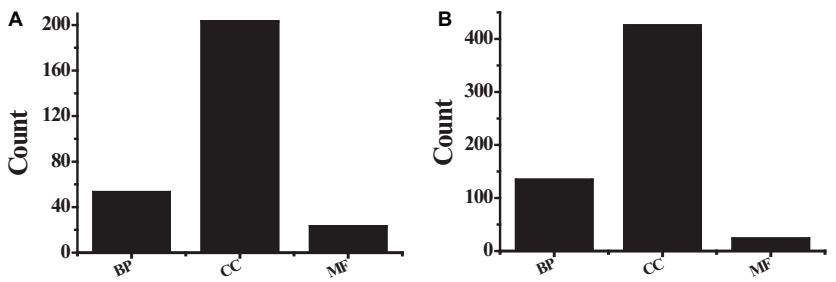


FIGURE 5 | The GO analysis results **(A)** the first 10 up-regulated genes; **(B)** the first 10 down-regulated genes.

p53 signaling pathway, starch and sucrose metabolism, and down-regulated genes were mainly enriched in chemokine signal transduction pathways, malaria, TNF (tumor necrosis factor) signaling pathway, cell adhesion molecules, African trypanosomiasis pathway.

PPI Analysis Results

A 292 nodes and 1,425 interaction networks were obtained from 291 up-regulated genes, and 529 nodes and 1,624 interaction networks were obtained from 584 down-regulated genes by analyzing the string tool. After processing with visualization software, the significant module in the protein-protein interaction relationship network in **Figure 7** was obtained, and the high expression in the center of the protein-protein interaction network was selected from the protein-protein interaction network. The gene (see **Figure 7A**), included TOP2A (Degree = 62), CCNB1 (Degree = 57), CCNA2 (Degree = 54), CDK1 (Degree = 55), and TTK (Degree = 51), all of which have larger mutual Acting relationship. A low-expression gene at the center of the protein-protein interaction network (see **Figure 7B**), including IL6 (Degree = 89), IL1B (Degree = 60), CCL1 (Degree = 58), EDN1 (Degree = 53), and FGF2 (Degree = 51) had a large interaction relationship. These highly expressed genes and low expressed genes may be key target genes related to lung cancer diseases.

Protein Expressions of High-Expressed and Low-Expressed Genes

Protein expressions of the high-expressed genes CCNB1 and TOP2A were illustrated in **Figure 8** below:

Afterward, the expression of messenger RNA corresponding to CCNB1 and TOP2A proteins was analyzed, and the results were shown in **Figure 9** below.

As shown in **Figures 8, 9**, CCNB1 and TOP2A proteins corresponding to the messenger RNA expression level in normal humans were around 10, while the CCNB1 and TOP2A protein corresponding to the messenger RNA expression level in the patient group both exceeded 35, indicating that CCNB1 and TOP2A proteins were highly expressed in patients.

TABLE 4 | The KEGG pathway analysis results of differently expressed up-regulated genes and differently expressed down-regulated genes.

Oncogenes	Pathway ID	Amount of genes	Names	Pathway description	*P* value
Up-regulated genes	00500	6	AMY1A, AMY1B, AMY1C, AMY2A, AMY2B, PGM2L1	Starch and sucrose metabolism	2.34e-4
	04110	11	BUB1B, CCNB1, CDC20, CDK1, CDKN2A, MAD2L1, MCM2, ORC6, PTTG1, SFN, TTK	Cell cycle	1.06e-5
	04115	7	CCNB1, CDK1, CDKN2A, IGFBP3, RRM2, SFN, STEAP3	p53 signaling pathway	9.95e-4
	04512	10	COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, COL11A1, COMP, HMMR, SPP1, THBS2	Extracellular matrix receptor interaction	2.01e-5
	04974	10	ACE2, COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, COL11A1, DPP4, KCNN4, KCNK5	Digestion and absorption of protein	2.78e-6
Down-regulated genes	04062	16	ARRB1, CCL2, CCL4, CCL14, CCL21, CCL23, CXCL3, CXCL12, CXCR2, CX3CL1, ELMO1, FGR, GNG11, PLCB4, PPBP, PREX1	Chemokine signaling transduction pathway	1.37e-5
	04514	12	CADM1, CDH5, CD274, CLDN5, CLDN18, CLDN22, ESAM, ICAM1, ICAM2, PECAM1, PTPRM, SELP	Cell adhesion molecule	1.46e-3
	04668	13	CCL2, CXCL3, CX3CL1, EDN1, FOS, ICAM1, IL1B, IL18R1, IL6, JUNB, MAP3K8, PTGS2, TNFAIP3	TNF signaling pathway	1.69e-6
	05143	7	ICAM1, IL1B, IL6, HBA1, HBA2, HBB, PLCB4	African trypanosomiasis	1.54e-4
	05144	14	CCL2, CD36, CSF3, GYPC, HBA1, HBA2, HBB, ICAM1, IL1B, IL6, KLRB1, PECAM1, SELE, SELP	Malaria	1.01e-11

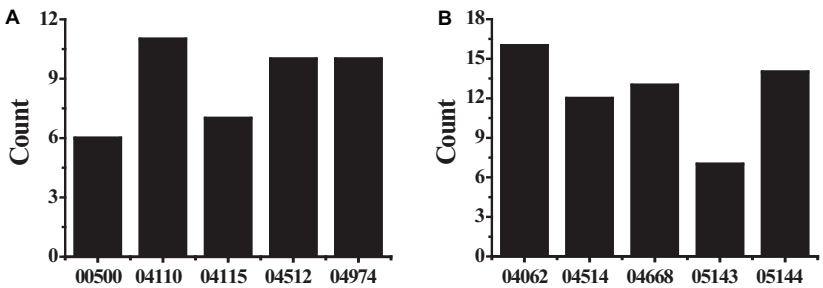


FIGURE 6 | The KEGG pathway analysis results **(A)** showed the differently expressed up-regulated genes; **(B)** showed the differently expressed down-regulated genes.

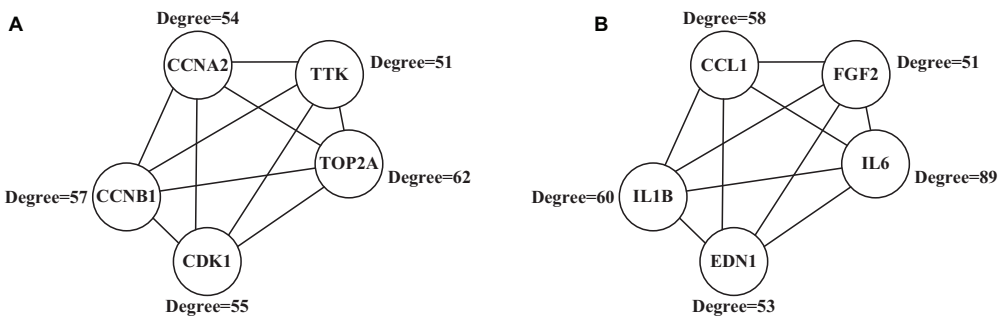


FIGURE 7 | Significant modules in the protein-protein interaction network **(A)** showed the highly expressed genes; **(B)** showed the low expressed genes.

Protein expressions of the low-expressed genes IL6 and IL1B were illustrated in **Figure 10** below:

Then, the messenger RNA expression of IL6 and IL1B proteins was analyzed (**Figure 11**).

As shown in **Figures 10**, **11**, protein expressions of the low-expressed genes IL6 and IL1B in patients were low. The messenger RNA expressions corresponding to IL6 and IL1B proteins in the control group were around 5, while they both exceeded 25 in the patient group. This suggested that even though IL6 and IL1B proteins were low-expressed in patients, they were still lung cancer-related target genes.

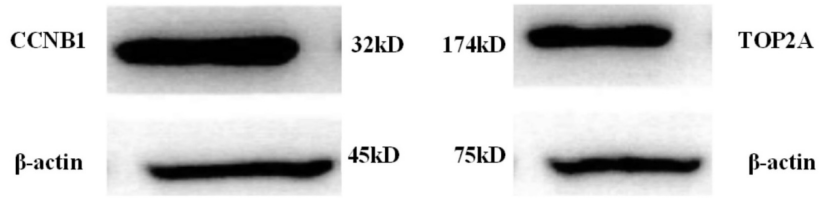


FIGURE 8 | Protein expressions of CCNB1 and TOP2A in lung cancer patients.

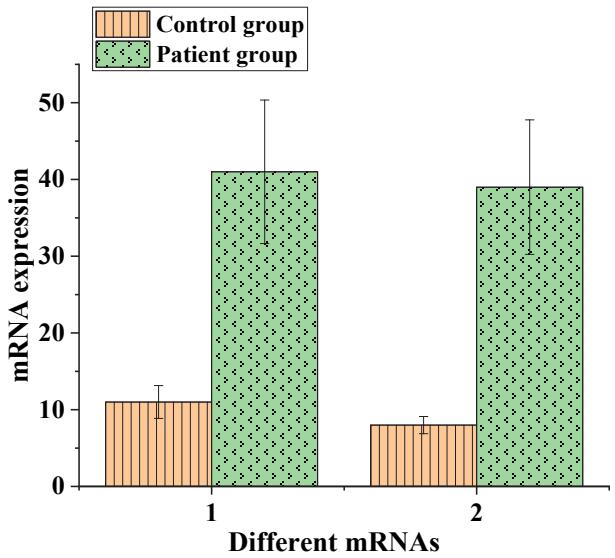


FIGURE 9 | The expression of messenger RNA corresponding to CCNB1 and TOP2A proteins in the control and experimental groups (1: CCNB1; 2: TOP2A).

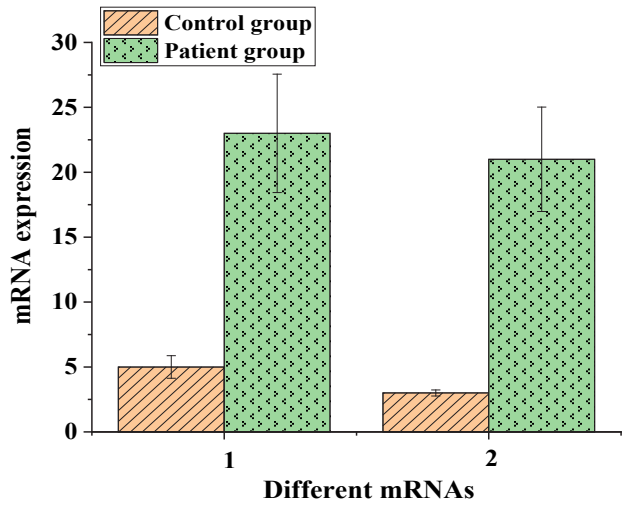


FIGURE 11 | The messenger RNA expressions of IL6 and IL1B proteins in the control and experimental groups (1: IL6; 2: IL1B).

CONCLUSION

This study attempts to reveal the molecular driving mechanism of lung cancer through signal pathway, biological function enrichment, protein interaction analysis, and gene target network analysis. A total of 875 differentially expressed genes were obtained by analyzing the samples. These genes are mainly involved in biological processes such as protein metabolism, protein hydrolysis, mitosis and cell division. TOP2A, CCNB1, CCNA2, CDK1, and TTK may be the key target genes of lung cancer. Exploring the changes of various genes and pathways in the pathogenesis of lung cancer provides reference for the molecular driving mechanism of lung cancer, and provide theoretical basis for molecular-targeted drug therapy and clinical nursing guidance of lung cancer. However, there are still some shortcomings. The selection number of up-regulated and down-regulated genes is limited, which cannot meet the huge molecular network analysis. In the later stage, the screening amount

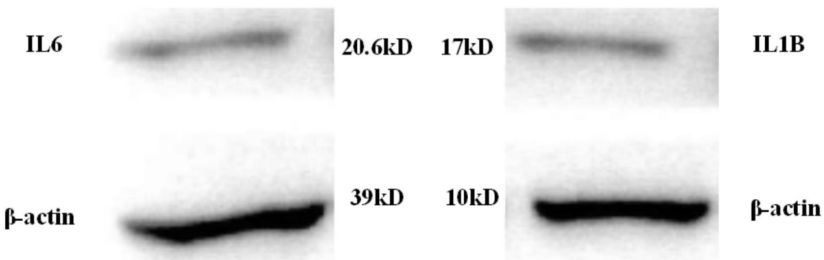


FIGURE 10 | Protein expressions of IL6 and IL1B in patients with lung cancer.

of up-regulated and down-regulated genes will be increased. The molecular driving mechanism of lung cancer was still in the preliminary stage. In the subsequent research, TOP2A with large interaction relations among the critical target genes related to lung cancer obtained by screening would be screened for drug resistance, providing assistance for the development of its inhibitors.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: the GEO (Gene Expression Omnibus) database.

AUTHOR CONTRIBUTIONS

RH: writing – original draft and conceptualization. XX: data curation and software. KZ: supervision and resources. YZ: formal analysis. CW: validation. GH: writing, review, editing, and methodology. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Natural Science Foundation of Zhejiang Province of China (LY21H160011).

REFERENCES

Allan-Blitz, L. T., Hemarajata, P., Humphries, R. M., Wynn, A., Segura, E. R., and Klausner, J. D. A. (2018). Cost analysis of gyrase a testing and targeted ciprofloxacin therapy versus recommended 2-drug therapy for *Neisseria gonorrhoeae* infection. *Sex. Transm. Dis.* 45, 87–91. doi: 10.1097/olq.0000000000000698

Horimasu, Y., Ishikawa, N., Taniwaki, M., Yamaguchi, K., Hamai, K., Iwamoto, H., et al. (2017). Gene expression profiling of idiopathic interstitial pneumonias (IIPs): identification of potential diagnostic markers and therapeutic targets. *BMC Med. Genet.* 18:88. doi: 10.1186/s12881-017-0449-9

Krasnov, G. S., Puzanov, G. A., Kudryavtseva, A. V., Dmitriev, A. A., Beniaminov, A. D., Kondratieva, T. T., et al. (2017). Differential expression of an ensemble of the key genes involved in cell-cycle regulation in lung cancer. *Mol. Biol.* 51, 740–747. doi: 10.1134/s0026893317050107

Masters, G. A., Johnson, D. H., and Temin, S. (2017). Systemic therapy for stage IV non-small-cell lung cancer: american society of clinical oncology clinical practice guideline update. *J. Oncol. Pract.* 33, 832–837. doi: 10.1200/jop.2017.026716

Shen, J., Hu, Y., Putt, K. S., Singhal, S., Han, H., Visscher, D. W., et al. (2018). Assessment of folate receptor alpha and beta expression in selection of lung and pancreatic cancer patients for receptor targeted therapies. *Oncotarget* 9, 4485–4495. doi: 10.18632/oncotarget.23321

Shi, Y., Sun, Y., Yu, J., Ding, C., Wang, Z., Wang, C., et al. (2017). China experts consensus on the diagnosis and treatment of advanced stage primary lung cancer (2016 version). *Zhongguo Fei Ai Za Zhi* 19, 1–15. doi: 10.21147/j.issn.1000-9604.2019.01.01

Wang, L., Qu, J., Liang, Y., Zhao, D., Rehman, F. U., Qin, K., et al. (2020). Identification and validation of key genes with prognostic value in non-small-cell lung cancer via integrated bioinformatics analysis. *Thorac. Cancer* 11, 851–866. doi: 10.1111/1759-7714.13298

Xu, R., Xu, R., Zhong, G., Huang, T., He, W., Kong, C., et al. (2018). Sequencing of circulating tumor DNA for dynamic monitoring of gene mutations in advanced non-small cell lung cancer. *Oncol. Lett.* 15, 3726–3734.

Zhang, Y., Luo, J., Wang, X., Wang, H. L., Zhang, X. L., Gan, T. Q., et al. (2018). A comprehensive analysis of the predicted targets of miR-642b-3p associated with the long non-coding RNA HOXA11-AS in NSCLC cells. *Oncol. Lett.* 15, 6147–6160.

Zhou, G., Ye, J., Fang, Y., Zhang, Z., Zhang, J., Sun, L., et al. (2017). Identification of DBCCR1 as a suppressor in the development of lung cancer that is associated with increased DNA methyltransferase 1. *Oncotarget* 8, 32821–32832. doi: 10.18632/oncotarget.15826

Zins, K., Heller, G., Mayerhofer, M., Schreiber, M., and Abraham, D. (2018). Differential prognostic impact of interleukin-34 mRNA expression and infiltrating immune cell composition in intrinsic breast cancer subtypes. *Oncotarget* 9, 23126–23148. doi: 10.18632/oncotarget.25226

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.

Copyright © 2021 Huang, Xiang, Zhang, Zheng, Wang and Hu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 November 2021
doi: 10.3389/fgene.2021.783128

Predicting Human Protein Subcellular Locations by Using a Combination of Network and Function Features

Lei Chen[1,2†], ZhanDong Li[3†], Tao Zeng[4†], Yu-Hang Zhang[5], ShiQi Zhang[6], Tao Huang[4,7] and Yu-Dong Cai[1]**

[1]School of Life Sciences, Shanghai University, Shanghai, China, [2]College of Information Engineering, Shanghai Maritime University, Shanghai, China, [3]College of Food Engineering, Jilin Engineering Normal University, Changchun, China, [4]Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [5]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [6]Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark, [7]CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

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

Reviewed by:
Yijie Ding,
University of Electronic Science and Technology of China, China
Xuekun Song,
Henan University of Traditional Chinese Medicine, China

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

†These authors have contributed equally to this work

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

Received: 25 September 2021
Accepted: 22 October 2021
Published: 05 November 2021

Citation:
Chen L, Li Z, Zeng T, Zhang Y-H, Zhang S, Huang T and Cai Y-D (2021) Predicting Human Protein Subcellular Locations by Using a Combination of Network and Function Features. Front. Genet. 12:783128. doi: 10.3389/fgene.2021.783128

Given the limitation of technologies, the subcellular localizations of proteins are difficult to identify. Predicting the subcellular localization and the intercellular distribution patterns of proteins in accordance with their specific biological roles, including validated functions, relationships with other proteins, and even their specific sequence characteristics, is necessary. The computational prediction of protein subcellular localizations can be performed on the basis of the sequence and the functional characteristics. In this study, the protein–protein interaction network, functional annotation of proteins and a group of direct proteins with known subcellular localization were used to construct models. To build efficient models, several powerful machine learning algorithms, including two feature selection methods, four classification algorithms, were employed. Some key proteins and functional terms were discovered, which may provide important contributions for determining protein subcellular locations. Furthermore, some quantitative rules were established to identify the potential subcellular localizations of proteins. As the first prediction model that uses direct protein annotation information (i.e., functional features) and STRING-based protein–protein interaction network (i.e., network features), our computational model can help promote the development of predictive technologies on subcellular localizations and provide a new approach for exploring the protein subcellular localization patterns and their potential biological importance.

Keywords: protein subcellular location, protein-protein interaction network, GO enrichment, KEGG enrichment, feature selection, classification algorithm

1 INTRODUCTION

Eukaryotic organisms, such as human beings, have complicated cell structures with delicate functional membrane structures surrounded by effective compartments (Thul et al., 2017; Tjondro et al., 2019). The complicated membrane structures in eukaryotic cells have generally divided the intercellular space into the cytoplasm and the nucleus through the nuclear membrane (Yeagle, 1989; Mangeat et al., 1999). Specific organelles, such as the mitochondria, have a specific and

independent membrane system (Set et al., 2019). The major components of these structures divide the intercellular space into different isolated rooms for independent biological reactions and interactions and restrict the intercellular localizations of proteins (Thul et al., 2017). For instance, the replication of DNA depends on various effective proteins and enzymes. However, some proteins, such as DNA polymerase and DNA ligase, are not synthesized in the nucleus, in which they function (Ganai and Johansson, 2016). Some proteins play a specific role in biological processes in the nucleus (Ganai and Johansson, 2016). Therefore, the subcellular localization controls the protein to some extent to act at the proper localization.

Given the limitation of technologies, the subcellular localizations of proteins are difficult to identify. Therefore, predicting the subcellular localization and the intercellular distribution patterns of proteins in accordance with their specific biological roles, including validated functions, relationships with other proteins, and even their specific sequence characteristics, is necessary. The computational prediction of protein subcellular localizations can be performed on the basis of the sequence and the functional characteristics. Sequence characteristics-based methods can be further divided into three kinds, namely, the N-terminal sorting method, amino acid composition-based prediction, and homology. The N-terminal sorting method is based on subcellular localization prediction. In 2006, researchers from Greece reported a subcellular localization predictor by using the N-terminal signaling sequence of the protein, resulting in a cross-validated accuracy of 87.1% in animals (Petsalaki et al., 2006). The amino acid composition of proteins is easy to determine and describe, but the models that use amino acid composition do not have good prediction performance. Therefore, amino acid compositions are generally used to accompany other characteristics, such as N-terminal sorting and homology. The homology considers another important feature subgroups of sequence characteristics. Predictors, such as the Proteome Analyst (Szafron et al., 2004) and the PairProSVM (Mak et al., 2008), have been validated to have a good performance for protein subcellular localization prediction. Recently, some advanced computational methods, such as deep learning, multiple kernel learning, etc. are adopted to learn features derived from protein sequence and set up prediction models (Wei et al., 2018; Ding et al., 2020).

Apart from the above sequence-based prediction methods, predicting the subcellular localization of proteins by using the functional annotation and correlations between proteins has attracted attention due to the accomplishment of human protein function annotation and the establishment of the protein–protein interaction (PPI) network. However, the extraction of protein functional features is quite difficult compared with extracting protein sequencing features. With the development of bioinformatics, the most widely used approaches have been established on the annotation and clustering of the gene ontology (GO) (Consortium, 2015) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Zhang and Wiemann, 2009). In these methods, the GO and the KEGG pathway terms are applied to describe and cluster proteins as optimal protein characteristics. GO has terms on cellular components that describe the general subcellular localization. Some predictors, such as the ProLoc-GO (Huang et al., 2008), the ILoc-Virus (Xiao et al., 2011), and the Cell-PLoc (Chou and Shen, 2008), combine the general description with the sequence characteristics, thereby establishing a novel and effective prediction method on subcellular localization. However, the functional annotation of proteins remains imperfect, and potentially new functions of proteins emerge. Therefore, additional methods should be presented to supplement current research.

In this study, the comprehensive PPI network provided by STRING (Szklarczyk et al., 2016) and GO/KEGG pathway annotations on proteins were employed to analyze the current proteins with known subcellular localizations. Qualitative and quantitative predictive models were established to identify the potential subcellular localizations of proteins on the basis of several machine learning algorithms, such as feature selection methods, classification algorithms. In addition to models, we also obtained some key proteins and functional terms that may provide important contributions for determining protein subcellular locations. As the first prediction model that used direct protein annotation information (i.e., functional features) and the STRING-based PPI network (i.e., network features), our computational model can help promote the development of predictive technologies on subcellular localizations and provide a new approach for exploring the protein subcellular localization patterns and their potential biological importance.

TABLE 1 | Number of proteins in each category.

Index	Category	Number of proteins
Class 1	Biological membrane	1,487
Class 2	Cell periphery	35
Class 3	Cytoplasm	506
Class 4	Cytoplasmic vesicle	70
Class 5	Endoplasmic reticulum	190
Class 6	Endosome	25
Class 7	Extracellular space or cell surface	649
Class 8	Flagellum or cilium	3
Class 9	Golgi apparatus	98
Class 10	Microtubule cytoskeleton	48
Class 11	Mitochondrion	345
Class 12	Nuclear periphery	33
Class 13	Nucleolus	112
Class 14	Nucleus	1,285
Class 15	Peroxisome	46
Class 16	Vacuole	54

2 MATERIALS AND METHODS

2.1 Data

The data used in this study were extracted from the Swiss-Prot (http://cn.expasy.org/, release 54.0) by searching the proteins annotated with "subcellular location". Initially, 53,427 protein sequences were downloaded. Proteins with length shorter than 50 amino acids (e.g., protein fragments) and those with length longer

than 5,000 amino acids (e.g., protein complexes) were excluded. Proteins containing unknown amino acid abbreviation, such as X, were also excluded. Protein sequences with high degree of similarity were also removed using the program CD-HIT (Li and Godzik, 2006) and a cutoff value of 0.7. Finally, only human proteins were studied. Thus, 4,986 protein sequences remained after these exclusions and were classified into 16 categories (**Table 1**).

2.2 Feature Representation

Good representation of proteins is very important to build efficient models for identification of human protein subcellular locations. In this study, each protein was represented by three groups of features, where one group was derived from PPI network, two groups were extracted from functional terms (GO and KEGG pathway). Their descriptions are as follows.

2.2.1 Network Features Derived From PPI Network

The initial PPI network was downloaded from STRING (version 9.0) (Szklarczyk et al., 2011) (http://string.embl.de/), which contained known and predicted protein interaction. The interaction network considers proteins as its nodes and has an edge between two proteins if they can interact with each other. Furthermore, each edge was assigned a weight, which was defined as the confidence score of the corresponding interaction. As such score was obtained by considering several aspects of proteins, it can widely measure the associations of proteins. Given a protein, a feature vector was constructed, where each component indicated a protein in the PPI network. Each component was defined as the confidence score of the interaction between the protein and the corresponding protein of such component. If such interaction did not exist, the component was set to zero. For an easy description, these features were called network features. As there were 20,770 proteins in the PPI network, 20,770 network features were generated for each protein.

2.2.2 Functional Features Derived From KEGG Pathway

The immediate neighborhood method is usually used for predicting the function of a query protein on the basis of the other proteins with known functions (Sharan et al., 2007). A query protein interacts with many neighboring proteins in the STRING network (Szklarczyk et al., 2011). With these neighboring proteins, we can assess the relationship between the query protein and one KEGG pathway. Let the neighboring proteins and the query protein constitute a protein set *PS*. For a KEGG pathway, proteins in such pathway comprised another protein set, denoted by *KP*. The relationship between the query protein and the KEGG pathway, called KEGG enrichment score, was defined as the –log10 of the hypergeometric test p value (Carmona-Saez et al., 2007; Cai et al., 2010) on above-constructed protein sets. All obtained enrichment scores on all KEGG pathways were collected in a vector, comprising the functional KEGG features of the protein. 297 KEGG pathways were considered, inducing 297 functional KEGG features.

2.2.3 Functional Features Derived From GO

Similarly, the relationship between the query protein and one GO term can be obtained. For a GO term, let *GP* be a protein set consisting of proteins annotated by such GO term. The relationship was defined as the –log10 of the hypergeometric test p value (Cai et al., 2010; Li et al., 2012) on *PS* and *GP*. The obtained value was called GO enrichment score. Likewise, GO enrichment scores on all GO terms were collected in a vector, constituting the functional GO features of the query protein. 20,681 GO terms were involved, generating 20,681 functional GO features.

2.3 Boruta Feature Filtering

The Boruta feature filtering (Kursa and Rudnicki, 2010) can screen features that are relevant to target sample labels on the basis of the random forest (RF) in a wrapper manner. The Boruta feature filtering iteratively identifies key features by comparing the importance scores that correspond to the real and the shuffled features. The Boruta approach has three steps: 1) copying the training data and shuffling the feature values for new shuffled data to be produced; 2) training the RF classifier on the produced shuffled data and calculating the importance score for each feature; and 3) evaluating the importance score of each feature in the original training data and removing the real features with remarkably lower importance scores than the shuffled features. By executing the above steps with a few iterations, Boruta approach selects the relevant features.

This study adopted the Boruta program retrieved from https://github.com/scikitlearn-contrib/boruta_py. For convenience, it was performed with its default parameters.

2.4 Minimum Redundancy Maximum Relevance

The mRMR (Peng et al., 2005) can select and rank informative features in accordance with the following assumptions. On the one hand, the mRMR selects features with minimum redundancy among themselves. On the other hand, the mRMR selects features with maximum relevance with class labels. Therefore, the mRMR only selects the features that satisfy minimum redundancy and maximum relevance simultaneously by using mutual information. These features are important in distinguishing the class labels for follow-up classification modeling. In fact, two feature lists can be obtained through the mRMR method. The MaxRel feature list ranks features based on their relevance to class labels, whereas the mRMR feature list sorts features by further considering the redundancies among features. Evidently, from the mRMR feature list, we can obtain a compact feature subspace for a given classification algorithm. Thus, this study only adopted the mRMR feature list.

The present study used the mRMR program downloaded from http://home.penglab.com/proj/mRMR/. Likewise, default parameters were adopted to execute such program.

2.5 Incremental Feature Selection

IFS, an ordered feature selection approach (Liu and Setiono, 1998), can determine the best number of selected features in an iteration manner. The IFS first constructs a series of feature subsets from the ranked features supplied by a feature ranking (e.g., mRMR feature list). For instance, the first feature subset consists of the top 10 features, and the second feature subset consists of the top 20 features, and so on. Next, the IFS trains a model on the training samples, which consist of features from each feature subset, based on a given classification algorithm. Such classification model performance is evaluated by 10-fold cross-validation (Kohavi, 1995). Finally, the model with the highest performance is found out, which was called the optimum model. The feature subset used in this model was called the optimum feature subset.

2.6 Classification Algorithm

To execute the IFS method, one classification algorithm is necessary. This study tried four classification algorithms: 1) RF (Breiman, 2001), 2) Support vector machine (SVM) (Cortes and Vapnik, 1995), 3) k-nearest neighbor (kNN) (Cover and Hart, 1967), 4) Decision tree (DT) (Swain and Hauska, 1977). These algorithms have been widely used to tackle various biological problems (Jia et al., 2020; Zhou et al., 2020; Chen et al., 2021; Pan et al., 2021; Yang and Chen, 2021; Zhang et al., 2021a; Zhang et al., 2021b; c).

2.6.1 Random Forest

RF builds an assemble classification algorithm depending on many tree classifiers. The predicted sample label/category of RF is determined using multiple tree classifiers by an aggregating vote. Notably, RF usually adopts the final consensus results in accordance with the average of all decision trees' predictions, aiming to avoid overfitting and improve the performance robustness of learned models because a subtle difference among decision trees exists in RF. To quickly implement RF, the tool "RandomForest" in Weka (https://www.cs.waikato.ac.nz/ml/weka/) (Hall et al., 2009) was employed. Default parameters were used.

2.6.2 Support Vector Machine

As a classification algorithm based on statistical learning theory, the SVM can map samples to a given category. The SVM transforms samples from a low-dimensional space to a high-dimensional space by using a kernel function (e.g., Gaussian kernel) and can divide samples of each label/category by maximizing the data interval in high-dimensional space. The SVM can further predict the test samples' label/category in accordance with the interval to which this new sample belongs. In this study, we used the SVM optimized by the sequence minimization optimization (SMO) (Platt, 1998a; Platt, 1998b) algorithm. This type of SVM is implemented by the tool "SMO" in the Weka (https://www.cs.waikato.ac.nz/ml/weka/) (Hall et al., 2009).

2.6.3 k-Nearest Neighbor

The kNN builds a classification model by using a voting scheme (Theilhaber et al., 2002; Zhang and Srihari, 2004; Yu et al., 2016). In the sample space, the class labels of the kNNs of a given sample were used to produce a predicted class label for a new sample. In the learning of kNN classification model, the nearest neighbors are selected from the training data, where k is a given parameter that usually ranges from 1 to 10. Briefly, the kNN includes several calculation steps: 1) calculating the distance between the test sample and all the training samples in the feature space; 2) ranking the training samples in accordance with their distance with the test sample; 3) selecting the k training samples with least distance to the test sample (i.e., *kNNs*); 4) determining the distribution of class labels of the k nearest training samples; and 5) using the class label with highest distribution frequency as the predicted class label for the test sample. The tool "IBK" in Weka (https://www.cs.waikato.ac.nz/ml/weka/) (Hall et al., 2009) implements the kNN algorithm, which was directly employed in this study.

2.6.4 Decision Tree

The DT can produce interpretative rules that easily explain the classification and the regression models for wide applications in many research fields. In brief, DT is a nonparametric supervised learning method and uses a white box model with the IF-TEHN format to provide definite indications of individual features for classification and regression. A common construction strategy of DT is greedy algorithm, which achieves satisfactory performance with reasonable computational cost. The corresponding pack collected in Scikit-learn (https://scikit-learn.org/stable/) (Pedregosa et al., 2011), which implements an optimized version of the CART algorithm with the Gini index, was used to build DT model in this study.

2.7 Synthetic Minority Oversampling Technique

Table 1 shows that the analyzed data were unbalanced numbers of samples with different labels (i.e., different classified categories). Thus, the SMOTE (Chawla et al., 2002) was applied. It can produce new samples for the minor sample classes iteratively until the sample numbers of these minor sample classes are equivalent to that of the major sample class. The balanced data can improve the construction of the classification models. In this study, we used the tool "SMOTE" in the Weka (https://www.cs.waikato.ac.nz/ml/weka/) (Hall et al., 2009), which implements SMOTE method. Samples generated by SMOTE were not used in the methods of Boruta and mRMR because these newly added samples may influence the results of these two methods, which cannot fully reflect actual distribution of subcellular locations of proteins.

2.8 Performance Evaluation

In this study, the Matthew correlation coefficients (MCC) (Matthews, 1975) within 10-fold cross-validation (Kohavi, 1995) was used to evaluate the prediction performance of each classification model. MCC is a commonly used measurement and ranges between −1 and +1, achieving +1 when the classification model has the best performance. The multiclass version of MCC

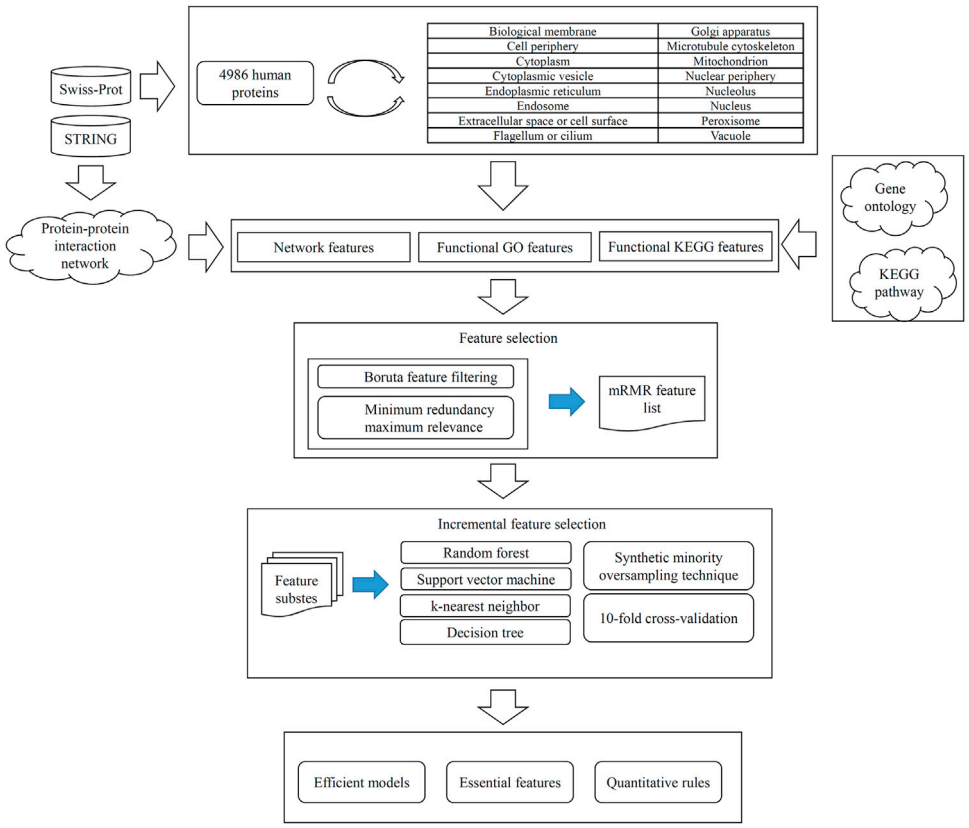


FIGURE 1 | Entire procedures for constructing and evaluating protein subcellular location prediction models. Human proteins and their subcellular location information are retrieved from Swiss-Prot. Each protein is represented by three feature groups: network features, functional KEGG features, and functional GO features. All features are analyzed by Boruta and minimum redundancy maximum relevance one by one, resulting in an mRMR feature list. Such list is fed into the incremental feature selection method, incorporating four classification algorithms, synthetic minority oversampling technique and 10-fold cross-validation, to build efficient models, extract essential features and access quantitative rules.

is proposed by Gorodkin (Gorodkin, 2004). Our analyzed data contained 16 categories, and MCC was calculated as follows:

$$MCC = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{cov}(X, X)\mathrm{cov}(Y, Y)}} \quad (1)$$

where $\mathrm{cov}(\cdot, \cdot)$ represents the covariance of two matrices, X is a 0–1 matrix that indicates the predicted class of each sample, and Y is a 0–1 matrix that represents the actual classes of all samples.

Besides, the performance of each constructed model was also evaluated by other measurements, including individual accuracy on each category and overall accuracy.

3 RESULTS

In this study, we conducted a computational investigation on identification of human protein subcellular locations. The entire procedures are illustrated in **Figure 1**. Detailed results were described in this section.

3.1 Results of Boruta and Minimum Redundancy Maximum Relevance Methods

As described in *Feature Representation*, each protein was represented by lots of network, functional KEGG and functional GO features. The Boruta method was first applied to analyze all features. Irrelevant features were discarded. 4,773 features remained, which are provided in **Supplementary Table S1**. Among these features, 399 were network features, 151 were functional KEGG features, and 4,223 were functional GO features, which are shown in **Figure 2A**. Evidently, functional GO features occupied most features (~88%).

For these 4,773 features, the mRMR method followed to analyze their importance. An mRMR feature list was generated, as listed in **Supplementary Table S1**. We counted ranks of features in each feature group and plotted a boxplot in **Figure 2B**. It can be observed that network features received many high ranks in the mRMR feature list although their quantity was not most. This suggested that network features can provide key contributions for determination of protein subcellular locations.

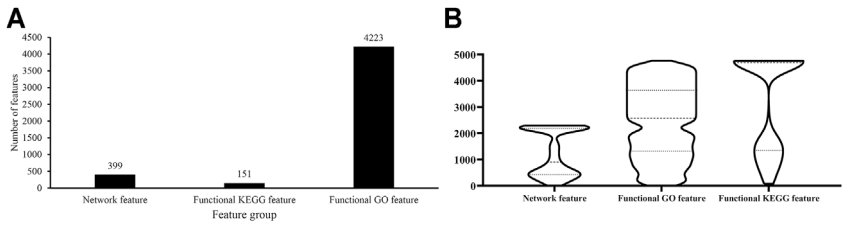


FIGURE 2 | Analysis of features selected by Boruta. **(A)** Distribution of features selected by Boruta on three feature groups; **(B)** Violin plot to show ranks of features in three feature groups which are obtained by mRMR method.

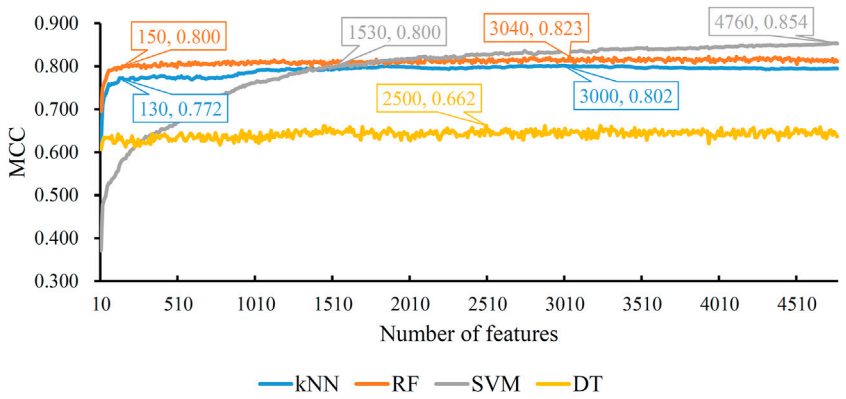


FIGURE 3 | IFS with four classification algorithms on mRMR feature list of network and functional features. The highest MCC values obtained by four classifications are 0.802, 0.823, 0.854, and 0.662, respectively. kNN, RF, and SVM can yield quite good performance when much less features are adopted.

TABLE 2 | Performance of key models for identification of human protein subcellular locations.

Classification algorithm	Number of features	Overall accuracy	MCC
k-nearest neighbor	3,000	0.830	0.802
	130	0.805	0.772
Random forest	3,040	0.852	0.823
	150	0.833	0.800
Support vector machine	4,760	0.879	0.854
	1,530	0.833	0.800
Decision tree	2,500	0.716	0.662

3.2 Results of IFS Method

Based on the mRMR feature list, the IFS method was executed. 477 feature subsets were constructed with step 10. On each feature subset, a model was built based on each of the four classification algorithms (RF, kNN, SVM, and DT). The model was further evaluated by 10-fold cross-validation. The evaluation results, including MCC, overall accuracy and individual accuracies on 16 categories, for RF, kNN and SVM are listed in **Supplementary Table S2**. For an easy observation, an IFS curve was plotted for each classification algorithm, which is shown in **Figure 3**. For kNN, the highest MCC was 0.802, which was obtained by using top 3,000 features in the mRMR feature list. Thus, we can construct an optimum kNN model with these features. The overall accuracy of such model was 0.830 (**Table 2**). For RF, it produced the highest MCC of 0.823 when top 3,040 features were adopted, thereby building the optimum RF model with these features. The overall accuracy of such model was 0.852 (**Table 2**). As for SVM, the highest MCC was 0.854. This performance was obtained by using top 4,760 features in the list. Accordingly, an optimum SVM model was set up with these features. Its overall accuracy was 0.879 (**Table 2**). Evidently, each optimum model provided good performance with MCC higher than 0.800, suggesting combination of network and functional features can really capture the essential properties of proteins.

Although three optimum models were set up as mentioned above, their efficiencies were not very high because lots of features were used. To build models with high efficiency, we carefully checked the performance of three classification algorithms on different feature subsets. Other three models using much less features were constructed, where the kNN model used the top 130 features, RF model adopted the top 150 features and SVM model used the top 1,530 features (**Figure 3**). Although these models adopted much less features, their performance was only a little lower than those of the optimum models. This fact can be concluded from **Table 2** and **Figure 4**. Thus, these models can be efficient tools for identification of protein subcellular locations.

For DT, we conducted the same IFS procedure. The IFS results are provided in **Supplementary Table S3**, which induced a curve, as shown in **Figure 3**. It can be observed that the highest MCC was 0.662 when top 2,500 features were adopted. Accordingly, we

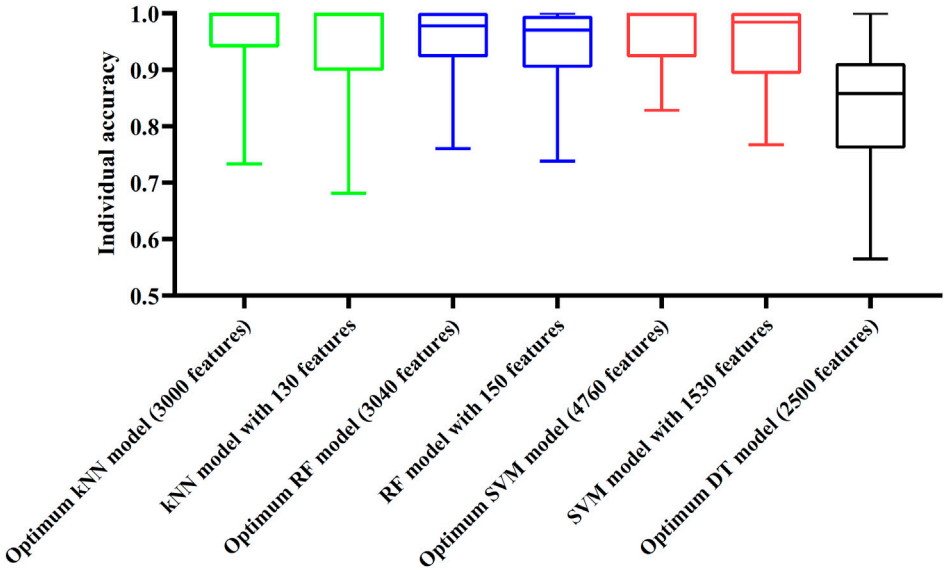


FIGURE 4 | Box plot to show performance of some models on 16 categories. For three classification algorithms (kNN, RF, and SVM), models with much less features can provide similar performance to the optimum models. Optimum DT model yields much lower performance.

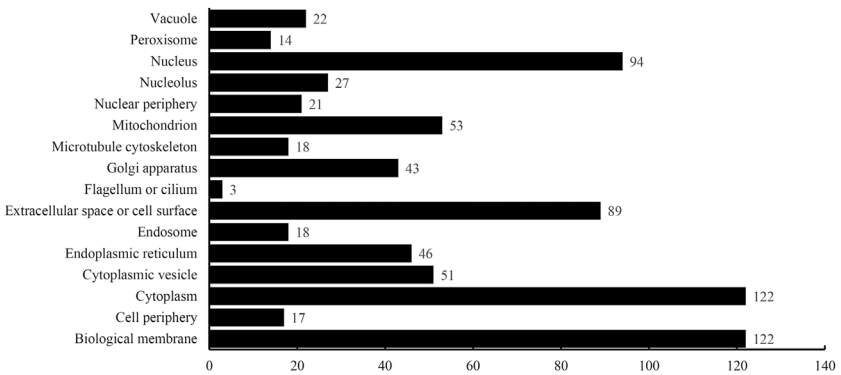


FIGURE 5 | Number of quantitative rules for each of 16 categories.

can set up an optimum DT model using these features. The overall accuracy was 0.716, as listed in **Table 2**. Evidently, such performance was much lower than that of the optimum kNN/RF/SVM model. It was also lower than those of the models with higher efficiency mentioned in the above paragraph. The individual accuracies on 16 categories yielded by this DT model were also obviously lower than those of other models, as shown in **Figure 4**. However, the utility of DT model was not to identify protein subcellular locations. Different from kNN, RF, and SVM, which were complete black-box algorithms, the classification procedures of DT were open. Thus, it can provide much more biological insights than other three classification algorithms.

3.3 Results of Quantitative Rules

The optimum DT model adopted the top 2,500 features in the mRMR feature list. Accordingly, DT was executed on the dataset containing all 4,986 proteins, thereby constructing a big tree. From this tree, 760 quantitative rules were extracted, which are provided in **Supplementary Table S4**. Each of 16 categories was assigned some rules. **Figure 5** shows the number of rules for each of 16 categories. Some categories (e.g., Class 1: Biological membrane, Class 3: Cytoplasm) received more than 100 rules, whereas there were only three rules for Class 8: Flagellum or cilium. In *Quantitative Rules That Contribute to Subcellular Localization Prediction*, several rules would be analyzed.

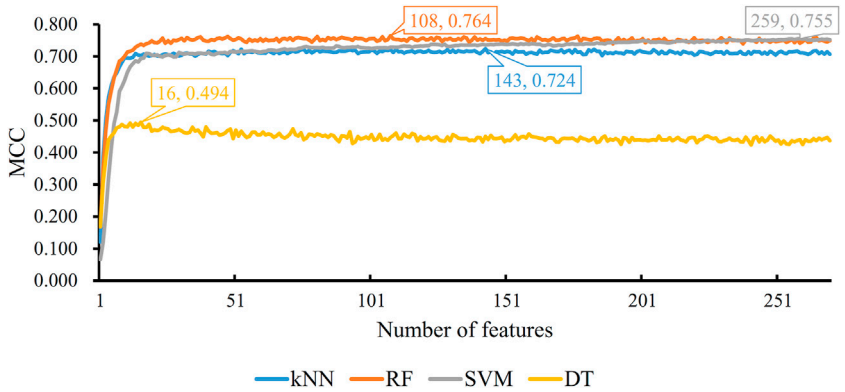


FIGURE 6 | IFS with four classification algorithms on mRMR feature list of PseAAC features. The highest MCC values obtained by four classification algorithms are 0.724, 0.764, 0.755, and 0.494, respectively.

TABLE 3 | Performance of the optimum models using PseAAC features.

Classification algorithm	Number of features	Overall accuracy	MCC
k-nearest neighbor	143	0.757	0.724
Random forest	108	0.803	0.764
Support vector machine	259	0.794	0.755
Decision tree	16	0.559	0.494

3.4 Comparison of the Classic Model

The pseudo-amino acid composition (PseAAC) (Chou, 2001) is a classic protein encoding scheme and has been widely adopted to build models for identification of protein subcellular locations (Cai and Chou, 2003; Pan et al., 2003; Lin et al., 2008; Shi et al., 2008; Liu et al., 2010). Here, we used such scheme to encode each protein mentioned in *Data* and further build models for the comparison of models proposed in this study.

Five physicochemical and biochemical properties of amino acids were employed to generate features, including codon diversity, electrostatic charge, molecular volume, polarity and secondary structure. The weight factor was set to 0.15 and Lambda parameter was set to 50. From each physicochemical and biochemical property, 50 features were extracted. Thus, 250 (50×5) features were obtained for each protein. Furthermore, 20 amino acid composition features were also employed. Accordingly, each protein was represented by 270 (250 + 20) features. These features were directly analyzed by mRMR method, resulting in a feature list. Such list was fed into the IFS method. Likewise, four classification algorithms: kNN, RF, SVM, and DT, were also tried in the IFS method. For each classification algorithm, MCC values obtained on all possible feature subsets are illustrated in **Figure 6**. It can be observed that the highest MCC values for four classification algorithms were 0.724, 0.764, 0.755, and 0.494, respectively, which are also listed in **Table 3**. The corresponding ACC values are also listed in this table. Compared with ACC and MCC values obtained by models using network and functional features (**Table 2**), with the same classification algorithm, our models were superior to models with PseAAC features. It was suggested that network and functional features were more efficient than PseAAC features for identification of protein subcellular locations. These features provided new directions for building more efficient protein subcellular location prediction models.

4 DISCUSSION

A group of effective proteins that may directly contribute to the identification and clustering of different subcellular localizations is screened by using some machine learning models. According to recent publications, the top optimal features have already been validated to contribute to the subcellular localization, validating the efficacy and the accuracy of our predictions. The detailed analyses and discussion can be seen below.

4.1 Features From Proteins That Contribute to Subcellular Localization Prediction

The first feature protein is **SUMO2** (ENSP00000405965). According to recent publications, this protein is a member of the small ubiquitin-like modifier family and contributes to ubiquitin-mediated post-translational modification system by acting as a signal for proteasomal degradation (Hecker et al., 2006; Tammsalu et al., 2014). In 2013, a research on testis functions confirmed that SUMO2 is specifically located in the nucleus region of the cell and is mediated by retinoic acid (Zhu et al., 2010). Therefore, this protein is a potential feature for specific subcellular regions.

The following feature protein is NDUFS3 (ENSP00000263774). As a specific iron–sulfur protein component of the mitochondrial NADH, this protein participates in the electron transport in the mitochondrion and contributes to energy-associated metabolisms in living cells (Benit et al., 2004). This protein is located in the mitochondrial and the nucleus regions (Vogel et al., 2007b; Taurino et al., 2012). Specifically, most of this protein is

directly located and functions in the inner mitochondrion membrane (Benit et al., 2004; Vogel et al., 2007a).

GRK3, the next predicted feature protein (ENSP00000317578), acts as a beta-adrenergic receptor kinase, contributes to the GPCR signaling pathway (Antony et al., 2009; Kumari et al., 2016), and participates in the CCR5 pathway in macrophages (Vroon et al., 2004; Balabanian et al., 2008). In general cells, GRK3 does not have a specific localization pattern. However, in macrophages, this protein merges with CXCR4 to form specific complexes in the cellular membrane system (Wang et al., 2001). Therefore, in these functional cells, our candidate protein GRK3 has a specific spatial distribution pattern and may contribute to the identification of the biological membrane region, validating the efficacy and the accuracy of our prediction.

BRIX1 (ENSP00000338862) is the specific regulator in the biogenesis of the 60S ribosomal subunit and is predicted to contribute to subcellular localization (Fromont-Racine et al., 2003; Strunk and Karbstein, 2009). According to recent publications, this protein is mostly located inside the nucleus and regulates ribosome biosynthesis (Zieve and Penman, 1976; Nguyen et al., 1998). According to the Human Protein Atlas (HPA), this protein is identified in the cytoplasm, but most of the protein is still located and functions in the nucleus, validating that the specific subcellular localization subgrouping is dependent on this protein (Pontén et al., 2008).

MDH2 (ENSP00000327070) contributes to the catalyzation of the reversible oxidation of malate to oxaloacetate and is predicted to help in the identification of a certain subcellular region (Pines et al., 1997; Shi and Gibson, 2011). According to HPA (Pontén et al., 2008), like NDUFS3, this protein is mostly identified in the mitochondrion. Recent publications also confirm that this protein can be identified in multiple intracellular organelles but is actually enriched in the mitochondria system (Lo et al., 2015) especially the mitochondria-associated ER membranes (Guardia-Laguarta et al., 2014; Lo et al., 2015). Moreover, this protein acts as a potential subcellular signature and corresponds with our prediction.

The H3-3B (ENSP00000254810) in our prediction list is the basic nuclear protein that contributes to the maintenance of the chromosomal fiber in eukaryotes (Frey et al., 2014). Therefore, this protein is definitely located in the nucleus region, thereby indicating subcellular localization. Similar with BRIX1, the protein NHP2 (ENSP00000274606) is a specific protein required for ribosome biogenesis (Vulliamy et al., 2008; Fumagalli et al., 2009) and telomere maintenance (Wong and Collins, 2003; Vulliamy et al., 2008). Therefore, this protein is also identified in the cytoplasm and the nucleus. This protein has potential to act as a subcellular localization signature because most of it is located in the nucleus (Pontén et al., 2008). Other feature proteins, e.g., CYC1 (ENSP00000317159) (Chen et al., 1994) and H2AZ2 (ENSP00000308405) (Eskandarian, 2013), have specific distribution patterns inside the cell, cytoplasm, and nucleus according to recent publications.

Overall, the feature proteins we analyzed have already been validated to contribute to the subcellular localization, validating the efficacy and the accuracy of our prediction. Thus, our newly presented computational method may be an effective tool for the prediction of subcellular localizations.

4.2 Features From Functions That Contribute to Subcellular Localization Prediction

The functional enrichment analysis is performed, and a group of effective GO (Consortium, 2015) and KEGG terms (Kanehisa, 2002) is screened to describe the core biological functions related to subcellular localization and further show the functional distribution pattern of feature proteins.

The top four GO terms in our prediction list describe specific subcellular localization or effective structures contributing to the distinction of different subcellular localization. These terms include GO:0070013 (describes the intracellular organelle lumen), GO:0031975 (describes the specific envelope structures in cells), GO:0031090 (describes the organelle membrane), and GO:0005887 (describes the integral component of the plasma membrane).

For example, the intracellular organelle lumen is a specific part of effective organelles, such as mitochondrion, peroxisomes, and Golgi apparatus (Lorenz et al., 2006a; Lorenz et al., 2006b; Masyuk et al., 2006), distinguishing perticular subcellular localization from the other ones. Therefore, GO:0070013 can contribute to subcellular localization. For GO:0031975, the envelope is a multilayered structure connected to the cell membrane or other membrane systems (Peabody et al., 2016). Therefore, this GO term is functionally correlated with the cell membrane and with various organelles with membrane-like mitochondrion and Golgi apparatus (Graham et al., 1991; Finnegan et al., 2001; Peabody et al., 2016). Other subcellular localization prediction algorithms also consider this term as a specific parameter for classification (Peabody et al., 2016). Similarly, GO:0031090 and GO:0005887 describe a part of the membrane system in cells.

4.3 Quantitative Rules That Contribute to Subcellular Localization Prediction

Apart from the qualitative analysis on specific GO or KEGG terms, a group of quantitative rules are established for the identification of different subcellular localizations. According to recent publications, these rules contribute to subcellular localization, thereby validating the efficacy and the accuracy of our prediction. Here, 16 typical rules referring to 16 clusters are chosen for detailed analyses.

The first rule is to identify the biological membrane subcellular localization (Class 1). According to the quantitative rules, the first parameter is GO:0031224. According to our prediction, the proteins enriched in this cellular component positively contribute to the biological membrane. Considering that GO: 0031224 describes the intrinsic component of membrane, this GO term is the first parameter to identify the proteins associated with the biological membrane, validating our prediction. Similarly, GO:0005886 describes the plasma membrane and may positively contribute to the identification of such

subcellular localization. Some terms negatively participate in this identification. For instance, the nuclear lumen described by GO: 0031981 located inside the nucleus is in our prediction list.

For the rules that contribute to the identification of cell periphery subcellular localization (Class 2), GO:0031224 is in this predictive parameter list. The specific GO term GO:0007043 highly enriches proteins associated with the identification of cell periphery subcellular localization. According to the GO annotation, this GO term describes the cell–cell junction assembly, which definitely occurs in the periphery subcellular regions (Setzer et al., 2004; Dawson et al., 2012), validating the efficacy and the accuracy of our prediction.

The third rule focuses on the identification of cytoplasm (Class 3). Specifically, wound healing (GO:0042060) is identified as a specific positive enrichment marker for this rule. The cytoplasm plays an essential role for wound healing (Jeon and Jeon, 1975). Therefore, the proteins that are located at the cytoplasm can be identified by a specific biological process (Jeon and Jeon, 1975; Gabbiani et al., 1978), such as wound healing.

Similar with that of the cytoplasm, a group of rules for the identification of cytoplasmic vesicle (Class 4) are identified. Among the rule parameters, the specific GO term GO:0070727 that describes the cellular macromolecule localization (Franklin and Baltimore, 1962) is a key feature that contributes to the identification of the cytoplasmic vesicle. According to recent publications, the cytoplasmic vesicle is a major transporter of macromolecules during synthesis and functioning (Franklin and Baltimore, 1962). Therefore, this GO term is a distinctive parameter for the sublocation of the cytoplasmic vesicle.

Furthermore, some specific rules are identified for endosome (Class 6), extracellular space or cell surface (Class 7), and flagellum or cilium (Class 8). Apart from some general GO terms, such as GO:0031224, the GO:1902115 is a specific parameter for the identification of endosome. Describing the assembly of effective intracellular organelles, this GO term contributes to the identification of endosome subcellular localization due to the tight correlation between endosome and organelle assembly (Kjeken et al., 2004; Kloer et al., 2010). For the identification of the extracellular space or the cell surface, apart from a series of GO terms like other predictive rules, the specific protein SDAD1 is obtained for the prediction of subcellular localization on the extracellular space or the cell surface. According to recent publications, this protein is located mostly inside the nucleus (Zeng et al., 2017) but not outside or on the biomembrane system. As for flagellum or cilium (Class 8), a specific parameter called GO:2000816 is positively correlated with the identification of this subcellular localization. This GO term describes the negative regulation of mitotic sister chromatid separation. Considering that mitotic separation is one of the major biological functions of the centriole–flagellum system (Wilson, 1969; Bettencourt-Dias et al., 2005), this parameter (biological process) is correlated with our predicted subcellular localization to a certain extent and definitely contributes to the identification of this cellular structure, thereby validating our predictions.

In addition, specific organelles, such as endoplasmic reticulum localization (Class 5), Golgi apparatus (Class 9), and mitochondrion (Class 11), can be identified and located by specific quantitative rules. The specific parameter GO:0005789 contributes to the localization of the endoplasmic reticulum. The GO:0005789 describes the endoplasmic reticulum membrane, validating the efficacy and the accuracy of our prediction. For the localization of Golgi apparatus, the specific parameter has00601 describing the glycosphingolipid biosynthesis is identified. Considering that glycosphingolipid biosynthesis is a typical biological process happening in the Golgi apparatus (Burger et al., 1996; Butters et al., 2000), this function is predicted as a quantitative parameter for the identification of Golgi apparatus subcellular localization. The mitochondrion is the next predicted subcellular localization with typical predictive parameters (such as GO:0031975), and the envelope is analyzed above (Peabody et al., 2016). This GO term is functionally correlated with the mitochondrion (Graham et al., 1991; Finnegan et al., 2001; Peabody et al., 2016), confirming our prediction.

Furthermore, the cell nucleus-associated locations, such as nuclear periphery (Class 12), nucleolus (Class 13), and nucleus (Class 14), can be quantitatively identified by our rules. For class 11, nuclear periphery regions are identified. Apart from the typical parameters, such as GO:0031981 and GO:0005654, the typical protein ENSP00000345895 or NUP50 is identified. According to recent publications, this protein is enriched in the periphery regions of the nucleus (Hajeri et al., 2010; Vaquerizas et al., 2010), thereby positively corresponding with our prediction. For the nucleolus, the specific biological process RNA surveillance (GO:0071025) is enriched in such rules as an effective parameter. Considering that RNA surveillance does occur in this region (Hernandez-Verdun et al., 2010), this GO term is a functional predictive parameter, validating the efficacy and the accuracy of our prediction. Similar with the nucleolus, the nucleus has its specific "biomarkers" in these quantitative rules. GO:0045596 describes the negative regulation of cell differentiation and is positively enriched in these rules. Considering that the physical plasticity of nucleus is quite important for cell differentiation (Pajerowski et al., 2007), this GO term is a positive parameter for subcellular localization, validating the efficacy and the accuracy of our prediction.

Three effective subcellular regions, namely, microtubule cytoskeleton (Class 10), peroxisome (Class 15), and vacuole (Class 16) remain. For the identification of microtubule cytoskeleton, the typical GO term GO:0044450 describes the obsolete microtubule organizing center part and is functionally and positively correlated with the microtubule system. Therefore, the predicted quantitative rules may be effective for the identification of the microtubule cytoskeleton's subcellular localization. Peroxisome identification requires the specific quantitative parameter GO:0031903, which describes the microbody membrane. According to recent publications, peroxisomes are major functional components of the microbody. Thus, this GO term is an effective parameter (Fahimi, 1969; Tolbert and Essner, 1981). The last subcellular localization is the vacuole. Similar with the peroxisomes' rules, a specific GO term describing only the vacuolar lumen, a part of the vacuole, is identified, thereby validating our prediction.

5 CONCLUSION

We identified a group of feature proteins that effectively contributes to intracellular subcellular localization and screened a series of qualitative functional enrichment patterns (i.e., GO and KEGG terms), revealing the functional distribution patterns of these proteins that contribute to subcellular localization identification. Combining proteins and functional annotations, a series of quantitative prediction rules was built for further analysis. Several screened feature proteins, functional annotation terms (i.e., GO or KEGG terms), and parameters of quantitative rules have been validated by recent publications. This study can provide a computational model for effective subcellular localization prediction and lay a solid foundation for further experimental research in such fields. The data as well as the information of used programs and software are available at https://github.com/chenlei1982/subc_prediction.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://cn.expasy.org/

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA26040304, XDB38050200), National Key R&D Program of China (2018YFC0910403), the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.783128/full#supplementary-material

REFERENCES

Antony, J., Kellershohn, K., Mohr-Andrä, M., Kebig, A., Prilla, S., Muth, M., et al. (2009). Dualsteric GPCR Targeting: a Novel Route to Binding and Signaling Pathway Selectivity. *FASEB j.* 23, 442–450. doi:10.1096/fj.08-114751

Balabanian, K., Levoye, A., Klemm, L., Lagane, B., Hermine, O., Harriague, J., et al. (2008). Leukocyte Analysis from WHIM Syndrome Patients Reveals a Pivotal Role for GRK3 in CXCR4 Signaling. *J. Clin. Invest.* 118, 1074–1084. doi:10.1172/JCI33187

Benit, P., Slama, A., Cartault, F., Giurgea, I., Chretien, D., Lebon, S., et al. (2004). Mutant NDUFS3 Subunit of Mitochondrial Complex I Causes Leigh Syndrome. *J. Med. Genet.* 41, 14–17. doi:10.1136/jmg.2003.014316

Bettencourt-Dias, M., Rodrigues-Martins, A., Carpenter, L., Riparbelli, M., Lehmann, L., Gatt, M. K., et al. (2005). SAK/PLK4 Is Required for Centriole Duplication and Flagella Development. *Curr. Biol.* 15, 2199–2207. doi:10.1016/j.cub.2005.11.042

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

Burger, K. N., Van Der Bijl, P., and Van Meer, G. (1996). Topology of Sphingolipid Galactosyltransferases in ER and Golgi: Transbilayer Movement of Monohexosyl Sphingolipids Is Required for Higher Glycosphingolipid Biosynthesis. *J. Cel. Biol.* 133, 15–28. doi:10.1083/jcb.133.1.15

Butters, T. D., Van Den Broek, L. A. G. M., Fleet, G. W. J., Krulle, T. M., Wormald, M. R., Dwek, R. A., et al. (2000). Molecular Requirements of Imino Sugars for the Selective Control of N-Linked Glycosylation and Glycosphingolipid Biosynthesis. *Tetrahedron: Asymmetry* 11, 113–124. doi:10.1016/s0957-4166(99)00468-1

Cai, Y.-D., and Chou, K.-C. (2003). Nearest Neighbour Algorithm for Predicting Protein Subcellular Location by Combining Functional Domain Composition and Pseudo-amino Acid Composition. *Biochem. Biophysical Res. Commun.* 305, 407–411. doi:10.1016/s0006-291x(03)00775-7

Cai, Y., He, J., Li, X., Feng, K., Lu, L., Feng, K., et al. (2010). Prediction of Protein Subcellular Locations with Feature Selection and Analysis. *Ppl* 17, 464–472. doi:10.2174/092986610790963654

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M., and Pascual-Montano, A. (2007). GENECODIS: a Web-Based Tool for Finding Significant Concurrent Annotations in Gene Lists. *Genome Biol.* 8, R3. doi:10.1186/gb-2007-8-1-r3

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *jair* 16, 321–357. doi:10.1613/jair.953

Chen, J., Ding, M., and Pederson, D. S. (1994). Binding of TFIID to the CYC1 TATA Boxes in Yeast Occurs Independently of Upstream Activating Sequences. *Proc. Natl. Acad. Sci.* 91, 11909–11913. doi:10.1073/pnas.91.25.11909

Chen, W., Chen, L., and Dai, Q. (2021). iMPT-FDNPL: Identification of Membrane Protein Types with Functional Domains and a Natural Language Processing Approach. *Comput. Math. Methods Med.* 2021, 7681497. doi:10.1155/2021/7681497

Chou, K.-C. (2001). Prediction of Protein Cellular Attributes Using Pseudo-amino Acid Composition. *Proteins* 43, 246–255. doi:10.1002/prot.1035

Chou, K.-C., and Shen, H.-B. (2008). Cell-PLoc: a Package of Web Servers for Predicting Subcellular Localization of Proteins in Various Organisms. *Nat. Protoc.* 3, 153–162. doi:10.1038/nprot.2007.494

Consortium, G. O. (2015). Gene Ontology Consortium: Going Forward. *Nucleic Acids Res.* 43, D1049–D1056. doi:10.1093/nar/gku1179

Cortes, C., and Vapnik, V. (1995). Support-vector Networks. *Mach Learn.* 20, 273–297. doi:10.1007/bf00994018

Cover, T., and Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Trans. Inform. Theor.* 13, 21–27. doi:10.1109/tit.1967.1053964

Dawson, J. C., Bruche, S., Spence, H. J., Braga, V. M. M., and Machesky, L. M. (2012). Mtss1 Promotes Cell-Cell junction Assembly and Stability through the Small GTPase Rac1. *PLoS One* 7, e31141. doi:10.1371/journal.pone.0031141

Ding, Y. J., Tang, J. 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

Eskandarian, H. A. (2013). *Characterization of Histone H3 Lysine 18 Deacetylation during Infection with Listeria Monocytogenes*. PhD Thesis. Paris: Universite Rene Descartes-Paris V.

Fahimi, H. D. (1969). Cytochemical Localization of Peroxidatic Activity of Catalase in Rat Hepatic Microbodies (Peroxisomes). *J. Cel. Biol.* 43, 275–288. doi:10.1083/jcb.43.2.275

Finnegan, C. M., Berg, W., Lewis, G. K., and Devico, A. L. (2001). Antigenic Properties of the Human Immunodeficiency Virus Envelope during Cell-Cell Fusion. *J. Virol.* 75, 11096–11105. doi:10.1128/jvi.75.22.11096-11105.2001

Franklin, R. M., and Baltimore, D. (1962). "Patterns of Macromolecular Synthesis in normal and Virus-Infected Mammalian Cells," in *Cold Spring Harbor Symposia on Quantitative Biology* (New York, NY: Cold Spring Harbor Laboratory Press), 175–198.

Frey, A., Listovsky, T., Guilbaud, G., Sarkies, P., and Sale, J. E. (2014). Histone H3.3 Is Required to Maintain Replication Fork Progression after UV Damage. *Curr. Biol.* 24, 2195–2201. doi:10.1016/j.cub.2014.07.077

Fromont-Racine, M., Senger, B., Saveanu, C., and Fasiolo, F. (2003). Ribosome Assembly in Eukaryotes. *Gene* 313, 17–42. doi:10.1016/s0378-1119(03)00629-2

Fumagalli, S., Di Cara, A., Neb-Gulati, A., Natt, F., Schwemberger, S., Hall, J., et al. (2009). Absence of Nucleolar Disruption after Impairment of 40S Ribosome Biogenesis Reveals an rpL11-translation-dependent Mechanism of P53 Induction. *Nat. Cel Biol* 11, 501–508. doi:10.1038/ncb1858

Gabbiani, G., Chaponnier, C., and Hüttner, I. (1978). Cytoplasmic Filaments and gap Junctions in Epithelial Cells and Myofibroblasts during Wound Healing. *J. Cel. Biol.* 76, 561–568. doi:10.1083/jcb.76.3.561

Ganai, R. A., and Johansson, E. (2016). DNA Replication-A Matter of Fidelity. *Mol. Cel.* 62, 745–755. doi:10.1016/j.molcel.2016.05.003

Gorodkin, J. (2004). Comparing Two K-Category Assignments by a K-Category Correlation Coefficient. *Comput. Biol. Chem.* 28, 367–374. doi:10.1016/j.compbiolchem.2004.09.006

Graham, L. L., Harris, R., Villiger, W., and Beveridge, T. J. (1991). Freeze-substitution of Gram-Negative Eubacteria: General Cell Morphology and Envelope Profiles. *J. Bacteriol.* 173, 1623–1633. doi:10.1128/jb.173.5.1623-1633.1991

Guardia-Laguarta, C., Area-Gomez, E., Rüb, C., Liu, Y., Magrané, J., Becker, D., et al. (2014). -Synuclein Is Localized to Mitochondria-Associated ER Membranes. *J. Neurosci.* 34, 249–259. doi:10.1523/jneurosci.2507-13.2014

Hajeri, V. A., Little, B. A., Ladage, M. L., and Padilla, P. A. (2010). NPP-16/Nup50 Function and CDK-1 Inactivation Are Associated with Anoxia-Induced Prophase Arrest in *Caenorhabditis elegans*. *MBoC* 21, 712–724. doi:10.1091/mbc.e09-09-0787

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software. *SIGKDD Explor. Newsl.* 11, 10–18. doi:10.1145/1656274.1656278

Hecker, C.-M., Rabiller, M., Haglund, K., Bayer, P., and Dikic, I. (2006). Specification of SUMO1- and SUMO2-Interacting Motifs. *J. Biol. Chem.* 281, 16117–16127. doi:10.1074/jbc.m512757200

Hernandez-Verdun, D., Roussel, P., Thiry, M., Sirri, V., and Lafontaine, D. L. (2010). The Nucleolus: Structure/function Relationship in RNA Metabolism. *Wiley Interdiscip. Rev. RNA* 1, 415–431. doi:10.1002/wrna.39

Huang, W.-L., Tung, C.-W., Ho, S.-W., Hwang, S.-F., and Ho, S.-Y. (2008). ProLoc-GO: Utilizing Informative Gene Ontology Terms for Sequence-Based Prediction of Protein Subcellular Localization. *BMC bioinformatics* 9, 80. doi:10.1186/1471-2105-9-80

Jeon, K., and Jeon, M. (1975). Cytoplasmic Filaments and Cellular Wound Healing in Amoeba proteus. *J. Cel. Biol.* 67, 243–249. doi:10.1083/jcb.67.1.243

Jia, Y., Zhao, R., and Chen, L. (2020). Similarity-Based Machine Learning Model for Predicting the Metabolic Pathways of Compounds. *IEEE Access* 8, 130687–130696. doi:10.1109/access.2020.3009439

Kanehisa, M. (2002). "The KEGG Database," in *Novartis Foundation Symposium* (Bridgewater: Wiley Online Library), 91–100.

Kjeken, R., Egeberg, M., Habermann, A., Kuehnel, M., Peyron, P., Floetenmeyer, M., et al. (2004). Fusion between Phagosomes, Early and Late Endosomes: a Role for Actin in Fusion between Late, but Not Early Endocytic Organelles. *MBoC* 15, 345–358. doi:10.1091/mbc.e03-05-0334

Kloer, D. P., Rojas, R., Ivan, V., Moriyama, K., Van Vlijmen, T., Murthy, N., et al. (2010). Assembly of the Biogenesis of Lysosome-Related Organelles Complex-3 (BLOC-3) and its Interaction with Rab9. *J. Biol. Chem.* 285, 7794–7804. doi:10.1074/jbc.m109.069088

Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *International Joint Conference on Artificial Intelligence* (London: Lawrence Erlbaum Associates), 1137–1145.

Kursa, M., and Rudnicki, W. (2010). Feature Selection with the Boruta Package. *J. Stat. Softw. Articles* 36, 1–13. doi:10.18637/jss.v036.i11

Li, B.-Q., Huang, T., Liu, L., Cai, Y.-D., and Chou, K.-C. (2012). Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network. *PLoS One* 7, e33393. doi:10.1371/journal.pone.0033393

Li, W., and Godzik, A. (2006). Cd-hit: a Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* 22, 1658–1659. doi:10.1093/bioinformatics/btl158

Lin, H., Ding, H., Guo, F.-B., Zhang, A.-Y., and Huang, J. (2008). Predicting Subcellular Localization of Mycobacterial Proteins by Using Chous Pseudo Amino Acid Composition. *Ppl* 15, 739–744. doi:10.2174/092986608785133681

Liu, H., and Setiono, R. (1998). Incremental Feature Selection. *Appl. Intelligence* 9, 217–230. doi:10.1023/a:1008363719778

Liu, T., Zheng, X., Wang, C., and Wang, J. (2010). Prediction of Subcellular Location of Apoptosis Proteins Using Pseudo Amino Acid Composition: an Approach from Auto Covariance Transformation. *Ppl* 17, 1263–1269. doi:10.2174/092986610792231528

Lo, Y. W., Lin, S. T., Chang, S. J., Chan, C. H., Lyu, K. W., Chang, J. F., et al. (2015). Mitochondrial proteomics with si RNA knockdown to reveal ACAT 1 and MDH 2 in the development of doxorubicin-resistant uterine cancer. *J. Cel. Mol. Med.* 19, 744–759. doi:10.1111/jcmm.12388

Lorenz, H., Hailey, D. W., and Lippincott-Schwartz, J. (2006a). Fluorescence Protease protection of GFP Chimeras to Reveal Protein Topology and Subcellular Localization. *Nat. Methods* 3, 205–210. doi:10.1038/nmeth857

Lorenz, H., Hailey, D. W., Wunder, C., and Lippincott-Schwartz, J. (2006b). The Fluorescence Protease protection (FPP) Assay to Determine Protein Localization and Membrane Topology. *Nat. Protoc.* 1, 276–279. doi:10.1038/nprot.2006.42

Man-Wai Mak, M.-W., Jian Guo, J., and Sun-Yuan Kung, S.-Y. (2008). PairProSVM: Protein Subcellular Localization Based on Local Pairwise Profile Alignment and SVM. *Ieee/acm Trans. Comput. Biol. Bioinf.* 5, 416–422. doi:10.1109/tcbb.2007.70256

Mangeat, P., Roy, C., and Martin, M. (1999). ERM Proteins in Cell Adhesion and Membrane Dynamics. *Trends Cell Biology* 9, 187–192. doi:10.1016/s0962-8924(99)01544-5

Masyuk, A. I., Masyuk, T. V., Splinter, P. L., Huang, B. Q., Stroope, A. J., and Larusso, N. F. (2006). Cholangiocyte Cilia Detect Changes in Luminal Fluid Flow and Transmit Them into Intracellular Ca2+ and cAMP Signaling. *Gastroenterology* 131, 911–920. doi:10.1053/j.gastro.2006.07.003

Matthews, B. W. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta (Bba) - Protein Struct.* 405, 442–451. doi:10.1016/0005-2795(75)90109-9

Nguyen, Y. H., Mills, A. A., and Stanbridge, E. J. (1998). Assembly of the QM Protein onto the 60S Ribosomal Subunit Occurs in the Cytoplasm. *J. Cel. Biochem.* 68, 281–285. doi:10.1002/(sici)1097-4644(19980201)68:2<281:aid-jcb14>3.0.co;2-i

Pajerowski, J. D., Dahl, K. N., Zhong, F. L., Sammak, P. J., and Discher, D. E. (2007). Physical Plasticity of the Nucleus in Stem Cell Differentiation. *Proc. Natl. Acad. Sci.* 104, 15619–15624. doi:10.1073/pnas.0702576104

Pan, X., Li, H., Zeng, T., Li, Z., Chen, L., Huang, T., et al. (2021). Identification of Protein Subcellular Localization with Network and Functional Embeddings. *Front. Genet.* 11, 626500. doi:10.3389/fgene.2020.626500

Pan, Y.-X., Zhang, Z.-Z., Guo, Z.-M., Feng, G.-Y., Huang, Z.-D., and He, L. (2003). Application of Pseudo Amino Acid Composition for Predicting Protein Subcellular Location: Stochastic Signal Processing Approach. *J. Protein Chem.* 22, 395–402. doi:10.1023/a:1025350409648

Peabody, M. A., Laird, M. R., Vlasschaert, C., Lo, R., and Brinkman, F. S. L. (2016). PSORTdb: Expanding the Bacteria and Archaea Protein Subcellular Localization Database to Better Reflect Diversity in Cell Envelope Structures. *Nucleic Acids Res.* 44, D663–D668. doi:10.1093/nar/gkv1271

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Machine Learn. Res.* 12, 2825–2830.

Peng, H., Fuhui Long, L., and Ding, C. (2005). Feature Selection Based on Mutual Information Criteria of max-dependency, max-relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Machine Intell.* 27, 1226–1238. doi:10.1109/tpami.2005.159

Petsalaki, E. I., Bagos, P. G., Litou, Z. I., and Hamodrakas, S. J. (2006). PredSL: a Tool for the N-Terminal Sequence-Based Prediction of Protein Subcellular Localization. *Genomics, Proteomics & Bioinformatics* 4, 48–55. doi:10.1016/s1672-0229(06)60016-8

Pines, O., Shemesh, S., Battat, E., and Goldberg, I. (1997). Overexpression of Cytosolic Malate Dehydrogenase (MDH2) Causes Overproduction of Specific Organic Acids in *Saccharomyces cerevisiae*. *Appl. Microbiol. Biotechnol.* 48, 248–255. doi:10.1007/s002530051046

Platt, J. (1998a). *Fast Training of Support Vector Machines Using Sequential Minimal Optimization* (Cambridge, MA: MIT Press).

Platt, J. (1998b). Sequential Minimal Optimizaton: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14.

Pontén, F., Jirström, K., and Uhlen, M. (2008). The Human Protein Atlas-A Tool for Pathology. *J. Pathol.* 216, 387–393. doi:10.1002/path.2440

Set, K. K., Sen, K., Huq, A. H. M., and Agarwal, R. (2019). Mitochondrial Disorders of the Nervous System: A Review. *Clin. Pediatr. (Phila)* 58, 381–394. doi:10.1177/0009922818821890

Setzer, S. V., Calkins, C. C., Garner, J., Summers, S., Green, K. J., and Kowalczyk, A. P. (2004). Comparative Analysis of Armadillo Family Proteins in the Regulation of A431 Epithelial Cell junction Assembly, Adhesion and Migration. *J. Invest. Dermatol.* 123, 426–433. doi:10.1111/j.0022-202x.2004.23319.x

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based Prediction of Protein Function. *Mol. Syst. Biol.* 3, 88. doi:10.1038/msb4100129

Shi, J.-Y., Zhang, S.-W., Pan, Q., and Zhou, G.-P. (2008). Using Pseudo Amino Acid Composition to Predict Protein Subcellular Location: Approached with Amino Acid Composition Distribution. *Amino Acids* 35, 321–327. doi:10.1007/s00726-007-0623-z

Shi, Q., and Gibson, G. E. (2011). Up-regulation of the Mitochondrial Malate Dehydrogenase by Oxidative Stress Is Mediated by miR-743a. *J. Neurochem.* 118, 440–448. doi:10.1111/j.1471-4159.2011.07333.x

Kumari, P., Srivastava, A., Banerjee, R., Ghosh, E., Gupta, P., Ranjan, R., et al. (2016). Functional Competence of a Partially Engaged GPCR β-Arrestin Complex. *Nat. Commun.* 7, 13416.

Strunk, B. S., and Karbstein, K. (2009). Powering through Ribosome Assembly. *Rna* 15, 2083–2104. doi:10.1261/rna.1792109

Swain, P. H., and Hauska, H. (1977). The Decision Tree Classifier: Design and Potential. *IEEE Trans. Geosci. Electron.* 15, 142–147. doi:10.1109/tge.1977.6498972

Szafron, D., Lu, P., Greiner, R., Wishart, D. S., Poulin, B., Eisner, R., et al. (2004). Proteome Analyst: Custom Predictions with Explanations in a Web-Based Tool for High-Throughput Proteome Annotations. *Nucleic Acids Res.* 32, W365–W371. doi:10.1093/nar/gkh485

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The STRING Database in 2011: Functional Interaction Networks of Proteins, Globally Integrated and Scored. *Nucleic Acids Res.* 39, D561–D568. doi:10.1093/nar/gkq973

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The STRING Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible. *Nucleic Acids Res.* 45, D362–D368. doi:10.1093/nar/gkw937

Tammsalu, T., Matic, I., Jaffray, E. G., Ibrahim, A. F. M., Tatham, M. H., and Hay, R. T. (2014). Proteome-wide Identification of SUMO2 Modification Sites. *Sci. Signal.* 7, rs2. doi:10.1126/scisignal.2005146

Taurino, F., Stanca, E., Siculella, L., Trentadue, R., Papa, S., Zanotti, F., et al. (2012). Mitochondrial Proteome Analysis Reveals Depression of the Ndufs3 Subunit and Activity of Complex I in Diabetic Rat Brain. *J. Proteomics* 75, 2331–2341. doi:10.1016/j.jprot.2012.02.002

Theilhaber, J., Connolly, T., Roman-Roman, S., Bushnell, S., Jackson, A., Call, K., et al. (2002). Finding Genes in the C2C12 Osteogenic Pathway by K-Nearest-Neighbor Classification of Expression Data. *Genome Res.* 12, 165–176. doi:10.1101/gr.182601

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., et al. (2017). A Subcellular Map of the Human Proteome. *Science* 356, eaal3321. doi:10.1126/science.aal3321

Tjondro, H. C., Loke, I., Chatterjee, S., and Thaysen-Andersen, M. (2019). Human Protein Paucimannosylation: Cues from the Eukaryotic Kingdoms. *Biol. Rev.* 94, 2068–2100. doi:10.1111/brv.12548

Tolbert, N. E., and Essner, E. (1981). Microbodies: Peroxisomes and Glyoxysomes. *J. Cel. Biol.* 91, 271s–283s. doi:10.1083/jcb.91.3.271s

Vaquerizas, J. M., Suyama, R., Kind, J., Miura, K., Luscombe, N. M., and Akhtar, A. (2010). Nuclear Pore Proteins Nup153 and Megator Define Transcriptionally Active Regions in the Drosophila Genome. *Plos Genet.* 6, e1000846. doi:10.1371/journal.pgen.1000846

Vogel, R. O., Dieteren, C. E. J., Van Den Heuvel, L. P. W. J., Willems, P. H. G. M., Smeitink, J. A. M., Koopman, W. J. H., et al. (2007a). Identification of Mitochondrial Complex I Assembly Intermediates by Tracing Tagged NDUFS3 Demonstrates the Entry point of Mitochondrial Subunits. *J. Biol. Chem.* 282, 7582–7590. doi:10.1074/jbc.m609410200

Vogel, R. O., Janssen, R. J. R. J., Van Den Brand, M. A. M., Dieteren, C. E. J., Verkaart, S., Koopman, W. J. H., et al. (2007b). Cytosolic Signaling Protein Ecsit Also Localizes to Mitochondria where it Interacts with Chaperone NDUFAF1 and Functions in Complex I Assembly. *Genes Dev.* 21, 615–624. doi:10.1101/gad.408407

Vroon, A., Heijnen, C. J., Lombardi, M. S., Cobelens, P. M., Mayor, F., Jr, Caron, M. G., et al. (2004). Reduced GRK2 Level in T Cells Potentiates Chemotaxis and Signaling in Response to CCL4. *J. Leukoc. Biol.* 75, 901–909. doi:10.1189/jlb.0403136

Vulliamy, T., Beswick, R., Kirwan, M., Marrone, A., Digweed, M., Walne, A., et al. (2008). Mutations in the Telomerase Component NHP2 Cause the Premature Ageing Syndrome Dyskeratosis Congenita. *Proc. Natl. Acad. Sci.* 105, 8073–8078. doi:10.1073/pnas.0800042105

Wang, J., Guan, E., Roderiquez, G., Calvert, V., Alvarez, R., and Norcross, M. A. (2001). Role of Tyrosine Phosphorylation in Ligand-independent Sequestration of CXCR4 in Human Primary Monocytes-Macrophages. *J. Biol. Chem.* 276, 49236–49243. doi:10.1074/jbc.m108523200

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

Wilson, H. J. (1969). Arms and Bridges on Microtubules in the Mitotic Apparatus. *J. Cel. Biol.* 40, 854–859. doi:10.1083/jcb.40.3.854

Wong, J. M., and Collins, K. (2003). Telomere Maintenance and Disease. *The Lancet* 362, 983–988. doi:10.1016/s0140-6736(03)14369-3

Xiao, X., Wu, Z.-C., and Chou, K.-C. (2011). iLoc-Virus: A Multi-Label Learning Classifier for Identifying the Subcellular Localization of Virus Proteins with Both Single and Multiple Sites. *J. Theor. Biol.* 284, 42–51. doi:10.1016/j.jtbi.2011.06.005

Yang, Y., and Chen, L. (2021). Identification of Drug–Disease Associations by Using Multiple Drug and Disease Networks. *Curr. Bioinformatics*. doi:10.2174/1574893616666210825115406

Yèagle, P. L. (1989). Lipid Regulation of Cell Membrane Structure and Function. *FASEB j.* 3, 1833–1842. doi:10.1096/fasebj.3.7.2469614

Yu, Z., Chen, H., Liuxs, J., You, J., Leung, H., and Han, G. (2016). Hybrid k -Nearest Neighbor Classifier. *IEEE Trans. Cybern.* 46, 1263–1275. doi:10.1109/tcyb.2015.2443857

Zeng, M., Zhu, L., Li, L., and Kang, C. (2017). miR-378 Suppresses the Proliferation, Migration and Invasion of colon Cancer Cells by Inhibiting SDAD1. *Cell Mol Biol Lett* 22, 12. doi:10.1186/s11658-017-0041-5

Zhang, B., and Srihari, S. N. (2004). Fast K-Nearest Neighbor Classification Using Cluster-Based Trees. *IEEE Trans. Pattern Anal. Machine Intell.* 26, 525–528. doi:10.1109/tpami.2004.1265868

Zhang, J. D., and Wiemann, S. (2009). KEGGgraph: a Graph Approach to KEGG PATHWAY in R and Bioconductor. *Bioinformatics* 25, 1470–1471. doi:10.1093/bioinformatics/btp167

Zhang, Y.-H., Li, H., Zeng, T., Chen, L., Li, Z., Huang, T., et al. (2021a). Identifying Transcriptomic Signatures and Rules for SARS-CoV-2 Infection. *Front. Cel Dev. Biol.* 8, 627302. doi:10.3389/fcell.2020.627302

Zhang, Y.-H., Li, Z., Zeng, T., Chen, L., Li, H., Huang, T., et al. (2021b). Detecting the Multiomics Signatures of Factor-specific Inflammatory Effects on Airway Smooth Muscles. *Front. Genet.* 11, 599970. doi:10.3389/fgene.2020.599970

Zhang, Y.-H., Zeng, T., Chen, L., Huang, T., and Cai, Y.-D. (2021c). Determining Protein-Protein Functional Associations by Functional Rules Based on Gene Ontology and KEGG Pathway. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1869, 140621. doi:10.1016/j.bbapap.2021.140621

Zhou, J.-P., Chen, L., Wang, T., and Liu, M. (2020). iATC-FRAKEL: a Simple Multi-Label Web Server for Recognizing Anatomical Therapeutic Chemical Classes of Drugs with Their Fingerprints Only. *Bioinformatics* 36, 3568–3569. doi:10.1093/bioinformatics/btaa166

Zhu, L., Doyle, T. J., and Kim, K. H. (2010). Retinoic Acid Modulates the Subcellular Localization of Small Ubiquitin-Related Modifier-2/3 (SUMO-2/3) in the Testis. *J. Androl.* 31, 406–418. doi:10.2164/jandrol.109.008763

Zieve, G., and Penman, S. (1976). Small RNA Species of the HeLa Cell: Metabolism and Subcellular Localization. *Cell* 8, 19–31. doi:10.1016/0092-8674(76)90181-1

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.

Copyright © 2021 Chen, Li, Zeng, Zhang, Zhang, Huang and Cai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 November 2021
doi: 10.3389/fgene.2021.751791

Research on Mental Stress Recognition of Depressive Disorders in Patients With Androgenic Alopecia Based on Machine Learning and Fuzzy K-Means Clustering

*Yulong Li[1], Baojin Wu[2], Xiujun Li[3], Qin Zhou[4], Xin Yang[5] and Yufei Li[4]**

[1]Department of Military Medical Psychology, Air Force Medical University, Xi'an, China, [2]Department of Plastic Surgery, Huashan Hospital, Fudan University, Shanghai, China, [3]College of Education, Shanghai Normal University, Shanghai, China, [4]Department of Plastic Surgery, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai, China, [5]Medical Cosmetic Center, Department of Dermatology, Tongji Hospital, School of Medicine, Tongji University, Shanghai, China

OPEN ACCESS

Edited by:
Arun Kumar Sangaiah, VIT University, India

Reviewed by:
Yingying Fan, Hebei GEO University, China
Yan Sun, Changchun Normal University, China
Haiyang Yu, Dalian Ocean University, China

***Correspondence:**
Yufei Li
lufylee@126.com

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

Received: *02 August 2021*
Accepted: *20 October 2021*
Published: *12 November 2021*

Citation:
Li Y, Wu B, Li X, Zhou Q, Yang X and Li Y (2021) Research on Mental Stress Recognition of Depressive Disorders in Patients With Androgenic Alopecia Based on Machine Learning and Fuzzy K-Means Clustering. Front. Genet. 12:751791. doi: 10.3389/fgene.2021.751791

Under the new trend of industry 4.0 software-defined network, the value of meta heuristic algorithm was explored in the recognition of depression in patients with androgenic alopecia (AGA), and there was an analysis on the effect of comprehensive psychological interventions in the rehabilitation of AGA patients. Based on the meta heuristic algorithm, the Filter and Wrapper algorithms were combined in this study to form a new feature selection algorithm FAW-FS. Then, the classification accuracy of FAW-FS and the ability to identify depression disorders were verified under different open data sets. 54 patients with AGA who went to the Medical Cosmetic Center of Tongji Hospital were selected as the research objects and rolled into a control group (routine psychological intervention) and an intervention group (routine + comprehensive psychological interventions) according to different psychological intervention methods, with 27 cases in each group. The differences of the self-rating anxiety scale (SAS), self-rating depression scale (SDS), Hamilton depression scale (HAMD), and physical, psychological, social, and substance function scores before and after intervention were compared between the two groups of AGA patients, and the depression efficacy and compliance of the two groups were analyzed after intervention. The results showed that the classification accuracy of FAW-FS algorithm was the highest in logistic regression (LR), decision tree (DT), K-nearest neighbor (KNN) algorithm, support vector machine (SVM) algorithm, and random forest (RF) algorithm, which was 80.87, 79.24, 80.42, 83.07, and 81.45%, respectively. The LR algorithm had the highest feature selection accuracy of 82.94%, and the classification accuracy of depression disorder in RF algorithm was up to 73.01%. Besides, the SDS, SAS, and HAMD scores of the intervention group were lower sharply than the scores of the control group ($p < 0.05$). The physical function, psychological function, social function, and substance function scores of the intervention group were higher markedly than those of the control group ($p < 0.05$). In addition, the proportions of cured, markedly effective, total effective, full compliance, and total compliance patients in the intervention group increased obviously in contrast to the proportions of the control group ($p < 0.05$). Therefore, it

indicated that the FAW-FS algorithm established in this study had significant advantages in the recognition of depression in AGA patients, and comprehensive psychological intervention had a positive effect in the rehabilitation of depression in AGA patients.

Keywords: meta heuristic algorithm, androgenetic alopecia, depressive disorder, psychological intervention, machine learning and fuzzy K-means clustering

INTRODUCTION

Androgenetic alopecia (AGA) is a kind of hair loss skin disease which is characterized by non-scarring and progressive hair follicle miniaturization. It is a common clinical skin disease, mainly characterized by shortening of hair follicle growth period, terminal hair follicle miniaturization, and progressive thinning of hair. The incidence of AGA was approximately 50% in white males over 40 years old and 32.2% in white females over 20 years old (Lolli et al., 2017). Psychosocial factors may aggravate or recur the condition of patients, and patients with hair loss are more likely to suffer from various physical and mental disorders such as anxiety and depression than normal people (Starace et al., 2020). Studies have shown that negative emotions such as anxiety and depression lead to a decline in the ability to deal with challenges and solve problems in patients with hair loss, which seriously affects their quality of life (Tanaka et al., 2018). A large number of investigation reports and meta-analysis prompts to inquire about the cognitive evaluation, emotional expression, and response of different patients to their diseases. Moreover, they have received the necessary psychological interventions, so as to establish scientific disease cognition and psychological behavioral responses. Its own adjustment is employed to promote patient adaptability and disease outcome, which is more significant than treating the disease itself (Rajabi et al., 2018; Völker et al., 2020).

Big data based on Industry 4.0 has the characteristics of large capacity, low signal-to-noise ratio, multiple types, high latitude, and fast access speed. Therefore, there are obvious differences in the methods of identification, analysis, and mining for industrial big data (Lake, 2019). Electroencephalogram (EEG) plays an important role in the diagnosis and recognition of depression. Deep learning can learn useful EEG signals automatically from the original data, to perform pattern recognition process, especially suitable for brain electrical signal recognition task. Many researchers will combine the deep learning algorithm with EEG, to operate EEG with feature extraction, selection, and classification, which can provide an auxiliary tool for the clinical diagnosis of depression (Craik et al., 2019). What's more, meta heuristic algorithm is a combination of random algorithm and local search algorithm, which is featured with self-organization, self-adaptation, and self-learning. It has been extensively applied in image recognition and classification (Munoz et al., 2018), and it has been also adopted in the recognition of depression (Phadikar et al., 2021). Eilbeigi et al. (2018) , (Eilbeigi and Setarehdan, 2018) used meta-heuristic algorithm to classify EEG data of patients with depression, with the highest accuracy of 78.24%. However, most of the current deep learning methods for depression recognition are to manually extract multiple features and simply combine the extracted features with traditional classification algorithms or neural network models. This method is time-consuming and laborious, so it is of great significance to explore an automatic computer-aided method for depression diagnosis.

To sum up, AGA patients have different degrees of depressive disorder. The meta heuristic algorithm has marked advantages in image recognition classification, but its classification accuracy needs to be further improved. In this study, a new depressive disorder recognition algorithm based on the meta heuristic algorithm was established and applied to AGA patients with depression, thereby evaluating the rehabilitation value of comprehensive psychological intervention for AGA patients, which can provide a reference for the diagnosis and treatment of AGA patients.

MATERIALS AND METHODS

Research Objects and Grouping

54 patients with AGA who were treated in the Medical Cosmetic Center of Tongji Hospital from January 2020 to October 2020 were selected as the research objects, and all agreed to receive treatment in this hospital for a long time. Among them, there were 31 males and 23 females. Besides, they were 18–60 years old, and the average age was 39.15 ± 4.07 years. The criteria for inclusion were defined to include patients who were older than or equaled to 18 years old, and conformed to AGA diagnostic criteria. The criteria for exclusion were defined to include patients who suffered from hair loss caused by resting period, physiological and postpartum hair loss, and other cause, had neuropsychiatric diseases, and were accompanied with other serious systemic diseases. In addition, they were grouped into the control group (n = 27) and the intervention group (n = 27) based on the different ways of psychological intervention. The process was approved by the ethics committee of Tongji Hospital, and all the research objects included in this study signed the informed consent forms.

Feature Selection Method Based on Meta Heuristic Algorithm

The optimization mathematical model of meta heuristic algorithm can be expressed as follows.

$$\min f(x), s.t.\mathrm{g}_{\mathrm{i}}(x) = 0, i = 1, 2, L, m; h_j(x) \geq 0, j = 1, 2, L, n. \tag{1}$$

In the **Eq. 1**, x stands for the decision variable, representing the p-dimensional vector, and its calculation method is

$x = \{x_1, x_2, \cdots, x_p\}^T \in \mathbb{R}^p$. Besides, $f(x)$ means the objective function, $g_i(x)$ indicates the equality constraint function, $h_j(x)$ represents the inequality constraint function, and *s.t.* expresses the abbreviation of "subject to," which means "restricted to."

The Filter algorithm in feature selection has a fast calculation speed, and the Wrapper algorithm has a higher calculation accuracy (Padfield et al., 2019). In this study, the Filter and Wrapper algorithms were combined to form a new feature selection algorithm, which was named FAW-FS. The two algorithms of analysis of variance (ANOVA) and mutual information were adopted to calculate the data to filter out the feature subset, thereby obtaining the union as the new feature space.

ANOVA (Peng et al., 2020) is a common special statistical hypothesis testing model in data analysis. The total variance (TV), total variance between groups (BGV), and variance within groups (WGV) of ANOVA are expressed as the following equations.

$$TV = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_i\right)^2 \tag{2}$$

$$BGV = \sum_i n_i \left(\bar{Y}_i - \bar{Y}_t\right)^2 \tag{3}$$

$$WGV = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_i\right)^2 \tag{4}$$

In the **Eqs 2–4**, i represents the group, and $i = 1, 2, \cdots, a$; Y_{ij} means the j-th eigenvalue in the i-th dimension feature; j stands for the subscript of the observation value; $\bar{Y}_t$ expresses the mean of all eigenvalues; n_i represents the total number of the i-th dimensional eigenvalues; $\bar{Y}_i$ indicates the mean of the i-th dimensional eigenvalues.

The mean square between groups (MSG) and mean square within groups (MSW) of ANOVA can be calculated as follows.

$$MSG = \frac{BGV}{k-1} = \frac{\sum_i n_i \left(\bar{Y}_i - \bar{Y}_t\right)^2}{k-1} \tag{5}$$

$$MSW = \frac{WGV}{N-k} = \frac{\sum_i \sum_j \left(Y_{ij} - \bar{Y}_i\right)^2}{N-k} \tag{6}$$

In the **Eq. 5** and **Eq. 6**, k and N stand for the dimension of the feature and the total number of eigenvalues, respectively.

Mutual information (MI) is mainly used to evaluate the joint probability distribution and marginal probability distribution between two variables (Wen et al., 2020). For discrete random variables, MI is defined as the following.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{7}$$

In the **Eq. 7**, $p(x,y)$ represents the joint probability distribution function between the two variables X and Y, $p(x)$ means the marginal probability distribution of X, and $p(y)$ shows the marginal probability distribution of Y.

For continuous random variables, MI can be defined as the following.

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy \tag{8}$$

In the **Eq. 8**, $p(x,y)$ is the joint probability density function between the two variables X and Y, $p(x)$ is the marginal probability density function of X, and $p(y)$ is the marginal probability density function of Y.

Search strategy is the core of Wrapper's selection method. In this study, the simulated annealing algorithm was introduced in the optimization process to improve the convergence of the Wrapper method and form a new genetic algorithm (GA). The simulated annealing algorithm can be expressed as follows.

$$p = \begin{cases} 1, & if \ e_1 < e_2 \\ e^{-\frac{E(e_1 - e_2)}{T}}, & if \ e_1 > e_2 \end{cases} \tag{9}$$

In the **Eq. 9**, T, p, E, e_1, and e_2 represent the temperature, the substitution probability, the internal energy, the objective function, and the objective function of the substitute object in turn.

Fitness is an important index to evaluate individual survivability in GA (Hasserjian, 2019). For the evaluation function $f(x)$, the fitness (*Fit*) function is $F(x)$, so the *Fit* of the individual x can be expressed as $Fit = F[f(x)]$. When the largest problem is solved, the *Fit* can be expressed as shown in the **Eq. 10**. What's more, D means the minimum estimate of $f(x)$.

$$Fit = F[f(x)] = \begin{cases} f(x) - D, & f(x) > D \\ 0, & f(x) \le D \end{cases} \tag{10}$$

After introducing the simulated annealing algorithm, GA is improved and optimized. For the optimized GA, the parameters should be set, including the number of iterations of the population, the number of local search iterations, the initial size, the crossover probability, the probability of mutation, and the temperature. Multiple suitable individuals are used as the initial population, and the fitness of individuals in the population is calculated. If the termination condition is satisfied and the output optimal solution satisfies the termination condition, the algorithm ends. For the individuals that do not meet the termination conditions, crossover operation is carried out for each pair of matching individuals in the population according to the specified selection operator, and new populations are generated according to the local search strategy. Then, it is further verified whether the individuals meet the termination conditions and enter the next cycle. The optimized GA flow chart is shown in **Figure 1**.

Establishment of Depression Recognition Method Based on Meta Heuristic Algorithm

During the EEG acquisition process, different types of noise will have a certain impact on EEG. In this study, a band-pass filter is used to filter the data, and the EEG data are removed by the combination of Kalman filter and wavelet transform (Cabrera

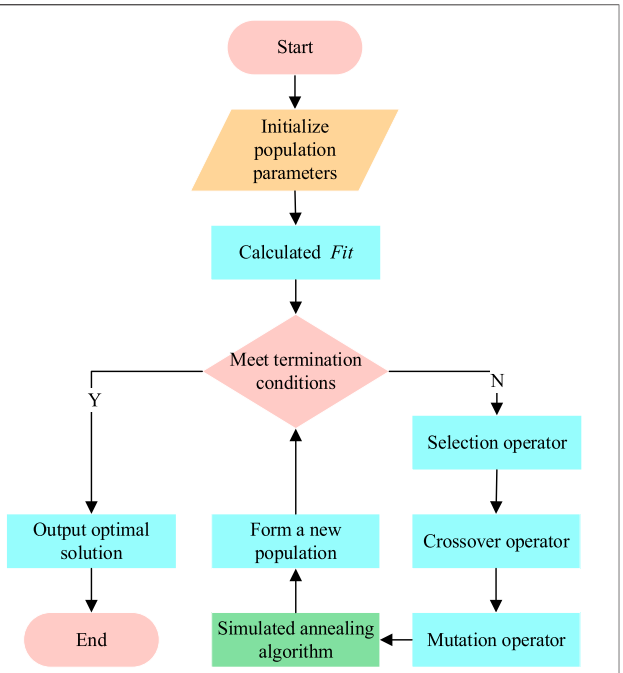


FIGURE 1 | Flow chart of optimized GA.

et al., 2018). The power spectrum entropy in nonlinear features is mainly applied to evaluate the strength of brain activity (Mendez-Balbuena et al., 2018). For the signal $X(w)$, the kilometer density is obtained after processing by the FAW-FS algorithm, and its power spectrum entropy can be expressed in the **Eq. 11**.

$$H_w = -\sum_{i=o}^{n} p_x(w_i)\log_2[p_x(w_i)] \tag{11}$$

Shannon entropy is employed to quantify EEG, and its calculation method is presented in the **Eq. 12**.

$$H_x = -x\log_2(x) - (1-x)\log_2(1-x) \tag{12}$$

The correlation dimension is mainly applied to describe the irregularity of EEG, and its calculation method is shown in the **Eq. 13**. In addition, $\ln C(\mathrm{r})$ stands for the correlation function.

$$C = \lim_{r\to 0}\frac{\ln C(\mathrm{r})}{\ln r} \tag{13}$$

Kolmogorov entropy describes the dynamic characteristics and signal complexity of the signal. The larger the Kolmogorov entropy, the more chaotic the dynamic characteristics, and the more complex the signal (Mutanen et al., 2018). The Kolmogorov calculation can be expressed in the **Eq. 14**.

$$K = \lim_{T\to 0}\lim_{\varepsilon\to 0}\lim_{N\to\infty}\frac{1}{NT}\sum_{n=0}^{N-1}(K_{n+1} - K_{\mathrm{n}}) \tag{14}$$

C0 complexity is adopted to evaluate the degree of randomness of EEG, which can be calculated in the **Eq. 15**. The greater the *C0* complexity value, the stronger the randomness of the EEG sequence.

$$C0 = \frac{\sum_{n=1}^{N}|X(\mathrm{n}) - Y(\mathrm{n})|^2}{\sum_{n=1}^{N}|X(\mathrm{n})|^2} \tag{15}$$

In the above equation, $X(\mathrm{n})$ represents the original EEG sequence, and $Y(\mathrm{n})$ means the EEG sequence after Fourier transform.

The collected EEG data are used for filtering and electro-oculogram operation through the band-pass filter, Kalman filter, and wavelet transform. Then, the current and nonlinear characteristic EEG data are extracted, and finally, the FAW-FS algorithm is employed to select the EEG features. The flow chart of depression recognition based on meta heuristic algorithm is shown in **Figure 2** below.

For each feature vector set output from the feature vector input module, it was first divided into a training set and a test set. Samples of the training set were derived from the public data set, with a sample size of 128. Each training set was divided into 1–5 of the 5 training subsets. Four training subsets out of the five training subsets were used for training in the deep forest each time, and the remaining one training subset was used as the verification set to verify the sub-model of training. The above sub-training process was repeated until every training subset in the whole training process made a verification set. After each verification of the trained sub-model, a set of feature vectors with a size of 12 was eventually obtained.

Assessment Method for Classification Accuracy of FAW-FS Algorithm

The data sets published in the public database (http://archive.ics.uci.edu/ml/index.php) were compared with the FAW-FS algorithm established in this study to verify the classification accuracy of the FAW-FS algorithm. The information of the 7 public data sets selected in this study was displayed in **Table 1**.

Accuracy was employed to evaluate the recognition results of depression EEG signals, and its specific calculation method was shown in the following equation.

$$Accuracy = \frac{\left|X:\ X\in D_{\mathrm{t}}\cap\bar{Y}(X) = Y(X)\right|}{|X:\ X\in D_{\mathrm{t}}|} \tag{16}$$

In the **Eq. 16**, D_{t} stood for the test data, X was the test sample, $Y(X)$ represented the real classification result of the test sample, and $\bar{Y}(X)$ indicated the classification result identified by the classification model.

Psychological Intervention Methods for Depression in Androgenic Alopecia Patients

The patients from both groups were given with paroxetine tablets (20 mg/time.d) for 3 weeks of drug treatment. On the basis of drug treatment, the control group was treated with routine psychological interventions, including sports, interest development, music listening, and social activities. The

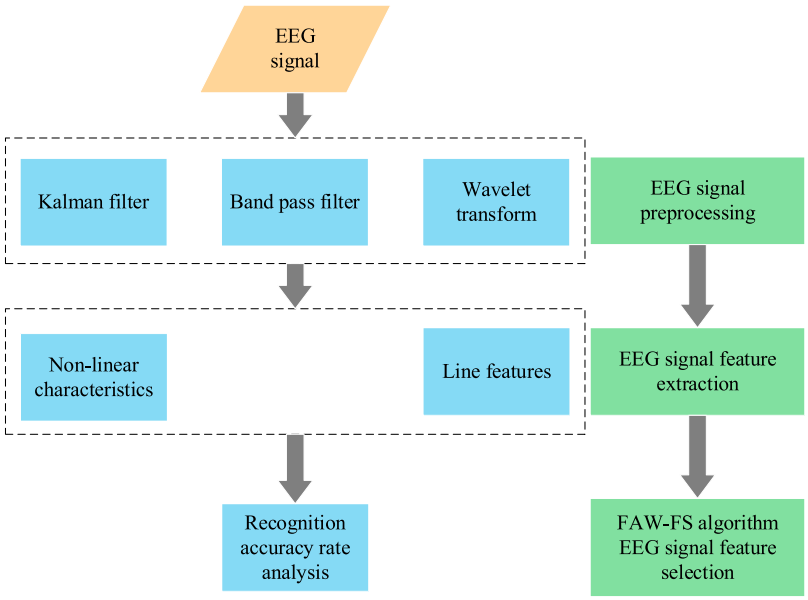


FIGURE 2 | Flow chart of depression recognition based on meta heuristic algorithm.

TABLE 1 | Relevant information of the public data sets.

Data set	Abbreviation	Sample size	Feature dimension	Class
Breast Cancer Wisconsin Data Set	BCW	569	30	2
Ionosphere Dataset	ION	351	33	2
Cryotherapy	CRY	90	7	2
SPECT Heart Data Set	SPE	267	22	2
Glass Identification	GI	241	9	2
Parkinson Multiple Sound Recording	PMS	1040	26	2
Connectionist Bench Data Set	CBD	208	60	2

intervention group underwent the comprehensive psychological nursing intervention on the basis of routine psychological intervention. The content of comprehensive psychological intervention mainly included the following. First, patients were guided to make psychological adjustments with psychological counseling, psychological care, and psychological support, so as to reduce their depression and build confidence in treatment. Second, the medical staff should explain the clinical manifestations, treatment, and prognosis of AGA to patients, thereby establishing a proactive cognitive model. Third, the medical staff needed to help patients establish support from family members and friends.

Psychological Intervention Observation Indexes for AGA Patients

The conditions of patients from the two groups were scored through the self-rating anxiety scale (SAS), self-rating depression scale (SDS), and Hamilton depression scale (HAMD) before and after treatment, respectively. SAS consists of 20 items in 1 dimension, which is scored from 1 to 4 levels; 50–59 points is considered as mild anxiety, 60–69 points as moderate anxiety, and 70 or above points as severe anxiety (Yue et al., 2020). There are also 20 items in 1 dimension of SDS, which are rated from 1 to 4 levels; 50–59 points is classified as mild depression, 60–69 points as moderate depression, and more than 70 points as severe depressions (Zou et al., 2016). The 1–4 levels were applied in the scoring of HAMD, with a total score of more than 35 points classified as severe depression; a score of 20–34 points indicates mild or moderate depression, and 8–20 points indicates mild depression (Zhao et al., 2019).

The differences of SAS, SDS, and HAMD scores before and after treatment were compared between the two groups. Besides, the changes in the depressive symptoms, treatment compliance, and quality of life of patients from the two groups were observed before and after treatment. The efficacy of depressive symptoms was evaluated by Jang et al. (2019). After treatment, the patient's HAMD score reduction rate was greater than 75%, which means that the patient was cured; 50% < HAMD score reduction rate ≤75% indicated that the efficacy was markedly effective; 25% < HAMD score reduction rate ≤50% showed effectiveness;

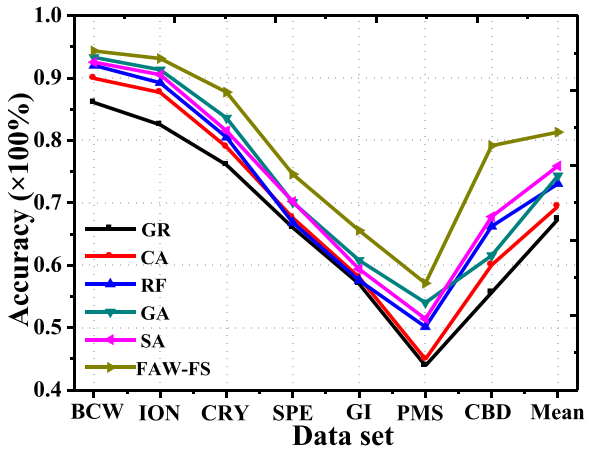


FIGURE 3 | Comparison on the classification accuracy of different algorithms under the logistic regression feature selection method.

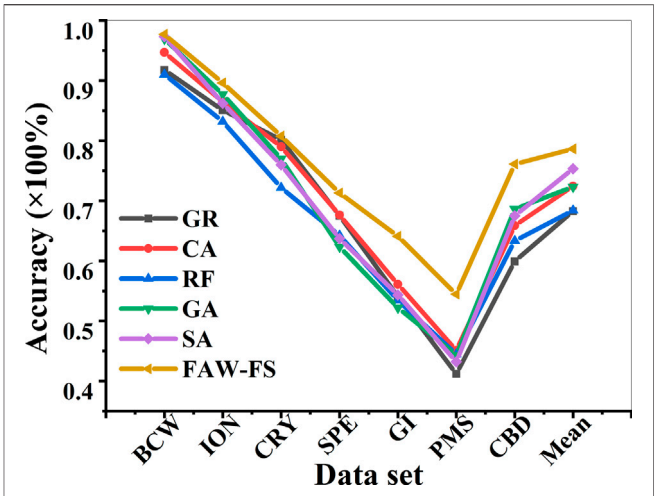


FIGURE 4 | Comparison on the classification accuracy of different algorithms under the classification features of DT.

HAMD score reduction rate was less than 25%, meaning that the efficacy was ineffective. In addition, the total effect included clinical recovery, marked effect, and effectiveness.

The method of Kraepelien et al. (2019) was referred to assess the compliance to the treatment of depression. Those who strictly followed the doctor's advice during treatment were complete compliance; those who basically followed the doctor's advice were basic compliance; those who often did not follow the doctor's advice or interrupt the treatment were regarded as non-compliance. Total compliance contained complete compliance and basic compliance.

Referring to the method of Teles et al. (2018), the quality of life of patients was evaluated before and after intervention for depression, and GQOLI-74 was adopted to analyze the 4 sub-items of the patient's body, psychology, society, and substance.

Statistical Methods

The experimental data were processed by SPSS19.0 statistical software, and the measurement data were expressed as mean ± standard deviation ($\bar{x} \pm s$). The count data were represented by percentage (%), and the χ^2 test was used. In addition, $p < 0.05$ indicated that the difference was statistically substantial.

RESULTS

Analysis of Classification Accuracy Based on Meta Heuristic Logistic Regression Algorithm

The classification accuracy of FAW-FS algorithm established in this study was compared with Correlation Attribute Eval (CA), Gain Ratio Attribute Eval (GR), Relief FAttribute Eval (RF), simulated annealing (SA) algorithm, and GA in the feature selection of logistic regression (**Figure 3**). In different public data sets, the classification accuracy of different algorithms changed in the same trend, while the classification accuracy of the same algorithm in different data sets varied greatly. In 7 different data sets, the classification accuracy of the FAW-FS algorithm was higher substantially than the accuracy of other algorithms, and its classification accuracy was 54.72–98.45%, with the mean classification accuracy of 80.87%.

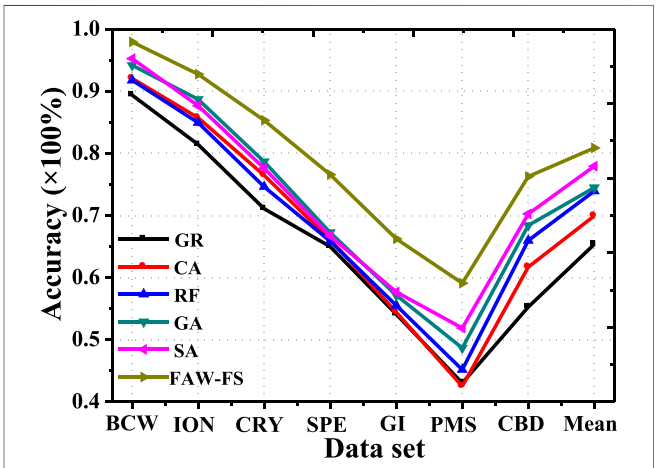


FIGURE 5 | Comparison on the classification accuracy of different algorithms under the features of K-nearest neighbor algorithm.

There was a comparison on the classification accuracy of the 6 algorithms under the classification features of DT (**Figure 4**). Among the 7 different data sets, all algorithms had the lowest classification accuracy in the PMS data set. The classification accuracy of FAW-FS algorithm rose obviously compared with other algorithms. Moreover, its classification accuracy was in the range of 43.28–98.81%, and the mean classification accuracy was 79.24%.

The classification accuracy of the 6 algorithms was compared under the K-nearest neighbor algorithm (**Figure 5**). In the 7 different data sets, all algorithms had the highest classification accuracy in the BCW data set. The classification accuracy of

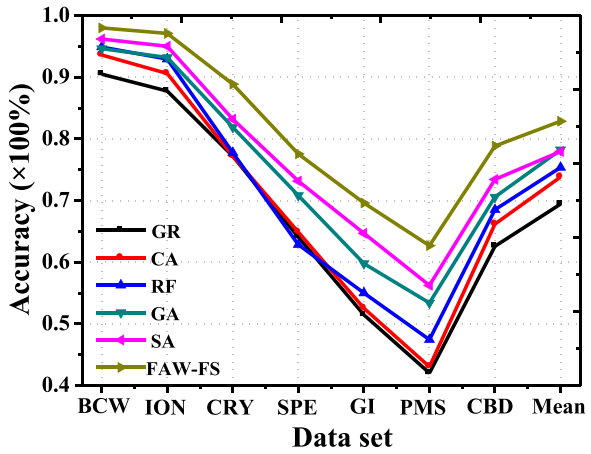


FIGURE 6 | Comparison on the classification accuracy of different algorithms under the feature of SVM.

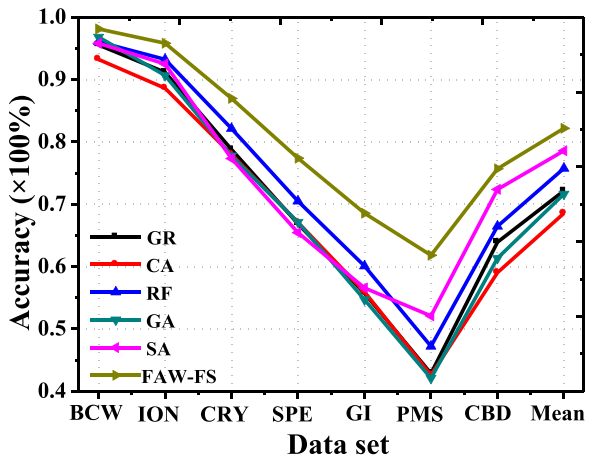


FIGURE 7 | Analysis on accuracy based on meta heuristic algorithm under the feature of RF.

FAW-FS algorithm elevated obviously in contrast to the accuracy of other algorithms. Its classification accuracy was distributed in the range of 42.94–99.12%, and the mean classification accuracy was 80.42%.

Figure 6 indicated that the classification accuracy of the 6 algorithms was compared under the features of SVM. In the 7 different data sets, the classification accuracy of the FAW-FS algorithm was higher hugely than that of other algorithms, and its classification accuracy was within 62.33–99.07%, with the mean classification accuracy of 83.07%.

The classification accuracy of the six algorithms was compared under the characteristics of RF, and the results were presented in **Figure 7**. In the seven different data sets, the classification accuracy of the FAW-FS algorithm was higher greatly than the accuracy of other algorithms, and its classification accuracy was 61.93–99.26%, with the mean classification accuracy of 81.45%.

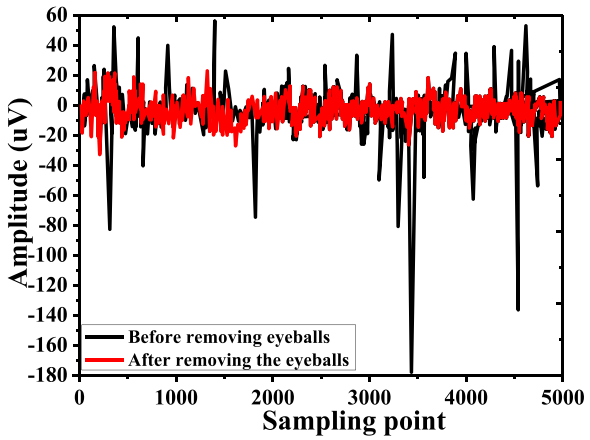


FIGURE 8 | EEG before and after removing electro-oculogram noise.

Analysis of Electroencephalogram de-noising Results Based on Meta Heuristic Algorithm

In this study, a combination of Kalman filter and wavelet transform was used to preprocess EEG to remove electro-oculogram noise before the FAW-FS algorithm was adopted to extract and select EEG features, and the results were shown in **Figure 8**. Before electro-oculogram noise was removed, EEG had more electro-oculogram artifacts. After removing electro-oculogram noise, a pure EEG was obtained.

During the processing of the EEG raw data (**Figure 9A**), the EEG data were initially processed with a down-sampling method of 1,000–250 Hz, so that the original signal was separated from the noise and the original data was enhanced (**Figure 9B**). A band-pass filter was applied to filter the data to remove the EEG artifacts in the EEG data (**Figure 9C**). Finally, the Kalman filter and wavelet transform were combined to remove the electro-oculogram artifacts in the EEG data, and the pure EEG data were obtained after extraction by the FAW-FS algorithm established in this study (**Figure 9D**).

Analysis on the Accuracy of Electroencephalogram Feature Selection Based on Meta Heuristic Algorithm

The accuracy of EEG feature selection of the FAW-FS algorithm under different data sets was analyzed under the resting state and five audio stimuli, as shown in **Figure 10**. It was found that among the five classification algorithms, the LR algorithm had the highest feature selection accuracy under the five audio types, and its mean feature selection accuracy was 82.94%, followed by KNN (73.72%) and RF (70.09%). The mean accuracy of feature selection for DT and SVM was 65.77 and 55.49%, respectively. The mean accuracy of feature selection of SVM was the lowest among the 5 algorithms. What's more, 5 different algorithms all had the highest mean

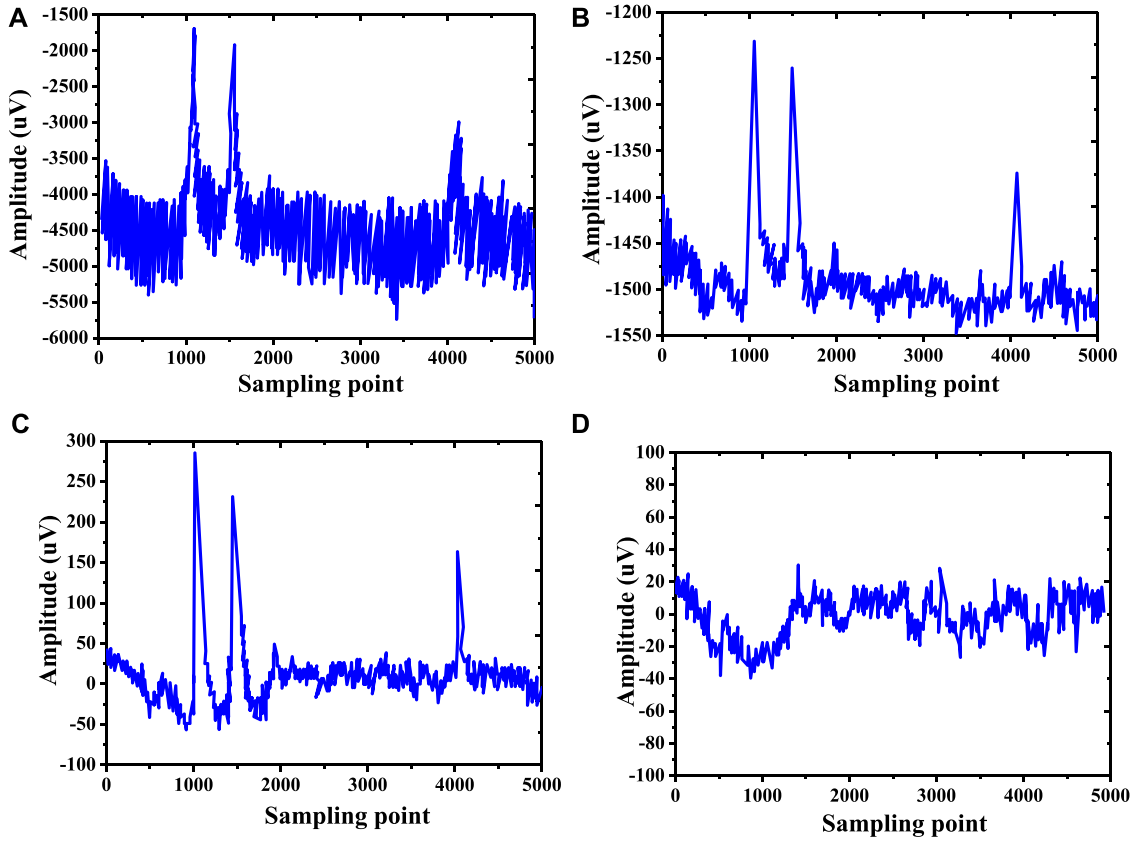


FIGURE 9 | FAW-FS algorithm EEG de-noising results. [Note: **(A)**: Original EEG image; **(B)**: EEG image after down-sampling processing; **(C)**: EEG image after band-pass filter; **(D)**: EEG image of FAW-FS algorithm feature extraction].

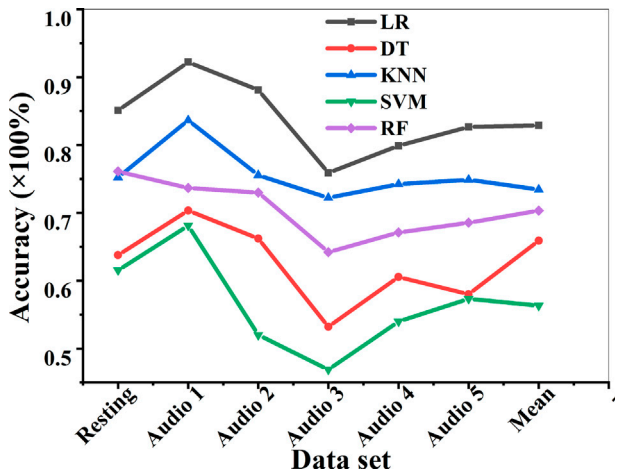


FIGURE 10 | Analysis on the accuracy of EEG feature selection under different conditions.

accuracy of feature selection on audio stimulus 1 in the 6 data sets of resting state EEG and audio stimulation EEG, and the lowest accuracy was on audio stimulus 3.

Analysis on the Accuracy of the Classification Algorithm on the Test Set

There was an analysis on the EEG accuracy of the FAW-FS algorithm under different data sets under the resting state and the five audio stimuli (**Figure 11**). It revealed that among the five classification algorithms, the RF algorithm had the highest classification accuracy under the five audio types, with a mean accuracy of 73.01%, followed by KNN (58.94%) and LR (52.76%). In addition, the mean accuracy of DT and SVM were 40.18 and 42.55% in turn. The mean accuracy of feature selection of DT was the lowest among the five algorithms. In the 6 data sets of resting state EEG and audio stimulation EEG, 5 different algorithms had the highest average accuracy on audio stimulation 1, and there was the lowest mean accuracy on audio stimulation 4.

Comparison on Basic Data of Patients From the Two Groups

The basic data of patients from the two groups were compared and analyzed, and the results were displayed in **Table 2**. There was no statistical difference in age, gender ratio, body mass index (BMI), weight, height, and course of disease between the two groups ($p > 0.05$).

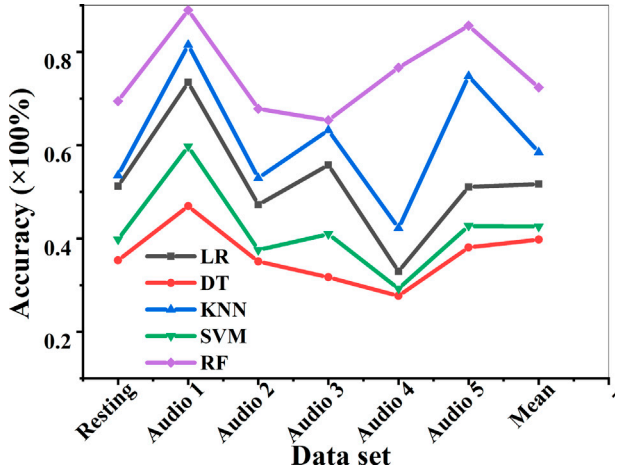


FIGURE 11 | The accuracy analysis of the classification algorithm on the test set.

Electroencephalogram Changes in Androgenic Alopecia Patients With Depression Before and After Psychological Intervention

The changes of EEG before and after psychological intervention in AGA patients with depression were analyzed (**Figure 12**). Before the intervention, the EEG power spectrum amplitude of AGA patients showed a smoothly downward trend with the continuous growth of the normalized frequency. The EEG power spectrum amplitude was distributed in the range of −0.5026–59.8248 dB, and the EEG mean power spectrum amplitude was 17.883 ± 8.190 dB. After the intervention, the EEG power spectrum amplitude of AGA patients rose first and then decreased with the continuous increase of the normalized frequency. The EEG power spectrum amplitude was distributed in the range of 21.0315–63.9881 dB, and the EEG mean power spectrum amplitude was 34.854 ± 3.465 dB.

Comparison on Depression Improvement Between the Two Groups of Patients Before and After Psychological Intervention

The scores of SDS, SAS, and HAMD scales before and after psychological intervention between the two groups were compared (**Figure 13**). There was no marked difference in SDS, SAS, and HAMD scale scores between the two groups of patients before the intervention ($p > 0.05$). After the intervention, the SDS and SAS scores of patients from the two groups were lower steeply than those before the treatment, and the difference was statistically obvious ($p < 0.05$). After the intervention, the HAMD scale scores of patients from the two groups were dramatically different from those before the treatment ($p < 0.01$). The SDS, SAS, and HAMD scores of the intervention group reduced sharply in contrast to the scores of the control group ($p < 0.05$).

There was a comparison on the quality of life scores of patients from the control group and the intervention group before and after psychological intervention (**Figure 14**). Before the intervention, there was no significant difference in the physical function, psychological function, social function, and substance function between the two groups of patients ($p > 0.05$). After the intervention, the physical function, mental function, social function, and substance function of patients from the two groups increased hugely compared with before the intervention, with a statistically huge difference ($p < 0.05$). The scores of physical function, mental function, social function, and substance function of the intervention group were higher markedly than the scores of the control group ($p < 0.05$).

Comparison on the Depression Efficacy and Compliance of Patients From the Two Groups After Treatment

Figure 15 showed the statistical analysis on the improvement of the efficacy of depression after treatment in patients from the two groups. In the control group, there were 4 cured cases (14.81%), 5 cases (18.52%) with marked effect, 7 effective cases (25.93%), and 11 cases (40.74%) with no effect after the intervention, and the total number of effective cases was 16 (59.26%). In the intervention group, 12 cases (44.44%) were cured, 8 cases (29.63%) were markedly effective, 5 cases (18.52%) were effective, and 2 cases (7.41%) were ineffective, so the total number of effective cases was 25 (92.59%). The proportion of cured, markedly effective, and total effective patients in the intervention group was higher greatly than the proportion of the control group ($p < 0.05$). After the intervention, there were 4 cases (14.81%) with complete compliance, 7 cases (25.93%) with basic compliance, and 16

TABLE 2 | Comparison on basic data of patients from the two groups.

Group	Control group (n = 27)	Intervention group (n = 27)	t value or χ^2 value	p Value
Age (years old)	40.12 ± 4.07	38.85 ± 5.03	0.554	0.121
Male [cases, (%)]	15 (55.56)	16 (59.26)	0.083	0.224
Female [case, (%)]	12 (44.44)	11 (40.74)	0.063	0.257
BMI (kg/m^2)	24.92 ± 2.07	26.03 ± 2.59	0.134	0.717
Weight (kg)	59.09 ± 11.98	62.07 ± 9.98	0.094	0.392
Height (cm)	162.95 ± 5.43	163.77 ± 5.19	0.141	0.762
Course of disease (years)	2.09 ± 1.28	2.26 ± 1.33	0.196	0.824

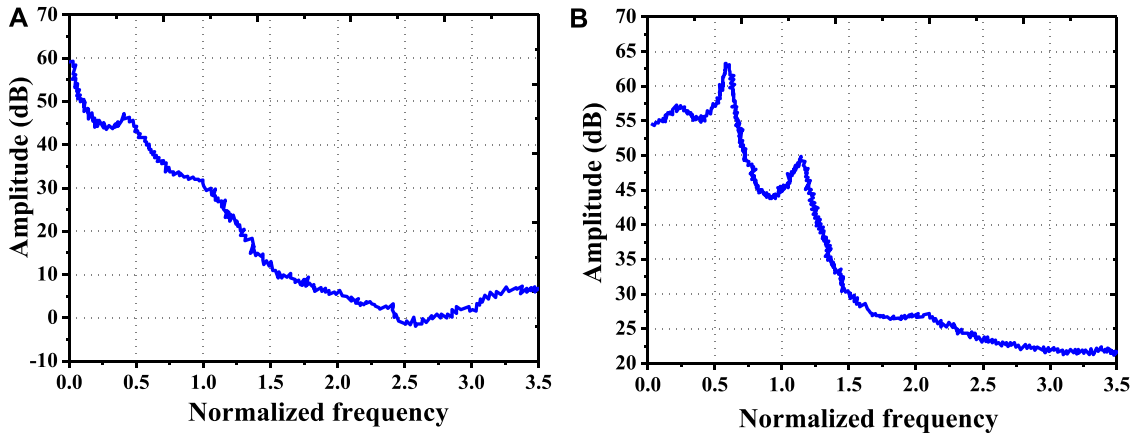


FIGURE 12 | EEG changes in AGA patients with depression before and after psychological intervention. [Note: **(A)**: EEG of patients with AGA before psychological intervention; **(B)**: EEG of patients with AGA after psychological intervention].

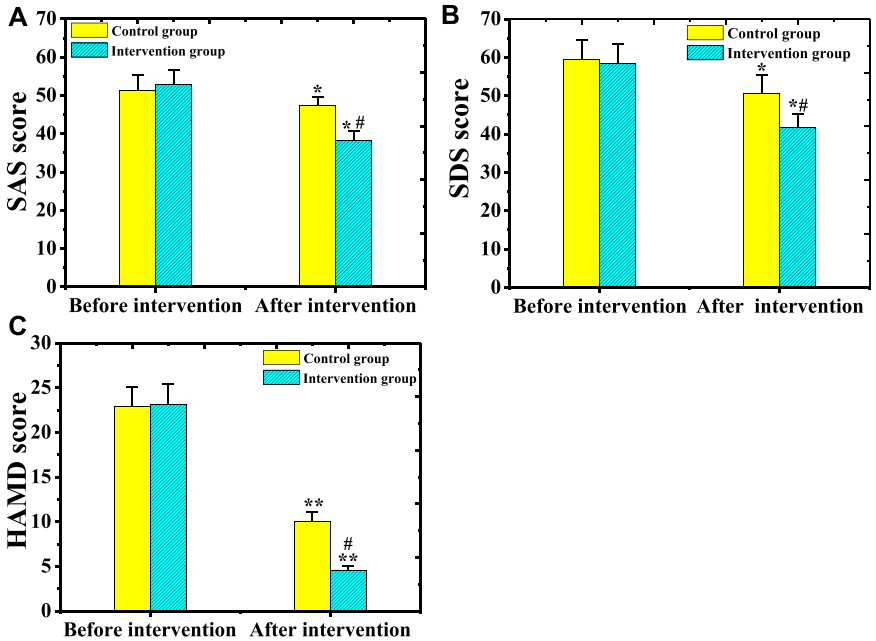


FIGURE 13 | Comparison on SDS, SAS, and HAMD scale scores between the two groups of patients before and after psychological intervention.[Note: **(A)**: Comparison of SAS scale scores before and after psychological intervention between the two groups of patients; **(B)**: Comparison of SDS scale scores before and after psychological intervention between the two groups of patients; **(C)**: Comparison of HAMD scale scores before and after psychological intervention between the two groups of patients; * indicated $p < 0.05$ compared with before the intervention; ** meant $p < 0.01$ compared with before the intervention; # showed $p < 0.05$ compared with the control group].

cases (59.26%) with no compliance in the control group, and the total number of cases with compliance was 11 (40.74%). In the intervention group, 15 patients (55.56%), 9 patients (33.33%), and 3 patients (11.112%) were completely compliant, basically compliant, and non-compliant, so there were 24 cases with compliance (88.89%). The proportion of patients with complete compliance and total compliance in the intervention group elevated substantially compared with the control group, and there was a significant difference between the two ($p < 0.01$).

DISCUSSION

In this study, FAW-FS algorithm was established based on meta-heuristic algorithm and applied to depression disorder

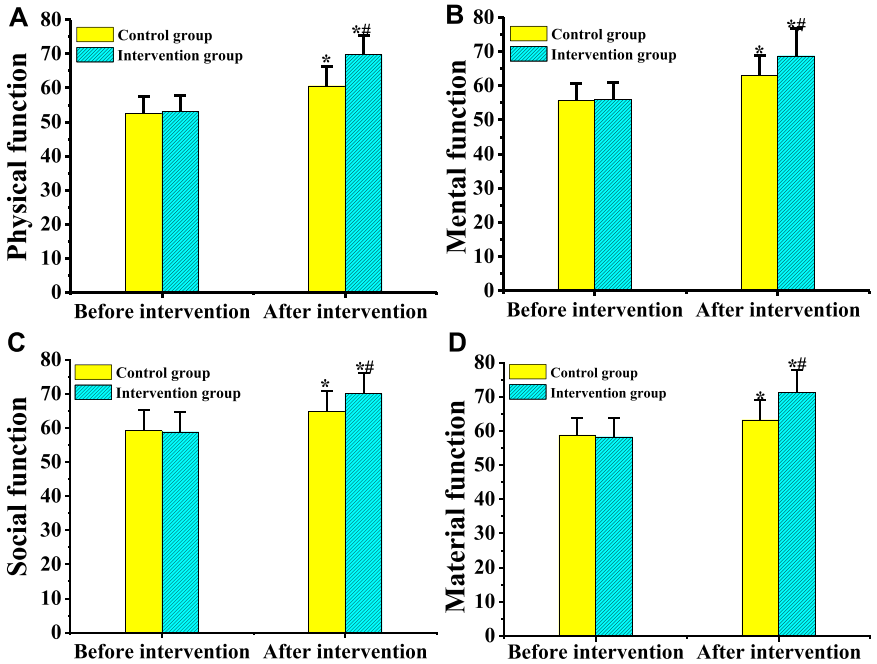


FIGURE 14 | Comparison on the quality of life of patients from the two groups before and after psychological intervention.[Note: **(A)**: Comparison of the physical function of the two groups of patients before and after psychological intervention; **(B)**: Comparison of the psychological function of the two groups of patients before and after psychological intervention; **(C)**: Comparison of the social function of the two groups of patients before and after psychological intervention; **(D)**: Comparison of the substance function of the two groups of patients before and after psychological intervention; * indicated $p < 0.05$ compared with before the intervention; # showed that $p < 0.05$ in contrast to the control group].

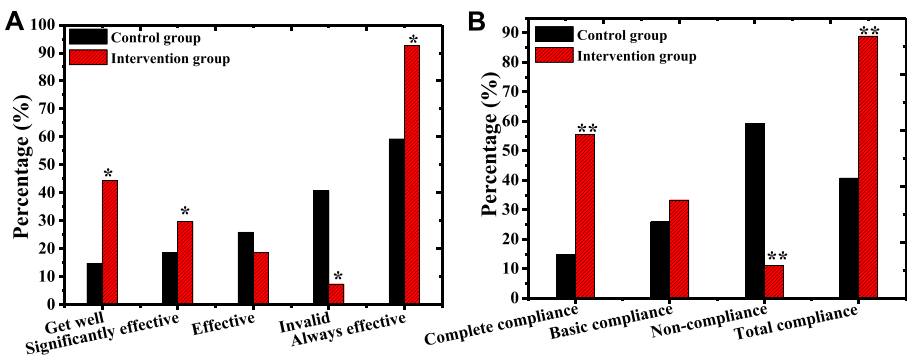


FIGURE 15 | Comparison on the efficacy and compliance of depression after psychological intervention between the two groups.[Note: **(A)**: Comparison on the efficacy of depression after intervention between the two groups of patients; **(B)**: Comparison of intervention compliance between the two groups of patients; * indicated $p < 0.05$ compared with the control group; ** meant $p < 0.01$ compared with the control group].

recognition. The classification accuracy of FAW-FS algorithm was compared with CA, GR, RF, SA, and GA algorithms in deep learning under different feature selection methods. CA, GR and RF algorithms are all classic Filter feature selection methods, and SA and GA algorithms are commonly used methods in Wrapper feature selection methods (Becerra-Sánchez et al., 2020). The results of this study showed that FAW-FS algorithm had the highest classification accuracy in LR, DT, KNN, SVM, and RF, which were 80.87, 79.24, 80.42, 83.07, and 81.45% respectively. These results indicated that the classification accuracy of FAW-FS established in this study was higher than that of the Filter feature selection method and Wrapper feature selection method. The reason was that the FAW-FS algorithm had the high accuracy of Wrapper and simplicity of Filter calculation (Albasri et al., 2019), and the two Filter feature selection algorithms through ANOVA and mutual information were used to initially screen the original data, forming a new feature space (Varsehi and Firoozabadi, 2021). In the calculation process, the local optimization method

of individuals in the population and the simulated annealing strategy were adopted to improve the premature convergence of GA, and finally, the classification accuracy of the FAW-FS algorithm was promoted. The results of this study suggested that the LR algorithm had the highest feature selection accuracy of 82.94% under five audio types. The five different algorithms all had the highest average feature selection accuracy on audio stimulus 1, and audio stimulus 3 had the lowest. In this study, the EEG data on the three electrode positions Fp1, Fp2, and Fpz were used to identify depression, and the highest recognition accuracy was 73.01%. Mohammadi et al. (2020) (Mohammadi and Moradi, 2021) used meta-heuristic algorithm to classify FP1, FP2, and FPZ EEG data of patients with depression, and the highest accuracy was 70.24%. Bachmann et al. (2018) also applied Fp1, Fp2, and Fpz to classify depression, and the highest accuracy was 71.29%. The classification accuracy of the FAW-FS algorithm in this study was higher substantially than these studies, indicating that the fusion feature selection algorithm FAW-FS had a certain practicability and generalization, which could improve the accuracy of depression recognition to a certain extent.

At present, a large number of research results disclose that necessary psychological intervention for patients with AGA is of great significance for disease control and treatment (Wang et al., 2018). In this study, patients from the two groups were treated with different psychological intervention methods to explore the influence of comprehensive psychological intervention methods on patients. The results showed that the scores of SDS, SAS, and HAMD of patients from the two groups dropped hugely after intervention compared with before treatment ($p < 0.05$). The scores of SDS, SAS, and HAMD in the intervention group were lower dramatically than the scores of the control group ($p < 0.05$). The scores of physical function, psychological function, social function, and substance function in patients from the intervention group were higher remarkably than those of the control group ($p < 0.05$). The proportion of cured, effective, and total effective patients from the intervention group elevated obviously compared with the control group ($p < 0.05$). The proportion of patients with complete compliance and total compliance in the intervention group was bigger substantially than the proportion of the control group ($p < 0.01$). Therefore, these results showed that comprehensive psychological intervention was more helpful to the recovery of patients with AGA, which was similar to the research findings of Gonzalez et al. (2010).

CONCLUSION

In this study, a depression EEG signal recognition model FAW-FS was established based on deep learning meta-heuristic algorithm, which was applied to the recognition of depression EEG signals in patients with AGA. A comprehensive psychological intervention method was adopted to intervene in patients with AGA. The results showed that the FAW-FS algorithm based on deep learning meta-heuristic algorithm could significantly improve the accuracy of depression disorder recognition, and comprehensive psychological intervention played a positive role in the rehabilitation of depression disorder in patients with AGA. However, there are still some shortcomings in this study. In this study, the EEG data on FP1, FP2, and FPZ are only classified, and the EEG characteristic data of Beta bands related to depression recognition are not analyzed. The value of FAW-FS algorithm in the classification of EEG characteristic data in the sub-bands will be further explored in the future work. In conclusion, the FAW-FS algorithm established based on meta-heuristic algorithm in this study can improve the accuracy of EEG signal recognition for depression disorders, and comprehensive psychological intervention plays a positive role in the rehabilitation of depression disorders in patients with AGA, thereby providing a reference basis for the diagnosis and treatment of AGA patients.

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.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of Shanghai East Hospital. The participants provided their written informed consent to participate in the study.

AUTHOR CONTRIBUTIONS

YL, BW and XL: Writing- Original draft preparation QZ, XY and YL: Resources; Supervision; data analysis, Investigation.

FUNDING

This work was supported by the east hospital affiliated to tongji university introduced talent research startup fund (DFRC2019008) and shanghai pudong new area health committee clinical characteristic discipline fund.

REFERENCES

Albasri, A., Abdali-Mohammadi, F., and Fathi, A. (2019). EEG Electrode Selection for Person Identification Thru a Genetic-Algorithm Method. *J. Med. Syst.* 43 (9), 297. doi:10.1007/s10916-019-1364-8

Bachmann, M., Päeske, L., Kalev, K., Aarma, K., Lehtmets, A., Ööpik, P., et al. (2018). Methods for Classifying Depression in Single Channel EEG Using Linear and Nonlinear Signal Analysis. *Computer Methods Programs Biomed.* 155, 11–17. doi:10.1016/j.cmpb.2017.11.023

Becerra-Sánchez, P., Reyes-Munoz, A., and Guerrero-Ibañez, A. (2020). Feature Selection Model Based on EEG Signals for Assessing the Cognitive Workload in Drivers. *Sensors* 20 (20), 5881. doi:10.3390/s20205881

Cabrera, A. F., Petersen, E. B., Graversen, C., Sorensen, A. T., Lunner, T., and Rank, M. L. (2018). Individual Classification of Single Trial EEG Traces to Discriminate Brain Responses to Speech with Different Signal-To-Noise Ratios. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2018, 987–990. doi:10.1109/EMBC.2018.8512491

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep Learning for Electroencephalogram (EEG) Classification Tasks: a Review. *J. Neural Eng.* 16 (3), 031001. doi:10.1088/1741-2552/ab0ab5

Eilbeigi, E., and Setarehdan, S. K. (2018). Global Optimal Constrained ICA and its Application in Extraction of Movement Related Cortical Potentials from Single-Trial EEG Signals. *Computer Methods Programs Biomed.* 166, 155–169. doi:10.1016/j.cmpb.2018.07.013

Gonzalez, M. E., Cantatore-Francis, J., and Orlow, S. J. (2010). Androgenetic Alopecia in the Paediatric Population: a Retrospective Review of 57 Patients. *Br. J. Dermatol.* 163 (2), 378–385. doi:10.1111/j.1365-2133.2010.09777.x

Hasserjian, R. P. (2019). Myelodysplastic Syndrome Updated. *Pathobiology* 86 (1), 7–13. doi:10.1159/000489702

Jang, H.-M., Lee, K.-E., and Kim, D.-H. (2019). The Preventive and Curative Effects of Lactobacillus Reuteri NK33 and Bifidobacterium Adolescentis NK98 on Immobilization Stress-Induced Anxiety/depression and Colitis in Mice. *Nutrients* 11 (4), 819. doi:10.3390/nu11040819

Kraepelien, M., Blom, K., Lindefors, N., Johansson, R., and Kaldo, V. (2019). The Effects of Component-specific Treatment Compliance in Individually Tailored Internet-Based Treatment. *Clin. Psychol. Psychother* 26 (3), 298–308. doi:10.1002/cpp.2351

Lake, F. (2019). From Industry 4.0 to Lab 4.0. *Biotechniques* 66 (6), 247. doi:10.2144/btn-2019-0061

Lolli, F., Pallotti, F., Rossi, A., Fortuna, M. C., Caro, G., Lenzi, A., et al. (2017). Androgenetic Alopecia: a Review. *Endocrine* 57 (1), 9–17. doi:10.1007/s12020-017-1280-y

Mendez-Balbuena, I., Arrieta, P., Huidobro, N., Flores, A., Lemuz-Lopez, R., Trenado, C., et al. (2018). Augmenting EEG-Global-Coherence with Auditory and Visual Noise. *Medicine (Baltimore)* 97 (35), e12008. doi:10.1097/MD.0000000000012008

Mohammadi, Y., and Moradi, M. H. (2021). Prediction of Depression Severity Scores Based on Functional Connectivity and Complexity of the EEG Signal. *Clin. EEG Neurosci.* 52 (1), 52–60. doi:10.1177/1550059420965431

Munoz, R., Olivares, R., Taramasco, C., Villarroel, R., Soto, R., Barcelos, T. S., et al. (2018). Using Black Hole Algorithm to Improve EEG-Based Emotion Recognition. *Comput. Intell. Neurosci.* 2018, 3050214. doi:10.1155/2018/3050214

Mutanen, T. P., Metsomaa, J., Liljander, S., and Ilmoniemi, R. J. (2018). Automatic and Robust Noise Suppression in EEG and MEG: The SOUND Algorithm. *Neuroimage* 166, 135–151. doi:10.1016/j.neuroimage.2017.10.021

Padfield, N., Zabalza, J., Zhao, H., Masero, V., and Ren, J. (2019). EEG-based Brain-Computer Interfaces Using Motor-Imagery: Techniques and Challenges. *Sensors* 19 (6), 1423. doi:10.3390/s19061423

Peng, G., Nourani, M., Harvey, J., and Dave, H. (2020). Feature Selection Using F-Statistic Values for EEG Signal Analysis. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2020, 5963–5966. doi:10.1109/EMBC44109.2020.9176434

Phadikar, S., Sinha, N., and Ghosh, R. (2021). Automatic Eyeblink Artifact Removal from EEG Signal Using Wavelet Transform with Heuristically Optimized Threshold. *IEEE J. Biomed. Health Inform.* 25 (2), 475–484. doi:10.1109/JBHI.2020.2995235

Rajabi, F., Drake, L. A., Senna, M. M., and Rezaei, N. (2018). Alopecia Areata: a Review of Disease Pathogenesis. *Br. J. Dermatol.* 179 (5), 1033–1048. doi:10.1111/bjd.16808

Starace, M., Orlando, G., Alessandrini, A., and Piraccini, B. M. (2020). Female Androgenetic Alopecia: an Update on Diagnosis and Management. *Am. J. Clin. Dermatol.* 21 (1), 69–84. doi:10.1007/s40257-019-00479-x

Tanaka, Y., Aso, T., Ono, J., Hosoi, R., and Kaneko, T. (2018). Androgenetic Alopecia Treatment in Asian Men. *J. Clin. Aesthet. Dermatol.* 11 (7), 32–35.

Teles, F., Amorim de Albuquerque, A. L., Freitas Guedes Lins, I. K., Carvalho Medrado, P., and Falcão Pedrosa Costa, A. (2018). Quality of Life and Depression in Haemodialysis Patients. *Psychol. Health Med.* 23 (9), 1069–1078. doi:10.1080/13548506.2018.1469779

Varsehi, H., and Firoozabadi, S. M. P. (2021). An EEG Channel Selection Method for Motor Imagery Based Brain-Computer Interface and Neurofeedback Using Granger Causality. *Neural Networks* 133, 193–206. doi:10.1016/j.neunet.2020.11.002

Völker, J. M., Koch, N., Becker, M., and Klenk, A. (2020). Caffeine and its Pharmacological Benefits in the Management of Androgenetic Alopecia: a Review. *Skin Pharmacol. Physiol.* 33 (3), 93–109. doi:10.1159/000508228

Wang, X., Xiong, C., Zhang, L., Yang, B., Wei, R., Cui, L., et al. (2018). Psychological Assessment in 355 Chinese College Students with Androgenetic Alopecia. *Medicine (Baltimore)* 97 (31), e11315. doi:10.1097/MD.0000000000011315

Wen, D., Yuan, J., Zhou, Y., Xu, J., Song, H., Liu, Y., et al. (2020). The EEG Signal Analysis for Spatial Cognitive Ability Evaluation Based on Multivariate Permutation Conditional Mutual Information-Multi-Spectral Image. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28 (10), 2113–2122. doi:10.1109/TNSRE.2020.3018959

Yue, T., Li, Q., Wang, R., Liu, Z., Guo, M., Bai, F., et al. (2020). Comparison of Hospital Anxiety and Depression Scale (HADS) and Zung Self-Rating Anxiety/Depression Scale (SAS/SDS) in Evaluating Anxiety and Depression in Patients with Psoriatic Arthritis. *Dermatology* 236 (2), 170–178. doi:10.1159/000498848

Zhao, B., Li, Z., Wang, Y., Ma, X., Wang, X., Wang, X., et al. (2019). Can Acupuncture Combined with SSRIs Improve Clinical Symptoms and Quality of Life in Patients with Depression? Secondary Outcomes of a Pragmatic Randomized Controlled Trial. *Complement. Therapies Med.* 45, 295–302. doi:10.1016/j.ctim.2019.03.015

Zou, Q., Jiang, Y., Mu, F., Shi, Y., and Fang, Y. (2016). Correlation of Axial Spondyloarthritis with Anxiety and Depression. *Med. Sci. Monit.* 22, 3202–3208. doi:10.12659/msm.897232

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.

Copyright © 2021 Li, Wu, Li, Zhou, Yang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 November 2021
doi: 10.3389/fgene.2021.784814

Improving the Classification of Alzheimer's Disease Using Hybrid Gene Selection Pipeline and Deep Learning

Nivedhitha Mahendran[1], *P. M. Durai Raj Vincent*[1]*, *Kathiravan Srinivasan*[2] and *Chuan-Yu Chang*[3]*

[1]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India, [2]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India, [3]Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan

Alzheimer's is a progressive, irreversible, neurodegenerative brain disease. Even with prominent symptoms, it takes years to notice, decode, and reveal Alzheimer's. However, advancements in technologies, such as imaging techniques, help in early diagnosis. Still, sometimes the results are inaccurate, which delays the treatment. Thus, the research in recent times focused on identifying the molecular biomarkers that differentiate the genotype and phenotype characteristics. However, the gene expression dataset's generated features are huge, 1,000 or even more than 10,000. To overcome such a curse of dimensionality, feature selection techniques are introduced. We designed a gene selection pipeline combining a filter, wrapper, and unsupervised method to select the relevant genes. We combined the minimum Redundancy and maximum Relevance (mRmR), Wrapper-based Particle Swarm Optimization (WPSO), and Auto encoder to select the relevant features. We used the GSE5281 Alzheimer's dataset from the Gene Expression Omnibus We implemented an Improved Deep Belief Network (IDBN) with simple stopping criteria after choosing the relevant genes. We used a Bayesian Optimization technique to tune the hyperparameters in the Improved Deep Belief Network. The tabulated results show that the proposed pipeline shows promising results.

Keywords: deep learning, Alzheimer's disease, gene selection, gene expression, molecular bio-markers

OPEN ACCESS

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

Reviewed by:
M.A. Khan,
HITEC University, Pakistan
Suresh Manic Kesavan,
National University of Science and Technology, Oman

***Correspondence:**
P. M. Durai Raj Vincent
pmvincent@vit.ac.in
Chuan-Yu Chang
chuanyu@yuntech.edu.tw

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

Received: *28 September 2021*
Accepted: *20 October 2021*
Published: *12 November 2021*

Citation:
Mahendran N, Vincent PMDR, Srinivasan K and Chang C-Y (2021) Improving the Classification of Alzheimer's Disease Using Hybrid Gene Selection Pipeline and Deep Learning. Front. Genet. 12:784814. doi: 10.3389/fgene.2021.784814

INTRODUCTION

Dementia is a broad term for a group of disorders with abnormal changes in the brain. The common forms of dementia interrupt the communication between the brain cells (Salat et al., 2001). When the communication between the cells is disrupted, the cognitive abilities, such as memory loss, feelings, thinking, and problem solving, behavior, and language proficiency of the individual will also be disrupted (Jo et al., 2019). Some of the common types of dementia are Parkinson's disease, Lewy body dementia, Alzheimer's disease (AD), Down's syndrome, vascular dementia, dementia caused by alcohol, and HIV. Among these, 60–70% is accounted for by AD (Lawrence and Sahakian, 1995). Recently, there are increasing researches in the field of gerontology, a study of the physical aspects of aging. One such neurological disorder that appears in the elderly is the AD. Our work in this paper focuses on AD. AD is known to humankind for more than 100 years, yet the molecular mechanism and pathogenesis is far from fully understood (Reitz et al., 2011).

As commonly said, AD is a progressive, neurodegenerative brain disease, which is irreversible. The term progressive concerning AD means it gets worse over time because of irreversible degeneration of neurons (neurodegenerative) (Nussbaum and Ellis, 2003). In other words, the pathological change of AD is a slowly accumulating process. AD affects the hippocampus and cortex regions of the brain primarily. The primary reason for developing AD symptoms is more than the required accumulation of proteins around the brain cells (Wenk, 2003). The high levels of proteins make the communication between brain cells tedious. The actual reason for the onset of AD is still uncertain. Yet, few hypotheses were framed over the years, such as the accumulation of Tau and amyloid proteins, cholinergic, and genetics (Citron, 2010). Even with dominant symptoms, the dysfunctions of AD take years to be noticed, decoded, and revealed.

The early diagnosis starts with recognizing the mild cognitive impairment (MCI), which has a high possibility of causing AD (Liu et al., 2014). The onset of AD is commonly found around 65 years; however, early onset at a younger age is rare. Even after thorough research, the cause and progression seem to be uncertain (Huber et al., 2018). The proper diagnosis can be made only after the autopsy, yet, with advanced technologies in clinical screening, such as cerebrospinal fluid analysis, imaging techniques have led to early AD diagnosis. These methods provide inaccurate results, which delays the treatment at times (Wang and Liu, 2019). The limitations in clinical screening have led to the molecular data-based analysis. Identifying molecular biomarkers offers promising results, as it establishes accurate relationships between the phenotype and genotype symptoms. The accurate and early diagnosis of AD will help patients have the awareness and indulge in preventive measures, for instance, medications and changes in their lifestyle.

Although the molecular biomarkers offer better results than the clinical screening, the environmental and genetic factors should be taken into account. There are more than 1,000 even 10,000 features generated through transcripts, genes, proteins, and their interaction with each other (Moradifard et al., 2018). It is a considerable challenge to find the AD causing biomarkers from such Big data. Thus, machine learning and Artificial intelligence-based methods are focused on these days to meet the challenges. There is another issue with the molecular biomarkers; more than the volume, the dimensionality of the dataset increases faster (Tanveer et al., 2020). Molecular data, such as gene expression, are ultra-high dimensional datasets. The dimensionality is achieving higher levels of thousands and hundreds of thousands.

Meanwhile, the sample size did not witness the same amount of growth. Such a situation is commonly known as the High Dimensionality Low Sample Size (HDLSS) problem or "curse of dimensionality". The machine learning techniques widely used are not suited for such cursed dimensional data (Lee and Lee, 2020). The inconsistent ratio between the number of features and the number of samples will lead to overfitting, incompatible algorithm, and extended computational time.

To solve the curse of dimensionality problem, feature selection is proposed as a solution. In this study, we develop a gene selection pipeline combining filter, wrapper, and unsupervised method to select the relevant features in causing AD. Later, the selected genes are passed through the Improved Deep Belief Network (IDBN), which is implemented to classify the AD and non-AD individuals. The selection of relevant features will make the classification of AD and non-AD individuals accurate and easy.

BACKGROUND AND MOTIVATION

Alzheimer's Disease and Machine Learning Algorithms

The most widely used technique in diagnosing AD is the clinical screening methods, such as brain imaging. At times, the clinical screening methods provide inaccurate results due to technical errors, which eventually delays the treatment. Hence, the research is gradually moving towards molecular data, for instance, microarray data. In the process of finding out differentially expressed genes, thousands of genes are captured and monitored to evaluate the effects of a disease or a treatment (Fung and Stoeckel, 2007). For detecting the expression of hundreds and thousands of genes simultaneously, microarray technology is used widely. In microarray, thousands of genes or DNA sequences are printed in already defined positions. The DNA microarray datasets have vast volumes of genes captured, which might not be relevant to the undertaken domain (treatment/disease) (Huang et al., 2018).

Considering the huge volume of features involved, machine learning-based methods help greatly in classifying AD patients from healthy individuals. Machine learning is a continuously growing area of research, advantageous in many domains, mainly in healthcare. Machine learning algorithms are trained on a set of data, learn from the data, find out the patterns, and predict the future possibilities without much human intervention (Orimaye et al., 2017). It is a part of Artificial Intelligence, assists in data analysis, and automates model building. There are three categories of machine learning algorithm based on the dataset used (Hutter et al., 2019): supervised learning, when the data are structured and attributes are labeled; unsupervised learning, when the data are unstructured and the attributes are unlabeled; and semi-supervised/semi-unsupervised learning, when the data are a combination of supervised and unsupervised categories. Although machine learning algorithms offer great assistance in finding patterns and classification, it is not suitable when the ratio of sample to feature is largely different. In that case, machine learning algorithms will have an overfitting problem.

Related Works

Artificial Intelligent models have been widely deployed in genetics research (Mahendran et al., 2020). Deep learning approaches remove certain data pre-processing, which is usually deployed in machine learning (Srinivasan et al., 2017; Agarwal et al., 2018; Chakriswaran et al., 2019; Khan et al., 2021a; Khan et al., 2021b; Khan et al., 2021c) (Sanchez-Riera et al., 2018; Srinivasan et al., 2020; Afza et al., 2021; Ashwini et al., 2021;

Attique Khan et al., 2021; Khan et al., 2021d; Mamdiwar et al., 2021; Srinivasan et al., 2021). AD is a neurological disorder identified through brain imaging, and there are many works focused on classifying AD through brain images with the help of machine learning or deep learning techniques. For instance, Convolutional Neural Network (CNN) and LeNet architecture is applied on the MRI data to classify AD (Sarraf and Tofighi, 2016a). There are many such works focused on classifying AD through the brain images (Sarraf and Tofighi, 2016b; Farooq et al., 2017; Ji et al., 2019; Ramzan et al., 2020; Tufail et al., 2020), though the imaging data provide inaccurate results at times. Thus, the focus recently is shifted to the molecular dataset such as the Gene Expression and DNA Methylation data, though the problem with such data is the dimensionality. There is a huge number of features, yet very small sample size.

Therefore, the research is focused more on the gene selection techniques to select the relevant features in classifying the AD. For instance, Park et al. (Park et al., 2020) implemented machine learning-based gene selection and a deep learning classifier combining the gene expression and DNA Methylation datasets. Also, the gene pair interaction-based research is done to identify the biomarkers accurately to classify the AD (Chen et al., 2019). Furthermore, there are approaches implemented to detect the possible progression of a dementia to AD with the help of machine learning techniques (Martínez-Ballesteros et al., 2017; Miao et al., 2017). Also, the artificial intelligence approaches are adopted in precision medicine to validate the drugs for AD.

Feature Selection

There are four mainly used feature selection techniques, Filter, Wrapper, Hybrid, and Ensemble (Bashir et al., 2019). Filter-based techniques are independent of the classifier model and computationally efficient at times (Acharya et al., 2019). The search for relevant features is isolated completely from the classifier model. The features with the lowest relevance score are eliminated. The filter methods are further classified into univariate and multivariate filters, where univariate treats and evaluates the features individually and multivariate evaluates the feature dependencies. The wrapper methods are implemented as a part of the classification model (Zhou et al., 2018). The feature subsets selected are validated through training and testing datasets. The features with maximum evaluation score are selected for the final classification. The wrapper method's major drawbacks are as follows: it demands high computational time, it is classifier dependent, and overfitting (Mirzaei et al., 2018).

The ensemble methods simultaneously build different feature subsets and combine the results using standard aggregate methods, such as majority voting, sum rule, mean rule, and weighted voting (Pes, 2019). The exponential growth of technologies across all the domains created a data explosion, which is continuously spreading at an unprecedented speed. The previously mentioned feature selection methods are not suitably designed for a dataset with HDLSS problem and unstable and not robust with changing inputs. Thus, ensemble methods are designed aiming to bring more robustness and stability to the model (Neumann et al., 2017). The main goal of the ensemble model is to attain a better trade-off between stability and predictive performance. The ensemble methods are generally grouped under homogeneous and heterogeneous methods. The homogeneous algorithms use selection algorithm with the varying dataset, for instance, boosting or bagging. The homogeneous ensembles handle the stability issues better. The heterogeneous ensembles implement different selection algorithms with the same dataset. In both cases, the output will be combined to a single feature set, which probably provides an optimal solution (Pes, 2019). Apart from homogeneous and heterogeneous methods, there is another group called the hybrid, which uses different selection algorithms with other datasets.

Though these three feature selection methods are needed, there are various reasons that make them unreliable, unstable, and sometimes ignore the algorithms' stability. However, there is a fourth method that is focused on much these days, the hybrid method. To solve the issues with respect to filter and wrapper methods, a hybrid method is introduced. It combines two or more feature selection techniques and produces a new method with added benefits. In most cases, wrapper and filter methods are made hybrid by combining their advantages (Hoque et al., 2018; Kollias et al., 2018; Thavavel and Karthiyayini, 2018). This study implemented a feature selection pipeline for selecting relevant genes from the raw Alzheimer's gene expression dataset. The filter method is simple and ignores the feature dependencies most of the time and also occasionally includes the redundant features. Wrapper methods are at high risk of overfitting and are stuck in the local optima. It is also computationally intensive. Ensemble methods are better than filter and wrapper; however, on the dataset with the High Dimensional and Low Sample Size (HDLSS) issue, it does not perform well. Thus, we desired to implement a feature selection method catering to the HDLSS issue. Hybrid methods are flexible and robust upon high-dimensional data. Also, they offer higher performance and better computational complexity than the filter and wrapper methods. The pipeline consists of a filter method, wrapper method, and unsupervised gene selection method.

DATASET AND RESOURCES

For a better treatment of AD, the gene expressions are captured preferentially during normal neurological aging (Lima et al., 2016; Carpenter and Huang, 2018). The data captured during the course of AD will improve the understanding of the underlying pathogenesis of AD. This practice will help in the early diagnosis and treatment of AD. The dataset (GSE5281) (Liang et al., 2007) used in this study is from one of the widely accessed data repository, Gene Expression Omnibus (GEO). The dataset consists of information about AD and normal aged brain with 161 samples and 54,675 features (gene expression). The gene expressions are captured from six brain regions of *Homo sapiens* using the LCM cells on the Affymetrix U133 plus 2.0 array with approximately 55,000 transcripts. Among the 161 records, 74 controls and 87 are affected. We have used RStudio for implementing the mentioned approaches in this study. To

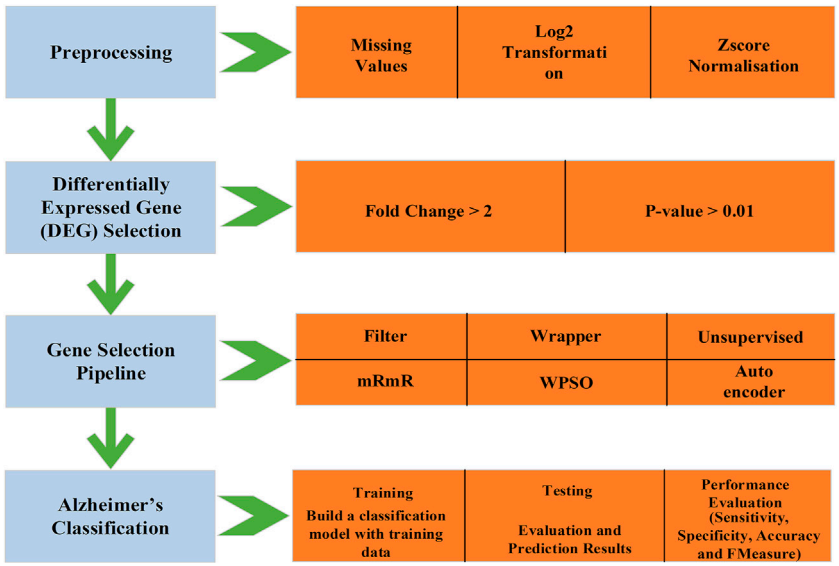


FIGURE 1 | Process flow diagram—proposed system.

analyze the gene expression dataset, R has many beneficial packages such as the Bio-conductor.

METHODOLOGY

We implemented a gene selection pipeline by combining a filter (mRmR), wrapper (Wrapper-based PSO), and unsupervised method (Autoencoder). The mRmR eliminates the genes with maximum redundancy (high correlation among themselves) and the selected genes are inputted to the Wrapper-based PSO, which has k-means as the wrapper method and selects the relevant genes. The selected genes are passed through an autoencoder for further compression. The compressed genes are used for classifying the AD and non-AD individuals using the IDBN. The process flow of the proposed framework is shown in **Figure 1**.

Minimum Redundancy and Maximum Relevance (mRmR)

mRmR is the most widely used practical statistical approach for feature selection. It was proposed by Peng et al. (Ding and Peng, 2005) initially for classifying patterns. The mRmR method strives to choose the feature subset that is highly relevant to the outcome (target class) and minimally redundant. In simple terms, the features are highly similar to the outcome class (relevance) and dissimilar among themselves (redundancy). The feature selection process in mRmR is carried out by adding the features with the highest feature importance to the feature list at each step (El Akadi et al., 2011; Billah and Waheed, 2020).

The aim of mRmR in gene selection is to select a gene subset, G_s, with $\{X_m\}$ features that are highly correlated with the target class T (output class). The mRmR involves three steps, finding the relevance, finding the redundancy, and combining the two to get the mRmR feature set.

Step 1. Maximum Relevance

The maximum relevance is calculated using the mean of Mutual Information of all the features in Xm with target class T. The Mutual Information between R and S random variables can be calculated by

$$MI(R,S) = \sum_{r\in\Omega R}\sum_{s\in\Omega S} p(r,s)log\frac{p(r,s)}{p(r)p(s)} \quad (1)$$

Where.

R—response variable.

S—number of features.

ΩR and ΩS—sample spaces with respect to R and S,

p (r, s)—joint probability density, and

p ()—marginal density function.

The maximum relevance is given by

$$Relevance(G_s) = \frac{1}{|G_s|}\sum_{X_m\in G_s} MI(T,X_m) \quad (2)$$

Where MI(T, X_m)—Mutual Information of feature X_m with class T.

Step 2. Minimum Redundancy

The minimum redundancy is calculated with the formula

$$Redundancy(G_s) = \frac{1}{|G_s|^2}\sum_{X_i,X_j\in G_s} MI\left(X_i,X_j\right) \quad (3)$$

Step 3. Combining the above two constraints

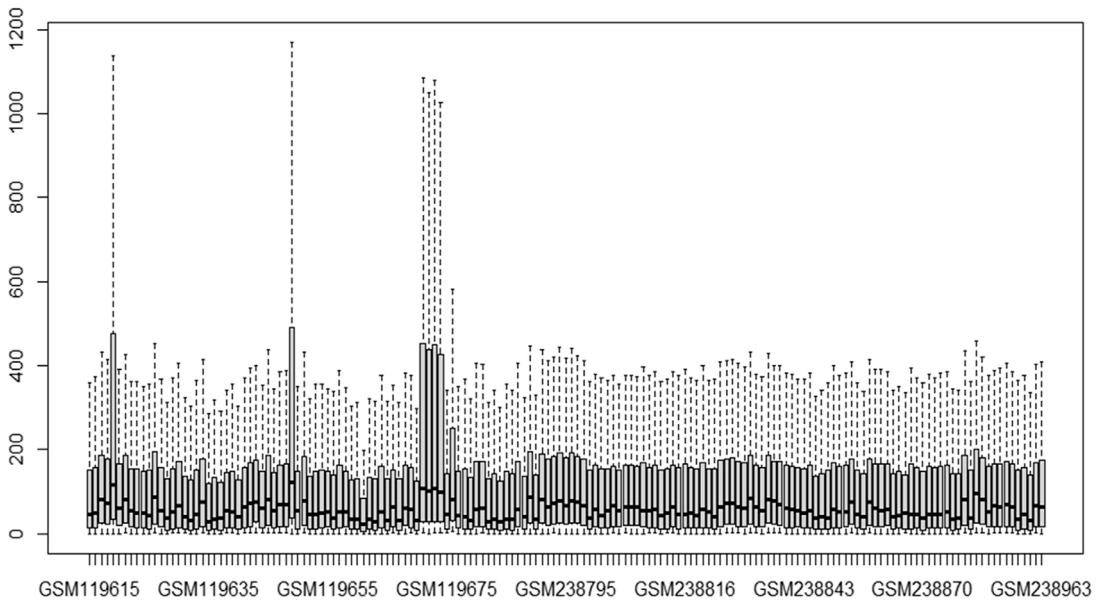


FIGURE 2 | Boxplot of gene expression data before transformation.

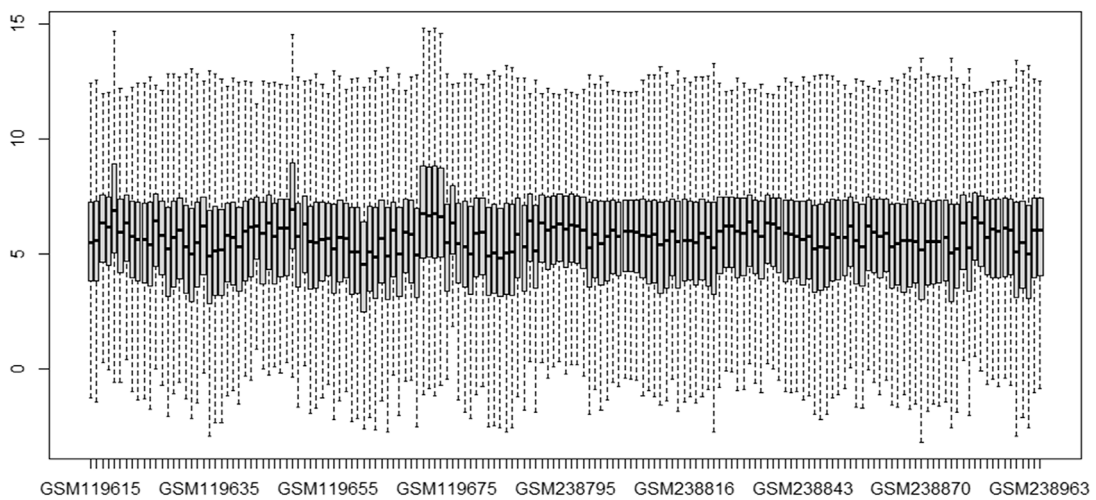


FIGURE 3 | Boxplot of gene expression data after transformation.

The maximum relevance and minimum redundancy are combined to form the mRmR using the formula

$$\max \alpha\left[\boldsymbol{Relevance}\left(\boldsymbol{G_s}\right), \boldsymbol{Redundancy}\left(\boldsymbol{G_s}\right)\right] \quad (4)$$

Where α = (Relevance (G_s) − Redundancy (G_s))

Particle Swarm Optimization (PSO)

PSO is a stochastic, metaheuristic algorithm inspired by the birds' swarming behavior. From the birds' flocking behavior, it is understood that each individual is affected by the leader or the global optima and the personal performance or the local optima (Deepthi and Thampi, 2015). The PSO is an optimization technique based on population proposed by Eberhart and Kennedy (Kennedy and Eberhart, 1995), successfully applied in many global search problems. It is considered in many feature selection problems as it is easy to implement, and has reasonable computational time, global search, and fewer parameters.

In PSO, the population is initialized with particles, each having its own position and velocity. The quality of the particles is estimated at each iteration with the help of a fitness function. Every particle in the search space will carry the present position

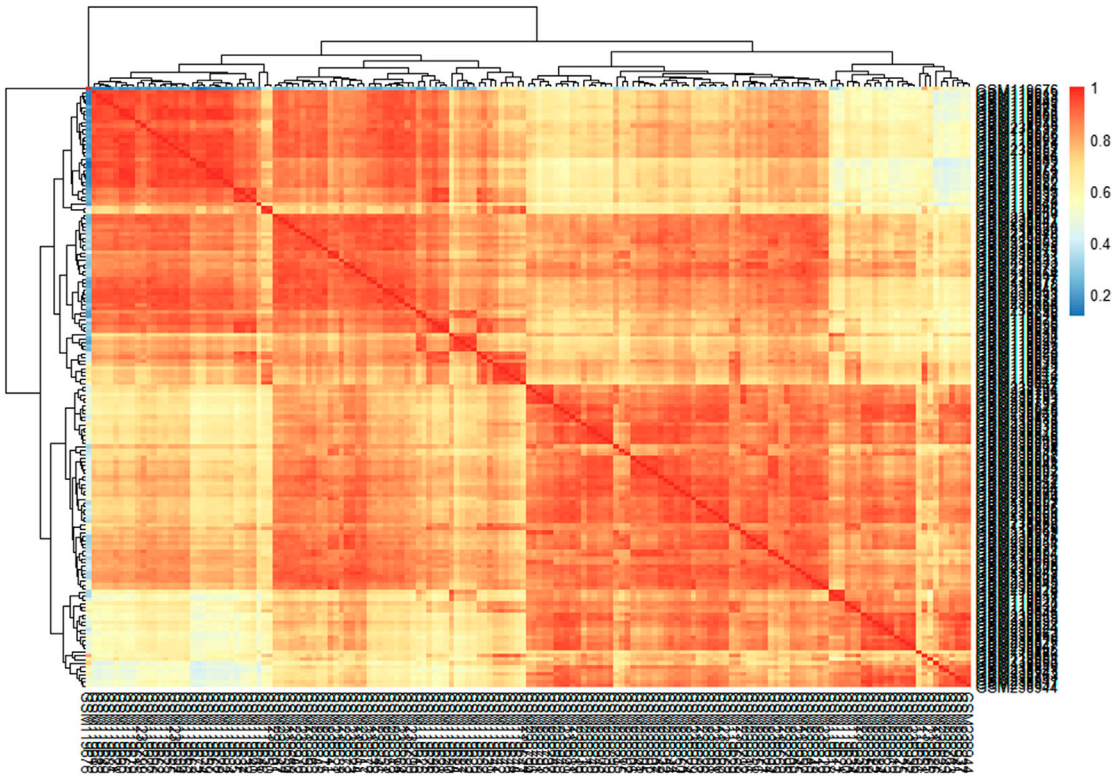


FIGURE 4 | Heat map of differentially expressed genes.

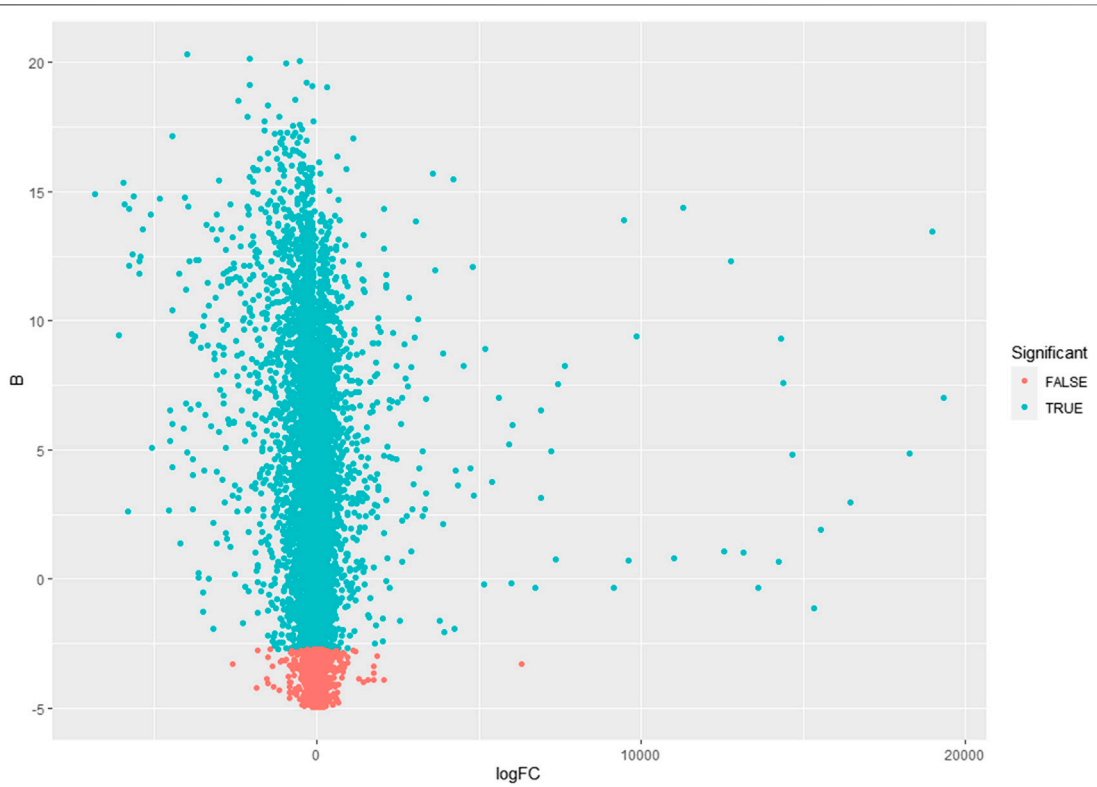


FIGURE 5 | *p*-value and fold change plot.

TABLE 1 | Results analysis.

Feature selection technique	Evaluation metrics	SVM	LDA	NB	MLP	IDBN
PCA	Sensitivity	86.61	81.28	83.50	89.35	94.68
	Specificity	82.57	83.07	88.96	87.48	92.38
	Accuracy	87.62	85.94	83.08	91.87	94.96
	FMeasure	88.42	83.26	88.78	90.05	91.20
CBFS	Sensitivity	78.65	76.27	75.58	80.26	85.46
	Specificity	77.57	78.31	76.27	79.56	83.87
	Accuracy	79.04	77.65	76.89	82.07	85.96
	FMeasure	76.32	76.07	75.63	80.10	82.08
mRmR	Sensitivity	83.61	82.51	80.51	85.40	89.36
	Specificity	83.45	80.87	81.02	85.07	87.12
	Accuracy	85.21	84.78	83.45	87.80	91.58
	FMeasure	84.93	81.02	81.07	86.18	90.55
mRmR-WPSO-AE	Sensitivity	92.43	88.71	85.64	91.51	94.54
	Specificity	91.07	89.36	86.97	92.87	96.17
	Accuracy	92.91	89.74	88.75	94.56	96.78
	FMeasure	90.64	87.98	85.50	92.06	95.09

(x_{ppos}), present velocity (v_{pvel}), and personal best (y_{pbest}). After every iteration, the velocity can be updated by

$$v_{pvel}(new) = i_w v_{pvel}(old) + a_{c1} r_1 \left(y_{pbest} - x_{ppos}(old) \right) + a_{c2} r_2 \left(\hat{y}_{gbest} - x_{ppos}(old) \right) \quad (5)$$

Where.
i_w—inertia weight,
a_{c1} and a_{c2}—acceleration constants,
r_1 and r_2—random numbers (range [0, 1]),
v_{pvel} (old)—present best solution of the particle,
y_{pbest}—personal best solution of the particle, and
$\hat{y}_{gbest}$—global best solution.
The new position of the particle can be determined by

$$x_{ppos}(new) = x_{ppos} + v_{pvel}(new) \quad (6)$$

The positions and the velocity of the particle are updated at every iteration using the formulas given. The process is stopped when certain minimized fitness function criteria are achieved or a particular predefined iteration is reached. For position and velocity updates, the particles use the knowledge of their own and that of other neighboring particles. The final output represents the optimal feature set. We implemented a wrapper-based PSO with the k-means algorithm as the wrapper method. This wrapper method will aid in overcoming the problem of reaching local optima.

The fitness function for each subset is calculated using the below equation,

$$Sum\ of\ Squared\ Error = \sum_{x=1}^{k} \sum_{i \in C_x} d^2(c_x, i) \quad (7)$$

Where.
k—number of clusters,
i—object in the cluster,
c_x—cluster centroid, and
d—Euclidean distance.

Autoencoders

Autoencoder is an artificial neural network with feed-forward processing. The autoencoder consists of input and output with one or more hidden layers, where the number of neurons (features) in the input and output layer is the same (Chicco et al., 2014). The autoencoder's main aim is to reconstruct the inputs such that the difference between the input and the output is minimized. The learning in autoencoder is compressed and distributed (encoding) (Ferri et al., 2021). The training of autoencoder involves three steps:

1. If "x" is the input and "x" is the output, the feed-forward pass is done to estimate the values of all the nodes in the hidden layers after applying the activation function. For an autoencoder with a single hidden layer, the hidden unit vector h_u is given by

$$h_u = a_{func}(W_e.x + bias_e) \quad (8)$$

Where.
h_u—hidden unit,
a_{func}—activation function.
W_e—parameter matrix (encoding),
x—input, and
$bias_e$—bias parameter vector (encoding).

2. Map the hidden representation into the space "x" with the help of the decoding function. The decoding function is given by

$$\hat{x} = a_{func}(W_d.h_u, +bias_d) \quad (9)$$

Where.
W_d—parameter matrix (decoding) and
$bias_d$—bias parameter vector (decoding).

3. Calculate back propagation error using the formula

$$MSE(x, \hat{x}) = \|x - \hat{x}\|_2^2 = \|x - (W_d.h_u + bias_d)\|_2^2 \quad (10)$$

Deep Belief Network

In Deep Belief Network (DBN), the Restricted Boltzmann Machines are stacked together to form a network (An et al., 2020). RBMs are

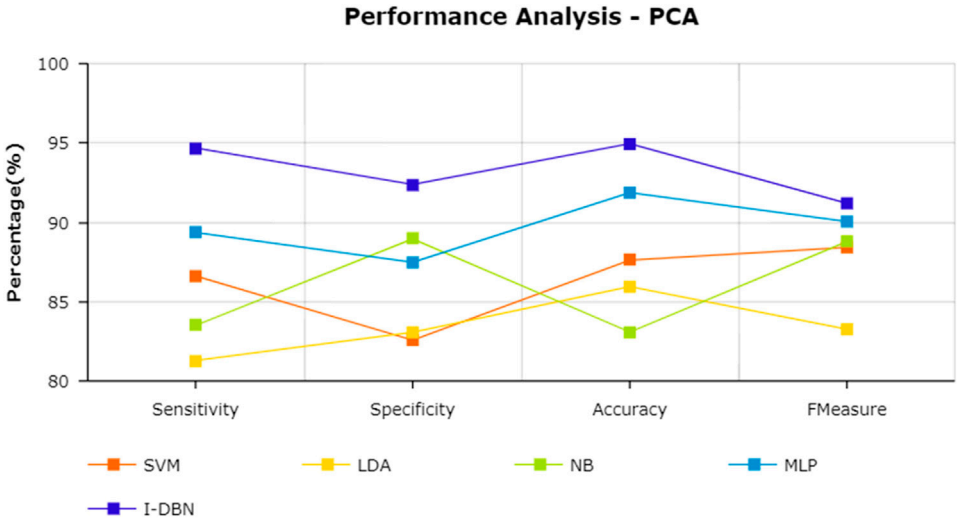


FIGURE 6 | Performance analysis—PCA.

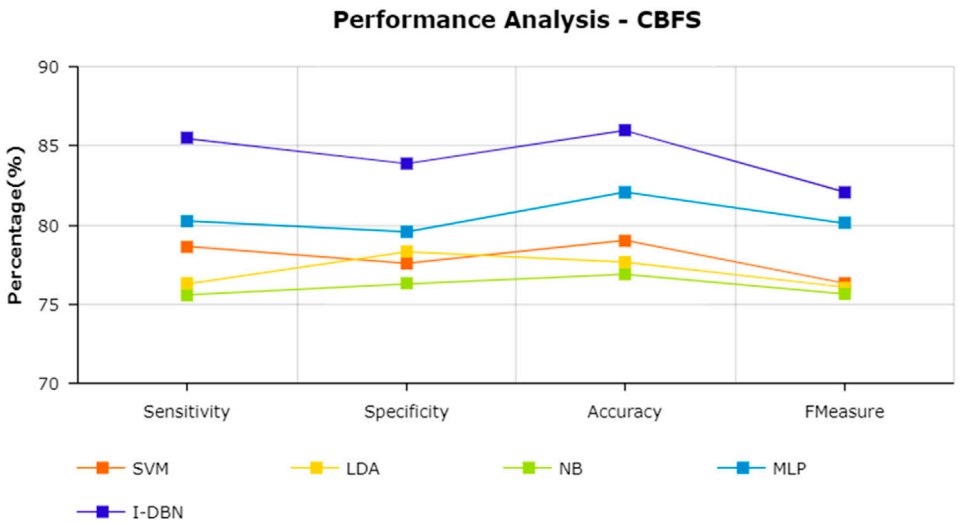


FIGURE 7 | Performance analysis—CBFS.

energy-based generative models with two layers, visible and hidden. Both the layers have nodes connected to each other (Mahendran et al., 2020; Sureshkumar et al., 2020). The major components in RBMs are bias, weight, and activation function (Le Roux and Bengio, 2008; Sekaran and Sudha, 2020). The output is produced after processing with the activation function. We implemented an IDBN with stopping criteria. We chose the hyperparameters using the Bayesian Optimization technique. The Bayesian approach for tuning the hyperparameter keeps past records and verifies the probability to select the next set of parameters. It takes informative decisions in choosing the parameters. The final values for the hyperparameters in IDBN are as follows: learning rate = 0.01, hidden layers = 9, number of nodes per layer = 342, and dropout rate = 0.85. We used the Rectified Linear Unit (RLu) as the activation function. To avoid the overfitting problem, we introduced a stopping criteria strategy. After every 40 epochs, the test accuracy of the last 10 epochs will be compared and checked for convergence, and the training accuracy will also be checked. If both the conditions are satisfied, the learning is ended.

Evaluation Metrics

For evaluating the results of the proposed model, we have used the standard evaluation metrics such as Accuracy, Sensitivity, Specificity, and FMeasure.

- Accuracy: It is a simple ratio between the correctly classified as AD and non-AD to the total number of samples.

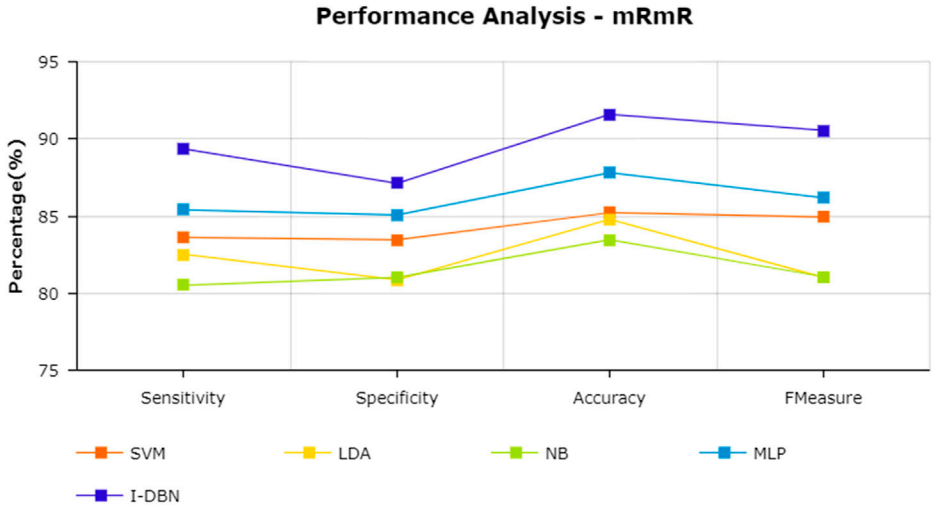


FIGURE 8 | Performance analysis—mRmR.

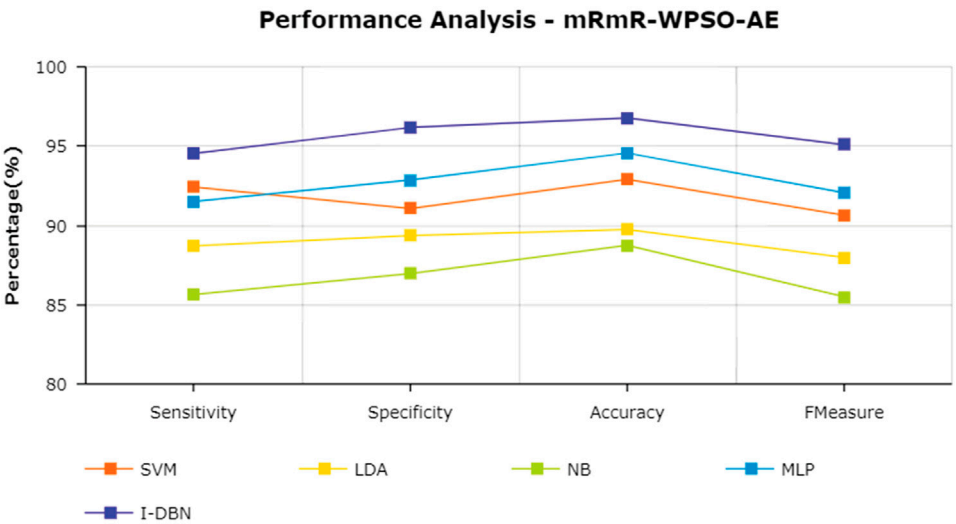


FIGURE 9 | Performance analysis—mRmR-WPSO-AE.

- Sensitivity: It is a measure to identify correctly those with AD.
- Specificity: It is a measure to identify correctly those without AD.
- FMeasure: It is the weighted average of recall and precision (the percentage of samples that are classified as AD positive and are actually positive).

Pseudocode

mRmR—WPSO—AE

mRmR.

Input: Candidates (set of initial genes).

Step 1: for genes g_i in candidates do.

Step 2: relevance = calculate the relevance score using **Eq. 2**;

Step 3: redundancy = 0;

Step 4: for genes g_j in candidates do.

Step 5: redundancy = calculate the redundancy score using **Eq. 3**;

Step 6: end for.

Step 7: *mrmr_values* [g_i] = **Eq. 4**;

Step 8: end for.

Step 9: selected_genes = take (number_of_genes_required);

WPSO.

Step 10: Initialize x random gene subsets from the selected_genes with y number of genes in each subset.

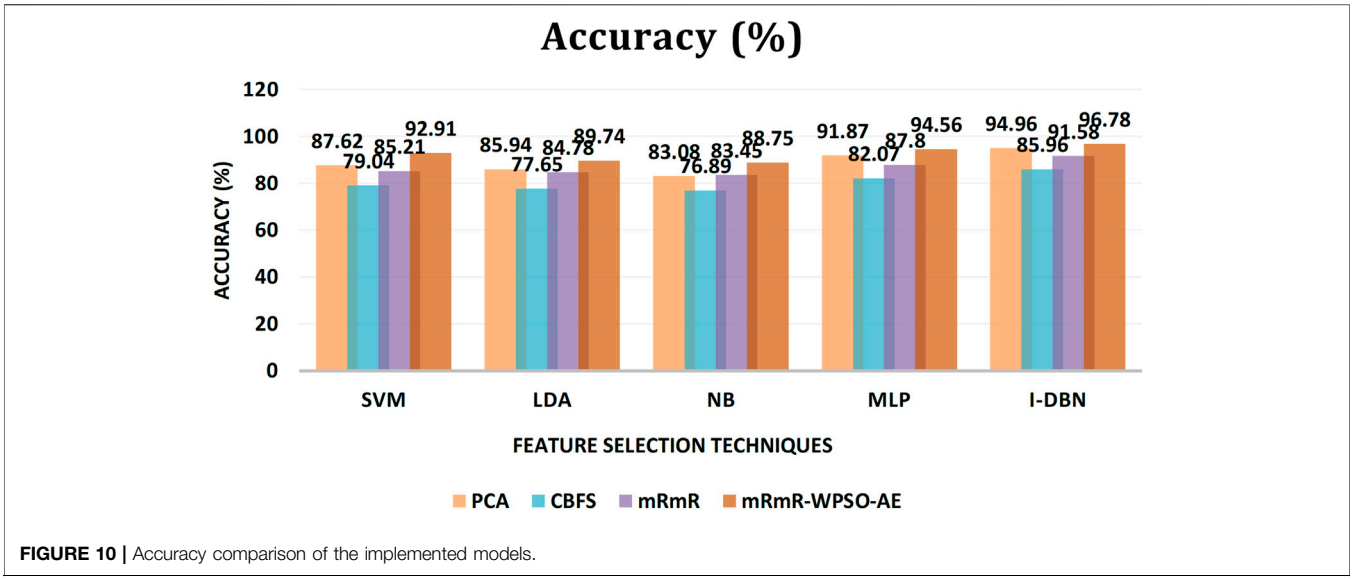


FIGURE 10 | Accuracy comparison of the implemented models.

Step 11: For every random subset *x*, initialize position and velocity vectors.

Step 12: Cluster initial subset with K = *k* using k-means clustering.

Step 13: Evaluate the fitness_value using **Eq. 7**

Step 14: Based on the fitness function, update the subset's *pbest* and *pbestloc*.

Step 15: repeat.

Step 16: if (*fitness_value* < *pbest*) then.

Step 17: update *pbest* and *pbestloc*;

Step 18: end if.

Step 19: Initialize *gbest* and *gbestloc* after finding the minimum *fitness_value* in all the subsets.

Step 20: for j = 0 to *swarm_size-1* do.

Step 21: Estimate velocity using **Eq. 5**;

Step 22: Update subset location using **Eq. 6**;

Step 23: end for.

Step 24: Set the *fitness_value* by computing the squared error using the present location of the gene subset.

Step 25: until predefined number of iterations reached;

Output: Best subset of genes (*gbestloc*).

Autoencoder.

Input: *gbestloc_matrix* (*GM*) $\in \{0, 1\}^{m \times n}$, where m and n are genes and features.

Step 26: Initialize hidden units (h_u), where $h_u < m$, and hidden layers (d).

Step 27: Training:

Step 28: for each GM_i (gene profile) of GM, where i $\in$ [1, m].

Step 29: for each hidden layer d.

Step 30: compute hidden activation function using **Eq. 8**

Step 31: reconstruct the output using **Eq. 9**

Step 32: evaluate the error using **Eq. 10**

Step 33: update the weight by back propagating the error.

Step 34: Testing:

Step 35: for each GM_i (gene profile) of GM, where i $\in$ [1, m].

Step 36: autoencode GM_i and produce. $\widehat{GM_i}$

Step 37: set $\widehat{GM_i}$ as *i*th row of the output matrix. $\widehat{GM}$

RESULTS AND DISCUSSION

This study's primary aim is to improve the classification accuracy of the model in classifying Alzheimer patients by selecting the most relevant feature subset. The dataset used in this study has 161 samples and 54,675 features. The raw gene expression level data are highly skewed, as can be seen in the box plot shown in **Figure 2**. Thus, we applied log2 transformation to make it symmetrical. The results after applying the log2 transformation can be seen in the box plot shown in **Figure 3**. We applied *Z*-score normalization on the transformed data to make it comparable across all the platforms. Once the data are normalized, the differentially expressed genes are identified with fold change and *p*-value. The threshold used for fold change and *p*-value is |FC| > 2 and *p*-value > 0.01. The heat map from **Figure 4** shows the levels of differentially expressed genes. The plot from **Figure 5** shows the *p*-value and fold change levels of the differentially expressed genes. False represents the expression levels that are below the threshold, and true represents the expression levels that are above the threshold. The respective genes are selected and carried forward for the next stage, which is feature selection.

We designed a feature selection pipeline with mRmR, PSO, and autoencoder. The mRmR selects the gene with maximum relevance and minimum redundancy. Then, we applied a wrapper-based PSO technique with k-means clustering as the wrapper to further select the candidate genes. The candidate genes selected are passed through an autoencoder to form the latent representation of the provided input, compress the data without much information loss, and then rebuild as output with as low error as possible. The primary goal of passing the genes through autoencoder is to make the data less sensitive to variations in training. After selecting the relevant features

(CTD-3092A11.2, CHGB, JPX, MAFF, AC004951.6, APLNR, MT1M, SST, PCYOX1L, PRO 1804, and SLC39A12), we implemented an IDBN.

We used the leave-one-out cross-validation method to evaluate the proposed model (Srinivasan et al., 2019). Leave-one-out validation is used because the sample size is less than the feature size. The metrics we used to evaluate the model are Sensitivity, Specificity, Accuracy, and FMeasure. We compared the results of the proposed feature selection pipeline with widely used Principal Component Analysis (PCA), Correlation-based Feature Selection (CBFS), and minimum Redundancy and maximum Relevance (mRmR). We also implemented two linear and two non-linear classifiers Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Naïve Bayes (NB), and Multi-Layer Perceptron (MLP), to compare the results with the IDBN.

The results are tabulated in **Table 1**. The tabulated results show that the proposed feature selection algorithm pipeline (mRmR-WPSO-AE), along with IDBN, classifies Alzheimer's slightly better than the other implemented models. The linear models SVM and LDA produce an accuracy of 92.91 and 89.74% with the proposed gene selection pipeline, which is better than the PCA's 87.62% (SVM), 85.94% (LDA), CBFS's 79.04% (SVM), 77.65% (LDA), and mRmR's 85.21% (SVM), 84.78% (LDA). Also, with the non-linear models, NB and MLP produce an accuracy of 88.75 and 94.56%, which is again better than the PCA's 83.08% (NB), 91.87% (MLP), CBFS's 76.89% (NB), 82.07% (MLP), and mRmR's 83.45% (NB), 87.80% (MLP). SVM performs better among the linear models than LDA along all the implemented gene selection methods, and MLP performs better than NB in the non-linear category. The combination of the proposed gene selection pipeline (mRmR-WPSO-AE) and IDBN shows the promising result with 96.78% accuracy in classifying Alzheimer's patients. From the plots shown in **Figures 6–9**, it is clear that IDBN shows slightly better results than the other implemented classification models. The plot from **Figure 10** shows the Accuracy comparison of the implemented models. The plot shows that the IDBN and mRmR-WPSO-AE have better accuracy than the other models.

CONCLUSION AND FUTURE WORK

Alzheimer's is a progressive degenerative brain disease in the elderly. It is difficult to diagnose even with dominant symptoms. The proper diagnoses are made only during an autopsy after the death of the individual. Recent advances have made it easy to be detected early, using clinical screening with technologies such as brain imaging. Although brain imaging proves effective in most cases, in some cases, the results are inaccurate. The inaccuracies in the results make it difficult for early diagnoses and appropriate treatment for the individual. Thus, the research now shifts to molecular biomarker identification, which helps to differentiate clearly between genotype and phenotype characteristics.

The molecular data-based research proves to be effective. Still, it generates huge volumes of data consisting of transcripts, transcriptomes, etc. It creates a "curse of dimensionality" problem. Thus, machine learning-based feature selection techniques are implemented to select only the relevant genes affecting the target class (outcome). We implemented one such gene selection method for choosing the relevant genes. We designed a hybrid gene selection pipeline by combining mRmR, WPSO, and AE. We compared the results with other commonly used feature selection techniques, such as PCA, CBFS, and mRmR. We compared the results by implementing two linear (SVM and LDA) and two non-linear (NB and MLP) machine learning classification algorithms. We also implemented the IDBN with simple criteria to avoid overfitting. The results show that the proposed pipeline and the IDBN perform slightly better than the linear and non-linear models implemented in this study. In the future, we would implement the proposed pipeline on SNP and DNA Methylation dataset to evaluate the model's generalization.

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

PV and C-YC did the conceptualization and supervised the research. C-YC carried out the funding acquisition. NM investigated the data, performed the methodology, and implemented the software code. C-YC and KS carried out the project administration and validated the results. NM wrote the manuscript. PV and KS reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Ministry of Science and Technology, Taiwan: MOST 109-2221-E-224-048-MY2.

REFERENCES

Acharya, U. R., Fernandes, S. L., WeiKoh, J. E., Ciaccio, E. J., Fabell, M. K. M., Tanik, U. J., et al. (2019). Automated Detection of Alzheimer's Disease Using Brain MRI Images- A Study with Various Feature Extraction Techniques. *J. Med. Syst.* 43 (9), 302–314. doi:10.1007/s10916-019-1428-9

Afza, F., Sharif, M., Mittal, M., Khan, M. A., and Hemanth, D. J. (2021). A Hierarchical Three-step Superpixels and Deep Learning Framework for Skin Lesion Classification. *Methods* 2021, 1046–2023. doi:10.1016/j.ymeth.2021.02.013

Agarwal, P., Wang, H.-C., and Srinivasan, K. (2018). Epileptic Seizure Prediction over EEG Data Using Hybrid CNN-SVM Model with Edge Computing Services. *MATEC Web Conf.* 210, 03016. doi:10.1051/matecconf/201821003016

An, N., Jin, L., Ding, H., Yang, J., and Yuan, J. (2020). A Deep Belief Network-Based Method to Identify Proteomic Risk Markers for Alzheimer Disease. *arXiv preprint arXiv:2003.05776*.

Ashwini, K., Srinivasan, K., and Chang, C. Y. (2021). Deep Learning Assisted Neonatal Cry Classification via Support Vector Machine Models. *Front. Public Health* 9, 670352. doi:10.3389/fpubh.2021.670352

Attique Khan, M., Sharif, M., Akram, T., Kadry, S., and Hsu, C. H. (2021). A Two-stream Deep Neural Network-based Intelligent System for Complex Skin Cancer Types Classification. *Int. J. Intell. Syst.* 2021, 1–29. doi:10.1002/int.22691

Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., and Bashir, K. (2019). "Improving Heart Disease Prediction Using Feature Selection Approaches," in 2019 16th international bhurban conference on applied sciences and technology (IBCAST) (Islamabad, Pakistan: IEEE), 619–623. doi:10.1109/ibcast.2019.8667106

Billah, M., and Waheed, S. (2020). Minimum Redundancy Maximum Relevance (mRMR) Based Feature Selection from Endoscopic Images for Automatic Gastrointestinal Polyp Detection. *Multimedia Tools Appl.* 79 (33), 23633–23643. doi:10.1007/s11042-020-09151-7

Carpenter, K. A., and Huang, X. (2018). Machine Learning-Based Virtual Screening and its Applications to Alzheimer's Drug Discovery: a Review. *Curr. Pharm. Des.* 24 (28), 3347–3358. doi:10.2174/1381612824666180607124038

Chakriswaran, P., Vincent, D. R., Srinivasan, K., Sharma, V., Chang, C.-Y., and Reina, D. G. (2019). Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues. *Appl. Sci.* 9, 5462. doi:10.3390/app9245462

Chen, H., He, Y., Ji, J., and Shi, Y. (2019). A Machine Learning Method for Identifying Critical Interactions between Gene Pairs in Alzheimer's Disease Prediction. *Front. Neurol.* 10, 1162. doi:10.3389/fneur.2019.01162

Chicco, D., Sadowski, P., and Baldi, P. (2014). "Deep Autoencoder Neural Networks for Gene Ontology Annotation Predictions," in Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics, Newport Beach, CA, September 20–23, 2014 (New York, NY: Association for Computing Machinery, 533–540. doi:10.1145/2649387.2649442

Citron, M. (2010). Alzheimer's Disease: Strategies for Disease Modification. *Nat. Rev. Drug Discov.* 9 (5), 387–398. doi:10.1038/nrd2896

Deepthi, P. S., and Thampi, S. M. (2015). "PSO Based Feature Selection for Clustering Gene Expression Data," in 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES) (Kozhikode, India: IEEE), 1–5. doi:10.1109/spices.2015.7091510

Ding, C., and Peng, H. (2005). Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *J. Bioinform. Comput. Biol.* 03 (02), 185–205. doi:10.1142/s0219720005001004

El Akadi, A., Amine, A., El Ouardighi, A., and Aboutajdine, D. (2011). A Two-Stage Gene Selection Scheme Utilizing MRMR Filter and GA Wrapper. *Knowl Inf. Syst.* 26 (3), 487–500. doi:10.1007/s10115-010-0288-x

Farooq, A., Anwar, S., Awais, M., and Rehman, S. (2017). "A Deep CNN Based Multi-Class Classification of Alzheimer's Disease Using MRI," in 2017 IEEE International Conference on Imaging systems and techniques (IST) (Beijing, China: IEEE), 1–6.

Ferri, R., Babiloni, C., Karami, V., Triggiani, A. I., Carducci, F., Noce, G., et al. (2021). Stacked Autoencoders as New Models for an Accurate Alzheimer's Disease Classification Support Using Resting-State EEG and MRI Measurements. *Clin. Neurophysiol.* 132 (1), 232–245. doi:10.1016/j.clinph.2020.09.015

Fung, G., and Stoeckel, J. (2007). SVM Feature Selection for Classification of SPECT Images of Alzheimer's Disease Using Spatial Information. *Knowl Inf. Syst.* 11 (2), 243–258. doi:10.1007/s10115-006-0043-5

Hoque, N., Singh, M., and Bhattacharyya, D. K. (2018). EFS-MI: an Ensemble Feature Selection Method for Classification. *Complex Intell. Syst.* 4 (2), 105–118. doi:10.1007/s40747-017-0060-x

Huang, X., Liu, H., Li, X., Guan, L., Li, J., Tellier, L. C. A. M., et al. (2018). Revealing Alzheimer's Disease Genes Spectrum in the Whole-Genome by Machine Learning. *BMC Neurol.* 18 (1), 5–8. doi:10.1186/s12883-017-1010-3

Huber, C. M., Yee, C., May, T., Dhanala, A., and Mitchell, C. S. (2018). Cognitive Decline in Preclinical Alzheimer's Disease: Amyloid-Beta versus Tauopathy. *J. Alzheimers Dis.* 61 (1), 265–281. doi:10.3233/JAD-170490

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated Machine Learning: Methods, Systems, Challenges*. Basingstoke, UK: Springer Nature, 219.

Ji, H., Liu, Z., Yan, W. Q., and Klette, R. (2019). "Early Diagnosis of Alzheimer's Disease Using Deep Learning," in Proceedings of the 2nd International Conference on Control and Computer Vision, Jeju, Republic of Korea, June 15–18, 2019 (New York, NY: Association for Computing Machinery), 87–91. doi:10.1145/3341016.3341024

Kennedy, J., and Eberhart, R. (1995). "Particle Swarm Optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, Perth, WA (Citeseer), Vol. 4, 1942–1948. doi:10.1109/ICNN.1995.488968

Jo, T., Nho, K., and Saykin, A. J. (2019). Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Front. Aging Neurosci.* 11, 220. doi:10.3389/fnagi.2019.00220

Khan, M. A., Akram, T., Zhang, Y.-D., and Sharif, M. (2021). Attributes Based Skin Lesion Detection and Recognition: A Mask RCNN and Transfer Learning-Based Deep Learning Framework. *Pattern Recognition Lett.* 143, 58–66. doi:10.1016/j.patrec.2020.12.015

Khan, M. A., Muhammad, K., Sharif, M., Akram, T., and Albuquerque, V. H. C. D. (2021). "Multi-Class Skin Lesion Detection and Classification via Teledermatology". In IEEE Journal of Biomedical and Health Informatics. doi:10.1109/JBHI.2021.3067789

Khan, M. A., Sharif, M., Akram, T., Damaševičius, R., and Maskeliūnas, R. (2021). Skin Lesion Segmentation and Multiclass Classification Using Deep Learning Features and Improved Moth Flame Optimization. *Diagnostics* 11, 811. doi:10.3390/diagnostics11050811

Khan, M. A., Zhang, Y.-D., Sharif, M., and Akram, T. (2021). Pixels to Classes: Intelligent Learning Framework for Multiclass Skin Lesion Localization and Classification. *Comput. Electr. Eng.* 90, 106956. doi:10.1016/j.compeleceng.2020.106956

Kollias, D., Tagaris, A., Stafylopatis, A., Kollias, S., and Tagaris, G. (2018). Deep Neural Architectures for Prediction in Healthcare. *Complex Intell. Syst.* 4 (2), 119–131. doi:10.1007/s40747-017-0064-6

Lawrence, A. D., and Sahakian, B. J. (1995). Alzheimer Disease, Attention, and the Cholinergic System. *Alzheimer Dis. associated Disord.* 9 (Suppl 2), 43–49. doi:10.1097/00002093-199501002-00008

Le Roux, N., and Bengio, Y. (2008). Representational Power of Restricted Boltzmann Machines and Deep Belief Networks. *Neural Comput.* 20 (6), 1631–1649. doi:10.1162/neco.2008.04-07-510

Lee, T., and Lee, H. (2020). Prediction of Alzheimer's Disease Using Blood Gene Expression Data. *Sci. Rep.* 10 (1), 1–13. doi:10.1038/s41598-020-60595-1

Liang, W. S., Dunckley, T., Beach, T. G., Grover, A., Mastroeni, D., Walker, D. G., et al. (2007). Gene Expression Profiles in Anatomically and Functionally Distinct Regions of the normal Aged Human Brain. *Physiol. Genomics* 28 (3), 311–322. doi:10.1152/physiolgenomics.00208.2006

Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., and Honorio, K. M. (2016). Use of Machine Learning Approaches for Novel Drug Discovery. *Expert Opin. Drug Discov.* 11 (3), 225–239. doi:10.1517/17460441.2016.1146250

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., and Feng, D. (2014). "Early Diagnosis of Alzheimer's Disease with Deep Learning," in 2014 IEEE 11th international symposium on biomedical imaging (ISBI) (Beijing, China: IEEE), 1015–1018.

Mahendran, N., Durai Raj Vincent, P. M., Srinivasan, K., and Chang, C.-Y. (2020). Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues, and Future Research Directions. *Front. Genet.* 11, 603808. doi:10.3389/fgene.2020.603808

Mamdiwar, S. D., R, A., Shakruwala, Z., Chadha, U., Srinivasan, K., and Chang, C.-Y. (2021). Recent Advances on IoT-Assisted Wearable Sensor Systems for Healthcare Monitoring. *Biosensors* 11, 372. doi:10.3390/bios11100372

Martínez-Ballesteros, M., García-Heredia, J. M., Nepomuceno-Chamorro, I. A., and Riquelme-Santos, J. C. (2017). Machine Learning Techniques to Discover Genes with Potential Prognosis Role in Alzheimer's Disease Using Different Biological Sources. *Inf. Fusion* 36, 114–129. doi:10.1016/j.inffus.2016.11.005

Miao, Y., Jiang, H., Liu, H., and Yao, Y.-d. (2017). An Alzheimers Disease Related Genes Identification Method Based on Multiple Classifier Integration. *Comput. Methods Programs Biomed.* 150, 107–115. doi:10.1016/j.cmpb.2017.08.006

Mirzaei, S., El Yacoubi, M., Garcia-Salicetti, S., Boudy, J., Kahindo, C., Cristancho-Lacroix, V., et al. (2018). Two-stage Feature Selection of Voice Parameters for

Early Alzheimer's Disease Prediction. *IRBM* 39 (6), 430–435. doi:10.1016/j.irbm.2018.10.016

Moradifard, S., Hoseinbeyki, M., Ganji, S. M., and Minuchehr, Z. (2018). Analysis of microRNA and Gene Expression Profiles in Alzheimer's Disease: a Meta-Analysis Approach. *Sci. Rep.* 8 (1), 1–17. doi:10.1038/s41598-018-20959-0

Neumann, U., Genze, N., and Heider, D. (2017). EFS: an Ensemble Feature Selection Tool Implemented as R-Package and Web-Application. *BioData Min* 10 (1), 21–29. doi:10.1186/s13040-017-0142-8

Nussbaum, R. L., and Ellis, C. E. (2003). Alzheimer's Disease and Parkinson's Disease. *N. Engl. J. Med.* 348 (14), 1356–1364. doi:10.1056/nejm2003ra020003

Orimaye, S. O., Wong, J. S., Golden, K. J., Wong, C. P., and Soyiri, I. N. (2017). Predicting Probable Alzheimer's Disease Using Linguistic Deficits and Biomarkers. *BMC bioinformatics* 18 (1), 34–13. doi:10.1186/s12859-016-1456-0

Park, C., Ha, J., and Park, S. (2020). Prediction of Alzheimer's Disease Based on Deep Neural Network by Integrating Gene Expression and DNA Methylation Dataset. *Expert Syst. Appl.* 140, 112873. doi:10.1016/j.eswa.2019.112873

Pes, B. (2019). Ensemble Feature Selection for High-Dimensional Data: a Stability Analysis across Multiple Domains. *Neural Comput. Appl.* 32, 1–23. doi:10.1007/s00521-019-04082-3

Ramzan, F., Khan, M. U. G., Rehmat, A., Iqbal, S., Saba, T., Rehman, A., et al. (2020). A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks. *J. Med. Syst.* 44 (2), 37–16. doi:10.1007/s10916-019-1475-2

Reitz, C., Brayne, C., and Mayeux, R. (2011). Epidemiology of Alzheimer Disease. *Nat. Rev. Neurol.* 7 (3), 137–152. doi:10.1038/nrneurol.2011.2

Salat, D. H., Kaye, J. A., and Janowsky, J. S. (2001). Selective Preservation and Degeneration within the Prefrontal Cortex in Aging and Alzheimer Disease. *Arch. Neurol.* 58 (9), 1403–1408. doi:10.1001/archneur.58.9.1403

Sanchez-Riera, J., Srinivasan, K., Hua, K.-L., Cheng, W.-H., Hossain, M. A., and Alhamid, M. F. (2018). Robust RGB-D Hand Tracking Using Deep Learning Priors. *IEEE Trans. Circuits Syst. Video Technol.* 28 (9), 2289–2301. doi:10.1109/TCSVT.2017.2718622

Sarraf, S., and Tofighi, G. (2016). Classification of Alzheimer's Disease Using Fmri Data and Deep Learning Convolutional Neural Networks. *arXiv preprint arXiv:1603.08631*.

Sarraf, S., and Tofighi, G. (2016). "Deep Learning-Based Pipeline to Recognize Alzheimer's Disease Using fMRI Data," in 2016 future technologies conference (FTC) (San Francisco, CA, USA: IEEE), 816–820. doi:10.1109/ftc.2016.7821697

Sekaran, K., and Sudha, M. (2020). Predicting Autism Spectrum Disorder from Associative Genetic Markers of Phenotypic Groups Using Machine Learning. *J. Ambient Intelligence Humanized Comput.* 12, 1–14. doi:10.1007/s12652-020-02155-z

Srinivasan, K., Ankur, A., and Sharma, A. (2017). "Super-resolution of Magnetic Resonance Images Using Deep Convolutional Neural Networks," in 2017 IEEE International Conference on Consumer Electronics, June 12–14, 2017 (Taipei, Taiwan: IEEE), 41–42. doi:10.1109/ICCE-China.2017.7990985

Srinivasan, K., Cherukuri, A., Vincent, D., Garg, A., and Chen, B. (2019). An Efficient Implementation of Artificial Neural Networks with K-fold Cross-Validation for Process Optimization. *J. Internet Technol.* 20 (4), 1213–1225. doi:10.3966/160792642019072004020

Srinivasan, K., Garg, L., Datta, D., A. Alaboudi, A., Agarwal, N. R., and George Thomas, A. (2021). Performance Comparison of Deep CNN Models for Detecting Driver's Distraction. *Comput. Mater. Continua* 68 (3), 4109–4124. doi:10.32604/cmc.2021.016736

Srinivasan, K., Mahendran, N., Vincent, D. R., Chang, C.-Y., and Syed-Abdul, S. (2020). Realizing an Integrated Multistage Support Vector Machine Model for Augmented Recognition of Unipolar Depression. *Electronics* 9, 647. doi:10.3390/electronics9040647

Sureshkumar, V., Chandrasekar, V., Venkatesan, R., and Prasad, R. K. (2020). Improved Performance Accuracy in Detecting Tumor in Liver Using Deep Learning Techniques. *J. Ambient Intelligence Humanized Comput.* 12, 1–8. doi:10.1007/s12652-020-02107-7

Tanveer, M., Richhariya, B., Khan, R. U., Rashid, A. H., Khanna, P., Prasad, M., et al. (2020). Machine Learning Techniques for the Diagnosis of Alzheimer's Disease. *ACM Trans. Multimedia Comput. Commun. Appl.* 16 (1s), 1–35. doi:10.1145/3344998

Thavavel, V., and Karthiyayini, M. (2018). Hybrid Feature Selection Framework for Identification of Alzheimer's Biomarkers. *Indian J. Sci. Technol.* 11 (22), 1–10. doi:10.17485/ijst/2018/v11i22/123310

Tufail, A. B., Ma, Y.-K., and Zhang, Q.-N. (2020). Binary Classification of Alzheimer's Disease Using sMRI Imaging Modality and Deep Learning. *J. Digit Imaging* 33 (5), 1073–1090. doi:10.1007/s10278-019-00265-5

Wang, L., and Liu, Z.-P. (2019). Detecting Diagnostic Biomarkers of Alzheimer's Disease by Integrating Gene Expression Data in Six Brain Regions. *Front. Genet.* 10, 157. doi:10.3389/fgene.2019.00157

Wenk, G. L. (2003). Neuropathologic Changes in Alzheimer's Disease. *J. Clin. Psychiatry* 64 (Suppl. 9), 7–10.

Zhou, K., He, W., Xu, Y., Xiong, G., and Cai, J. (2018). Feature Selection and Transfer Learning for Alzheimer's Disease Clinical Diagnosis. *Appl. Sci.* 8 (8), 1372. doi:10.3390/app8081372

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.

Copyright © 2021 Mahendran, Vincent, Srinivasan and Chang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 November 2021
doi: 10.3389/fgene.2021.784863

Prediction of Protein–Protein Interaction Sites Based on Stratified Attentional Mechanisms

Minli Tang[1,2†], *Longxin Wu*[1†], *Xinyu Yu*[1], *Zhaoqi Chu*[3], *Shuting Jin*[1,4*] *and Juan Liu*[3*]

[1]Department of Computer Science and Technology, Xiamen University, Xiamen, China, [2]School of Big Data Engineering, Kaili University, Kaili, China, [3]Department of Instrumental and Electrical Engineering, School of Aerospace Engineering, Xiamen University, Xiamen, China, [4]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China

Proteins are the basic substances that undertake human life activities, and they often perform their biological functions through interactions with other biological macromolecules, such as cell transmission and signal transduction. Predicting the interaction sites between proteins can deepen the understanding of the principle of protein interactions, but traditional experimental methods are time-consuming and labor-intensive. In this study, a new hierarchical attention network structure, named HANPPIS, by adding six effective features of protein sequence, position-specific scoring matrix (PSSM), secondary structure, pre-training vector, hydrophilic, and amino acid position, is proposed to predict protein–protein interaction (PPI) sites. The experiment proved that our model has obtained very effective results, which was better than the existing advanced calculation methods. More importantly, we used the double-layer attention mechanism to improve the interpretability of the model and to a certain extent solved the problem of the "black box" of deep neural networks, which can be used as a reference for location positioning on the biological level.

Keywords: protein–protein interaction, multilevel attention mechanism, feature fusion, deep learning, protein features

OPEN ACCESS

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

Reviewed by:
Yansen Su,
Anhui University, China
Wenbin Liu,
Guangzhou University, China

***Correspondence:**
Shuting Jin
stjin@stu.xmu.edu.cn
Juan Liu
cecyliu@xmu.edu.cn

†These authors have contributed equally to this work

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

Received: *28 September 2021*
Accepted: *08 October 2021*
Published: *22 November 2021*

Citation:
Tang M, Wu L, Yu X, Chu Z, Jin S and Liu J (2021) Prediction of Protein–Protein Interaction Sites Based on Stratified Attentional Mechanisms. Front. Genet. 12:784863. doi: 10.3389/fgene.2021.784863

INTRODUCTION

Proteins participate in various biological processes in organisms. They usually do not play a single role but interact with other biological macromolecules to perform biological functions (Geng et al., 2015). Protein–protein interactions (PPIs) refer to the process in which two or more protein molecules form a protein complex through non-covalent bonds. Protein interactions play an extremely important role in most biochemical functions (Bradford and Westhead, 2005; Nilofer et al., 2020). The identification of protein interaction sites can help researchers understand how proteins perform their biological functions (Ofran and Rost, 2003; Nilofer et al., 2020), and it can also help design new antibacterial drugs (Gainza et al., 2020). Conventional biological experimental methods, such as two-hybrid screening, affinity purification, and mass spectrometry, can be used to identify protein interaction sites (Chung et al., 2007; Gainza et al., 2020). Biological experimental methods have disadvantages of being expensive and time-consuming. Therefore, it is of great value for biologists to develop accurate calculation methods to predict protein interaction sites.

In order to solve the problem concerning expenses, many non-biochemical experimental methods have been developed (Li et al., 2021), and most of the calculation methods are based on machine learning. Zhang and Kurgan (2019) evaluated a large number of functional features that could be

used, such as position-specific scoring matrix (PSSM), evolutionary conservation (ECO), and relative solvent accessibility (RSA). In the protein interaction site prediction methods designed by the predecessors, the high score fragment pairs (HSP) as (Li et al., 2021) and the one-hot (Yang et al., 2016; Wang et al., 2019) and amino acid-embedding representations (Zeng et al., 2020) were used to characterize protein sequences as model input features. Wang et al. (2010) proposed a new method for predicting protein interaction sites in hybrids by using the radial basis function neural network (RBFNN) model. This method only used the evolutionary conservation information of the protein and the spatial sequence profile and has achieved good prediction results.

Zhou and Shan (2001) proposed a neural network-based prediction method, taking the sequence distribution of adjacent amino acids and solvent exposure as input. Ofran and Rost (2006) proposed a neural network model for the interaction sites identified from sequence (ISIS), which were trained based on sequence contours and structural features predicted by the sequence. Porollo and Meller (2007) proposed a method named SPPIDER based on the support vector machine, neural network, and linear discriminant analysis, which used 19 features extracted from the sequence. Mizuguchi and Mizuguchi (2010) developed a predictor called PSIVER, which is a naive Bayes classifier based on a position-specific scoring matrix (PSSM), predicting relative solvent accessibility and kernel density estimation. Dhole et al. (2014) proposed a logistic regression classifier LORIS that uses L1 regularization. In addition, Dhole et al. (2015) proposed a new artificial neural network prediction method that used PSSM features, average cumulative hydrophilicity, and predicted relative solvent accessibility to train SPRINGS.

In these studies, a large number of features extracted from protein sequences are used. The commonly used features include evolutionary information and secondary structure (Murakami and Mizuguchi, 2010). In addition to these commonly used features, there are some other physical, chemical, biological, and statistical features, such as the accessible surface area of the protein, protein size, backbone flexibility, and sequence specificity, which have been used for protein interaction site prediction. However, existing methods tend to pay too much attention to protein sequence information, ignoring the characteristics of proteins at the biological level, and most machine learning methods are inexplicable.

In order to solve the above problems, we propose a double-layer attention mechanism prediction model based on graph convolution that uses multidimensional features as input. The main contributions are as follows:

1) For paying more attention to the features at the biological level, we add six effective features of proteins as the input of the model, which can dig out more potential information.
2) The use of the double-layer attention mechanism improves the performance and interpretability of the model and solves the "black box" problem of deep neural networks to a certain extent.

METHODS

Data

In this experiment, we used three benchmark data sets, namely Dset_186, Dset_72 (Murakami and Mizuguchi, 2010), and Dset_164 (Geng et al., 2015). Dset_186 is constructed from the Protein Data Ban (PDB) database, which is dedicated to the three-dimensional structure of proteins and nucleic acids. Dset_186 is composed of 186 protein sequences, and their sequence homology is less than 25%, and through X-ray crystallography, their resolution is found to be less than 3 Å. The structure of Dset_72 and Dset_164 is the same as that of Dset_186. Dset_72 contains 72 protein sequences, and Dset_164 consists of 164 protein sequences. Therefore, we have a total of 422 different protein sequences. In this study, if an amino acid has an absolute solvent proximity less than $1Å^2$ before and after binding with other proteins, then it is defined as the interaction site; otherwise, it is defined as the non-reciprocal site of action.

Dset_186, Dset_72, and Dset_164 contain 1,923, 5,517 and 6,096 active sites and 16,217, 30,702 and 27,585 non-interactive sites, respectively. Although the protein sequences in the three data sets are not duplicated, the three data sets are from different research groups. So, in order to ensure that the training set and the test set have the same distribution, we integrated the three datasets into a fusion data set. Next, we divided the fused data set into a training set (approximately 80% of the randomly selected protein sequences) and a test set (the remaining 20% of the protein sequences). In the end, we obtained 350 protein sequences in the training set and 70 protein sequences in the test set. Among them, we deleted two protein sequences without defined secondary structure of proteins (DSSP).

Feature Generation

Feature generation is a key step in the deep learning framework. Excellent features can perfectly represent the various properties of the protein, and features with insufficient expression ability will reduce the accuracy of the deep learning model. In order to better obtain the global features of the protein, we combined six effective features of the protein amino acid encoding, sequence, and structure as input vectors for training. These features include protein sequence, PSSM matrix, secondary structure, pre-training vector, hydrophilicity, and amino acid location.

Amino Acid Encoding

One-hot encoding. One-hot encoding is one of the simplest but very effective features, because the original protein sequence can accurately represent each amino acid and its position. Most proteins are composed of 20 different amino acids, so we use 20-dimensional one-hot codes to represent the types of various amino acids in the protein.

Sequence Features

PSSM matrix. The evolutionary information in PSSM (Jeong et al., 2011) has been proven to be effective for PPI site prediction. We run the PSI-BLAST algorithm and search NCBI's non-redundant sequence database with three iterations and a

threshold of 0.001 to generate the PSSM matrix. Each amino acid in the protein sequence is encoded as a vector with 20 numbers, which represents the probability of these 20 amino acids appearing at that position.

Hydrophilic characteristics. The hydrophilic characteristics of amino acids are determined through experiments, and this characteristic determines the free energy of the transfer of each amino acid. In detail, it is determined by the amount of change in the free energy of amino acids when they move from water to organic solvents. It can be measured by solubility in water and organic solvents. It also contains energetic information about protein interactions and is a very important feature in protein.

Structure Features

Protein secondary structure. Secondary structure features are often used in protein prediction. We use the DSSP program to generate secondary structure information. It encodes the structural information of amino acids and uses it to predict protein interaction sites. In this article, we use eight types of secondary structure states (G (3_10-helix), H (α-helix), I (π-helix), B (isolated bridge), E (extended sheet), T (β-turn), S (bend), and other states). Considering that some amino acids have no secondary structure status in the DSSP file, we use a one-hot vector of dimension 9 to encode them. The first eight dimensions indicate the state of each amino acid, and the last dimension indicates whether there is information about the state of the related secondary structure.

Pre-training vector based on SeqVec. In this experiment, we use the pre-training model, that is, SeqVec to obtain the pre-training vector. SeqVec is a protein sequence pre-training model trained using deep unsupervised learning (Villegas-Morcillo et al., 2020). It is based on the model ELMo (Heinzinger et al., 2019) and consists of a character-level convolutional neural network (char-CNN) and two-layer bidirectional long short-term memory (LSTM). The CNN embeds each amino acid in a latent space, generates the corresponding feature vector, and then uses LSTM to model the context of the surrounding amino acids. The model adds two LSTM layers to provide the final context-aware embedding. These embeddings indicate excellent performance in protein classification tasks, such as inferring protein secondary structure, structural category, disordered regions, and cell location. At the level of each amino acid, the predicted secondary structure and the regions with inherent disorder are significantly better than one-hot encoding or the method generated by Word2vec. The generation of protein embedding representations is rapid, and it only takes 0.03 s for SeqVec to generate the evolution information of the target protein. We choose to use SeqVec to represent each amino acid in the sequence as a feature vector with a dimension of 1,024 for subsequent training.

Residue location characteristics. The DeepPPISP model proposed by Zeng et al. (2020) shows that the global information of proteins helps predict protein interaction sites. We use the position information of each residue as the input feature because it provides global position information, and it can also make up for the defect that the attention model cannot capture in position information. The position of the residue in the protein is between 1 and L (protein length). We divide the position by the length of the protein so that a final value between 0 and 1 is obtained and then use this value as the residue position feature and input the model for training.

Model Structure

The hierarchical attention network (HAN) model (Yang et al., 2016) uses a multilevel attention mechanism to classify documents and has achieved good results. The HAN model has two notable features: 1) a hierarchical thinking to represent documents is used. The document is regarded as composed of sentences, and the sentences are regarded as composed of words and 2) the HAN model applies two attention mechanisms, which are used in documents. At the document level and sentence level, the attention weight of the words is calculated to obtain the representation of the sentence, and then the attention weight of the sentence is calculated to obtain the representation of the document. The abovementioned mechanism enables the HAN to give different sentences and words at different degrees of importance.

Inspired by the HAN, we applied it to the task of predicting PPI sites. The structure of our model is shown in **Figure 1**. First, we use a sliding window to obtain protein sequence fragments representing protein interaction sites and then divide the fragments into smaller fragments through K-mers. Next, we compare individual amino acids to words in the document. K-mers are analogous to sentences in documents, and the entire protein sequence fragment is analogous to documents. Then, the hierarchical attention model is used for training and prediction, and the final model is obtained. The model obtained in this way can identify the contribution degree of a single amino acid to K-mers and the contribution degree of K-mers to the entire protein sequence fragment. By further analysis, we can deduce which amino acids contribute to the target amino acid as the binding site of protein interaction, what is the specific contribution, and which amino acids may be invalid or can be obtained. We could also know the important characteristics of the amino acids that become the interaction sites of proteins.

In the model, we use a sliding window to integrate the features of the neighboring amino acids. We divide the fixed-length protein sequence into multiple fragments using K-mers and then use the previously introduced method to vectorize the amino acids in the fragments. After obtaining the vector representation of all amino acids, we use the Bi-GRU to encode each amino acid and then use the attention mechanism to calculate the importance of each amino acid for K-mers. Then, we obtain the vector of each K-mer after weighing and summing. We use the same method to encode the K-mer vector, obtain the vector representation of the protein sequence through the attention layer, and finally use Softmax for classification.

The features we use include the 20-dimensional one-hot amino acid feature, the PSSM matrix feature with the same dimension of 20, the 9-dimensional secondary structure feature, the 1-dimensional hydrophilic feature, the 1-dimensional amino acid position feature, and 1,024-

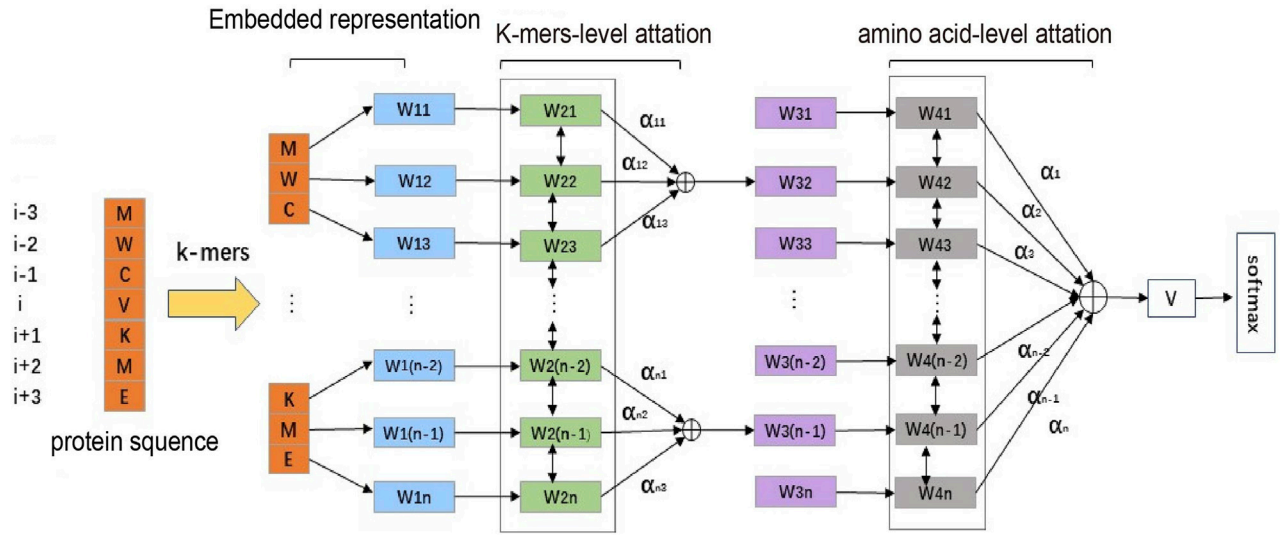


FIGURE 1 | Structure of HANPPIS. It consists of three steps, including embedded representation, amino acid–level attention and K-mers–level attention. We obtain vector representations of protein sequence fragments through multidimensional features. The vector representation of the protein fragment is the input to the first layer of the attention mechanism, and then the vector representation of the protein sequence is obtained through the second layer of attention and finally input to the prediction layer.

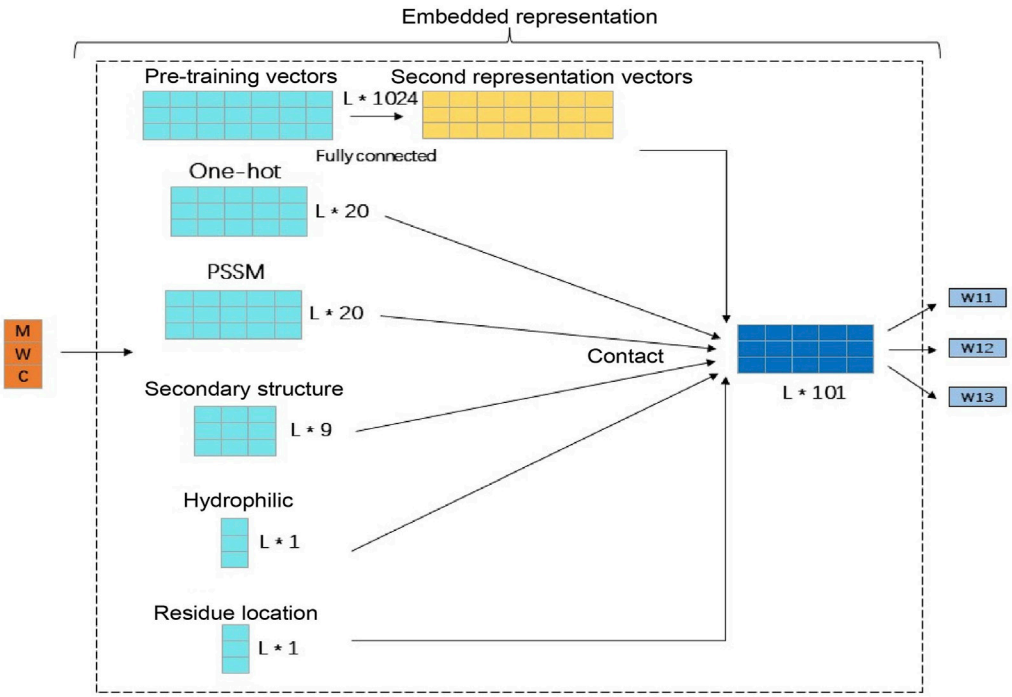


FIGURE 2 | Amino acid feature generation and expression. This figure illustrates the specific details of the amino acid signature generation. Among them, because the pre-training vector feature dimension is too large, a layer of feedforward neural network is used to reduce the dimension to 50 dimensions. Then, the remaining five features are spliced and finally the 101-dimensional amino acid feature vector as the input of the entire model is obtained.

dimensional pre-training vector feature. In order to prevent the pre-training vector feature dimensionality from being too large and affecting the other four features, we used a layer of feedforward neural network to reduce its dimensionality, reducing it to 50 dimensions, and then splicing the other five features to finally get the 101-dimensional amino acid feature

TABLE 1 | Model compares the experimental results on the test set.

Model	Accuracy	Precision	Recall	F1
SPPIDER	0.622	0.209	0.459	0.287
ISIS	0.694	0.211	0.362	0.267
RF_PPI	0.598	0.173	0.512	0.258
PSIVER	0.653	0.253	0.468	0.328
SPRINGS	0.631	0.248	0.598	0.350
GCNPPIS	0.623	0.233	0.395	0.293
HANPPIS	0.631	0.291	0.605	0.393

TABLE 2 | Results of ablation experiments.

Model	Accuracy	Precision	Recall	F1
Model_Dpsf	0.554	0.237	0.580	0.352
Model_Dpm	0.605	0.269	0.635	0.378
Model_Dssf	0.594	0.263	0.612	0.374
Model_Dhf	0.632	0.282	0.591	0.381
Model_Dlf	0.622	0.295	0.588	0.379
Model_Dptv	0.578	0.275	0.510	0.319
HANPPIS	0.631	0.291	0.605	0.393

representation vector which is used as the input of the entire model. The processing process is shown in **Figure 2**.

Model Training Settings

Our deep learning framework is implemented through Keras. The loss function we use is the cross-entropy loss function, which is defined as follows:

$$Loss = -\frac{1}{n}\sum\left[ylog\left(y_{pred}\right) + (1-y)\log\left(1-y_{pred}\right)\right],$$

where n is the number of all training data, y is the real label, and y_{pred} is the predicted label.

Our model uses Adam as the optimizer and the following formula to update the weights:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \varepsilon},$$

where θ_{t+1} is the updated parameter, α is the learning rate, ε is the constant added to maintain numerical stability, and$\widehat{m}_t$ and $\hat{v}_t$ are the first and second moments after deviation correction, respectively.

In order to extract the contextual sequence features of amino acids at protein interaction sites, we set the sliding window length to 7 and the protein sequence length to 500. Protein sequences longer than 500 will be truncated. For the deep learning model, we set the training batch size to 3, the number of neurons in the LSTM layer and the attention layer in the double-layer attention are both set to 86, and the fully connected layer connected by the pre-training vector has 50 neurons. The positive and negative samples of the training set are not uniformly distributed, so we set the sample weight at about 1:7, which allows the model to pay more attention to the positive samples during training and improve the performance of the model.

RESULTS AND DISCUSSIONS

Comparison With the Benchmark Method

To evaluate the performance of HANPPIS in predicting protein interaction sites, we compared HANPPIS with seven competing methods. These six competitive methods all use machine learning or deep learning methods as model training. SPPIDER (Porollo and Meller, 2007) uses an alternative machine learning technology, which combines fingerprints with other sequence and structural information to predict PPI sites. ISIS (Ofran and Rost, 2006) uses a shallow neural network to combine predicted structural features with evolutionary information to predict PPI sites. RF_PPI was developed by Hou et al. (2017). This algorithm uses various protein functions and characteristics and applies it to the random forest algorithm to predict protein interaction sites. PSIVER (Murakami and Mizuguchi, 2010) used sequence features (PSSM matrix and predicted accessibility) and then used a naive Bayes classifier to predict PPI sites. SPRINGS (Dhole et al., 2015) used a shallow neural network algorithm based on evolutionary information, average cumulative hydrophilicity, and predictive relative solvent accessibility to predict PPI sites. In addition, we used graph CNNs to predict PPI sites (GCNPPIS) as a comparative experimental model. The comparison results are shown in **Table 1**.

Table 1 shows the results of HANPPIS and other seven competitive methods on the test set. It is not difficult to find that most of the evaluation indicators measured by HANPPIS are higher than other competitive methods. Although the accuracy rate of HANPPIS is not the highest, other evaluation indicators are higher than competitive methods. Since protein interaction site prediction is an unbalanced learning issue, the ratio of positive and negative data samples is about 1:5.5, so we pay more attention to F1 in the evaluation indicators. Among all existing methods, HANPPIS has the highest F1 value, surpassing existing models.

Influence of Different Input Features

Obviously, different types of features (original protein sequence, PSSM matrix, secondary structure, hydrophilicity, positional features, pre-training vectors) play different roles in the model. In order to evaluate the importance of each feature, we delete each input feature of HANPPIS separately in the ablation experiment. Specifically, we compared the performance of different models that delete the original protein sequence, that is, PSSM matrix, secondary structure features, hydrophilic features, location features, and pre-training vectors. In order to distinguish between different models, we concluded the following definitions:

1) Model_Dpsf: Delete original protein sequence features
2) Model_Dpm: Delete the PSSM matrix
3) Model_Dssf: Delete secondary structure features
4) Model_Dhf: Delete hydrophilic features
5) Model_Dlf: Delete location features
6) Model_Dptv: Delete the pre-training vector

TABLE 3 | Effect of the sliding window on the model.

Window size	Accuracy	Precision	Recall	F1
7	0.631	0.291	0.605	0.393
9	0.638	0.298	0.551	0.381
11	0.601	0.281	0.593	0.387
13	0.654	0.300	0.540	0.379
15	0.578	0.275	0.690	0.389

The results of the ablation experiment are shown in **Table 2**. The results show that deleting the one-hot feature and the pre-training vector feature has the greatest impact on the model. After deletion, all indicators of the model are reduced at the same time, and the F1 value is as low as 0.319. When deleting several other features, the performance of the model also drops slightly. The experimental results show that comprehensive consideration of these features can obtain more comprehensive protein sequence information, which is helpful to improve the performance of the model and obtain better prediction results.

The Effect of Sliding Window Size

In addition to testing different feature inputs, we also studied the impact of different sizes of sliding windows on the model. Specifically, we use sliding windows of different lengths, that is, 7, 9, 11, 13, and 15 to observe the performance of HANPPIS. The results in **Table3** show that the model has the highest F1 value when the length of the sliding window is 7.

The sliding window has less impact on model performance. It may be because the task is to classify specific amino acids, and the surrounding amino acids are only used as a context to assist. In order to further verify our conjecture, we performed a visual analysis of the attention weight, and the details can be seen in the *Interpretability of the model* section.

Interpretability of the Model

In order to overcome the common "black box" problem of deep learning and understand the role of contextual amino acids, we randomly selected a sample, which is the 135th amino acid of the protein 1Z0J_A in Dset186. The sample sequence is "RDAKDYA", and the target amino acid is "K". We visualized the attention distribution of K-mers–level and amino acid–level to view the contribution of each amino acid or K-mers to protein interaction sites. The experimental results are shown in **Figure 3**.

As shown in **Figure 3**, in the K-mers–level attention, it can be seen that the central position of K-mers has the highest attention ratio, and the surrounding K-mers attention weight is less than half of the central position. In the amino acid–level attention, the "K" in the center position also is assigned the highest attention weight. This is consistent with the situation of the protein interaction site task. Because the prediction is whether the central amino acid is the binding site and the surrounding amino acids exist as auxiliary information of the central amino acid, the model's attention to the center position will be larger, and the proportion of the two sides will pay less attention.

It can be seen from **Figure 3B** that the attention weights of amino acids "R" and "A" are both less than 0.05, which is probably the reason why the sliding window does not have a high degree of influence on the model because the farther away from the center, the lower the weight of the amino acid. Therefore, even if the sliding window is enlarged, it has little effect on the amino acids in the central position.

In general, we verified that HANPPIS is suitable for discovering important patterns in protein sequences, and the

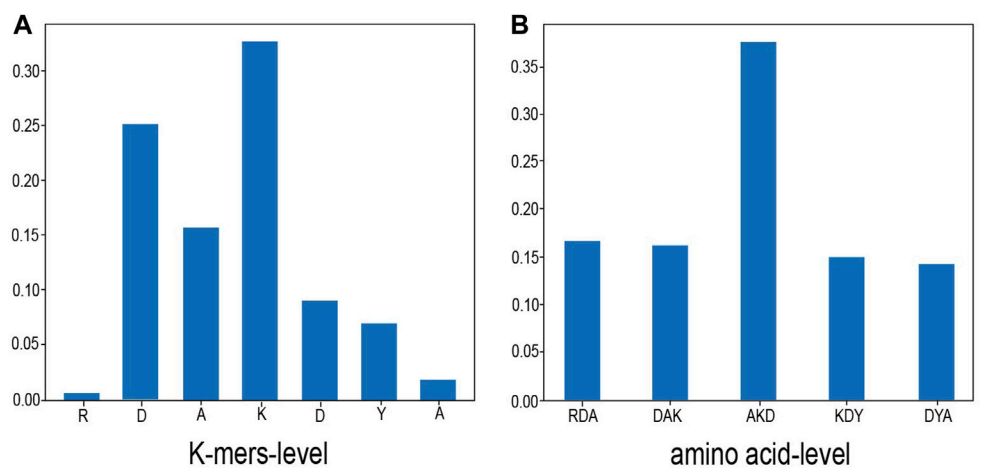


FIGURE 3 | Attention distribution. This figure shows the attention visualization result of one of the samples (from the 135th amino acid of protein 1Z0J_A in Dset186, the sample sequence is "RDAKDYA" and the target amino acid is "K"). **(A)** shows the proportion of K-mers–level attention distribution and **(B)** shows the distribution of amino acid–level attention. As shown in the figure, the center position has the highest proportion of attention, which is also consistent with the task of protein interaction sites.

attention mechanism can understand its relationship in context, which greatly increases the interpretability of the model.

CONCLUSION

The accurate prediction of protein interaction sites can promote the understanding of protein biological functions. In this article, we propose a deep learning framework HANPPIS to predict protein interaction sites at the amino acid level. The difference between HANPPIS and other existing methods is that the model uses hierarchical attention combined with neural networks to predict protein interaction sites. HANPPIS captures global sequence features through Bi-GRU, so that it can easily simulate the relationship between the target amino acid and the entire protein sequence. After Bi-GRU processing, the attention layer is used to let the model assign higher weights to the parts that need attention, and further follow-up results can be obtained. The experiment was repeated twice to generate attention weights for amino acids and K-mers and finally classify and output them through Softmax. The results show that HANPPIS basically surpasses the existing competitive methods in the task of predicting protein interaction sites. Sequence-based protein interaction site prediction is still a challenging problem, and one of the reasons is that there are no unique attributes in the sequence to directly analyze the protein sequence. But in this study, we showed that hierarchical attention can be used for protein interaction site prediction, and more important parts of disordered protein sequences can be found. The multiple experimental results also demonstrated the crucial role of attention mechanism that can increase the interpretability of the model and provided the possibility and direction for further exploration of the mystery of proteins.

But our method has some limitations, such as HANPPIS requires the multidimensional features of the protein as input. Obviously, these features may be missing in some data sets. In addition, the samples for testing attention visualization in the experiment are not enough. In future studies, we would be committed to use fewer features to obtain better performance and improve the interpretability of the model.

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

Conceptualization, JL, SJ, and MT; Methodology, JL, SJ, and MT; Investigation, JL, SJ, MT, XY, and ZC; Writing the original draft, SJ, MT, and XY; Writing, reviewing, and editing, JL, MT, and XY; Funding acquisition, JL; Resources, MT, JL, and SJ; Supervision, JL and SJ.

FUNDING

This study is supported by the National Key R&D Program of China (2017YFE0130600), the National Natural Science Foundation of China (Grant Nos 61772441, 61872309, 62072384, and 62072385), and the Basic Research Program of Science and Technology of Shenzhen (JCYJ20180306172637807).

REFERENCES

Bradford, J. R., and Westhead, D. R. (2005). Improved Prediction of Protein-Protein Binding Sites Using a Support Vector Machines Approach. *Bioinformatics* 21, 1487–1494. doi:10.1093/bioinformatics/bti242

Chung, J.-L., Wang, W., and Bourne, P. E. (2007). High-Throughput Identification of Interacting Protein-Protein Binding Sites. *BMC Bioinformatics* 8, 223–235. doi:10.1186/1471-2105-8-223

Dhole, K., Singh, G., Pai, P., and Mondal, S. (2015). SPRINGS: Prediction of Protein-Protein Interaction Sites Using Artificial Neural Networks [J]. *Peerj* 1, 1–15. doi:10.7287/PEERJ.PREPRINTS.266V2

Dhole, K., Singh, G., Pai, P. P., and Mondal, S. (2014). Sequence-Based Prediction of Protein-Protein Interaction Sites with L1-Logreg Classifier. *J. Theor. Biol.* 348, 47–54. doi:10.1016/j.jtbi.2014.01.028

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., et al. (2020). Deciphering Interaction Fingerprints from Protein Molecular Surfaces Using Geometric Deep Learning. *Nat. Methods* 17 (2), 184–192. doi:10.1038/s41592-019-0666-6

Geng, H., Lu, T., Lin, X., Liu, Y., and Yan, F. (2015). Prediction of Protein-Protein Interaction Sites Based on Naive Bayes Classifier. *Biochem. Res. Int.* 2015, 1–7. doi:10.1155/2015/978193

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., et al. (2019). Modeling Aspects of the Language of Life through Transfer-Learning Protein Sequences. *BMC Bioinformatics* 20 (1), 723–738. doi:10.1186/s12859-019-3220-8

Hou, Q., De Geest, P. F. G., Vranken, W. F., Heringa, J., and Feenstra, K. A. (2017). Seeing the Trees through the forest: Sequence-Based Homo- and Heteromeric Protein-Protein Interaction Sites Prediction Using Random Forest. *Bioinformatics* 33 (10), 1479–1487. doi:10.1093/bioinformatics/btx005

Jeong, J. C., Lin, X., and Chen, X. (2011). On Position-Specific Scoring Matrix for Protein Function Prediction. *Ieee/acm Trans. Comput. Biol. Bioinf.* 8 (2), 308–315. doi:10.1109/tcbb.2010.93

Li, Y., Golding, G. B., and Ilie, L. (2021). DELPHI: Accurate Deep Ensemble Model for Protein Interaction Sites Prediction. *Bioinformatics* 37 (7), 896–904. doi:10.1093/bioinformatics/btaa750

Mizuguchi, K., and Mizuguchi, K. (2010). Applying the Naïve Bayes Classifier with Kernel Density Estimation to the Prediction of Protein–Protein Interaction Sites. *Bioinformatics* 26 (15), 1841–1848. doi:10.1093/bioinformatics/btq302

Murakami, Y., and Mizuguchi, K. (2010). Applying the Naïve Bayes Classifier with Kernel Density Estimation to the Prediction of Protein-Protein Interaction Sites. *Bioinformatics (Oxford, England)* 26, 1841–1848. doi:10.1093/bioinformatics/btq302

Nilofer, C., Sukhwal, A., Mohanapriya, A., Sakharkar, M. K., and Kangueane, P. (2020). Small Protein-Protein Interfaces Rich in Electrostatic Are Often Linked to Regulatory Function. *J. Biomol. Struct. Dyn.* 38 (11), 3260–3279. doi:10.1080/07391102.2019.1657040

Ofran, Y., and Rost, B. (2006). ISIS: Interaction Sites Identified from Sequence. *Bioinformatics* 23, e13–6. doi:10.1093/bioinformatics/btl303

Ofran, Y., and Rost, B. (2003). Predicted Protein-Protein Interaction Sites from Local Sequence Information. *FEBS Lett.* 544 (3), 236–239. doi:10.1016/s0014-5793(03)00456-3

Porollo, A., and Meller, J. (2007). Prediction-Based Fingerprints of Protein-Protein Interactions. *Proteins* 66 (3), 630–645. doi:10.1002/prot.21248

Villegas-Morcillo, A., Makrodimitris, S., van Ham, R. C. H. J., Gomez, A. M., Sanchez, V., and Reinders, M. J. T. (2020). Unsupervised Protein Embeddings Outperform Hand-Crafted Sequence and Structure Features at Predicting Molecular Function. *Bioinformatics* 37 (2), 162–170. doi:10.1093/bioinformatics/btaa701

Wang, B., Chen, P., Wang, P., Zhao, G., and Zhang, X. (2010). Radial Basis Function Neural Network Ensemble for Predicting Protein-Protein Interaction Sites in Heterocomplexes. *Protein Pept. Lett.* 17 (9), 1111–1116. doi:10.2174/092986610791760397

Wang, X., Yu, B., Ma, A., Chen, C., Liu, B., and Ma, Q. (2019). Protein-Protein Interaction Sites Prediction by Ensemble Random Forests with Synthetic Minority Oversampling Technique. *Bioinformatics* 35 (14), 2395–2402. doi:10.1093/bioinformatics/bty995

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). *Hierarchical Attention Networks for Document Classification*. San Diego, California: Association for Computational Linguistics, 1480–1489.

Zeng, M., Zhang, F., Wu, F. X., Li, Y., Wang, J., and Li, M. (2020). Protein-Protein Interaction Site Prediction through Combining Local and Global Features with Deep Neural Networks. *Bioinformatics* 36 (4), 1114–1120. doi:10.1093/bioinformatics/btz699

Zhang, J., and Kurgan, L. (2019). SCRIBER: Accurate and Partner Type-Specific Prediction of Protein-Binding Residues from Proteins Sequences. *Bioinformatics* 35 (14), i343–i353. doi:10.1093/bioinformatics/btz324

Zhou, H.-X., and Shan, Y. (2001). Prediction of Protein Interaction Sites from Sequence Profile and Residue Neighbor List. *Proteins* 44 (3), 336–343. doi:10.1002/prot.1099

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.

Copyright © 2021 Tang, Wu, Yu, Chu, Jin and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 November 2021
doi: 10.3389/fgene.2021.797641

Prediction of Hormone-Binding Proteins Based on K-mer Feature Representation and Naive Bayes

Yuxin Guo[1,2,3,4†], *Liping Hou*[5†], *Wen Zhu*[1,3,4*] *and Peng Wang*[1,3,4]

[1]Key Laboratory of Computational Science and Application of Hainan Province, Haikou, China, [2]Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Quzhou, China, [3]Key Laboratory of Data Science and Intelligence Education, Hainan Normal University, Ministry of Education, Haikou, China, [4]School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [5]Beidahuang Industry Group General Hospital, Harbin, China

Hormone binding protein (HBP) is a soluble carrier protein that interacts selectively with different types of hormones and has various effects on the body's life activities. HBPs play an important role in the growth process of organisms, but their specific role is still unclear. Therefore, correctly identifying HBPs is the first step towards understanding and studying their biological function. However, due to their high cost and long experimental period, it is difficult for traditional biochemical experiments to correctly identify HBPs from an increasing number of proteins, so the real characterization of HBPs has become a challenging task for researchers. To measure the effectiveness of HBPs, an accurate and reliable prediction model for their identification is desirable. In this paper, we construct the prediction model HBP_NB. First, HBPs data were collected from the UniProt database, and a dataset was established. Then, based on the established high-quality dataset, the k-mer (K = 3) feature representation method was used to extract features. Second, the feature selection algorithm was used to reduce the dimensionality of the extracted features and select the appropriate optimal feature set. Finally, the selected features are input into Naive Bayes to construct the prediction model, and the model is evaluated by using 10-fold cross-validation. The final results were 95.45% accuracy, 94.17% sensitivity and 96.73% specificity. These results indicate that our model is feasible and effective.

Keywords: hormone binding protein, feature selection, protein classification, k-mer, naive Bayes model

OPEN ACCESS

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

Reviewed by:
Chunyan Ao,
Xidian University, China
Ru Xiaoqing,
University of Tsukuba, Japan

***Correspondence:**
Wen Zhu
syzhuwen@163.com

†These authors have contributed equally to this work

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

Received: *19 October 2021*
Accepted: *05 November 2021*
Published: *23 November 2021*

Citation:
Guo Y, Hou L, Zhu W and Wang P (2021) Prediction of Hormone-Binding Proteins Based on K-mer Feature Representation and Naive Bayes. Front. Genet. 12:797641. doi: 10.3389/fgene.2021.797641

INTRODUCTION

With the rapid development of society, people have higher and higher requirements for medical and health care (Lin, 2020). Therefore, it is urgent to learn more about the structure and function of proteins in order to explain more of the meaning of life and promote the development of biomedicine and other fields (Wang et al., 2020a; Qu et al., 2021). However, there is a difficulty in the current research, that is, how to use its sequence information to predict proteins effectively. Although effective prediction of protein sequences can be made using physical, chemical and biological experiments, these methods are costly and time consuming.

Hormone binding proteins (HBPs) are carrier proteins that bind specifically to targeted hormones and were first identified in the plasma of pregnant mice, rabbits and humans (Mortezaeefar et al., 2019; Niu et al., 2021a). They are involved in hormonal regulation in living organisms. HBPs not only regulate the amount of hormones reaching the target cell to produce the desired effect (Wang et al., 2018) but also regulate non-protein-binding or free-circulating active steroid hormones, which are

thought to be the main gatekeepers of steroid effects. Sexual HBPs, mainly produced in the liver, combine with sexual steroid hormones to regulate their bioavailability. The incorrect expression of HBPs, however, can cause various diseases (Tan et al., 2019).

Therefore, understanding the function and regulatory mechanism of HBPs has become very important. Accurately identifying HBPs is the first step in studying their function. Traditional HBPs identification methods involve wet biochemical experiments, such as immunoprecipitation, chromatography, or cross-linking (Sohm et al., 1998; Zhang and Marchant, 1999; Einarsdóttir et al., 2014; Cheng et al., 2016; Fang et al., 2019). These experimental methods are time-consuming and expensive, and with the discovery of a large number of protein sequences, it is difficult to determine HBPs through biochemical experiments. Therefore, it is necessary to establish an effective recognition model to identify HBPs (Akbar et al., 2020). The description of the characteristics of the protein sequence method contains a lot of information, such as the chemical and physical properties of amino acids, sequence characteristics, feature extraction algorithm for classification algorithm which has great impact on the design and the classification of results. Generally, prediction techniques based on machine learning consist of three steps: feature extraction, construction of predictors, and performance evaluation (Liu, 2017; Wang et al., 2018; Zhang et al., 2019). In 2018, Tang et al. (Hua et al., 2018). developed a method based on support vector machines to identify HBPs, which uses the optimal characteristic coding protein obtained by using the optimized dipeptide composition. Subsequently, Basith et al. developed the computational predictor iGHBP, which combined the dipeptide composition and the value of the amino acid index to obtain the optimal selection and predict the construction model (Basith et al., 2018). In this paper, we constructed a prediction model, HBP_NB, to correctly identify HBPs. First, the k-mer (Liu et al., 2008; Christopher et al., 2013; Liu et al., 2015a; Manavalan et al., 2019) method was used to obtain the frequency characteristics of protein sequences, and then the F-score value method was used to select the feature subset. Finally, input the obtained features into Naive Bayes (Gong and Tian, 2010; He et al., 2010; Gumus et al., 2014; Hu et al., 2020; Hu et al., 2021a; Hu et al., 2021b) to construct the prediction model.

MATERIALS AND METHODS

Main Process of the Article

Machine learning frameworks have been used to identify multiple protein types, such as DNA binding proteins (Zeng et al., 2015; Qu et al., 2017; Shen and Zou, 2020), RNA binding proteins (Xiao et al., 2017; Lei et al., 2021), lncRNA interacting proteins (Zhang et al., 2017; Liu, 2020), and drug targets (Yan et al., 2016; Wang et al., 2020b; Wang et al., 2020c). Since the recognition of protein sequences includes two important steps sequence feature extraction and classifier selection the effective combination of feature extraction algorithms and classifiers has also been extensively studied (Zhang et al., 2016). In this paper, we propose a predictive model for identifying hormone-binding proteins based on Naïve Bayes.

HBPs prediction analysis was carried out through the following five steps: 1) HBPs and non-HBPs were searched and downloaded from UniProt, and the similarity threshold of protein sequences was set by the CD-HIT program to construct a high-quality dataset (Zou et al., 2020); 2) feature extraction of protein sequences was performed using the k-mer feature coding method; 3) the extracted features were selected to improve the accuracy of classification; 4) different classification methods were used to classify and predict the selected feature subset and select the best classification methods; and 5) Performance evaluation. **Figure 1** shows the structural framework for identifying HBPs in this paper. This section will introduce dataset establishment, feature selection methods and classification methods in detail.

Dataset

It is necessary to collect sufficient correlation function data as the basis of statistical model prediction. Therefore, it is first necessary to construct an objective dataset to ensure the effectiveness and robustness of the model. Therefore, we adopt the benchmark dataset constructed by Tang et al. (Tang et al., 2018). To build this dataset, follow these steps. The first step was to search and collect HBPs from UniProt (Bairoch et al., 2009; Schneider, 2012) and to generate the original HBPs dataset by selecting the hormone binding keywords in the molecular function items of the gene body (Ashburner et al., 2000). Consequently, 357 HBPs with manual annotation and review were selected. In the second step, to avoid the high similarity of protein sequences affecting the results, we used the CD-HIT (Li and Godzik, 2006; Fu et al., 2012) program to set the truncation threshold to 0.6 to remove highly similar HBPs sequences. In the third step, when the protein sequence in the dataset contains unknown residues (such as "X," "Z," and "B"), it will affect the model prediction results, so protein sequences containing unknown residues need to be excluded. After the above steps, a total of 122 HBPs were obtained, which were regarded as positive data. As a control, 121 non-HBPs were randomly selected from UniProt as negative data using a similar selection strategy. The data of the model can be freely download from https://github.com/GUOYUXINXIN/-. The benchmark dataset can be expressed as:

$$D = D_p \cup D_n \tag{1}$$

Among them, subset $D_p$contains 122 HBPs, and subset $D_n$contains 121 non-HBPs.

Feature Extraction

Protein sequence is a string generated by the permutation and combination of 20 English letters with different lengths. Currently, general machine learning algorithms can only deal with feature vectors, so when machine learning methods are used, protein sequences need to be transformed into numerical vectors representing the characteristics of protein sequences. As the first step in building a biological sequence analysis model, feature extraction is an important part of correctly predicting protein sequences, an efficient feature extraction method can obtain a

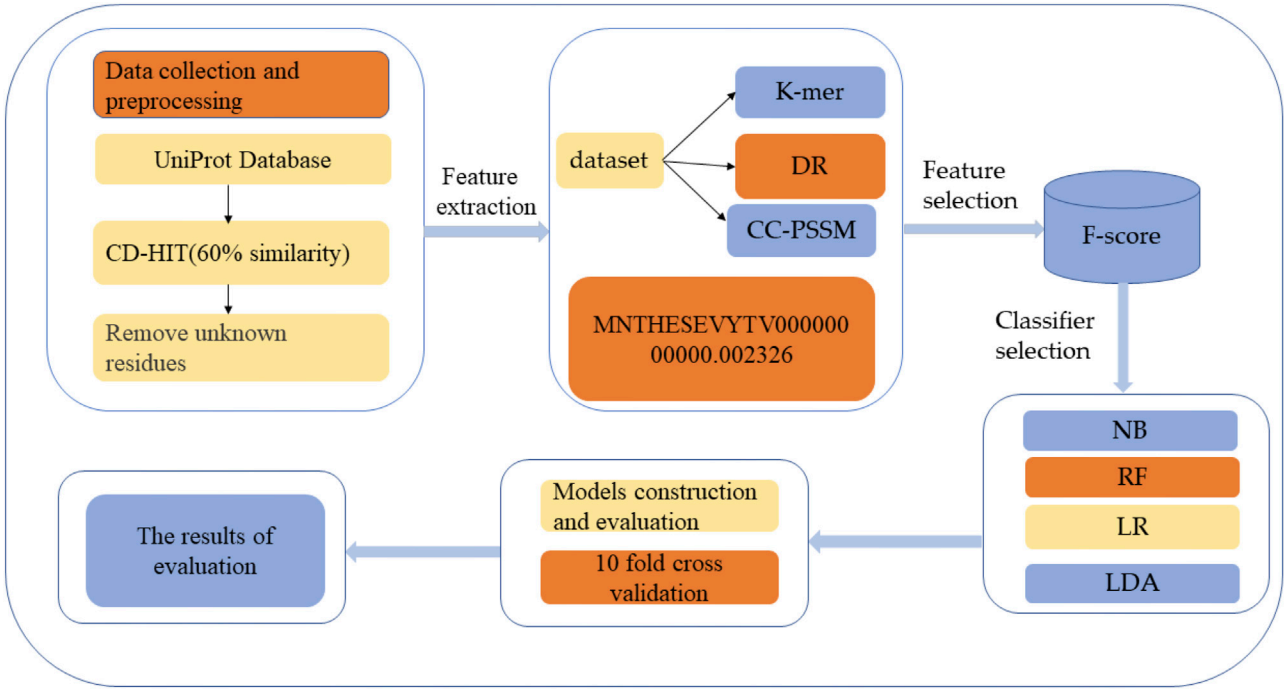


FIGURE 1 | Structure flow chart. The first step is to search and download HBPs and non-HBPs from the protein resource database and then use CD-HIT to perform protein de-redundancy operations. The threshold is set to 60%. Finally, protein sequences containing unknown residues are removed to generate the final protein dataset. The second step is to extract features of the protein, and the third step is to use different classification methods to classify the selected features.

high performance classification model. The extracted features should not only retain the protein sequence information to the maximum extent, but also have a greater correlation with protein classification. Given a protein sequence, express it as:

$$P = R_1 R_2 R_3 \cdots R_L \tag{2}$$

where Pstands for protein sequence, $R_i$represents the$i$th$i$amino acid residue of protein$P\,(i = 1, 2, \cdots, L)$.

K-Mer

K-mer (Liu et al., 2015b; Niu et al., 2021b) is the most basic method of expressing protein sequences as digital vectors (Liu et al., 2016), in which k-mer frequency coding refers to the occurrence frequency of all possible nucleotide sequences with k length in a given sequence (Liu et al., 2015c; Bin et al., 2017). The k-mer feature extraction algorithm is used to convert the protein sequence into a vector with a fixed length, which is used as the input vector of the machine learning classifier. For example, setting k to 2 produces a 400-dimensional vector $(AA, AC, AD, \cdots, AY, YA, YC, \cdots, YY)$. To avoid the problem of overfitting, we generally set$k < 4$ because when$k > 4$, more dimensions will be generated, resulting in dimension disaster (Wei et al., 2019). Therefore, we set k to 3 so that the input protein sequence could be converted into a vector with 8,000 dimensions of fixed length.

Distance-Based Residual

DR (Liu et al., 2014) is a feature expression method based on protein sequences that uses the distance between residue pairs to represent the feature vector of the protein. The feature vector is expressed by calculating the number of occurrences of residual pairs within a certain distance threshold. The feature vector dimension obtained by the DR feature extraction method is $20 + 20 \times 20 \times d_{MAX}$dimensions, where in 20 in $20 + 20 \times 20 \times d_{MAX}$represents the types of amino acids that make up the protein; $d_{MAX}$is a distance threshold that can be set manually, which represents the maximum distance between pairs of amino acid residues.

Profile-Based Cross-Covariance

Since machine learning-based technologies such as random forest (RF) and logistic regression (LR) require the input of fixed-length vectors as input vectors for training, it is necessary to convert protein sequences of different lengths into fixed length vectors as input vector machine learning. Because each residue in a protein has many physical and chemical properties, protein sequences can be regarded as time series with similar properties. Therefore, CC-PSSM (Dong et al., 2009) is used in this article to convert protein sequences of different lengths into fixed length vectors. PSSM algorithm is a common algorithm in the field of bioinformatics, known as the "position-specific scoring matrix," which can store the evolutionary information of protein sequences so that it can be used for protein prediction. It is a matrix that calculates the percentage of different residues at each position in a multi sequence alignment, the matrix size is $L \times 20$ (L for protein sequence length). Among them, CC is a measure of correlation between

two different properties of amino acid residues and can be calculated using the following equation:

$$CC(i1,i2,lag) = \sum_{j=1}^{L-lag} \left(S_{i1,j} - \bar{S}_{i1}\right)\left(S_{i2,j+lag} - \bar{S}_{i2}\right)/(L-lag) \quad (3)$$

where$i1, i2$represents amino acids, and $\bar{S}_{i1}, \bar{S}_{i2}$ represents the average score of $i1, i2$along the protein sequence. LAG is the maximum lag, lag is an integer value from 1 to LAG, and the total number of CC variables is $380 \times LAG$. In this paper, we set the value ofLAG to 2 to obtain a $720\,(380 \times 2)$-dimensional vector.

Feature Selection

When the feature size is large, there may be irrelevant features or inter-dependence between features, which will easily affect the accuracy of the prediction results. In particular, the more feature dimensions, the more likely it is to lead to "dimension disaster," model complexity and model generalization ability decline. Therefore, removing irrelevant or redundant features through feature selection can improve the accuracy of classification performance and reduce the running time of the model (Polat and Güneş, 2009; Quan et al., 2016; Zou et al., 2016; Guohua and Jincheng, 2018; Wei et al., 2018; Riaz and Li, 2019; He et al., 2020). In this paper, the F-score value is used to select the optimal feature (Chen and Lin, 2008; Cheng et al., 2019; Wei et al., 2019), which is a method to measure the distinguishing ability of features between the two categories, and the most effective feature selection can be achieved through this method. Therefore, we can use (**Eq. 4**) to describe the contribution of each feature and perform feature selection:

$$F(i) = \frac{s_b^2(i)}{s_w^2(i)} \quad (4)$$

where$F(i)$ is the score of theith feature of the F-score. Generally, the larger the value of $F(i)$ is, the stronger the ability to recognize samples.$s_w^2(i)$ is the intragroup variance, and$s_b^2(i)$ is the intergroup variance. Their calculation formula is as follows:

$$\begin{cases} s_b^2(i) = \dfrac{ss_b(i)}{K-1} \\ s_w^2(i) = \dfrac{ss_w(i)}{N-K} \end{cases} \quad (5)$$

where$ss_b(i)$is the sum of squares between groups; $ss_w(i)$is the sum of squares within the group; $K$is the total number of classes; and$N$is the total number of samples.

Classifier

In this paper, Naive Bayes, Random forests, logistic regression, linear discriminant and other classification algorithms are used to predict HBPs.

Na ve Bayes

The Naive Bayes method is a classification method based on Bayes' theorem and the assumption of the independence of characteristic conditions. It is characterized by combining prior probability and posterior probability and a very widely used algorithm. The main idea of the naive Bayes classifier is to solve the posterior probability $P(Y|X)$ through joint probability modeling and use Bayes' theorem. Then, the category corresponding to the largest posterior probability is used as the predicted category. Suppose there is a sample dataset $D = \{d_1, d_2, \cdots, d_n\}$, the feature dataset corresponding to the sample dataset is $X = \{x_1, x_2, \cdots, x_d\}$, features are independent and random, and the class variable is $Y = \{y_1, y_2, \cdots y_m\}$. According to the Naive Bayes algorithm, the posterior probability of the sample belonging to category$y$can be expressed as:

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} \quad (6)$$

Where$P(Y)$is the prior probability, Naive Bayes is based on the independence of each feature. In the case of a given category, **Eq. 6** can be further expressed as the following equation:

$$P(X|Y=y) = \prod_{i=1}^{d} P(x_i|Y=y) \quad (7)$$

The posterior probability can be calculated from the above two **Eqs 6**, **7**:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(x_i|Y)}{P(X)} \quad (8)$$

Since the magnitude of $P(X)$is fixed, when comparing the posterior probability, only the molecular part of the above equation can be compared. Therefore, a naive Bayesian calculation of sample data belonging to category y_i can be obtained:

$$P(y_i|x_1, x_2, \cdots, x_d) = \frac{P(y_i)\prod_{j=1}^{d} P(x_j|y_i)}{\prod_{j=1}^{d} P(x_j)} \quad (9)$$

Random Forests

RF is a flexible, easy-to-use machine learning algorithm that contains multiple decision trees. It is an optimized version of bagging (Su et al., 2019; Zeng et al., 2020). The idea of bagging is to vote on the results of multiple weak classifiers to combine them into a strong classifier, thereby improving the prediction accuracy of the model. In the training phase, RF uses the bootstrap sampling method to collect multiple different subsets from the input training dataset and then uses the different collected subsets to train the internal decision tree. Then, in the prediction phase, RF votes for the prediction results of multiple internal decision trees and then outputs the prediction results. Its advantages are as follows: 1) it can process high-dimensional data without feature selection; 2) accuracy can be maintained even if many of the features are missing; and 3) it has a fast training speed (Jiao et al., 2021).

Logistic Regression

As a classification model, LR can deal with the 0/1 classification problem because of the nonlinear factor introduced by the

sigmoid function. The image of the logical function is an S-shaped curve with values between (0, 1). The farther away from 0 a function is, the closer to 0 or 1 the value of the function will be. Therefore, this feature can be used to solve the problem of binary classification. The function formula is as follows:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{10}$$

Among them, $z = \theta^T x = \sum_{i=0}^{n} \theta_i x_i = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$; therefore, the predictive function of logistic regression can be expressed as:

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

Linear Discriminant Analysis

LDA is a classical linear learning method, also known as "Fisher" discriminant analysis in dichotomies. Unlike the perception machine, the principle of LDA is dimension reduction. In other words, given a set of training samples, the article tries to sample projections to a straight line, keeping the points with the same classification as close as possible and the classification of different points as far apart as possible, i.e., maximizing and minimizing the variance between variance. LDA can, therefore, make use of sample points in the projection line (or projection location) to determine the type of sample.

Performance Evaluation

In this article, we use the specificity (SP), sensitivity (SN), accuracy (ACC) (Yang et al., 2021) and Matthews correlation coefficient (MCC) to evaluate our proposed method (Snow et al., 2005; Cheng et al., 2018), which can be expressed as:

1. Accuracy: ACC represents the probability that all positive and negative samples will be correctly predicted.

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{12}$$

2. Sensitivity: SN represents the probability that the actual hormone-binding protein is predicted to be a hormone-binding protein.

$$SN = \frac{TP}{TP + FN} \tag{13}$$

3. Specificity: SP represents the probability that a non-hormone-binding protein is predicted to be a non-hormone-binding protein.

$$SP = \frac{TN}{TN + FP} \tag{14}$$

4. MCC: MCC represents the reliability of the algorithm results.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

5 Precision: Indicates how many of the samples predicted to be positive are true positive samples.

$$pre = \frac{TP}{TP + FP} \tag{16}$$

6. F1-Score: The F1 score is balanced by taking into account both accuracy and recall, so that both are maximized at the same time.

$$F1 - Socre = \frac{2 \times pre \times recell}{pre + recell} \tag{17}$$

Where, the recall rate is: $recell = \frac{TP}{TP+FN}$

7. The ROC curve: Receiver operating characteristic curve (the area under the curve is AUROC), *X*-axis is false positive rate (FPR), *Y*-axis is true positive rate (TPR):

$$TPR = \frac{TP}{TP + FN} \tag{18}$$

$$FPR = \frac{FP}{FP + TN} \tag{19}$$

8. PRC: PRC takes precision rate as *Y*-axis and recall rate as *X*-axis.

Where *TP*refers to the model correctly predicting positive category samples; *FP*refers to the model incorrectly predicting negative category samples as positive category; *TN* refers to the model correctly predicting negative category samples; and *FN*refers to the model incorrectly predicting positive category samples as negative category (Ding et al., 2020a; Ding et al., 2020b).

In machine learning, a test set is needed to test the model and describe its generalization ability. However, in practical applications, due to the limited number of datasets, cross validation is used as a test method. There are three types of cross validation: K-fold cross validation, fold cross validation and independent data verification. In this article, we use K-fold cross-validation to test the constructed model. K-fold cross-validation divides the training data into K parts, of which (K-1) pieces of data are used to train the model, and the remaining 1 piece of data is used to evaluate the quality of the model. This process is cycled K times, and the K evaluation results obtained are combined, such as averaging or voting. The flow chart of K-fold cross verification is shown in **Figure 2**.

RESULTS AND DISCUSSION

In machine learning, the predicted results of the model can be tested through cross-validation. In this article, we use 10-fold cross-validation to evaluate the built model.

Performance Comparison of Different Feature Expression Methods

According to the feature extraction part, protein sequences are transformed into feature vectors of different sizes through different feature extraction methods. Therefore, in this study we tested the performance of three feature extraction methods: k-mer (K = 2), k-mer (K = 3), DR and CC-PSSM.

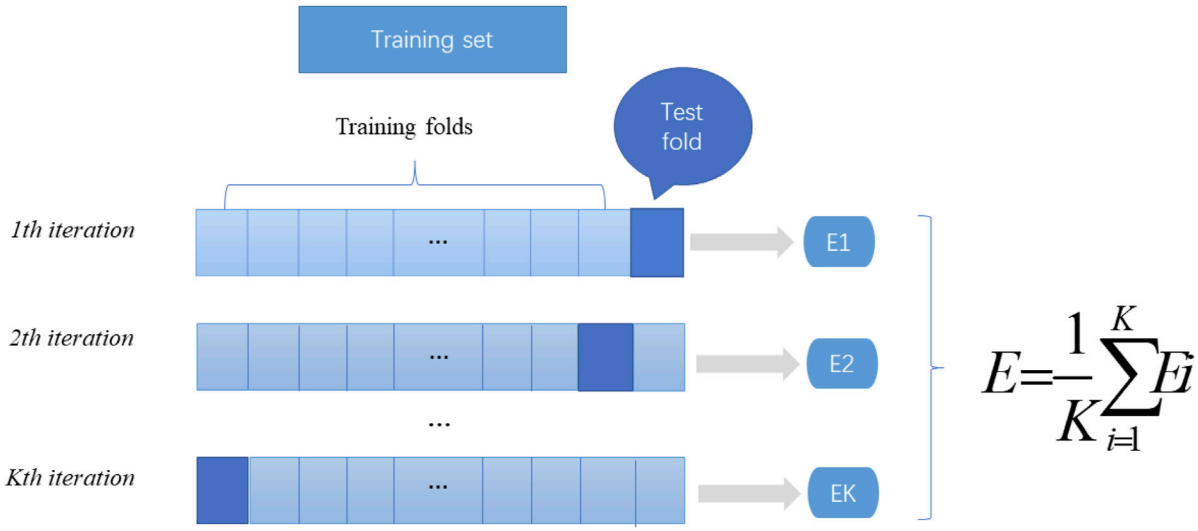


FIGURE 2 | K-fold cross-validation diagram. Divide the data into K parts, where k-1 parts are used as the training dataset, and the remaining part is used as the test set. The mean value of the results of the k groups is calculated as the performance index of the current k-fold cross-validation evaluation model.

TABLE 1 | Prediction results of different feature extraction algorithms based on the Bayesian classifier.

Feature extraction	SN(%)	SP(%)	ACC(%)	MCC(%)	AUROC(%)	PRC(%)
K-mer(k = 3)	**94.17**	**96.73**	**95.45**	**91.36**	**95.17**	**96.55**
K-mer(k = 2)	65.51	78.46	71.96	44.50	77.89	76.97
DR	83.46	37.12	60.39	25.64	66.35	75.99
CC-PSSM	64.10	80.13	72.09	45.29	78.24	80.27

First, use the F-score feature selection method to reduce the dimensionality of the feature vectors obtained by different feature extraction methods to 250 dimensions, then use the selected best feature vector as the input vector of the naive Bayes algorithm and perform 10-fold cross-validation, and finally draw forecast results. The prediction results are shown in **Table 1** (the maximum value is in bold). As shown in **Table 1**, the k-mer (k = 3) feature extraction algorithm used in this model performs best in all indicators, among which the values of ACC, MCC, SP and SN are, respectively, 95.45,91.36, 96.73, and 94.17%. These results prove the validity of our model.

Comparison With Other Classifiers

To show the superiority of naive Bayes in HBPs recognition, we can compare the HBPs recognition performance of different classification algorithms based on the same feature subset (i.e. 250 optimal features). In this paper, we used the constructed HBP_NB model to perform performance comparison with RF, LDA, Logistic regression and other models under the condition of 10-fold cross-validation, and the comparison results are shown as follows. **Table 2** shows the specific values of different classification models under SN, SP, ACC, MCC and other indicators (the maximum values are in bold). As can be seen from **Table 2**, HBP_NB prediction model achieved better results than other classification algorithms in identifying hormone-binding proteins, in which ACC, MCC, SN and SP were 95.45, 91.36, 94.17 and 96.73%, respectively. **Figures 3**, **4** respectively show the boxplot diagram of different models, ROC and PRC curves schematic diagram. These results show that our model has good classification ability. Therefore, we construct the final model based on naive Bayes. Where, the line in the middle of the box in the boxplot is the median of the data, representing the average level of the sample data; The top of the box represents the upper quartile and the bottom quartile represents the lower quartile, which means the box contains 50% of the data, so the width of the box reflects, to some extent, how much the data fluctuates; at the same time, the lines above and below the box represent the maximum and minimum values of data. The ROC curve is a curve that evaluates the effect of binary model on positive

TABLE 2 | Performance comparison of different classifiers under 10-fold cross validation

Classifier	SN(%)	SP(%)	ACC(%)	MCC(%)	AUROC(%)	PRC(%)
NB	94.17	**96.73**	**95.45**	**91.36**	**95.17**	**96.55**
RF	77.95	87.57	82.71	66.26	89.45	91.19
LDA	72.24	70.13	71.20	43.08	94.53	95.32
LR	**96.92**	17.50	57.00	14.42	76.35	79.43

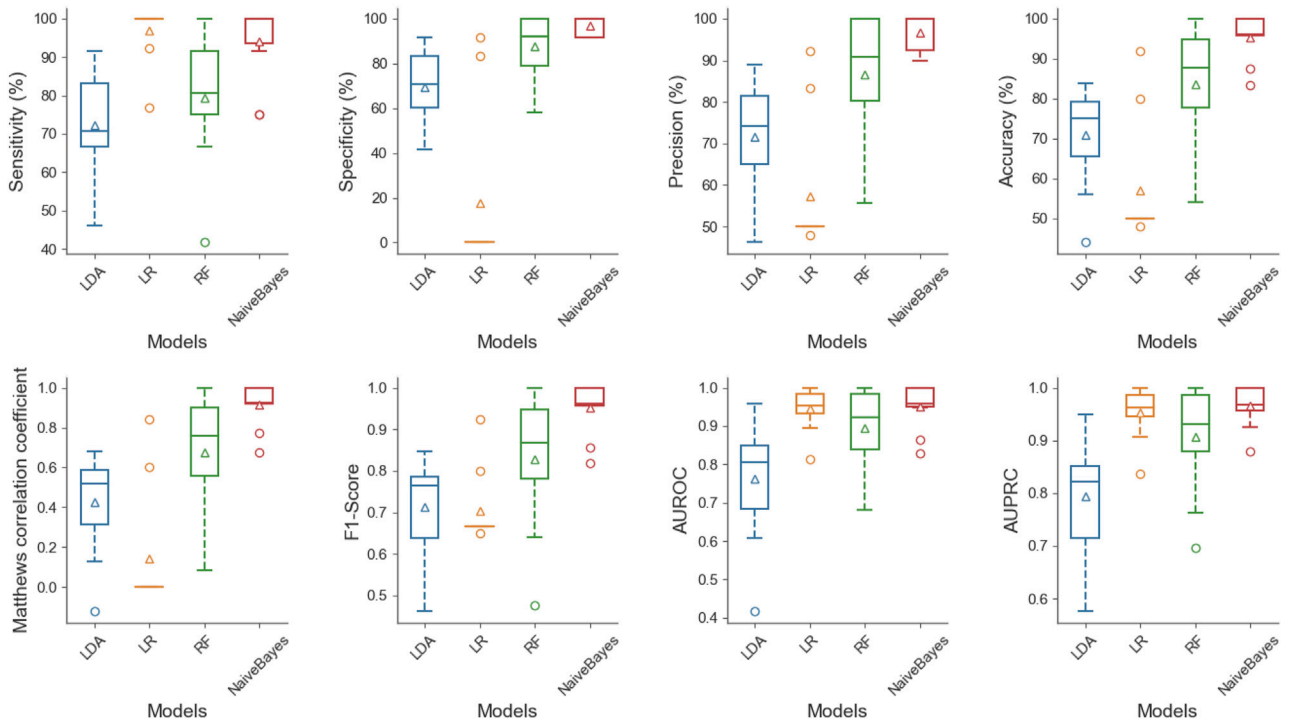


FIGURE 3 | Boxplot diagram of different classification models, this figure shows the distribution of LDA, LR, RF and NB under SN, SP, accuracy, ACC, MCC, F1-Score, AUROC and AUPRC successively from left to right and from top to bottom. At the same time, it can be seen from the figure that NB can achieve good results under different indicators.

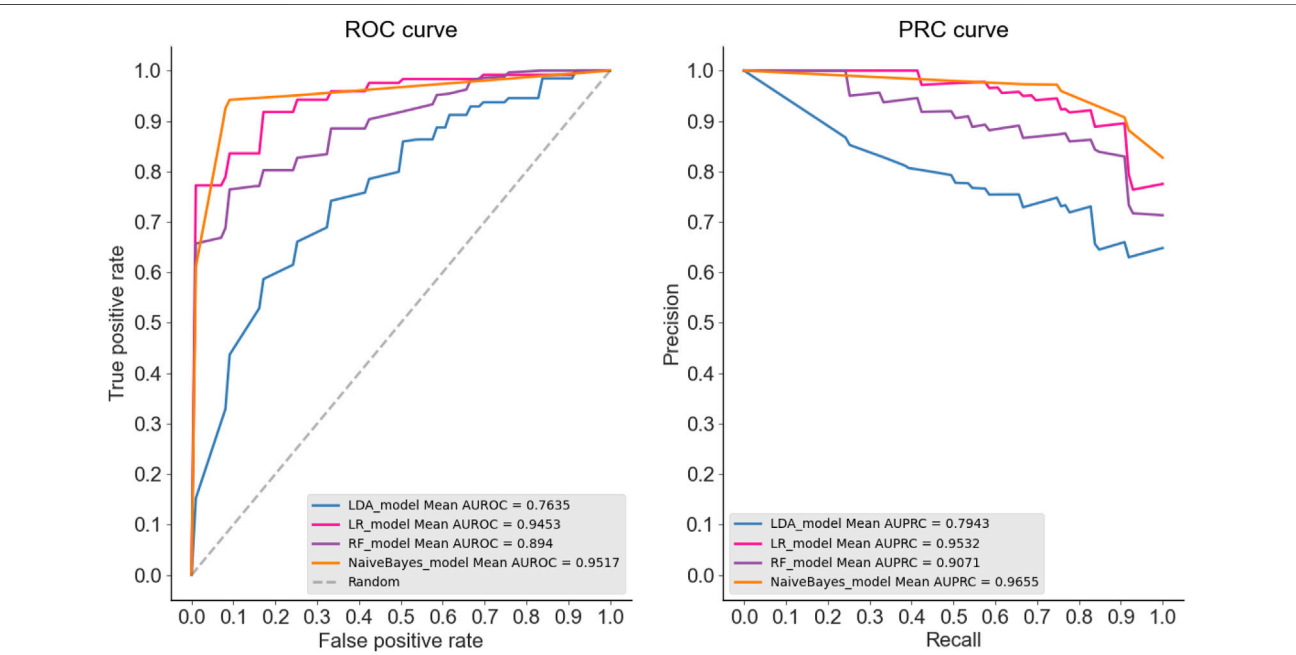


FIGURE 4 | As can be seen from the ROC curves and PRC curves of different classification models, the ROC curves of LDA, RF, LR and NB are 0.7635, 0.894 and 0.9453, respectively. The dotted line represents the ROC curve of a completely random classifier, and the ROC curve of a good classifier should be as far away from the dotted line as possible, as close as possible to the upper left corner; The PRC curve values of LDA, RF, LR and NB were 0.7943, 0.9071, 0.9532 and 0.9655, respectively, the closer the curve was to the upper right corner, the better the model classification ability was. Therefore, we constructed the final model based on NB.

TABLE 3 | Comparison of our method with other published methods.

Methods	SN(%)	SP(%)	ACC(%)
HBPred Hua et al. (2018)	80.43	56.52	68.48
iGHBP Basith et al. (2018)	86.96	47.83	67.39
HBP_NB	**94.17**	**96.73**	**95.45**

category prediction. *X*-axis is false positive rate (FPR), *Y*-axis is true positive rate (TPR), which indicates that the optimal classifier with the best performance is located in the upper left corner of the image (coordinate 0,1), and the area under its ROC curve is AUROC, with an area value between 0,1. PRC takes presion rate as *Y*-axis and recall rate as *X*-axis, and lines are drawn according to changes in the value of probability threshold. The ideal model would be at the point (1,1). The model with excellent performance is as close to this point as possible.

Performance Comparison With the Existing Optimal Algorithm

This section compares the model constructed in the article with other existing methods, in which the results of HBPred (Hua et al., 2018) and iGHBP (Basith et al., 2018) are directly obtained from the literature. The comparison results are shown in **Table 3** (the maximum value is in bold). As seen from **Table 3**, the HBP_NB model constructed in this paper has the best performance in all indicators, among which ACC, SP and SN have reached maximum values of 95.45, 96.73 and 94.17%, respectively. The effect is significantly better than that of the other two methods, which also proves the effectiveness of the HBP_NB model constructed in this paper.

CONCLUSION

As a carrier protein related to the regulation of hormones in the circulatory system, HBPs can cause various diseases when they are abnormally expressed. Therefore, it is very important to understand their function and regulatory mechanism, and the correct identification of HBPs is the first step in understanding their biological process and is necessary to further study their function. There is growing evidence that it is crucial to develop an efficient computational model to identify hormone-binding proteins. In this study, we used a reliable predictive model for HBP_NB to identify HBPs. First, the model uses the k-mer feature extraction method to extract the features of HBPs. Then, to remove redundancy and noise and improve the accuracy of model prediction, the F-score value is used to sort the features and select the optimal features. Secondly, the reduced feature set is input into naive Bayes classifier and the 10-fold cross validation is used to judge the quality of the prediction model. Finally, the accuracy, sensitivity and specificity of the HBP_NB model reached 95.45, 94.17 and 96.73%, respectively, in 10-fold cross validation. The feasibility and validity of our model are illustrated.

However, there is room for improvement in our current approach. Since the data set selected in this experiment is small, we will collect more data for model training and independent test set experiments in the future to improve the model's robustness and generalization ability. At the same time, we will further learn more effective feature representation methods and classification algorithms to gain an in-depth understanding of machine learning and establish a more stable model. In addition, we also hope that our work can help scholars to study hormone binding proteins, to promote research on hormone-binding protein drugs.

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

Conceptualization, WZ and PW; data collection or analysis, YG and PW; validation, YG and LH; writing—original draft preparation, YG and LH; writing—review and editing, YG and PW. All authors have read and agreed to the published version of the article.

FUNDING

This work was supported by the National Nature Science Foundation of China (Grant Nos 61,863,010, 11926205, 11926412, and 61873076), National Key R&D Program of China (No.2020YFB2104400) and Natural Science Foundation of Hainan, China (Grant Nos. 119MS036 and 120RC588), and Hainan Normal University 2020 Graduate Student Innovation Research Project (hsyx 2020–41), The Special Science Foundation of Quzhou (2020D003)

ACKNOWLEDGMENTS

Thanks to the guidance of my tutor and the joint efforts of other authors, the success of this article is the result of everyone's joint efforts.

REFERENCES

Akbar, S., Khan, S., Ali, F., Hayat, M., Qasim, M., and Gul, S. (2020). iHBP-DeepPSSM: Identifying Hormone Binding Proteins Using PsePSSM Based Evolutionary Features and Deep Learning Approach. *Chemometrics Intell. Lab. Syst.* 204, 104103. doi:10.1016/j.chemolab.2020.104103

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25 (1), 25–29. doi:10.1038/75556

Bairoch, A., Bougueleret, L., Altairac, S., Amendolia, V., and Zhang, J. (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* 37 (Suppl. 1), D169–D174. doi:10.1093/nar/gkn664

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

Bin, L., Hao, W., and Kuo-Chen, C. (2017). Pse-in-One 2.0: An Improved Package of Web Servers for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Nat. Sci.* 9 (4), 67–91. doi:10.4236/ns.2017.94007

Chen, Y. W., and Lin, C. J. (2008). *Combining SVMs with Various Feature Selection Strategies Feature Extraction*. Taipei, Taiwan: Studies in Fuzziness and Soft Computing.

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., 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., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019). MetSigDis: a Manually Curated Resource for the Metabolic Signatures of Diseases. *Brief Bioinform* 20 (1), 203–209. doi:10.1093/bib/bbx103

Christopher, F. B., Dongwon, L., Mccallion, A. S., and Beer, M. A. (2013). Kmer-SVM: a Web Server for Identifying Predictive Regulatory Sequence Features in Genomic Data Sets. *Nucleic Acids Res.* W1, W544–W556. doi:10.1093/nar/gkt519

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

Dong, Q., Zhou, S., and Guan, J. (2009). A New Taxonomy-Based Protein Fold Recognition Approach Based on Autocross-Covariance Transformation. *Bioinformatics* 25 (20), 2655–2662. doi:10.1093/bioinformatics/btp500

Einarsdóttir, I. E., Gong, N., Jnsson, E., Sundh, H., Hasselberg-Frank, L., Nilsen, T. O., et al. (2014). Plasma Growth Hormone-Binding Protein Levels in Atlantic salmonSalmo Salarduring Smoltification and Seawater Transfer. *J. Fish Biol.* 85 (4), 1279–1296. doi:10.1111/jfb.12473

Fang, S., Pan, J., Zhou, C., Tian, H., He, J., Shen, W., et al. (2019). Circular RNAs Serve as Novel Biomarkers and Therapeutic Targets in Cancers. *Cgt* 19 (2), 125–133. doi:10.2174/1566523218666181109142756

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data. *Bioinformatics* 28 (23), 3150–3152. doi:10.1093/bioinformatics/bts565

Gong, Z., and Tian, Y. (2010). "Chinese Web Text Classification System Model Based on Naive Bayes," in International Conference on E-product E-service & E-entertainment, Henan, China, 7-9 Nov. 2010.

Gumus, F., Sakar, C. O., Erdem, Z., and Kursun, O. (2014). "Online Naive Bayes Classification for Network Intrusion Detection," in IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining, Beijing, China, 17-20 Aug. 2014.

Guohua, H, and Jincheng, L (2018). Feature Extractions for Computationally Predicting Protein Post-Translational Modifications. *Curr. Bioinformatics* 12 (4), 387–395. doi:10.2174/1574893612666170707094916

He, S., Guo, F., Zou, Q., and Ding, H. (2020). MRMD2.0: A Python Tool for Machine Learning with Feature Ranking and Reduction. *Curr. Bioinformatics* 15 (10), 1213–1221.

He, X., Wang, S., Wang, R., and Dan, Z. (2010). "Research of P2P Traffic Identification Based on Naive Bayes and Decision Tables Combination Algorithm," in Seventh International Conference on Fuzzy Systems & Knowledge Discovery, Yantai, China, 10-12 Aug. 2010.

Hu, Y., Qiu, S., and Cheng, L. (2021). Integration of Multiple-Omics Data to Analyze the Population-specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* 2021, 7036592. doi:10.1155/2021/7036592

Hu, Y., Sun, J. Y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (2021). Variant Associates with Alzheimer's Disease and Regulates TMEM106B Expression in Human Brain Tissues. *BMC Med.* 19 (1), 11. doi:10.1186/s12916-020-01883-5

Hu, Y., Zhang, H., Liu, B., Gao, S., Wang, T., Han, Z., et al. (2020). rs34331204 Regulates TSPAN13 Expression and Contributes to Alzheimer's Disease with Sex Differences. *Brain* 143 (11), e95. doi:10.1093/brain/awaa302

Hua, T., Zhao, Y. W., Ping, Z., Zhang, C. M., Rong, C., Huang, P., et al. (2018). HBPred: a Tool to Identify Growth Hormone-Binding Proteins. *Int. J. Biol.* 14 (8), 957–964. doi:10.7150/ijbs.24174

Jiao, S., Zou, Q., Guo, H., and Shi, L. (2021). iTTCA-RF: a Random forest Predictor for Tumor T Cell Antigens. *J. Transl Med.* 19 (1), 449. doi:10.1186/s12967-021-03084-x

Lei, X., Shanshan, J., Jin, W., and Quan, Z. (2021). An In Silico Approach to Identification, Categorization and Prediction of Nucleic Acid Binding Proteins. *Brief. Bioinform.* 22 (3), bbaa171. doi:10.1093/bib/bbaa171

Lin, H. (2020). Development and Application of Artificial Intelligence Methods in Biological and Medical Data. *Cbio* 15 (6), 515–516. doi:10.2174/1574893615062006101123 45

Liu, B., Fang, L., Wang, S., Wang, X., Li, H., and Chou, K. C. (2015). Identification of microRNA Precursor with the Degenerate K-Tuple or Kmer Strategy. *J. Theor. Biol.* 385, 153–159. doi:10.1016/j.jtbi.2015.08.025

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K. C. (2015). Pse-in-One: a Web Server for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Nucleic Acids Res.* W1, W65–W71. doi:10.1093/nar/gkv458

Liu, B., Long, R., and Chou, K. C. (2016). iDHS-EL: Identifying DNase I Hypersensitive Sites by Fusing Three Different Modes of Pseudo Nucleotide Composition into an Ensemble Learning Framework. *Bioinformatics* 32 (16), 2411–2418. doi:10.1093/bioinformatics/btw186

Liu, B., Wang, X., Lin, L., Dong, Q., and Wang, X. (2008). A Discriminative Method for Protein Remote Homology Detection and Fold Recognition Combining Top-N-Grams and Latent Semantic Analysis. *BMC Bioinformatics* 9, 510. doi:10.1186/1471-2105-9-510

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a Web Server for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Nucleic Acids Res.* 43, W65–W71. doi:10.1093/nar/gkv458

Liu, B., Xu, J., Zou, Q., Xu, R., Wang, X., and Chen, Q. (2014). Using Distances between Top-N-Gram and Residue Pairs for Protein Remote Homology Detection. *Bmc Bioinformatics* 15 (S2), S3. doi:10.1186/1471-2105-15-s2-s3

Li, W, and Godzik, A(2006). Cd-hit A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* 22 (13), 1658. doi:10.1093/bioinformatics/btl158

Liu, Z.-P. (2020). Predicting lncRNA-Protein Interactions by Machine Learning Methods: A Review. *Curr. Bioinformatics* 15 (8), 831–840.

Liu, B (2017). BioSeq-Analysis: a Platform for DNA, RNA and Protein Sequence Analysis Based on Machine Learning Approaches. *Brief. Bioinform.* 20 (4), 4. doi:10.1093/bib/bbx165

Manavalan, B., Basith, S., Shin, T. H., Lee, D. Y., Wei, L., Lee, G., et al. (2019). 4mCpred-EL: An Ensemble Learning Framework for Identification of DNA N4-Methylcytosine Sites in the Mouse Genome. *Cells* 8 (11), 1332. doi:10.3390/cells8111332

Mortezaeefar, M., Fotovat, R., Shekari, F., and Sasani, S. (2019). Comprehensive Understanding of the Interaction Among Stress Hormones Signalling Pathways by Gene Co-expression Network. *Cbio* 14 (7), 602–613. doi:10.2174/1574893614666190226160742

Niu, M., Wu, J., Zou, Q., Liu, Z., and Xu, L. (2021). rBPDL: Predicting RNA-Binding Proteins Using Deep Learning. *IEEE J. Biomed. Health Inform.* (99), 1. doi:10.1109/jbhi.2021.3069259

Niu, M., Zou, Q., and Zou, Q. (2021). SgRNA-RF: Identification of SgRNA On-Target Activity with Imbalanced Datasets. *Ieee/acm Trans. Comput. Biol. Bioinf.* 105 (16), 1. doi:10.1109/tcbb.2021.3079116

Polat, K., and Güneş, S. (2009). A New Feature Selection Method on Classification of Medical Datasets: Kernel F-Score Feature Selection. *Expert Syst. Appl.* 36 (7), 10367–10373. doi:10.1016/j.eswa.2009.01.041

Qu, K., Han, K., Wu, S., Wang, G., and Wei, L. (2017). Identification of DNA-Binding Proteins Using Mixed Feature Representation Methods. *Molecules* 22 (10), 1602. doi:10.3390/molecules22101602

Qu, K., Zou, Q., and Shi, H. (2021). Prediction of Diabetic Protein Markers Based on an Ensemble Method. *Front. Bioscience-Landmark* 26 (7), 207–221. doi:10.52586/4935

Quan, Z., Zeng, J., Cao, L., and Ji, R. (2016). A Novel Features Ranking Metric with Application to Scalable Visual and Bioinformatics Data Classification. *Neurocomputing* 173, 346–354. doi:10.1016/j.neucom.2014.12.123

Riaz, F., and Li, D. (2019). Non-coding RNA Associated Competitive Endogenous RNA Regulatory Network: Novel Therapeutic Approach in Liver Fibrosis. *Cgt* 19 (5), 305–317. doi:10.2174/1566523219666191107113046

Schneider, M. (2012). *The Annotation of Plant Proteins in UniProtKB*. California: Plant & Animal Genome.

Shen, Z., and Zou, Q. (2020). Basic Polar and Hydrophobic Properties Are the Main Characteristics that Affect the Binding of Transcription Factors to Methylation Sites. *Bioinformatics* 36 (15), 4263–4268. doi:10.1093/bioinformatics/btaa492

Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y., and Hay, S. I. (2005). The Global Distribution of Clinical Episodes of Plasmodium Falciparum Malaria - Supplementary Information. *Nature* 434, 214–217. doi:10.1038/nature03342

Sohm, F., Manfroid, I., Pezet, A., Rentier-Delrue, F., Rand-Weaver, M., Kelly, P. A., et al. (1998). Identification and Modulation of a Growth Hormone-Binding Protein in Rainbow trout (*Oncorhynchus mykiss*) Plasma during Seawater Adaptation. *Gen. Comp. Endocrinol.* 111 (2), 216–224. doi:10.1006/gcen.1998.7106

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A Deep forest Model to Predict Anti-cancer Drug Response. *Methods* 166, 91–102. doi:10.1016/j.ymeth.2019.02.009

Tan, J.-X., Li, S. H., Li, S.-H., Zhang, Z.-M., Chen, C.-X., Chen, W., et al. (2019). Identification of Hormone Binding Proteins Based on Machine Learning Methods. *Math. biosciences Eng. MBE* 16 (4), 2466–2480. doi:10.3934/mbe.2019123

Tang, H., Zhao, Y.-W., Zou, P., Zhang, C.-M., Chen, R., Huang, P., et al. (2018). HBPred: a Tool to Identify Growth Hormone-Binding Proteins. *Int. J. Biol. Sci.* 14 (8), 957–964. doi:10.7150/ijbs.24174

Wang, J., Shi, Y., Wang, X., and Chang, H. (2020). A Drug Target Interaction Prediction Based on LINE-RF Learning. *Cbio* 15 (7), 750–757. doi:10.2174/1574893615666191227092453

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

Wang, K., Li, S., Wang, Q., and Hou, C. (2018). Identification of Hormone-Binding Proteins Using a Novel Ensemble Classifier. *Computing* 101 (6), 693–703. doi:10.1007/s00607-018-0682-x

Wang, X.-F., Gao, P., Liu, Y.-F., Li, H.-F., and Lu, F. (2020). Predicting Thermophilic Proteins by Machine Learning. *Cbio* 15 (5), 493–502. doi:10.2174/1574893615666200207094357

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019). Iterative Feature Representations Improve N4-Methylcytosine Site Prediction. *Bioinformatics* 35 (23), 4930–4937. doi:10.1093/bioinformatics/btz408

Wei, L., Su, R., Wang, B., Li, X., and Zou, Q. (2018). Integration of Deep Feature Representations and Handcrafted Features to Improve the Prediction of N 6 -methyladenosine Sites. *Neurocomputing* 324, S0925231218306325. doi:10.1016/j.neucom.2018.04.082

Xiao, Y., Zhang, J., and Deng, L. (2017). Prediction of lncRNA-Protein Interactions Using HeteSim Scores Based on Heterogeneous Networks. *Sci. Rep.* 7 (1), 3664. doi:10.1038/s41598-017-03986-1

Yan, XY, Zhang, SW, and Zhang, SY. 2016. Prediction of Drug-Target Interaction by Label Propagation with Mutual Interaction Information Derived from Heterogeneous Network. *Mol. Biosyst.* 12, 520–531. doi:10.1039/c5mb00615e

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

Zeng, X., Yuan, S., Huang, X., and Zou, Q. (2015). Identification of Cytokine via an Improved Genetic Algorithm. *Front. Comp. Sci.* 9 (004), 643–651. doi:10.1007/s11704-014-4089-3

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

Zhang, J., Ju, Y., Lu, H., Xuan, P., and Zou, Q. (2016). Accurate Identification of Cancerlectins through Hybrid Machine Learning Technology. *Int. J. Genomics* 2016 (7-13), 1–11. doi:10.1155/2016/7604641

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2017). The Linear Neighborhood Propagation Method for Predicting Long Non-coding RNA–Protein Interactions. *Neurocomputing* 273 (jan.17), 526–534. doi:10.1016/j.neucom.2017.07.065

Zhang, X., Shi, S., Shen, J., Zhao, M., and He, Q. (2019). Functional Immunoregulation by Heme Oxygenase 1 in Juvenile Autoimmune Diseases. *Cgt* 19 (2), 110–116. doi:10.2174/1566523219666190710092935

Zhang, Y., and Marchant, T. (1999). Identification of Serum GH-Binding Proteins in the Goldfish (Carassius auratus) and Comparison with Mammalian GH-Binding Proteins. *J. Endocrinol.* 161 (2), 255–262. doi:10.1677/joe.0.1610255

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 (4), 114. doi:10.1186/s12918-016-0353-5

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (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.

Copyright © 2021 Guo, Hou, Zhu and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

METHODS
published: 29 November 2021
doi: 10.3389/fgene.2021.811158

KK-DBP: A Multi-Feature Fusion Method for DNA-Binding Protein Identification Based on Random Forest

Yuran Jia[1], *Shan Huang*[2]* *and Tianjiao Zhang*[1]*

[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2]Department of Neurology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China

DNA-binding protein (DBP) is a protein with a special DNA binding domain that is associated with many important molecular biological mechanisms. Rapid development of computational methods has made it possible to predict DBP on a large scale; however, existing methods do not fully integrate DBP-related features, resulting in rough prediction results. In this article, we develop a DNA-binding protein identification method called KK-DBP. To improve prediction accuracy, we propose a feature extraction method that fuses multiple PSSM features. The experimental results show a prediction accuracy on the independent test dataset PDB186 of 81.22%, which is the highest of all existing methods.

Keywords: DNA-binding protein, position specificity score matrix, random forest, feature extraction, multi-feature fusion

OPEN ACCESS

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

Reviewed by:
Yijie Ding,
University of Electronic Science and Technology of China, China
Fei Guo,
Tianjin University, China
Lihong Peng,
Hunan University of Technology, China

***Correspondence:**
Shan Huang
hmuhuangshan@163.com
Tianjiao Zhang
tianjiaozhang@nefu.edu.cn

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

Received: *08 November 2021*
Accepted: *15 November 2021*
Published: *29 November 2021*

Citation:
Jia Y, Huang S and Zhang T (2021) KK-DBP: A Multi-Feature Fusion Method for DNA-Binding Protein Identification Based on Random Forest. Front. Genet. 12:811158. doi: 10.3389/fgene.2021.811158

INTRODUCTION

Proteins are spatially structured substances formed by the complex folding of amino acids into polypeptide chains through dehydration and condensation. Proteins are the material basis of life and they are required for every vital activity. Given the vast number of proteins and their roles, protein classification has always been central to the study of proteomics. DNA-binding proteins (DBP) are a very specific class of proteins whose specific binding to DNA guarantees the accuracy of biological processes and whose nonspecific binding to DNA guarantees the high efficiency of biological processes (Gao et al., 2008). DNA-protein interactions, such as gene expression and transcriptional regulation, occur ubiquitously throughout the biological activities of living bodies (Liu et al., 2019; Shen and Zou, 2020; Xu et al., 2021a). All of these interactions are tightly linked to DBP, where the fraction of DNA-binding proteins in eukaryotic genes is approximately 6–7%.

The role of DBP in biological activities has gained a lot of attention in recent years, as various large genome projects and research on DBP identification have rapidly progressed. However, identifying DBP using traditional biochemical analyses is inefficient and expensive (Li and Li, 2012; Xu et al., 2021b). In recent years, machine learning methods have been widely used in the field of bioinformatics (Jiang et al., 2013; Geete and Pandey, 2020; Tao et al., 2020; Wang et al., 2021a; Long et al., 2021). Using machine learning methods for DNA-binding protein identification can enable rapid and accurate prediction of DBP from a large number of proteins, while drastically reducing prediction costs (Fu et al., 2018). Because the number of proteins is large and promiscuous, overcoming every classification prediction problem with one method is difficult, if not impossible (Wang et al., 2021b). Therefore, we must continue to propose effective methods for high-quality DBP

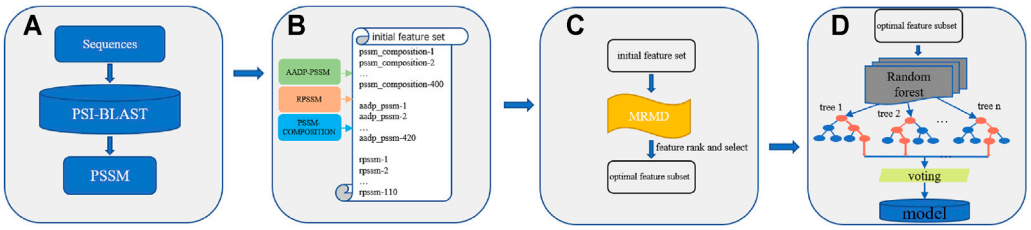


FIGURE 1 | Framework of KK-DBP. **Step A**: Construction of Position Specificity Score Matrices for protein sequences. **Step B:** Extraction of three features: AADP-PSSM, PSSM-COMPOSITION, and RPSSM as the initial feature set for a single sample. **Step C:** Feature ranking and selection using the MRMD algorithm. **Step D:** Identification of DBP using random forests.

prediction and identification in order to understand the significance of more vital activities and to promote further progress within the bioinformatics field.

Feature extraction methods can be broadly classified into two categories: those based on structural information and those based on sequence information (Kim et al., 2004; Meng and Kurgan, 2016; Qu et al., 2019; Ao et al., 2021a; Lv et al., 2021a; Liu et al., 2021; Tang et al., 2021; Wu and Yu, 2021); (Stawiski et al., 2003) proposed a model based on protein structure that utilises a neural network approach incorporating information like residue and hydrogen bond potential. Liu et al. (Liu et al., 2014) developed a model called IDNA-prot|dis, based on the pseudo amino acid composition (PseAAC) of protein sequence information. iDNAPro-PseAAC (Liu et al., 2015), which uses a similar feature extraction method, adopts a prediction model based on a support vector machine to predict DBP. IDNA-prot (Lin et al., 2011) was constructed based on physicochemical properties and random forest (RF) classification. In addition, a support vector machine model based on k-mer and autocovariance transformation was proposed by Dong et al. (Liu et al., 2016). Local-DPP (Wei et al., 2017a) used random forests based on PSE-PSSM features to predict DBP. MK-FSVM-SVDD is a multiple kernel SVM prediction tool based on the heuristic kernel alignment developed by Ding et al. (Zou et al., 2021) to identify DBP. In addition, two models for predicting DBP were developed: DNA-prot (Kumar et al., 2009) and DNAbinder (Kumar et al., 2007). Lu et al. (Lu et al., 2020) developed a prediction model for DBP based on support vector machines using Chou's five-step rule.

Currently, a number of DNA-binding protein prediction methods based on different strategies exist. Unfortunately, most of these DBP prediction methods fail to extract features based on evolutionary information, so their robustness and prediction accuracy have much room for improvement. To address these issues, more research is needed with regard to feature extraction and the selection of classifiers (Zuo et al., 2017; Zheng et al., 2019).

In this paper, we propose a new DNA-binding protein prediction method called KK-DBP. We first obtained the position specificity score matrix (PSSM) of the protein sequence for each sample used to train the model. PSSM information was then used to extract three features of each sample: PSSM-COMPOSITION (Zou et al., 2013), RPSSM (Ding et al., 2014) and AADP-PSSM (Liu et al., 2010), which were combined to form the initial feature set of each sample. The final initial feature set of each sample reached 930 dimensions. To avoid feature redundancy and improve prediction accuracy, KK-DBP used the max relevance max distance (MRMD) (Zou et al., 2016) feature ordering method to establish the optimal feature subset for model training. Finally, a new DBP prediction model was constructed using the random forest learning method. The complete method framework is shown in **Figure 1**:

TABLE 1 | benchmark datasets used in this paper.

Data set	PDB1075	PDB186
Positive	525	93
Negative	550	93
Total	1075	186

MATERIALS AND METHODS

Dataset

The dataset is one of the key factors determining the quality of the predictive model and is the cornerstone of machine learning algorithm learning, which directly affects the final effect of the model, so dataset construction is meticulous and complex (Liang et al., 2017; Su et al., 2021). Other researchers have proposed many prediction models for DNA-binding proteins that have been pertinent to objectively comparing existing data. In the present study, we have used protein sequences from the PDB database as our training dataset and test dataset. **Table 1** shows the contents of the dataset:

The training set PDB1075 contained 525 DNA-binding proteins and 550 non-DNA-binding proteins, and the test set PDB186 contained 93 DNA-binding proteins and 93 non-DNA-binding proteins. The dataset construction rules are as follows:

$$S = S^{+} \cup S^{-} \tag{1}$$

where S^{+} is the positive subset containing only DNA-binding proteins, and S^{-} is the negative subset containing only non-DNA-binding proteins.

Feature Extraction

Feature extraction is very important to modeling sequence classifications, which directly affect the accuracy of predictive models (Zhang et al., 2020a; Lv et al., 2021b). Evolutionary

information is among the most important information we have regarding protein function and genetics (Zuo et al., 2014). Position specificity score matrices (PSSM) can intuitively display protein evolutionary information. Thus, the feature extraction method based on PSSM is widely used in protein classification.

Position specificity Score Matrices

In 1997, Altschul et al. (Altschul et al., 1990) proposed the BLAST algorithm. When given a protein sequence, BLAST can represent the evolutionary information of a protein by aligning it with data in a specific database and extracting a position specific score matrix (PSSM). To improve the prediction accuracy of proteins, our method predominantly utilises protein evolution information to extract features. For the training and test sets used in our method, the PSSM matrices for each sequence were generated by three PSI-BLAST iterations with an E-value of 0.001. The PSSM is a matrix of size L × 20, where L is the length of the protein sequence and 20 is the number of amino acids. Coordinates (i, j) in the position specificity score matrix. (PSSM) represent the log score for the amino acid at position i being replaced by the log score of the amino acid at position j. When the coordinate value is greater than 0, it indicates that during the alignment, there is as large probability that the amino acid at the corresponding position in the sequence is mutated to 20 native amino acids. The higher the value is when the number is a negative integer, the less prone it is to alteration. This numerical pattern indicates the probability of the mutation of a residue in a given protein sequences. Its matrix form behaves as follows:

$$PSSM_{L\times 20} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,20} \\ \vdots & \vdots & \cdots & \vdots \\ p_{L,1} & p_{L,2} & \cdots & p_{L,20} \end{bmatrix} \quad (2)$$

Reduced Position Specificity Score Matrices and Position Specificity Score Matrices-Composition

PSSM-COMPOSITION is generated by adding the same amino acid rows in the original PSSM matrix, dividing by the sequence length and scaling to [-1,1]. For each protein sequence PSSM matrix, a 400-dimensional vector feature$\{d_1, d_2, d_3, ..., d_{400}\}$ is generated.

Li et al. (Li et al., 2003) first proposed that 10 might be the minimum number of residue types (letters) needed to construct a reasonably folded model. Reduced PSSM (RPSSM) borrowed this idea and simplified the original PSSM matrix with form L × 20 to one with form L × 10.

$a_1a_2 \ldots a_L$ is a protein in the dataset, a_i is assumed to be mutated to s, and $p_{i,s}$ represents the pseudo composition component of amino acid a_i. The pseudo composition of all amino acids in protein $a_1a_2 \ldots a_L$ is defined as:

$$D_s = \frac{1}{L}\sum_{i=1}^{L}\left(p_{i,s} - \frac{1}{L}\sum_{i=1}^{L}p_{i,s}\right)^2 \quad s = 1, 2, ...10; \quad i = 1, 2, ..., L \quad (3)$$

The dipeptide composition was later incorporated into the RPSSM method in order to overcome its inability to extract full sequence information. Assuming that a_{i+1} is replaced by 't', the dipeptide pseudocomposition of a_ia_{i+1} is defined as:

$$x_{i,i+1} = \frac{\left(p_{i,s} + p_{i+1,t}\right)^2}{2} \quad s, t = 1, 2, \ldots 10;\ i = 1, 2, \ldots, L-1 \quad (4)$$

where $x_{i,i+1}$ represents the difference of $p_{i,s}$ and $p_{i+1,t}$ from their mean values. Finally, because each protein sequence in the dataset will consist of the pseudo composition of all of its dipeptides, we can generate a 110-dimensional vector feature of RPSSM, defined as follows:

$$D_{s,t} = \frac{1}{L-1}\sum_{i=1}^{L-1}x_{i,i+1} = \frac{1}{L-1}\sum_{i=1}^{L-1}\frac{\left(p_{i,s} + p_{i+1,t}\right)^2}{2} \quad s, t = 1, 2, \ldots 10 \quad (5)$$

AADP-Position Specificity Score Matrices

A protein's structure is closely related to its amino acid composition. For every amino acid sequence in the dataset, AADP-PSSM produces a vector with dimensions 20 + 400 = 420. AADP-PSSM is divided into two parts. The amino acid composition is first extracted from its PSSM matrix: the average value of the PSSM matrix column of length 20 is called AAC-PSSM, where x_i is the type of amino acid in the PSSM matrix and represents the average fraction of amino acid mutations during evolution. It is defined as follows:

$$x_j = \frac{1}{L}\sum_{i=1}^{L}p_{i,j}\ (j = 1, 2, \ldots, 20) \quad (6)$$

The traditional dipeptide composition was later extended to PSSM and represented with DPC-PSSM to avoid the loss of information due to an X in the protein, which was defined as a vector of 400 dimensions:

$$y_{i,j} = \frac{1}{L-1}\sum_{K=1}^{L-1}P_{k,i} \times P_{k+1,j} \quad (1 \le i, j \le 20) \quad (7)$$

Feature Selection

Feature redundancy or dimensionality disasters often occur during feature extraction. Feature selection not only reduces the risk of overfitting but also improves the model's generalization ability and computational efficiency (Guo et al., 2020; Yang et al., 2021a; Ao et al., 2021b; Zhao et al., 2021). In the present paper, we use the max relevance max distance (MRMD) feature selection method to reduce the dimensions of the initial feature set (He et al., 2020).

In MRMD, feature selection is based primarily on the correlation between the subset and the target vector and the redundancy of the subset. When measuring correlations, MRMD used the Pearson correlation coefficient, which is defined as:

$$PCC\left(\vec{X},\vec{Y}\right) = \frac{\frac{1}{N-1}\sum_{k=1}^{N}\left(x_k-\frac{1}{N}\sum_{k=1}^{N}x_k\right)\left(y_k-\frac{1}{N}\sum_{k=1}^{N}y_k\right)}{\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(x_k-\frac{1}{N}\sum_{k=1}^{N}x_k\right)^2}\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(y_k-\frac{1}{N}\sum_{k=1}^{N}y_k\right)^2}} \quad (8)$$

where $\vec{X}$ and $\vec{Y}$ are two vectors, x_k and y_k are the kth elements in $\vec{X}$ and $\vec{Y}$, and N is the total sample number. The initial feature set constructed using this method is $F = \{f_1, f_2, f_3, \ldots, f_{930}\}$. The maximum correlation value $maxMR_i$ between feature f_i and target class vector C is defined as:

$$maxMR_i = \left|PCC\left(\overrightarrow{f_i},\overrightarrow{C_i}\right)\right| (1 \le i \le M) \quad (9)$$

where M is the initial feature set dimension, $\overrightarrow{f_i}$ is the vector composed of the ith feature of each instance, and $\overrightarrow{C_i}$ is the vector composed of the target category of each instance.

When evaluating the similarity between two vectors, MRMD uses the distance functions Euclidean distance (ED), cosine similarity (COS) and Tanimoto coefficient (TC) to measure:

$$ED\left(\vec{X},\vec{Y}\right) = \sqrt{\sum_{k=1}^{N}\left(x_k-y_k\right)^2} \quad (10)$$

$$COS\left(\vec{X},\vec{Y}\right) = \frac{\sum_{k=1}^{N}x_k y_k}{\sqrt{\sum_{k=1}^{N}x_k^2}\cdot\sqrt{\sum_{k=1}^{N}y_k^2}} \quad (11)$$

$$TC\left(\vec{X},\vec{Y}\right) = \frac{\sum_{k=1}^{N}x_k y_k}{\sum_{k=1}^{N}x_k^2+\sum_{k=1}^{N}y_k^2-\sum_{k=1}^{N}x_k y_k} \quad (12)$$

We use the mean of the three above as the maximum distance $maxMD_i$ for feature i:

$$ED_i = \frac{1}{M-1}\sum ED\left(\overrightarrow{f_i},\overrightarrow{f_k}\right) (1 \le k \le M, k \ne i) \quad (13)$$

$$COS_i = \frac{1}{M-1}\sum COS\left(\overrightarrow{f_i},\overrightarrow{f_k}\right) (1 \le k \le M, k \ne i) \quad (14)$$

$$TC_i = \frac{1}{M-1}\sum TC\left(\overrightarrow{f_i},\overrightarrow{f_k}\right) (1 \le k \le M, k \ne i) \quad (15)$$

$$maxMD_i = \frac{1}{3}\left(ED_i + COS_i + TC_i\right) (1 \le i \le M) \quad (16)$$

The MRMD values of all the features are calculated with the above two constraints. The PageRank algorithm is used to sort the initial feature set from high importance. One feature is added to the feature subset at a time and is used to train the model to determine which subset is the best.

Classification Algorithm

Protein prediction is usually described as a binary classification problem (Zhai et al., 2020; Zhang et al., 2021; Zulfiqar et al., 2021). We selected the random forest learning method for prediction modelling in the present study. Because the random forest method randomly extracts features and samples during construction of a decision tree set, it is more suitable to addressing the problem of high feature dimensions. By using RandomizedSearchCV and GridSearchCV for parameter selection, the random forest model constructed finally includes 800 subtrees, in which each tree has no limit, and a single decision tree is allowed to use all features. The maximum depth of each decision tree is 50.

RESULTS

Measurements

We selected four different performance measures, accuracy (ACC), specificity (SP), sensitivity (SN) and Matthew's correlation coefficient (MCC), to evaluate the methodology used by this study to demonstrate the predictive ability of the model used (Wei et al., 2014; Wei et al., 2017b; Manavalan et al., 2019a; Manavalan et al., 2019b; Jin et al., 2019; Su et al., 2019; Li et al., 2020a; Liu et al., 2020a; Ao et al., 2020; Li et al., 2020b; Zhang et al., 2020b; Yu et al., 2020; Zhao et al., 2020; Wang et al., 2021c; Zhu et al., 2021). The equations for determining these four parameters are shown below:

$$ACC = \frac{TN+TP}{TN+FP+FN+FP} \times 100\% \quad (17)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)\times(TN+FN)\times(TP+FP)\times(TN+FP)}} \quad (18)$$

$$SN = \frac{TP}{TP+FN} \times 100\% \quad (19)$$

$$SP = \frac{TN}{TN+FP} \times 100\% \quad (20)$$

Where TP represents positive samples predicted to be positive by the model, FP represents negative samples predicted to be positive by the model, and TN represents negative samples predicted to be negative by the model. FN represents positive samples predicted to be negative by the model. Removing the above four performance measures, the ROC curve will also be used to assess the effect of our predictions.

Experimental Results and Analysis

Performance of Different Features on Training Set PDB1075

A large amount of information on homologous proteins is contained in evolutionarily informative features based on the PSSM matrix. In our method, we selected the evolutionary information-based features PSSM-COMPOSITION, RPSSM, and AADP-PSSM for experimentation. To better show the efficiency of prediction models under different combinations of features, the receiver operating characteristic (ROC) curve was used for analysis. The closer the curve is to the y-axis, the better the classification results will be. The area under the curve (AUC) is defined as the area under the ROC curve enclosed by the coordinate axis. The closer the area is to 1, the better the prediction model will be Random forests can achieve better prediction performance when dealing with high-dimensional features. In this section, we use random forests with default hyperparameters on the training set

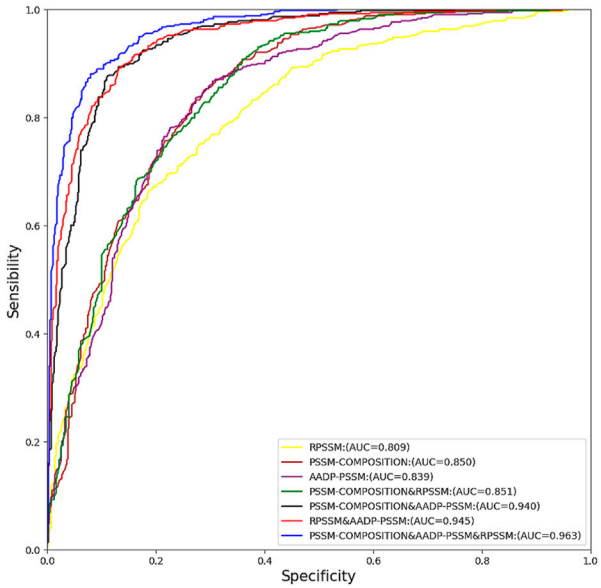


FIGURE 2 | ROC curves with different combinations of features on PDB1075.

pdb1075 for 10-fold cross validation of different feature fusion schemes and find out the feature fusion method that can maximize the area of AUC. As shown in **Figure 2**, the prediction performance of RF was the best after fusing the three features, and its AUC area reached 0.963. In addition, we also tested the predictive performance of SVM and KNN under different feature fusion schemes, and their optimal feature fusion schemes had AUC areas of 0.828 and 0.790, respectively. The ROC curve details of SVM and KNN are given in **Figure 1** and **Figure 2** of **supplementary material** respectively.

Performance After Feature Selection

For the 930-dimensional features of the initial vector set, we ranked all features from high to low based on MRMD scores. After obtaining the final feature ranking results, we took the first feature as the feature subset and utilised random forest to check the performance of the selected feature subset in 10-fold cross validation on PDB1075. Subsequently, we added one feature in the feature subset, one at a time, according to the feature sorting order. Then we repeated the above process until all the features in the initial feature set were included in the feature subset. Finally, we determined the best predictive accuracy and the optimal feature subset. The results are shown in **Figure 3**. The feature subset achieves the best accuracy when it contains 267-dimensional features, so the optimal feature subset we used for training models is 267-dimension. The optimal feature subset contains 98-dimensional AADP-PSSM features, 142-dimensional PSSM-COMPOSITION features, and 27-dimensional RPSSM features. The details of the optimal feature subset are given in the **supplementary materials**. From the distribution of the optimal feature subset, it can be found that the distribution difference of amino acid pairs is the key to identify DBP from massive proteins.

Performance of Different Classification Algorithms

To determine the prediction model with the best performance, we put the best feature subset into four powerful classification algorithms with default hyperparameters, KNN, SVM, RF and naïve Bayes, and we used 10-fold cross validation to compare performance. Experimental results show that the random forest method demonstrates the best classification performance (**Figure 4**).

We use ACC, Sn, SP, MCC and AUC to evaluate the performance. As shown in **Figure 4**, the five indicators of KNN are 78.6, 76.8, 80.1%, 0.571 and 0.785, respectively. The ACC, Sn, SP, MCC and AUC of SVM were 81.6, 88.2, 75.4%, 0.641 and 0.812, respectively. The ACC, Sn, SP, MCC and AUC of

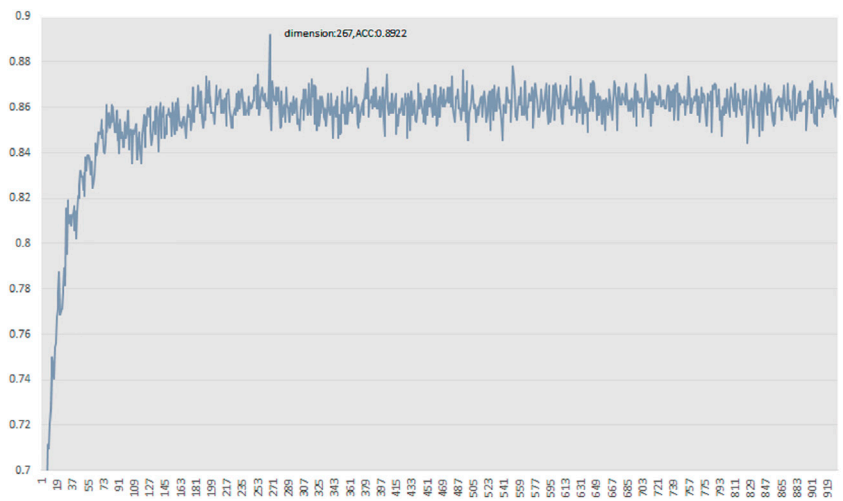


FIGURE 3 | Prediction accuracy curve of feature subset.

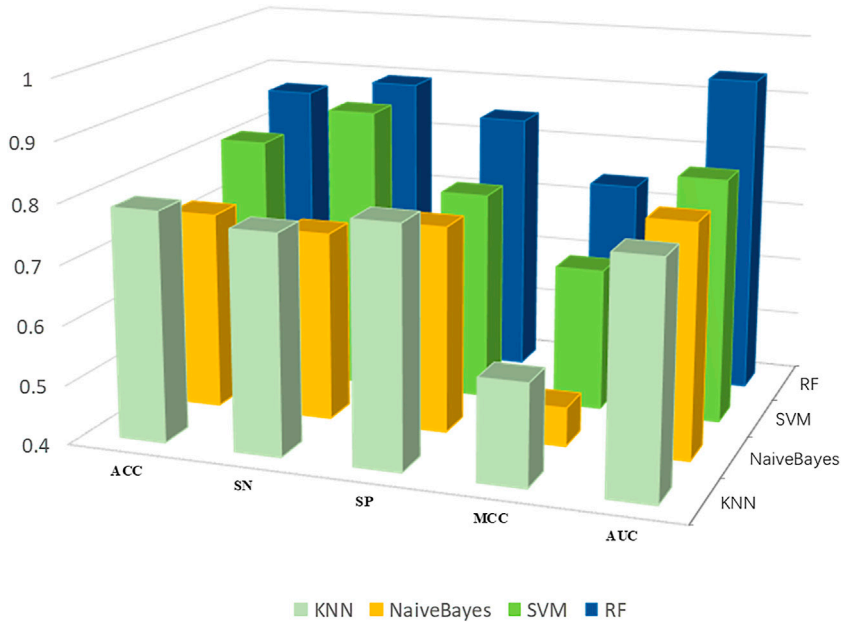


FIGURE 4 | Performance of training set PDB1075 on different classifiers.

TABLE 2 | Performance of this method and other existing methods on PDB186.

Methods	ACC (%)	MCC	SN (%)	SP (%)
IDNA-Prot|dis	72.0	0.445	79.5	64.5
DBPPred	76.9	0.538	79.6	74.2
IDNA-Prot	67.2	0.344	67.7	66.7
DNA-Prot	61.8	0.240	69.9	53.8
DNAbinder	60.8	0.216	57.0	64.5
iDNAPro-PseAAC	71.5	0.442	82.8	60.2
Kmer1+ACC	71.0	0.431	82.8	59.1
Local-DPP	79.0	0.625	92.5	65.6
SVM-based method	75.3	0.560	96.8	53.8
KK-DBP	81.2	0.661	97.8	64.5

Naïve Bayes were 73.3, 71.8, 74.7%, 0.465 and 0.789, respectively. Finally, the performance of RF in the above evaluation indexes are 86.9, 89.6, 84.5%, 0.741 and 0.941, respectively. The experimental results show that RF can yield better prediction results, which proves that RF is the best classification algorithm for Establishing DNA-binding protein prediction model.

Performance of Different Methods on Test Set PDB186

To evaluate the generalization ability of the prediction model proposed in this paper, we tested the model independently using dataset PDB186. **Table 2** compares the performance of this study to other prediction methods on the dataset PDB186.

From **Table 2**, we can see that on the independent test set PDB186, the ACC, SN, SP of KK-DBP reach 81.2, 97.8 and 64.5%, respectively. In terms of prediction accuracy, KK-DBP is higher than other existing methods. Compared with the current method with the highest accuracy Local-DPP, KK-DBP was improved by 2.2 and 5.3% on the ACC and SN, respectively. SP is slightly lower than Local-DPP and IDNA-Prot. The results of independent verification experiments confirm that KK-DBP has reliable predictive performance and can recognize DBP from a large number of unknown proteins more accurately than existing DBP recognition methods.

DISCUSSION AND CONCLUSION

A large number of studies have shown that the classification of DNA-binding proteins has important theoretical and practical significance for future genomics and proteomics research. This paper proposes a DNA-binding protein prediction method, called KK-DBP, that is based on multi-feature fusion and improves the feature extraction method in DNA-binding protein prediction. This method uses PSSM features that contain dipeptide composition information for multi-feature fusion to construct the initial feature set, and it obtains the optimal feature subset for modeling by the maximum correlation maximum distance method. Finally, PDB186 was used as an independent test to further evaluate the effectiveness of our method. On the independent test set, the prediction accuracy, sensitivity and specificity of the model reached 81.2, 97.8 and 64.5%, respectively. KK-DBP surpasses existing methods in prediction accuracy, confirming that our method can identify DBP more accurately than existing methods.

Although our method improves the prediction accuracy of DNA-binding proteins, we still do not know how to construct a better feature extraction algorithm based on sequence and structure information. Therefore, our future research direction will be towards finding more distinguishable feature extraction algorithms (Ding et al., 2016; Zeng et al., 2020a; Yang et al., 2021b;

Wang et al., 2021d; Jin et al., 2021) and more suitable classifiers (Ding et al., 2019; Ding et al., 2020a; Ding et al., 2020b; Yang et al., 2021c; Guo et al., 2021) and prediction models (Liu et al., 2020b; Zeng et al., 2020b; Chen et al., 2021; Xu et al., 2021c; Song et al., 2021; Xiong et al., 2021) to better recognise DNA-binding proteins.

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

YJ conceived the algorithm, performed the experiments, analyzed the data, and drafted the manuscript. TZ designed the experiments and revised the manuscript. YJ, SH, and TZ provided suggestions for the study design and the writing of the manuscript. All authors approved the final manuscript.

FUNDING

This work was supported by the Fundamental Research Funds for the Central Universities (2572021BH01) and the National Natural Science Foundation of China (62172087, 62172129).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.811158/full#supplementary-material

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

Ao, C., Zhou, W., Gao, L., Dong, B., and Yu, L. (2020). *Prediction of Antioxidant Proteins Using Hybrid Feature Representation Method and Random forest*. San Diego, CA: Genomics.

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

Ao, C., Yu, L., and Zou, Q. (2021). Prediction of Bio-Sequence Modifications and the Associations with Diseases. *Brief. Funct. genomics* 20 (1), 1–18. doi:10.1093/bfgp/elaa023

Chen, Y., Ma, T., Yang, X., Wang, J., Song, B., and Zeng, X. (2021). MUFFIN: Multi-Scale Feature Fusion for Drug-Drug Interaction Prediction. *Bioinformatics* 37, 2651–2658. doi:10.1093/bioinformatics/btab169

Ding, S., Li, Y., Shi, Z., and Yan, S. (2014). A Protein Structural Classes Prediction Method Based on Predicted Secondary Structure and PSI-BLAST Profile. *Biochimie* 97, 60–65. doi:10.1016/j.biochi.2013.09.013

Ding, Y., Tang, J., and Guo, F. (2019). Identification of Drug-Side Effect Association via Multiple Information Integration with Centered Kernel Alignment. *Neurocomputing* 325 (24), 211–224. doi:10.1016/j.neucom.2018.10.028

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. Applic* 32, 10303–10319. doi:10.1007/s00521-019-04569-z

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

Fu, X., Zhu, W., Liao, B., Cai, L., Peng, L., and Yang, J. (2018). *Improved DNA-Binding Protein Identification by Incorporating Evolutionary Information into the Chou's PseAAC*. IEEE Access, 1.

Gao, M., Skolnick, J., and Dbd-Hunter (2008). DBD-Hunter: a Knowledge-Based Method for the Prediction of DNA-Protein Interactions. *Nucleic Acids Res.* 36 (12), 3978–3992. doi:10.1093/nar/gkn332

Geete, K., and Pandey, M. (2020). Robust Transcription Factor Binding Site Prediction Using Deep Neural Networks. *Curr. Bioinformatics* 15 (10), 1137–1152.

Guo, X., Zhou, W., Shi, B., Wang, X., Du, A., Ding, Y., et al. (2021). An Efficient Multiple Kernel Support Vector Regression Model for Assessing Dry Weight of Hemodialysis Patients. *Cbio* 16 (2), 284–293. doi:10.2174/1574893615999200614172536

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

He, S., Guo, F., Zou, Q., and Ding, H. (2020). MRMD2.0: A Python Tool for Machine Learning with Feature Ranking and Reduction. *Curr. Bioinformatics* 15 (10), 1213–1221.

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, Q., Meng, Z., Pham, T. D., Chen, Q., Wei, L., Su, R., et al. (2019). DUNet: A Deformable Network for Retinal Vessel Segmentation. *Knowledge-Based Syst.* 178, 149–162. doi:10.1016/j.knosys.2019.04.025

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, D. E., Chivian, D., and Baker, D. (2004). Protein Structure Prediction and Analysis Using the Robetta Server. *Nucleic Acids Res.* 32, W526–W531. Web Server issue. doi:10.1093/nar/gkh468

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, 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* 22 (3), bbaa159. doi:10.1093/bib/bbaa159

Li, J., Pu, Y., Tang, J., Zou, Q., Guo, F., and DeepAVP (2020). DeepAVP: A Dual-Channel Deep Neural Network for Identifying Variable-Length Antiviral Peptides. *IEEE J. Biomed. Health Inform.* 24 (10), 3012–3019. doi:10.1109/jbhi.2020.2977091

Li, T., Fan, K., Wang, J., and Wang, W. (2003). Reduction of Protein Sequence Complexity by Residue Grouping. *Protein Eng. Des. Selection* 16 (5), 323–330. doi:10.1093/protein/gzg044

Li, T., and Li, Q.-Z. (2012). Annotating the Protein-RNA Interaction Sites in Proteins Using Evolutionary Information and Protein Backbone Structure. *J. Theor. Biol.* 312, 55–64. doi:10.1016/j.jtbi.2012.07.020

Liang, Z. Y., Lai, H. Y., Yang, H., Zhang, C. J., Yang, H., Wei, H. H., et al. (2017). Pro54DB: a Database for Experimentally Verified Sigma-54 Promoters. *Bioinformatics* 33 (3), 467–469. doi:10.1093/bioinformatics/btw630

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., 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., 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., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., et al. (2014). iDNA-Prot|dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition. *PloS one* 9 (9), e106691. doi:10.1371/journal.pone.0106691

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Mol. Ther. - Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

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

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-Of-Origin. *Front. Cel Dev. Biol.* 9, 619330. doi:10.3389/fcell.2021.619330

Liu, J., Lian, X., Liu, F., Yan, X., Cheng, C., Cheng, L., et al. (2020). Identification of Novel Key Targets and Candidate Drugs in Oral Squamous Cell Carcinoma. *Cbio* 15 (4), 328–337. doi:10.2174/1574893614666191127101836

Liu, T., Zheng, X., and Wang, J. (2010). Prediction of Protein Structural Class for Low-Similarity Sequences Using Support Vector Machine and PSI-BLAST Profile. *Biochimie* 92 (10), 1330–1334. doi:10.1016/j.biochi.2010.06.013

Long, J., Yang, H., Yang, Z., Jia, Q., Liu, L., Kong, L., et al. (2021). Integrated Biomarker Profiling of the Metabolome Associated with Impaired Fasting Glucose and Type 2 Diabetes Mellitus in Large-Scale Chinese Patients. *Clin. Transl Med.* 11 (6), e432. doi:10.1002/ctm2.432

Lu, W., Song, Z., Ding, Y., Wu, H., Cao, Y., Zhang, Y., et al. (2020). Use Chou's 5-Step Rule to Predict DNA-Binding Proteins with Evolutionary Information. *Biomed. Res. Int.* 2020, 6984045. doi:10.1155/2020/6984045

Lv, H., Dao, F. Y., Zulfiqar, H., and Lin, H. (2021). DeepIPs: Comprehensive Assessment and Computational Identification of Phosphorylation Sites of SARS-CoV-2 Infection Using a Deep Learning-Based Approach. *Brief. Bioinformatics* 22 (6). bbab244. doi:10.1093/bib/bbab244

Lv, H., Dao, F. Y., Zulfiqar, H., Su, W., Ding, H., Liu, L., et al. (2021). A Sequence-Based Deep Learning Approach to Predict CTCF-Mediated Chromatin Loop. *Brief. Bioinformatics* 22 (5), bbab031. doi:10.1093/bib/bbab031

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

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

Meng, F., and Kurgan, L. (2016). DFLpred: High-Throughput Prediction of Disordered Flexible Linker Regions in Protein Sequences. *Bioinformatics* 32 (12), i341–i350. doi:10.1093/bioinformatics/btw280

Qu, K., Wei, L., and Zou, Q. (2019). A Review of DNA-Binding Proteins Prediction Methods. *Cbio* 14 (3), 246–254. doi:10.2174/1574893614666181212102030

Shen, Z., and Zou, Q. (2020). Basic Polar and Hydrophobic Properties Are the Main Characteristics that Affect the Binding of Transcription Factors to Methylation Sites. *Bioinformatics* 36 (15), 4263–4268. doi:10.1093/bioinformatics/btaa492

Song, B., Huang, S., and Zeng, X. (2021). The Computational Power of Monodirectional Tissue P Systems with Symport Rules. *Inf. Comput.*, 104751. doi:10.1016/j.ic.2021.104751

Stawiski, E. W., Gregoret, L. M., and Mandel-Gutfreund, Y. (2003). Annotating Nucleic Acid-Binding Function Based on Protein Structure. *J. Mol. Biol.* 326 (4), 1065–1079. doi:10.1016/s0022-2836(03)00031-7

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A Deep forest Model to Predict Anti-cancer Drug Response. *Methods* 166, 91–102. doi:10.1016/j.ymeth.2019.02.009

Su, W., Liu, M.-L., Yang, Y.-H., Wang, J.-S., Li, S.-H., Lv, H., et al. (2021). PPD: A Manually Curated Database for Experimentally Verified Prokaryotic Promoters. *J. Mol. Biol.* 433 (11), 166860. doi:10.1016/j.jmb.2021.166860

Tang, X., Cai, L., Meng, Y., Gu, C., Yang, J., and Yang, J. (2021). A Novel Hybrid Feature Selection and Ensemble Learning Framework for Unbalanced Cancer Data Diagnosis with Transcriptome and Functional Proteomic. *IEEE Access* 9, 51659–51668. doi:10.1109/access.2021.3070428

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, D., Zhang, Z., Jiang, Y., Mao, Z., Wang, D., Lin, H., et al. (2021). DM3Loc: Multi-Label mRNA Subcellular Localization Prediction and Analysis Based on Multi-Head Self-Attention Mechanism. *Nucleic Acids Res.* 49 (8), e46. doi:10.1093/nar/gkab016

Wang, H., Ding, Y., 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 (1), 56. doi:10.1186/s12864-020-07347-7

Wang, X., Yang, Y., Liu, J., and Wang, G. (2021). The Stacking Strategy-Based Hybrid Framework for Identifying Non-coding RNAs. *Brief Bioinform* 22 (5), bbab023. doi:10.1093/bib/bbab023

Wang, Z., Liu, D., Xu, B., Tian, R., and Zuo, Y. (2021). Modular Arrangements of Sequence Motifs Determine the Functional Diversity of KDM Proteins. *Brief Bioinform* 22 (3). doi:10.1093/bib/bbaa215

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., 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.

Xiong, G., Wu, Z., Yi, J., Fu, L., Yang, Z., Hsieh, C., et al. (2021). ADMETlab 2.0: an Integrated Online Platform for Accurate and Comprehensive Predictions of ADMET Properties. *Nucleic Acids Res.* 49 (W1), W5–W14. doi:10.1093/nar/gkab255

Xu, B., Liu, D., Wang, Z., Tian, R., and Zuo, Y. (2021). Multi-substrate Selectivity Based on Key Loops and Non-homologous Domains: New Insight into ALKBH Family. *Cell. Mol. Life Sci.* 78 (1), 129–141. doi:10.1007/s00018-020-03594-9

Xu, H., Zeng, W., Zeng, X., and Yen, G. G. (2021). A Polar-Metric-Based Evolutionary Algorithm. *IEEE Trans. Cybern.* 51, 3429–3440. doi:10.1109/TCYB.2020.2965230

Xu, L., Jiang, S., Wu, J., and Zou, Q. (2021). An In Silico Approach to Identification, Categorization and Prediction of Nucleic Acid Binding Proteins. *Brief Bioinform* 22 (3), bbaa171. doi:10.1093/bib/bbaa171

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, 11387–11399. doi:10.1007/s00521-020-05573-4

Yang, H., Ding, Y., Tang, J., and Guo, F. (2021). Drug-disease Associations Prediction via Multiple Kernel-Based Dual Graph Regularized Least Squares. *Appl. Soft Comput.* 112, 107811. doi:10.1016/j.asoc.2021.107811

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

Yu, L., Xu, F., and Gao, L. (2020). Predict New Therapeutic Drugs for Hepatocellular Carcinoma Based on Gene Mutation and Expression. *Front. Bioeng. Biotechnol.* 8, 8. doi:10.3389/fbioe.2020.00008

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., Zhu, S., Hou, Y., Zhang, P., Li, L., Li, J., et al. (2020). Network-based Prediction of Drug-Target Interactions Using an Arbitrary-Order Proximity Embedded Deep forest. *Bioinformatics* 36 (9), 2805–2812. doi:10.1093/bioinformatics/btaa010

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, 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, D., Xu, Z. C., Su, W., Yang, Y. H., Lv, H., Yang, H., et al. (2020). iCarPS: a Computational Tool for Identifying Protein Carbonylation Sites by Novel Encoded Features. *Bioinformatics* 37 (2), 171–177. doi:10.1093/bioinformatics/btaa702

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*, 1. doi:10.1109/TCBB.2020.2968419

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

Zhu, Y., Li, F., Xiang, D., Akutsu, T., Song, J., and Jia, C. (2021). Computational Identification of Eukaryotic Promoters Based on Cascaded Deep Capsule Neural Networks. *Brief Bioinform* 22 (4). doi:10.1093/bib/bbaa299

Zou, L., Nan, C., and Hu, F. (2013). Accurate Prediction of Bacterial Type IV Secreted Effectors Using Amino Acid Composition and PSSM Profiles. *Bioinformatics* 29 (24), 3135–3142. doi:10.1093/bioinformatics/btt554

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A Novel Features Ranking Metric with Application to Scalable Visual and Bioinformatics Data Classification. *Neurocomputing* 173, 346–354. doi:10.1016/j.neucom.2014.12.123

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

Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Sun, Z.-J., Dao, F.-Y., Yu, X.-L., et al. (2021). Identification of Cyclin Protein Using Gradient Boost Decision Tree Algorithm. *Comput. Struct. Biotechnol. J.* 19, 4123–4131. doi:10.1016/j.csbj.2021.07.013

Zuo, Y.-C., Peng, Y., Liu, L., Chen, W., Yang, L., and Fan, G.-L. (2014). Predicting Peroxidase Subcellular Location by Hybridizing Different Descriptors of Chou' Pseudo Amino Acid Patterns. *Anal. Biochem.* 458, 14–19. doi:10.1016/j.ab.2014.04.032

Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a Flexible Web Server for Generating Pseudo K-Tuple Reduced Amino Acids Composition. *Bioinformatics* 33 (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.

Copyright © 2021 Jia, Huang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 08 December 2021
doi: 10.3389/fgene.2021.813873

Bioinformatics Analysis of Long Non-coding RNA and Related Diseases: An Overview

Yuxin Gong[1,2,3,4], *Wen Zhu*[1], *Meili Sun*[5]* *and Lei Shi*[6]*

[1]School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [2]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [3]Key Laboratory of Computational Science and Application of Hainan Province, Haikou, China, [4]Key Laboratory of Data Science and Smart Education, Hainan Normal University, Ministry of Education, Haikou, China, [5]Beidahuang Industry Group General Hospital, Harbin, China, [6]Department of Spine Surgery, Changzheng Hospital, Naval Medical University, Shanghai, China

Long non-coding RNAs (lncRNAs) are usually located in the nucleus and cytoplasm of cells. The transcripts of lncRNAs are >200 nucleotides in length and do not encode proteins. Compared with small RNAs, lncRNAs have longer sequences, more complex spatial structures, and more diverse and complex mechanisms involved in the regulation of gene expression. LncRNAs are widely involved in the biological processes of cells, and in the occurrence and development of many human diseases. Many studies have shown that lncRNAs can induce the occurrence of diseases, and some lncRNAs undergo specific changes in tumor cells. Research into the roles of lncRNAs has covered the diagnosis of, for example, cardiovascular, cerebrovascular, and central nervous system diseases. The bioinformatics of lncRNAs has gradually become a research hotspot and has led to the discovery of a large number of lncRNAs and associated biological functions, and lncRNA databases and recognition models have been developed. In this review, the research progress of lncRNAs is discussed, and lncRNA-related databases and the mechanisms and modes of action of lncRNAs are described. In addition, disease-related lncRNA methods and the relationships between lncRNAs and human lung adenocarcinoma, rectal cancer, colon cancer, heart disease, and diabetes are discussed. Finally, the significance and existing problems of lncRNA research are considered.

Keywords: lncRNA, database, mechanism of action, recognition methods, disease

OPEN ACCESS

Edited by:
Dariusz Mrozek,
Silesian University of Technology, Poland

Reviewed by:
Wang-Ren Qiu,
Jingdezhen Ceramic Institute, China
Yongchun Zuo,
Inner Mongolia University, China

***Correspondence:**
Meili Sun
1700271639@qq.com
Lei Shi
slspine@163.com

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

Received: *12 November 2021*
Accepted: *26 November 2021*
Published: *08 December 2021*

Citation:
Gong Y, Zhu W, Sun M and Shi L (2021) Bioinformatics Analysis of Long Non-coding RNA and Related Diseases: An Overview. Front. Genet. 12:813873. doi: 10.3389/fgene.2021.813873

INTRODUCTION

A long transcription product was discovered and identified by Okazaki in 2002 (Okazaki et al., 2003) when sequencing a mouse cDNA library. The transcript was called long non-coding RNA (lncRNA). LncRNAs are >200 nucleotides long and similar in structure to messenger RNAs, but they lack an open reading frame. Mainly distributed in the nucleus and cytoplasm of cells, lncRNAs are the transcription products of RNA polymerase II. LncRNAs have been classified based on the lengths of the coded transcripts as lncRNA, long-intergenic non-coding RNA, very long-intergenic non-coding RNA, macroRNA, and promoter-associated long RNA. They have also been classified according to the position of the lncRNA in the genome relative to the target protein-coding gene as 1) antisense lncRNA, which is partially or completely complementary to the transcription product on the opposite strand; 2) enhancer lncRNA, which is produced from the enhancer region of a protein coding gene; 3) bidirectional lncRNA, which shares the same promoter with protein-coding genes,

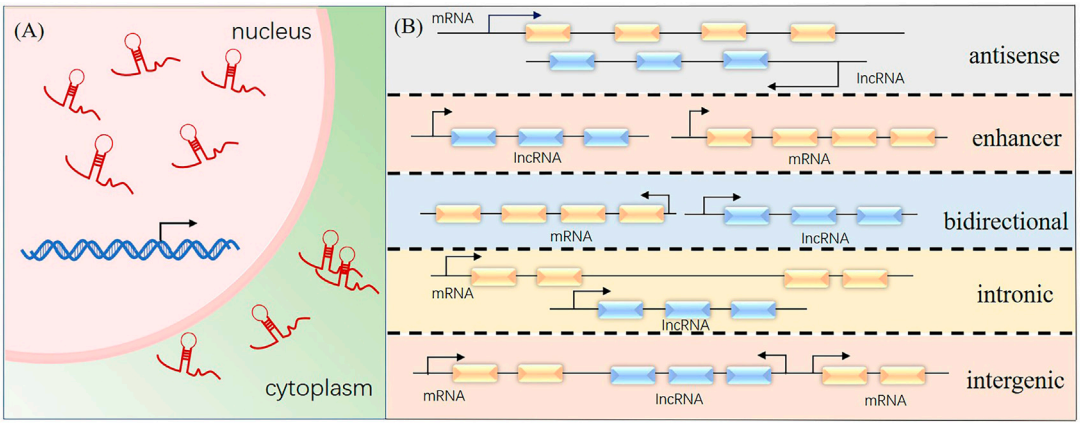


FIGURE 1 | Cellular localization and classification of long non-coding RNAs (lncRNA). **(A)** Cellular distribution of lncRNAs. **(B)** Positions of lncRNAs in the genome relative to the target protein-coding gene.

but the transcription direction is opposite; 4) intronic lncRNA, which is produced by introns of genes; and 5) large intergenic non-coding RNA, which is independently transcribed from sequences located between protein-coding genes (**Figure 1**).

Because non-coding RNAs (ncRNAs) do not encode proteins, they were thought to have no biological function and were regarded as "transcriptional noise" for a long time after they were discovered. Although, research on short ncRNAs, including microRNAs, short interfering RNAs, small nucleolar RNAs, and Piwi-interacting RNAs, has flourished, lncRNAs have been neglected because of their long sequences and limitations of the research methods (Liu, 2021).

Initially, lncRNAs were considered to be the noise of genome transcription and a by-product of polymerase II transcription. However, after lncRNAs were found to have conserved secondary structures, spliced forms, and subcellular localizations, it was realized that they may be functional. Indeed, it is now recognized that lncRNAs have functions that are essential for many biological processes, including epigenetic regulation, cell signal transduction, immune response, and cell proliferation and differentiation (Heo et al., 2013; Zhao et al., 2021; Yang et al., 2020; Hu et al., 2021; Hu et al., 20216222), and the abnormal expression of lncRNAs can result in a variety of complex diseases. Moreover, some lncRNAs can act as precursors of some functional short ncRNAs to indirectly participate in the regulation of target genes. LncRNAs regulate gene expression to exert these functions, including *cis*-regulation and *trans*-regulation.

According to data from the Encyclopedia of DNA Elements (ENCODE) project (2015), approximately 15,941 lncRNA loci have been identified in the human genome (Jalali et al., 2016). The discovery and research of a large number of lncRNAs have greatly promoted the development of RNA biology research. Analysis of these lncRNAs showed that they had important regulatory functions at the epigenetic, transcription, and post-transcriptional levels (Mercer et al., 2009; Hu et al., 2020; Shen et al., 2021). Until now, research on the relationship between lncRNA and disease has been focused mainly on tumors; however, a small number of differentially expressed lncRNAs associated with obesity, diabetes, hypertension, and other diseases have been found, but the mechanism of action is still unclear. LncRNA research suffers from problems such as few available resources, relatively independent research results, and lack of systematicness (Losko et al., 2016). In-depth analysis of disease-related lncRNAs will help in finding disease biomarkers and provide insights into the diagnosis, treatment, prognosis, and prevention of diseases (Tang et al., 2018; Sun et al., 2021). In this review, bioinformatics approaches for the study of lncRNAs and diseases are summarized, including mainly databases related to lncRNAs, the mechanisms and modes of lncRNA action, methods for identifying disease-related lncRNAs, and the relationships between lncRNAs and human metabolic syndrome, cancer, leukemia, heart disease, and neurodegenerative diseases.

DATABASES RELATED TO LNCRNAS

A large number of biological datasets related to lncRNA have been generated and many lncRNA-oriented databases have been built to store, manage, and integrate comprehensive lncRNA functional information, lncRNA structure and genome mutations, lncRNA expression analysis data, and lncRNA–disease associations. Here, the commonly used lncRNA databases are briefly described.

1) ChIPBase (Zhou et al., 2016) provides comprehensive identification and annotation of lncRNA expression profiles and transcriptional regulation data. The lncRNA expression profiles obtained by RNA sequencing and transcription factor binding sites identified by ChIP-Seq (chromatin immunoprecipitation followed by sequencing) are included in this database (http://rna.sysu.edu.cn/chipbase3/index.php).
2) lncRNAdb (Quek et al., 2015) contains comprehensive annotations of lncRNAs with biological functions, including gene expression, functional evidence, disease-

related lncRNAs, pathogen-induced or derived lncRNAs, and sequence information (http://www.lncrnadb.org/).

3) NRED (Dinger et al., 2009) provides the expression information of thousands of lncRNAs of humans and mice from microarray and *in situ* hybridization data, as well as auxiliary information such as secondary structure evidence, antisense relationships, evolutionary conservation, and genome-related text links (http://jsm-research.imb.uq.edu.au/nred/).
4) NONCODE (Fang et al., 2018) provides comprehensive lncRNA annotations, including expression information and functions predicted by the ncFANs software. NONCODE is widely used for ncRNA research (http://www.noncode.org).
5) LncRNADisease (Bao et al., 2019) contains annotations of disease-related lncRNAs reported in the literature. For the lncRNAs, basic information such as name, chromosome location, species, transcript number, and sequence is provided. For the related diseases, name, literature, and other information are given (http://www.rnanut.net/lncrnadisease/).
6) lncRNASNP2 (Miao et al., 2017) provides resources of single nucleotide polymorphisms (SNPs) in lncRNAs of humans and mice. The database has the following browse and search functions: the influence of SNPs in lncRNAs on their own genes, the binding of microRNAs (miRNAs) in lncRNAs and the influence of SNPs on binding, the mutation and expression of lncRNAs in The Cancer Genome Atlas, variations in lncRNAs in the COSMIC database (large databases of cancer-related somatic mutation sites), and the effect of mutations on lncRNAs (http://bioinfo.life.hust.edu.cn/lncRNASNP#!/).
7) StarBase v2.0 (Li et al., 2013) can search for lncRNA based on miRNA-lncRNA interaction, as well as competitive endogenous RNA regulatory molecules and binding proteins related to the specified lncRNAs (http://starbase.sysu.edu.cn/index.php).

MECHANISMS AND MODES OF ACTION OF LNCRNAS HEADINGS

Compared with small RNAs, lncRNAs are longer, have more complex spatial structure, and more diverse and complex mechanisms for expression regulation. LncRNAs are involved in the regulation of development, differentiation, and metabolism. LncRNAs can regulate gene expression at the epigenetic (Mercer and Mattick, 2013), transcriptional (Bonasio and Shiekhattar, 2014), and post-transcriptional (Yoon et al., 2013) levels. They also participate in important regulatory processes such as X chromosome silencing, genome imprinting, chromatin modification, transcription activation and inhibition, and nuclear transport. LncRNAs are also closely related with the occurrence, development, and prevention of human diseases. The mechanisms of lncRNA action discovered so far include X chromosome inactivation and genome imprinting, chromatin modification, cell cycle regulation and apoptosis, mRNA decay, and protein translation regulation (Zou et al., 2019).

The modes of lncRNA action are signal, decoy, guide, and scaffold (Wang and Chang, 2011) as shown in **Figure 2**. The molecular functions of lncRNAs can be explained based on the mode of action. In the signal mode, lncRNAs participate in gene imprinting processes. For example, the lncRNAs Kcnq1ot1 and Xist both function in the signal mode. The mode of action of Kcnq1ot1 and Xist are similar. Kcnq1ot1 binds to the chromosome and inhibits the expression of Kcnq1 by recruiting H3K9- and H3K27-specific histone methyltransferases and polycomb repressive complex 2 (PRC2) complexes. This function is hereditary (Pandey et al., 2008). In cells with two X chromosomes, one of the X chromosomes is suppressed. Xist regulates the process by which the X chromosome is selectively suppressed and maintains this phenotype to the next generation (Plath et al., 2002).

In the decoy mode, lncRNAs bind to proteins that have transcriptional regulatory functions (e.g., transcription factors and chromosome folding proteins), thereby regulating the transcriptional activation and inhibition of related genes by controlling the activity of molecules and signal pathways. LncRNA Gas5 functions in the decoy mode. It binds to the DNA-binding domain of the glucocorticoid receptor through the RNA motif to inhibit the physiological function of the receptor (Kino et al., 2010). LncRNA can also be used as a molecular decoy for miRNAs and splicing factors, inhibiting their functions. MiRNAs can promote the formation of protein complexes, hence playing an important role in gene regulation. LncRNA PTENP1 can inhibit human tumors. In the decoy mode of action, PTENP1 binds to a group of miRNAs that act on the PTEN 3′ untranslated region, thereby regulating the expression of PTEN (Poliseno et al., 2010).

In the guide mode, lncRNAs bind to proteins to guide the protein complex to a specific DNA sequence, thereby regulating the transcription of downstream molecules. The guiding can be either cis or trans. One *cis*-regulatory mechanism of lncRNA involves the inactive center of the X chromosome, which controls the silence of the maternal X chromosome. A *trans*-regulatory mechanism included reducing the expression of tumor-associated lncRNA HOTAIR and decreasing cell invasiveness.

In the scaffold mode, lncRNAs simultaneously bind multiple related transcription factors to provide a platform for interaction. For example, lncRNA HOTAIR splices and bridges between the PRC2 and LSD1 complexes. The 5′ end of HOTAIR combines with the PRC2 complex (acting on H3 and H27 to methylate them) to promote gene expression (Rinn et al., 2007), whereas the 3′ end of HOTAIR combines with LSD1 (acting on H3K4 to demethylate it) to antagonize gene expression activation (Tsai et al., 2010), thereby inducing the interaction between the PRC2 and LSD1 complexes.

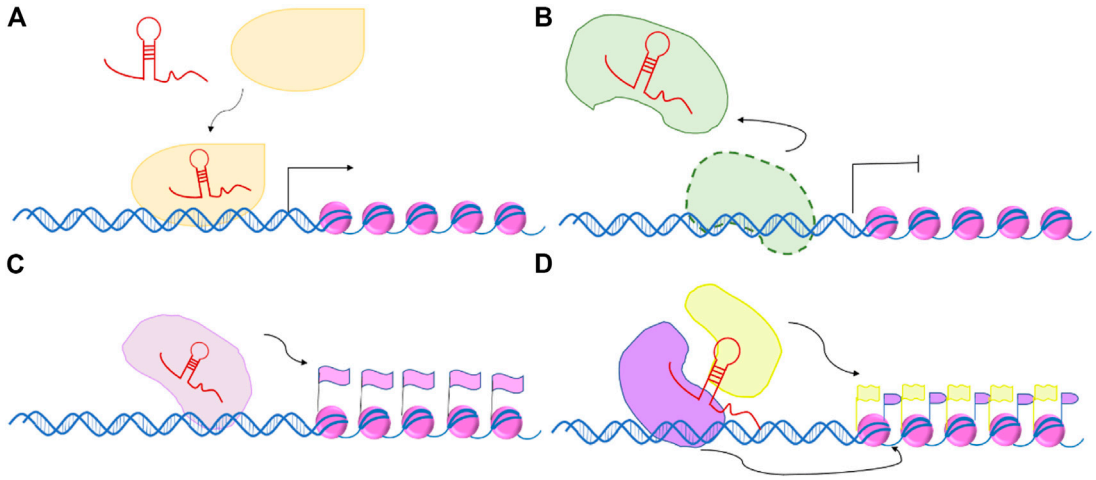


FIGURE 2 | Modes of action long non-coding RNAs. **(A)** Signal, **(B)** Decoy, **(C)** Guide, and **(D)** Scaffold.

METHODS TO IDENTIFY DISEASE-RELATED LNCRNAS

LncRNAs by far outnumber protein-coding genes, and, unlike protein-coding genes, lncRNAs are usually not conserved in sequence fragments and secondary structures, which make lncRNA functional prediction difficult. Two main approaches have been used to identify disease-related lncRNAs, methods based on biological experiments and methods based on computational predictions. The former generally produces more reliable results, but the cost is high and the efficiency is low; the latter uses multi-source biological data, such as disease-related genomes, transcriptomes, and proteomes, to predict disease-related lncRNAs. Bioinformatics methods have been applied to predict disease-related lncRNAs based on the analysis of multi-source biological data (Cao et al., 2021).

Identification of Disease-Related lncRNA by Biological Experiments

Large-scale biological experimental research is often limited by ethical factors that govern the collection of experimental samples. However, the influence of interference factors on the experimental results can often be largely controlled, and therefore the results are likely to be objective and highly reliable. Two examples of lncRNA research using biological experiments are briefly described.

Nakagawa et al. (Nakagawa et al., 2014) found that approximately 50% of lncRNA Neat1 knockout mice with abnormal ovulation did not become pregnant, and this outcome seemed to be random. Subsequently, corpus luteum dysfunction and the accompanying low progesterone were found to contribute most to the decline in fertility. Unlike the weak expression of Neat1 observed in most adult tissues, in the infertile Neat1 knockout mice, Neat1 was highly expressed in the corpus luteum and the formation of the corpus luteum was severely impaired. These results indicated that Neat1 may be closely related to the formation of the corpus luteum and some forms of infertility in humans.

Zhang et al. (Zhang et al., 2018) screened and identified a new type of lncRNA, HOXC-AS3, using publicly available gastric cancer expression profile data and integrated bioinformatics analysis. They found that the expression of HOXC-AS3 was highly up-regulated in gastric cancer tissues and was related to clinicopathological factors such as histological grade, depth of tumor invasion, lymph node metastasis, and poor prognosis. They performed chromatin immunoprecipitation assays to explore the mechanisms involved in the high HOXC-AS3 expression, and found that HOXC-AS3 was partially activated by H3K4me3 and H3K27ac in cells and tissues. Overexpression and knockout of HOXC-AS3 were used to detect cell apoptosis and proliferation. They found that overexpression of HOXC-AS3 promoted the proliferation of cancer cells, and knockout of HOXC-AS3 induced apoptosis of cancer cells. To further explore the mechanism of action of HOXC-AS3, the transcription factor YBX1 was selected, and three independent RNA pull-down mass spectrometry analyses were performed. The results showed that HOXC-AS3 interacted with YBX1. This result combined with the results of the immunoprecipitation assays, confirmed that YBX1 was involved in HOXC-AS3-mediated gene transcription regulation in gastric cancer.

Clearly, traditional biological experiments are not only time-consuming but also expensive. Computational models are less time-consuming and less expensive, therefore they have attracted more and more attention as a solution that can predict lncRNA functions on a large scale. Models can be used to predict the

possible functions of lncRNAs according to related priorities, and the predictions can be verified experimentally. This process effectively promotes the functional recognition of lncRNAs.

Identification of Disease-Related LncRNAs by Computational Prediction

Because biological experiments are costly and time-consuming, the use of bioinformatics calculations to predict disease-related lncRNAs has become the mainstream. In recent years, many lncRNA-disease association prediction (LDAP) models have been proposed, including models based on biological networks, models that do not rely on known lncRNA–disease associations, and models based on machine learning algorithms. The three types of LDAP models are briefly described.

LDAP Models Based on Biological Networks

These models integrate biological networks such as the disease similarity network, lncRNA similarity network, and lncRNA–disease association network to construct an LDAP model. In 2014, Sun et al. (Sun et al., 2014) proposed an LDAP model based on the functionally similar network of lncRNAs and random walks with restart. In 2015, Chen et al. (Weng et al., 2015) developed the KATZLDA prediction model by fusing lncRNA–disease association, lncRNA expression profile, lncRNA functional similarity, and disease semantic similarity data. In 2017, Yu et al. (Yu et al., 2017) developed the BRWLDA prediction model based on double random walks, and Gu et al. (Gu et al., 2017) developed a GRWLDA prediction model based on global network random walk. In 2018, Ping et al. (Ping et al., 2019) proposed a prediction model based on the known lncRNA–disease association network to infer potential lncRNA–disease associations. In 2019, Fan et al. (Fan et al., 2019) proposed an LDAP model based on multiple heterogeneous information networks and random walks with restart. Xie et al. (Xie et al., 2019) proposed an SFK-LDA prediction model based on similarity nuclear fusion. In 2020, Zhou et al. (Zhou et al., 2021) built a heterogeneous network by integrating various associations between diseases and miRNAs, lncRNAs, proteins, and drugs, and trained a LDAP model with the rotating forest classifier, and Zhang et al. (Zhang et al., 2020) proposed an LDAP model based on network feature similarity and gradient boosting. In 2021, Liu et al. (Liu et al., 2021) proposed an LDAP model based on the weighted graph regularized collaborative matrix factorization.

LDAP Models That do not Rely on Known LncRNA–Disease Associations

In these models, the expression and regulatory relationship between disease-related genes or miRNAs and lncRNAs are used to predict potential lncRNA–disease associations. In 2014, Liu et al. (Zhao et al., 2015) developed the first LDAP model that did not rely on known lncRNA–disease associations by integrating lncRNA expression profiles, gene expression profiles, and disease-related gene data. In 2015, Chen et al. (Chen, 2015) developed an LDAP model HGLDA based on hypergeometric distribution by integrating miRNA–disease associations and ncRNA–miRNA interactions. In 2016, Cheng et al. (Cheng et al., 2016) proposed the IntNetLncSim computing framework, which inferred the functional similarity of lncRNAs and predicted new lncRNA–disease associations by modeling the information flow in an integrated network that contained lncRNA transcription and post-transcription information. In 2017, Wang et al. (Wang et al., 2017) mapped lncRNAs to their functional genomic context based on the theory of competing endogenous RNAs to predict new lncRNA–disease associations. Fu et al. (Fu et al., 2018) proposed a matrix decomposition-based LDAP model MFLDA, which decomposed the data matrix of heterogeneous data sources into low-rank matrices through matrix decomposition to explore and use their internal and shared structure. In 2018, Ding et al. (Ding et al., 2018) proposed an LDAP model based on a lncRNA–disease–gene network that integrated gene–disease and lncRNA–disease associations. In 2020, Xiao et al. (Xiao et al., 2020) proposed an LDAP model that used both direct and indirect features of lncRNA–disease relationship pairs, and Tang et al. proposed a hierarchical extended LDAP model based on a Boolean matrix (Tang et al., 2020).

LDAP Models Based on Machine Learning Algorithms

These models integrate biological data and use various machine learning algorithms to predict disease-related lncRNAs. In 2013, Chen et al. (Chen and Yan, 2013) developed a semi-supervised learning framework LRLSLDA based on Laplace regularization least squares by integrating lncRNA expression profiles and known lncRNA–disease associations. In 2015, Liu et al. (Liu et al., 2015) developed an LDAP model based on the naive Bayes classifier to identify lncRNAs related to cancer by integrating genome, regulatory factors, and transcriptome data. In 2017, Lan et al. (Lan et al., 2017) proposed an LDAP model based on support vector machines. In 2018, Yu et al. (Yu et al., 2018) proposed the NBCLDA model based on the naive Bayes classifier. In 2019, Guo et al. (Guo et al., 2019) proposed two LDAP models, one based on rotating forest and neural network and another based on a random forest classifier. Sheng et al. (Sheng et al., 2021) proposed a series of LDAP models based on convolutional neural networks, including CNNLDA, as well as an attention multi-level representation coding model based on convolution and variance autoencoders. In 2020, Zeng et al. (Zeng et al., 2020) proposed SDLDA, an LDAP model based on singular value decomposition and deep learning. Fan et al. (Fan et al., 2020) proposed IDSSIM, a calculation model of lncRNA functional similarity based on improved disease semantic similarity. Tan et al. (Tan et al., 2020) proposed a multi-view consensus graph learning model to predict lncRNA–disease association. Wei et al. (Wei et al., 2021) proposed a convolutional neural network model fused with multiple biological characteristics to predict lncRNA–disease association.

LNCRNAS AND RELATED DISEASES

LncRNAs induce the occurrence of disease by regulating disease-related protein coding genes, thus leading to improper expression of lncRNAs or altering the chromatin that contain disease-related gene polymorphisms and non-coding regions. Therefore, the expression of lncRNAs is important in the diagnosis, occurrence, development, and treatment of diseases. In recent years, the associations of lncRNAs with cancer, leukemia, cardiovascular and cerebrovascular diseases, diabetes, and other diseases have been a focus of study. Future clinical applications of disease-related lncRNAs are very likely.

The roles of lncRNAs in cancer: Li et al. (Li et al., 2016) found that up-regulation of lncRNA MALAT1 was related to tumor size and lymph node metastasis, and to the shorter overall survival of patients with lung adenocarcinoma. *In vivo* and *in vitro* experiments showed that MALAT1 promoted epithelial–mesenchymal transition and metastasis of lung adenocarcinoma cells. Numerous lncRNAs have been found to encode small proteins or micropeptides, some of which play roles in diseases. Meng et al. (Meng et al., 2020) found that lncRNA LOC90024 encodes a splicing regulatory small protein that induces the formation of Sp4 transcription factor splice variants, thereby promoting the occurrence and development of advanced rectal cancer tumors. Zhu et al. (Zhu et al., 2020) found that lncRNA LINC00266-1 encodes an RNA-binding regulatory peptide that, when highly expressed in patients with colon cancer, leads to a poor prognosis. The oncogenic peptide encoded by LINC00266-1 exerts its carcinogenic function by enhancing the recognition of N6-methyladenosine of RNA.

The roles of lncRNAs in leukemia: Garzon et al. (Garzon et al., 2014) developed a prognostic scoring system to determine if lncRNAs were associated with cytogenetically normal acute myeloid leukemia (CN-AML) clinical features and recurrent mutations in patients older than 60 years. First, 48 lncRNAs most relevant to prognosis were identified. Then, patients with CN-AML were divided into two groups, those with good prognostic scores and those with poor prognostic scores, based on the 48 lncRNAs. The prognostic scores were verified in an independent matched group of patients with CN-AML who received the same treatment. The comparative analysis showed that the lncRNA expression profile was closely related to the recurrent mutation and expression of AML, implying that some of 48 lncRNAs may have a functional role in the development of leukemia. These lncRNAs are good candidates as biomarkers for the prognosis of AML.

The roles of lncRNAs in heart disease: Han et al. (Han et al., 2014) developed a new lncRNA–chromatin mechanism to treat heart failure. A lncRNA transcript Mhrt779 from myosin heavy chain 7 loci was found to be specifically expressed in cardiomyocytes and to gradually increase with the development of embryos, especially after birth. Furthermore, the Brg1–Hdac–Parp chromatin inhibitory complex was activated by pathological stress and lncRNA Mhrt transcription was inhibited in the heart, thereby protecting the heart from hypertrophy and failure. These results show that there is a conserved lncRNA mechanism in human cardiomyopathy, and also establish a new paradigm for lncRNA–chromatin interactions.

The roles of lncRNAs in neurodegenerative diseases: Alzheimer's disease is a progressively developing neurodegenerative disease with insidious onset that is believed to be caused by a large amount of amyloid β-protein (Aβ) expression, which leads to pathological changes in patients. Aβ is hydrolyzed from β-amyloid precursor protein, and excessive Aβ deposition can cause degenerative diseases related to neurons. BACE1 is a key enzyme in the production of Aβ. Faghihi et al. (Faghihi and Wahlestedt, 2009) found that lncRNA BACE1-AS (antisense transcript of BACE1) increased the stability of BACE1 mRNA through a mechanism that involved the formation of RNA duplexes, and this in turn facilitated the accumulation of Aβ in patients with Alzheimer's disease.

The roles of lncRNAs in diabetes: LncRNAs in human pancreatic β-cells exhibit dynamic regulation during differentiation or when glucose concentrations change. Akerman et al. (Akerman et al., 2017) studied the functions of β-cell-specific lncRNAs and transcription factors using transcript knockdown and co-expression network analysis strategies. They found that lncRNAs and transcription factors acted synergistically to regulate the specific transcription network of β-cells. LncRNA PLUTO affected local three-dimensional chromatin structure and transcription of PDX1, which encodes a key β-cell transcription factor. PLUTO and PDX1 were both down-regulated in islets from donors with type 2 diabetes or impaired glucose tolerance. These results indicate the role of lncRNAs in β-cell gene regulation and diabetes.

CHALLENGES AND RESEARCH PROSPECTS

Compared with protein-coding sequences and small RNA molecules, lncRNA-related research is insufficient and there are many problems still to be solved. The major ones are listed here.

1) No standardized naming of lncRNAs. Until now, lncRNAs have been named according to their functions, structural characteristics, or modes of action. Therefore, it is difficult to understand their roles and functions from the name.

2) Unannotated and unbalanced lncRNA data. Compared with other ncRNA databases, the annotation information in lncRNA databases is insufficient, and disease-related lncRNAs that are included in multi-source data have problems such as serious imbalances of information.

3) Lack of lncRNA-specific technologies. Because of the diverse types and functions of lncRNAs, more effective methods are needed for systematically studying the biological functions of lncRNAs and for identifying disease-related lncRNAs. The process of combining multi-data to predict disease-related lncRNAs has problems, such as high-dimensional feature space, high

noise, and redundant feature interference, that seriously affect the accuracy of the predictions.

4) Research fields need to be expanded. Current research on lncRNAs has focused mainly on tumors, nerves, and development. However, the genetic characteristics of cancer-related lncRNAs and the mechanism of action of complex diseases are still unclear. More areas of disease research related to lncRNAs need to be developed.

Despite these problems, lncRNA research has continuously advanced the understanding of lncRNAs. lncRNAs not only exert their biological functions in a variety of mechanisms in different organisms, but their dysfunction can lead to the occurrence and development of many diseases. Undoubtedly, new technologies and new methods will be developed for use in lncRNA bioinformatics research. Such developments will help to further analyze the functions and regulation mechanisms of lncRNAs, as well as the pathological mechanisms associate with the development of diseases.

DISCUSSION

LncRNAs are closely related to cell cycle and differentiation, aging and human diseases. Therefore, the research on their functions and mechanisms is also constantly deepening. This review summarized the following key points:1) according to the four modes of action of lncRNA, the corresponding molecular functions were described respectively. 2) The identification methods of lncRNA related to diseases were summarized into two parts: lncRNA identification research based on biological experiments and computational prediction. 3) The relationship between lncRNA and various human diseases was expounded.

Although the research technology of lncRNA is constantly developing, there are still a number of limitations:1) The low abundance of lncRNAs and lack of annotation information lead to inaccurate positioning. 2) The data set of lncRNAs in the database is not perfect. 3) The types of diseases associated with lncRNA are limited. Therefore, in the future, researchers need to continue to dig out the functional information of unknown lncRNAs and develop new lncRNA recognition models, which will help to enhance the scientific understanding of more human diseases.

AUTHOR CONTRIBUTIONS

YG and WZ designed the study. WZ and LS collected and analyzed the literatures. YG and MS analyzed the literatures and made the figures. All authors participated in writing the draft and approved the final manuscript.

FUNDING

The work was supported by the National Key R&D Program of China (No.2020YFB2104400), and the Special Science Foundation of Quzhou (2021D004).

ACKNOWLEDGMENTS

We thank Margaret Biswas, PhD, from Liwen Bianji (Edanz) (www.liwenbianji.cn/), for editing the English text of a draft of this manuscript.

REFERENCES

Akerman, I., Tu, Z., Beucher, A., Rolando, D. M. Y., Sauty-Colace, C., Benazra, M., et al. (2017). Human Pancreatic β Cell lncRNAs Control Cell-specific Regulatory Networks. *Cel Metab.* 25, 400–411. doi:10.1016/j.cmet.2016.11.016

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: an Updated Database of Long Non-coding RNA-Associated Diseases. *Nucleic Acids Res.* 47, D1034–d1037. doi:10.1093/nar/gky905

Bonasio, R., and Shiekhattar, R. (2014). Regulation of Transcription by Long Noncoding RNAs. *Annu. Rev. Genet.* 48, 433–455. doi:10.1146/annurev-genet-120213-092323

Cao, C., Wang, J., Kwok, D., Cui, F., Zhang, Z., Zhao, D., et al. (2021). webTWAS: a Resource for Disease Candidate Susceptibility Genes Identified by Transcriptome-wide Association Study. *Nucleic Acids Res.*, gkab957. doi:10.1093/nar/gkab957

Chen, X. (2015). Predicting lncRNA-Disease Associations and Constructing lncRNA Functional Similarity Network Based on the Information of miRNA. *Sci. Rep.* 5, 13186. doi:10.1038/srep13186

Chen, X., and Yan, G.-Y. (2013). Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics* 29, 2617–2624. doi:10.1093/bioinformatics/btt426

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, 47864–47874. doi:10.18632/oncotarget.10012

Ding, L., Wang, M., Sun, D., and Li, A. (2018). TPGLDA: Novel Prediction of Associations between lncRNAs and Diseases *via* lncRNA-Disease-Gene Tripartite Graph. *Sci. Rep.* 8, 1065. doi:10.1038/s41598-018-19357-3

Dinger, M. E., Pang, K. C., Mercer, T. R., Crowe, M. L., Grimmond, S. M., and Mattick, J. S. (2009). NRED: a Database of Long Noncoding RNA Expression. *Nucleic Acids Res.* 37, D122–D126. doi:10.1093/nar/gkn617

Faghihi, M. A., and Wahlestedt, C. (2009). Regulatory Roles of Natural Antisense Transcripts. *Nat. Rev. Mol. Cel Biol* 10, 637–643. doi:10.1038/nrm2738

Fan, W., Shang, J., Li, F., Sun, Y., Yuan, S., and Liu, J.-X. (2020). IDSSIM: an lncRNA Functional Similarity Calculation Model Based on an Improved Disease Semantic Similarity Method. *BMC Bioinformatics* 21, 339. doi:10.1186/s12859-020-03699-9

Fan, X.-N., Zhang, S.-W., Zhang, S.-Y., Zhu, K., and Lu, S. (2019). Prediction of lncRNA-Disease Associations by Integrating Diverse Heterogeneous Information Sources with RWR Algorithm and Positive Pointwise Mutual Information. *BMC Bioinformatics* 20, 87. doi:10.1186/s12859-019-2675-y

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., et al. (2018). NONCODEV5: a Comprehensive Annotation Database for Long Non-coding RNAs. *Nucleic Acids Res.* 46, D308–d314. doi:10.1093/nar/gkx1107

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix Factorization-Based Data Fusion for the Prediction of lncRNA-Disease Associations. *Bioinformatics* 34, 1529–1537. doi:10.1093/bioinformatics/btx794

Garzon, R., Volinia, S., Papaioannou, D., Nicolet, D., Kohlschmidt, J., Yan, P. S., et al. (2014). Expression and Prognostic Impact of lncRNAs in Acute Myeloid

Leukemia. *Proc. Natl. Acad. Sci. USA* 111, 18679–18684. doi:10.1073/pnas.1422050112

Gu, C., Liao, B., Li, X., Cai, L., Li, Z., Li, K., et al. (2017). Global Network Random Walk for Predicting Potential Human lncRNA-Disease Associations. *Sci. Rep.* 7, 12442. doi:10.1038/s41598-017-12763-z

Guo, Z.-H., You, Z.-H., Wang, Y.-B., Yi, H.-C., and Chen, Z.-H. (2019). A Learning-Based Method for LncRNA-Disease Association Identification Combing Similarity Information and Rotation Forest. *iScience* 19, 786–795. doi:10.1016/j.isci.2019.08.030

Han, P., Li, W., Lin, C.-H., Yang, J., Shang, C., Nurnberg, S. T., et al. (2014). A Long Noncoding RNA Protects the Heart from Pathological Hypertrophy. *Nature* 514, 102–106. doi:10.1038/nature13596

Heo, J. B., Lee, Y.-S., and Sung, S. (2013). Epigenetic Regulation by Long Noncoding RNAs in Plants. *Chromosome Res.* 21, 685–693. doi:10.1007/s10577-013-9392-6

Hu, Y., Qiu, S., and Cheng, L. (2021). Integration of Multiple-Omics Data to Analyze the Population-specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* 2021, 1–11. doi:10.1155/2021/7036592

Hu, Y., Sun, J.-y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (202162220). rs1990622 Variant Associates with Alzheimer's Disease and Regulates TMEM106B Expression in Human Brain Tissues. *BMC Med.* 19, 11. doi:10.1186/s12916-020-01883-5

Hu, Y., Zhang, H., Liu, B., Gao, S., Wang, T., Han, Z., et al. (2020). rs34331204 Regulates TSPAN13 Expression and Contributes to Alzheimer's Disease with Sex Differences. *Brain* 143, e95. doi:10.1093/brain/awaa302

Jalali, S., Gandhi, S., and Scaria, V. (2016). Navigating the Dynamic Landscape of Long Noncoding RNA and Protein-Coding Gene Annotations in GENCODE. *Hum. Genomics* 10, 35. doi:10.1186/s40246-016-0090-2

Kino, T., Hurt, D. E., Ichijo, T., Nader, N., and Chrousos, G. P. (2010). Noncoding RNA Gas5 Is a Growth Arrest- and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Sci. Signal.* 3, ra8. doi:10.1126/scisignal.2000568

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F.-X., Pan, Y., et al. (2017). LDAP: a Web Server for lncRNA-Disease Association Prediction. *Bioinformatics* 33, btw639–460. doi:10.1093/bioinformatics/btw639

Li, J., Wang, J., Chen, Y., Li, S., Jin, M., Wang, H., et al. (2016). LncRNA MALAT1 Exerts Oncogenic Functions in Lung Adenocarcinoma by Targeting miR-204. *Am. J. Cancer Res.* 6, 1099–1107.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2013). starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data. *Nucl. Acids Res.* 42, D92–D97. doi:10.1093/nar/gkt1248

Liu, J.-X., Cui, Z., Gao, Y.-L., and Kong, X.-Z. (2021). WGRCMF: A Weighted Graph Regularized Collaborative Matrix Factorization Method for Predicting Novel LncRNA-Disease Associations. *IEEE J. Biomed. Health Inform.* 25, 257–265. doi:10.1109/jbhi.2020.2985703

Liu, Y., Zhang, R., Qiu, F., Li, K., Zhou, Y., Shang, D., et al. (2015). Construction of a lncRNA-PCG Bipartite Network and Identification of Cancer-Related lncRNAs: a Case Study in Prostate Cancer. *Mol. Biosyst.* 11, 384–393. doi:10.1039/c4mb00439f

Liu, Z.-P. (2021). Predicting lncRNA-Protein Interactions by Machine Learning Methods: A Review. *Cbio* 15, 831–840. doi:10.2174/1574893615666200224095925

Losko, M., Kotlinowski, J., and Jura, J. (2016). Long Noncoding RNAs in Metabolic Syndrome Related Disorders. *Mediators Inflamm.* 2016, 1–12. doi:10.1155/2016/5365209

Meng, N., Chen, M., Chen, D., Chen, X. H., Wang, J. Z., Zhu, S., et al. (2020). Small Protein Hidden in lncRNA LOC90024 Promotes "Cancerous" RNA Splicing and Tumorigenesis. *Adv. Sci.* 7, 1903233. doi:10.1002/advs.201903233

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long Non-coding RNAs: Insights into Functions. *Nat. Rev. Genet.* 10, 155–159. doi:10.1038/nrg2521

Mercer, T. R., and Mattick, J. S. (2013). Structure and Function of Long Noncoding RNAs in Epigenetic Regulation. *Nat. Struct. Mol. Biol.* 20, 300–307. doi:10.1038/nsmb.2480

Miao, Y.-R., Liu, W., Zhang, Q., and Guo, A.-Y. (2017). lncRNASNP2: an Updated Database of Functional SNPs and Mutations in Human and Mouse lncRNAs. *Nucleic Acids Res.* 46, D276–D280. doi:10.1093/nar/gkx1004

Nakagawa, S., Shimada, M., Yanaka, K., Mito, M., Arai, T., Takahashi, E., et al. (2014). The lncRNA Neat1 Is Required for Corpus Luteum Formation and the Establishment of Pregnancy in a Subpopulation of Mice. *Development* 141, 4618–4627. doi:10.1242/dev.110544

Okazaki, N., Kikuno, R., Ohara, R., Inamoto, S., Aizawa, H., Yuasa, S., et al. (2003). Prediction of the Coding Sequences of Mouse Homologues of KIAA Gene: II. The Complete Nucleotide Sequences of 400 Mouse KIAA-Homologous cDNAs Identified by Screening of Terminal Sequences of cDNA Clones Randomly Sampled from Size-Fractionated Libraries. *DNA Res.* 10, 35–48. doi:10.1093/dnares/10.1.35

Pandey, R. R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., et al. (2008). Kcnq1ot1 Antisense Noncoding RNA Mediates Lineage-specific Transcriptional Silencing through Chromatin-Level Regulation. *Mol. Cel* 32, 232–246. doi:10.1016/j.molcel.2008.08.022

Ping, P., Wang, L., Kuang, L., Ye, S., Iqbal, M. F. B., and Pei, T. (2019). A Novel Method for LncRNA-Disease Association Prediction Based on an lncRNA-Disease Association Network. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16, 688–693. doi:10.1109/tcbb.2018.2827373

Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A., and Panning, B. (2002). Xist RNA and the Mechanism of X Chromosome Inactivation. *Annu. Rev. Genet.* 36, 233–278. doi:10.1146/annurev.genet.36.042902.092433

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A Coding-independent Function of Gene and Pseudogene mRNAs Regulates Tumour Biology. *Nature* 465, 1033–1038. doi:10.1038/nature09144

Quek, X. C., Thomson, D. W., Maag, J. L. V., Bartonicek, N., Signal, B., Clark, M. B., et al. (2015). lncRNAdb v2.0: Expanding the Reference Database for Functional Long Noncoding RNAs. *Nucleic Acids Res.* 43, D168–D173. doi:10.1093/nar/gku988

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., et al. (2007). Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* 129, 1311–1323. doi:10.1016/j.cell.2007.05.022

Shen, L., Shen, G., Lu, X., Ding, G., and Hu, X. (2021). Co-expression Network Analysis Revealing the Potential Regulatory Roles of LncRNAs in Atrial Fibrillation. *Cbio* 15, 980–987. doi:10.2174/1574893614666191210142141

Sheng, N., Cui, H., Zhang, T., and Xuan, P. (2021). Attentional Multi-Level Representation Encoding Based on Convolutional and Variance Autoencoders for lncRNA-Disease Association Prediction. *Brief Bioinform* 22, 22. doi:10.1093/bib/bbaa067

Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., et al. (2014). Inferring Novel lncRNA-Disease Associations Based on a Random Walk Model of a lncRNA Functional Similarity Network. *Mol. Biosyst.* 10, 2074–2081. doi:10.1039/c3mb70608g

Sun, Y., Chen, L., Zhang, Y., Zhang, J., and Tiwari, S. R. (2021). Genome-wide Identification of Differently Expressed lncRNAs, mRNAs, and circRNAs in Patients with Osteoarthritis. *Cbio* 15, 1222–1230. doi:10.2174/1574893615999200706002907

Tan, H., Sun, Q., Li, G., Xiao, Q., Ding, P., Luo, J., et al. (2020). Multiview Consensus Graph Learning for lncRNA-Disease Association Prediction. *Front. Genet.* 11, 89. doi:10.3389/fgene.2020.00089

Tang, L., Liang, Y., Jin, X., Liu, L., and Zhou, W. (2020). Hierarchical Extension Based on the Boolean Matrix for LncRNA-Disease Association Prediction. *Cmm* 20, 452–460. doi:10.2174/1566524019666191119104212

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor Origin Detection with Tissue-specific miRNA and DNA Methylation Markers. *Bioinformatics* 34, 398–406. doi:10.1093/bioinformatics/btx622

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., et al. (2010). Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science* 329, 689–693. doi:10.1126/science.1192002

Wang, K. C., and Chang, H. Y. (2011). Molecular Mechanisms of Long Noncoding RNAs. *Mol. Cel* 43, 904–914. doi:10.1016/j.molcel.2011.08.018

Wang, P., Guo, Q., Gao, Y., Zhi, H., Zhang, Y., Liu, Y., et al. (2017). Improved Method for Prioritization of Disease Associated lncRNAs Based on ceRNA Theory and Functional Genomics Data. *Oncotarget* 8, 4642–4655. doi:10.18632/oncotarget.13964

Wei, H., Liao, Q., and Liu, B. (2021). iLncRNAdis-FB: Identify lncRNA-Disease Associations by Fusing Biological Feature Blocks through Deep Neural Network. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18, 1946–1957. doi:10.1109/tcbb.2020.2964221

Weng, W., Lin, C., Zhu, S., Ma, Y., Zou, Q., and Chen, Z. (2015). Community Discovery Algorithm with the Label Propagation Based on the Initial Division. *J. Xiamen Univ. Nat. Sci.* 54, 107–113.

Xiao, Y., Xiao, Z., Feng, X., Chen, Z., Kuang, L., and Wang, L. (2020). A Novel Computational Model for Predicting Potential LncRNA-Disease Associations Based on Both Direct and Indirect Features of LncRNA-Disease Pairs. *BMC Bioinformatics* 21, 555. doi:10.1186/s12859-020-03906-7

Xie, G., Meng, T., Luo, Y., and Liu, Z. (2019). SKF-LDA: Similarity Kernel Fusion for Predicting lncRNA-Disease Association. *Mol. Ther. - Nucleic Acids* 18, 45–55. doi:10.1016/j.omtn.2019.07.022

Yang, Q., Wu, J., Zhao, J., Xu, T., Han, P., and Song, X. (2020). The Expression Profiles of lncRNAs and Their Regulatory Network during Smek1/2 Knockout Mouse Neural Stem Cells Differentiation. *Cbio* 15, 77–88. doi:10.2174/1574893614666190308160507

Yoon, J.-H., Abdelmohsen, K., and Gorospe, M. (2013). Posttranscriptional Gene Regulation by Long Noncoding RNA. *J. Mol. Biol.* 425, 3723–3730. doi:10.1016/j.jmb.2012.11.024

Yu, G., Fu, G., Lu, C., Ren, Y., and Wang, J. (2017). BRWLDA: Bi-random Walks for Predicting lncRNA-Disease Associations. *Oncotarget* 8, 60429–60446. doi:10.18632/oncotarget.19588

Yu, J., Ping, P., Wang, L., Kuang, L., Li, X., and Wu, Z. (2018). A Novel Probability Model for LncRNA-Disease Association Prediction Based on the Naïve Bayesian Classifier. *Genes* 9, 345. doi:10.3390/genes9070345

Zeng, M., Lu, C., Zhang, F., Li, Y., Wu, F.-X., Li, Y., et al. (2020). SDLDA: lncRNA-Disease Association Prediction Based on Singular Value Decomposition and Deep Learning. *Methods* 179, 73–80. doi:10.1016/j.ymeth.2020.05.002

Zhang, E., He, X., Zhang, C., Su, J., Lu, X., Si, X., et al. (2018). A Novel Long Noncoding RNA HOXC-AS3 Mediates Tumorigenesis of Gastric Cancer by Binding to YBX1. *Genome Biol.* 19, 154. doi:10.1186/s13059-018-1523-0

Zhang, Y., Ye, F., Xiong, D., and Gao, X. (2020). LDNFSGB: Prediction of Long Non-coding Rna and Disease Association Using Network Feature Similarity and Gradient Boosting. *BMC Bioinformatics* 21, 377. doi:10.1186/s12859-020-03721-0

Zhao, X., Zou, Q., Liu, B., and Liu, X. (2015). Exploratory Predicting Protein Folding Model with Random Forest and Hybrid Features. *Cp* 11, 289–299. doi:10.2174/1570164611041501211151 54

Zhao, Z., Zhang, C., Li, M., Yu, X., Liu, H., Chen, Q., et al. (2021). Integrative Analysis of miRNA-Mediated Competing Endogenous RNA Network Reveals the lncRNAs-mRNAs Interaction in Glioblastoma Stem Cell Differentiation. *Cbio* 15, 1187–1196. doi:10.2174/1574893615999200511074226

Zhou, J.-R., You, Z.-H., Cheng, L., and Ji, B.-Y. (2021). Prediction of lncRNA-Disease Associations *via* an Embedding Learning HOPE in Heterogeneous Information Networks. *Mol. Ther. - Nucleic Acids* 23, 277–285. doi:10.1016/j.omtn.2020.10.040

Zhou, K.-R., Liu, S., Sun, W.-J., Zheng, L.-L., Zhou, H., Yang, J.-H., et al. (2016). ChIPBase v2.0: Decoding Transcriptional Regulatory Networks of Non-coding RNAs and Protein-Coding Genes from ChIP-Seq Data. *Nucleic Acids Res.* 45, D43–D50. doi:10.1093/nar/gkw965

Zhu, S., Wang, J.-Z., Chen, D., He, Y.-T., Meng, N., Chen, M., et al. (2020). An Oncopeptide Regulates m6A Recognition by the m6A Reader IGF2BP1 and Tumorigenesis. *Nat. Commun.* 11, 1685. doi:10.1038/s41467-020-15403-9

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA. *RNA* 25, 205–218. doi:10.1261/rna.069112.118

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.

Copyright © 2021 Gong, Zhu, Sun and Shi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 December 2021
doi: 10.3389/fgene.2021.801261

Testing Gene-Gene Interactions Based on a Neighborhood Perspective in Genome-wide Association Studies

Yingjie Guo[1,2], Honghong Cheng[3], Zhian Yuan[4], Zhen Liang[5], Yang Wang[1] and Debing Du[6]**

[1]School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China, [2]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [3]School of Information, Shanxi University of Finance and Economics, Taiyuan, China, [4]Research Institute of Big Data Science and Industry, Shanxi University, Taiyuan, China, [5]School of Life Science, Shanxi University, Taiyuan, China, [6]Beidahuang Industry Group General Hospital, Harbin, China

Unexplained genetic variation that causes complex diseases is often induced by gene-gene interactions (GGIs). Gene-based methods are one of the current statistical methodologies for discovering GGIs in case-control genome-wide association studies that are not only powerful statistically, but also interpretable biologically. However, most approaches include assumptions about the form of GGIs, which results in poor statistical performance. As a result, we propose gene-based testing based on the maximal neighborhood coefficient (MNC) called gene-based gene-gene interaction through a maximal neighborhood coefficient (GBMNC). MNC is a metric for capturing a wide range of relationships between two random vectors with arbitrary, but not necessarily equal, dimensions. We established a statistic that leverages the difference in MNC in case and in control samples as an indication of the existence of GGIs, based on the assumption that the joint distribution of two genes in cases and controls should not be substantially different if there is no interaction between them. We then used a permutation-based statistical test to evaluate this statistic and calculate a statistical *p*-value to represent the significance of the interaction. Experimental results using both simulation and real data showed that our approach outperformed earlier methods for detecting GGIs.

Keywords: genome-wide association studies, qualitative traits, gene-gene interactions, maximal neighborhood coefficient, gene-based testing

OPEN ACCESS

Edited by:
Dariusz Mrozek, Silesian University of Technology, Poland

Reviewed by:
Yang Hu, Harbin Institute of Technology, China
Lumeng Chao, Inner Mongolia Agricultural University, China

****Correspondence:***
Yingjie Guo
yjguo0625@gmail.com
Debing Du
dudebing_hrb@163.com

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

Received: *25 October 2021*
Accepted: *15 November 2021*
Published: *08 December 2021*

Citation:
Guo Y, Cheng H, Yuan Z, Liang Z, Wang Y and Du D (2021) Testing Gene-Gene Interactions Based on a Neighborhood Perspective in Genome-wide Association Studies. Front. Genet. 12:801261. doi: 10.3389/fgene.2021.801261

1 INTRODUCTION

Genome-wide association studies (GWAS) has been used to investigate the associations between genetic variants and complex disorders with great success. Researchers have discovered more than 71,000 unique single nucleotide polymorphisms (SNPs) associated to diseases throughout the last decade (Hindorff et al., 2009; Zhang et al., 2016; Zeng et al., 2017; Guo et al., 2018; Buniello et al., 2019; Loos, 2020; Li et al., 2021). Traditional GWAS, on the other hand, concentrated on the independent, additive, and cumulative effects of individual SNPs on specific diseases. The majority of associated SNPs are common genetic variants with small effects that only explain a portion of complex disease heritability. Many genes, environmental variables, and interactions play a crucial role in the underlying genetic architecture of complex diseases (Cordell, 2009; Moore et al., 2010; Jiang et al., 2018; Liu et al., 2018; Liu et al., 2019a; Zhang et al., 2019; Chen et al., 2020; Luo et al., 2020; Liu et al., 2021; Shao et al., 2021; Su et al., 2021; Wang et al., 2021). As a result, genetic interactions are

thought to enlighten studies into "missing heritability" (Manolio et al., 2009; Fang et al., 2019; Young, 2019; Tang et al., 2020; Song et al., 2021) and give important knowledge for constructing topologies for complex disease-related pathway.

Genetic interaction was originally explored at the SNP level, named epistasis. Methods (Li et al., 2015a; Ritchie and Van Steen, 2018; Lyu et al., 2020) can be classified into three categories based on their search strategy: exhaustive methods, searching methods, and machine learning-based methods, such as statistics based on entropy (Dong et al., 2008) and odds-ratios (Emily, 2012); MDR (Ritchie et al., 2003), BEAM (Zhang and Liu, 2007), BOOST (Wan et al., 2010), Epi-GTBN (Guo et al., 2019), GenEpi (Chang et al., 2020), and some accelerate methods (Nobre et al., 2021). For example, a logistic regression analysis revealed a significant interaction between the genes ERAP1 (rs27524) and HLA-C (rs10484554) in psoriasis ($p = 6.95 \times 10^{-6}$), indicating that ERAP1 SNP was effective only in individuals who had at least one copy of the HLA-C SNP risk allele (Képíró et al., 2021). The statistical weakness of high-order or pairwise tests, which come from enormous multiple testing corrections over all pairs of SNPs, is one of the general problems of these marker-based approaches. Instead, we explored the interaction of two genes in a single gene-based interaction detection by treating SNPs inside a gene as a group.

The effectiveness of gene-based methods in GWAS marginal association studies should be extended to the study of gene-gene interaction (GGIs) (Emily, 2018; Emily et al., 2020). This strategy offers a number of possible benefits. For starters, it often has substantially fewer genes than SNPs, which dramatically decreases the number of pairwise testing. To discover GGIs in pair of 20,000 genes, for example, $\sim 2 \times 10^8$ tests are necessary. However, for three million SNPs in a marker-based interaction, more than 5×10^{12} tests are required. Second, gene-based methods are more powerful statistically because a gene carries more information than individual SNP and genes interact in a variety of ways (Liu et al., 2010; Li et al., 2011; Jiang et al., 2017; Su et al., 2019; Hu et al., 2020; Hu et al., 2021a; Hu et al., 2021b; Guo et al., 2021). Furthermore, these methods can include biological prior knowledge (e.g., information about known gene association within protein-protein interactions (PPIs) or pathways) (Wei et al., 2017a; Wei et al., 2017b; Wei et al., 2018; Liu et al., 2019b; Wei et al., 2019; Zeng et al., 2019; Cai et al., 2020; Zhai et al., 2020; Zhu et al., 2020). Finally, gene-based outcomes stand out for their better interpretability and crucial biological consequences.

Many statistical and computational approaches for detecting gene-based GGIs have been established. Peng et al.(Peng et al., 2010) introduced the canonical correlation-based U statistic (CCU). They calculated canonical correlation of two genes in both cases and controls. They next used CCU to calculate the difference in correlation, which revealed the presence of GGIs between the two genes. However, this strategy only considered linear correlation in the study. CCU was then expanded to Kernelized CCU (KCCU) (Yuan et al., 2012; Larson et al., 2013), where the kernel discovered a nonlinear relationship. Emily (Emily, 2016) recently introduced AGGrGATOr, a method that combines *p*-values of interaction tests at the marker-level to assess how a pair of genes interacted, which was a strategy that Ma et al. (Ma et al., 2013) previously utilized to discover interactions under quantitative traits. GBIGM is a non-parametric entropy-based approach suggested by Li et al. (Li et al., 2015b).

In this paper, we propose a new approach called gene-based, gene-gene interaction through a maximal neighborhood coefficient (GBMNC), which uses the maximal neighborhood coefficient (MNC) (Cheng et al., 2020) to identify gene-gene interaction of complex diseases at the gene-level in case-control studies. MNC measures a wide variety of dependence with no bias toward relationship types between two random vectors of arbitrary, but not necessarily equal, dimensions; this is superior to Pearson's correlation, which only consider linear correlations. We introduced a statistic that uses the difference of MIC in cases and controls as an indicator of occurrence of GGIs, bases on the assumption that the joint distribution of two genes should not be significantly different in case and in control samples if there is no interaction between them (i.e. independent) under complex diseases. In simulation studies, our method exhibited an outstanding performance in recognizing the underlying GGIs at the gene level under a variety of conditions. Its application using real data sets showed accurate identification of GGIs.

2 MATERIALS AND METHODS

The statistical procedure for GBMNC is described in depth in this section. We give different parameter settings for simulation studies to evaluate the power to identify GGIs and the ability to control type-I error. Then, we adopted a real-world Rheumatoid Arthritis data set from the WTCCC (Wellcome Trust case Control Consortium) database to evaluate out method's effectiveness in a real situation.

2.1 GBMNC

2.1.1 Preliminaries and Notation

Here, we take genes, a couple of SNPs, as the basic unit. Suppose that we have n random samples:

$$\left(G_{1,i}, G_{2,i}\right) \in \mathcal{R}^{p+q}, i = 1, 2, \ldots, n \tag{1}$$

where

$$G_{1,i} = \left(g_{1,i,1}, g_{1,i,2}, \ldots, g_{1,i,p}\right), G_{2,i} = \left(g_{2,i,1}, g_{2,i,2}, \ldots, g_{2,i,q}\right),\ i = 1, 2, \ldots, n$$

and G_1 and G_2 represent two genes each with p and q SNPs, independently. In the case-control studies, $y_i \in \{0, 1\}$ is a categorical label where 0 is a control subject and one is a case subject. $g_{k,i,j} \in \{0, 1, 2\}$ represents the copy number of the minor alleles of SNP j in gene k for sample i.

In this work, to investigate whether there is a statistical interaction between two genes in a qualitative phenotype, we designed a statistic based on the maximal neighborhood coefficient to characterize the GGI intensity. We applied a permutation strategy to estimate the distribution of the statistic. Our approach was based on the intuition that, if there was no interaction between two genes, then, if they were independent of the case set, they should be independent of the control set; if they were dependent on the case set, they should be dependent on the control set as well, and the "strength" of such dependence should be the same for the case and control sets. Pearson's correlation

coefficient measures the degree of dependence between two random variables. However, it can only measure linear dependency and not nonlinear dependency, and it is not very convenient for random variables that take a value in $\mathcal{R}^n$. Therefore, we proposed to measure dependency between random variables by the maximal neighborhood coefficient (MNC) instead.

2.1.2 Maximal Neighborhood Coefficient

MNC is an association measure that decipher the potential complex associations from neighborhood insight. It assumes that if a relationship exists between two variables, the samples of each variable will appear to have a similar neighborhood tendency to approximate that relationship, and MNC can find those common neighborhood structures by exploring the possible neighborhoods of each variable. By introducing a k-NN granule to reconstruct samples, and a novel neighborhood mutual information (NMI) to measure the certainty information of one variable from another under a fixed (k_x, k_y) neighborhood combination, MNC enables us to detect more complex associations.

Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathcal{R}^2$ be a finite set that is sampled from a joint distribution (X, Y), and $S_X = \{x_1, \ldots, x_n\}$ and $S_Y = \{y_1, \ldots, y_n\}$ represents samples from marginal variables X and Y, respectively. Given a designated neighborhood combination (k_x, k_y) (a pairwise positive integer), $N_X^{k_x}(x) = \{x_{j_1}, \ldots, x_{j_{k_x}}\}$ designed as the k_x-NN granule of x, where the subscript sequence $j_1 < j_2 < \ldots < j_{k_x}$ is obtained by $d(x, x_{j_i}) = \|x - x_{ji}\|_2$. All samples of k_x-NN granules form a cover of S_X, that is $\cup_{i=1}^{n} N_X^{k_x}(x_i) = US_X$. At the same time, there exists a cover for S_Y, $\cup_{i=1}^{n} N_Y^{k_y}(y_i) = S_Y$. The cover of samples S under (k_x, k_y) is recorded as C_{k_x,k_y}. Let $S|_{C_{k_x,k_y}}$ represents the distribution of S on the cover C_{k_x,k_y}, and different neighborhood combinations produce different distributions.

MNC is defined based on the neighborhood characteristic matrix (NM) of a sample set S. Given a finite data set S and a neighborhood combination (k_x, k_y), the element of NM of S is:

$$NM(S)_{k_x,k_y} = \frac{NMI\left(S\big|_{C_{k_x,k_y}}\right)}{log\frac{n}{\max(k_x,k_y)}} \tag{2}$$

$NMI(S|_{C_{k_x,k_y}})$ denotes the neighborhood mutual information of distribution $S|_{C_{k_x,k_y}}$. The neighborhood mutual information of (x_i, y_i) is defined as follow:

$$NMI_{C_{k_x,k_y}}(x_i, y_i) = -log\frac{n\left|N_X^{k_x}(x_i) \cap N_Y^{k_y}(y_i)\right|}{k_x k_y} \tag{3}$$

Based on the equation above, the neighborhood mutual information of (X, Y) is defined as:

$$NMI_{C_{k_x,k_y}}(X, Y) = -\frac{1}{n}\sum_{i=1}^{n} log\frac{n\left|N_X^{k_x}(x_i) \cap N_Y^{k_y}(y_i)\right|}{k_x k_y} \tag{4}$$

With the definition of $NM(S)$ in **Eq. 2**, NMC is defined as:

$$NMC(S) = \max_{1 \le k_x k_y \le NB(n)} \left\{NM(S)_{k_x,k_y}\right\} \tag{5}$$

where $NB(n)$ is the search range, and $1 \le k_x k_y \le O(n^\alpha)$ for some $0 < \alpha < 1$. It also naturally extends to the case of two random vectors with arbitrary, but not necessarily equal, dimensions.

MNC Satisfies the Following Properties

1) Symmertry: $MNC(X, Y) = MNC(Y, X)$;
2) Comparability: $MNC \in [0, 1]$, $MNC = 0$ denotes that two variables are statistically independent; $MNC = 1$ implies a strong association between two variables.
3) Generality: MNC captures comprehensive range relationships.
4) Equitability: MNC is robust to noisy relationships. It provides similar scores to the equally noisy relationships of different types.

2.1.3 Illustration of the GBMNC Workflow

Assume there are n_1 control samples and n_2 case samples in a case-control study for a pair of genes such that G_1 has p SNPs and G_2 has q SNPs. Let $MNC_n(G_1, G_2)$ be the sample association score between G_1 and G_2. First, we calculate the $MNC_{n_1}^C(G_1, G_2)$ for control samples and $MNC_{n_2}^D(G_1, G_2)$ for case samples. Second, we design a statistic $\Delta MNC = \frac{|MNC_{n_1}^C(G_1,G_2) - MNC_{n_2}^D(G_1,G_2)|}{MNC_{n_2}^D(G_1,G_2)}$ to measure the difference in MNC between cases and controls. ΔMNC represents how different the two joint distributions (G_1^C, G_2^C) and (G_1^D, G_2^D) are. The larger the ΔMNC, the higher the probability that G_1 and G_2 interact.

To get a p-value, we needed to estimate the distribution of ΔMNC^0 under the null hypothesis. Here, we used a non-parametric strategy based on permutation: we shuffled the label y randomly m times, calculated ΔMNC using the same procedure above, and used the resulting empirical distribution as an estimate for the distribution of ΔMNC under the null hypothesis. Let the result of these m permutations be $\Delta MNC^1, \ldots, \Delta MNC^m$, then an estimated p-value for the null hypothesis is

$$p = \frac{\left|\{i\colon \Delta MNC^i \ge \Delta MNC^0\}\right|}{m} \tag{6}$$

We summarized the process of GBMNC in the algorithm below (Algorithm 1) and presented the overall workflow (**Figure 1**).

Algorithm 1. GBMNC

Data: Genotype G_1, G_2, Phenotype y, permutation times m
Result: significant p-value for interaction between G_1, G_2
1 Calculate $MNC_{n_1}^C(G_1, G_2)$ and $MNC_{n_2}^D(G_1, G_2)$ for both (G_1^C, G_2^C) and (G_1^D, G_2^D) by **Eq. 5**;
2 Calculate the difference ΔMNC^0 between $MNC_{n_1}^C(G_1, G_2)$ and $MNC_{n_2}^D(G_1, G_2)$;
3 **for** $i = 1$ to m **do**
4 Randomly permute label y, and generate the new data set;
5 Repeat Steps 1 and 2;
6 **end**
7 Estimated p-value of ΔMNC^0 is the number of ΔMNC^i, $i = 1, \ldots, m$, which are larger than ΔMNC^0, divided by m.

2.2 Simulation Study

To assess the performance of GBMNC to control type I error and the power to detect GGIs, we compared GBMNC with KCCA (Larson et al., 2013), GBIGM (Li et al., 2015b), and AGGrEGATOr (Emily, 2016).

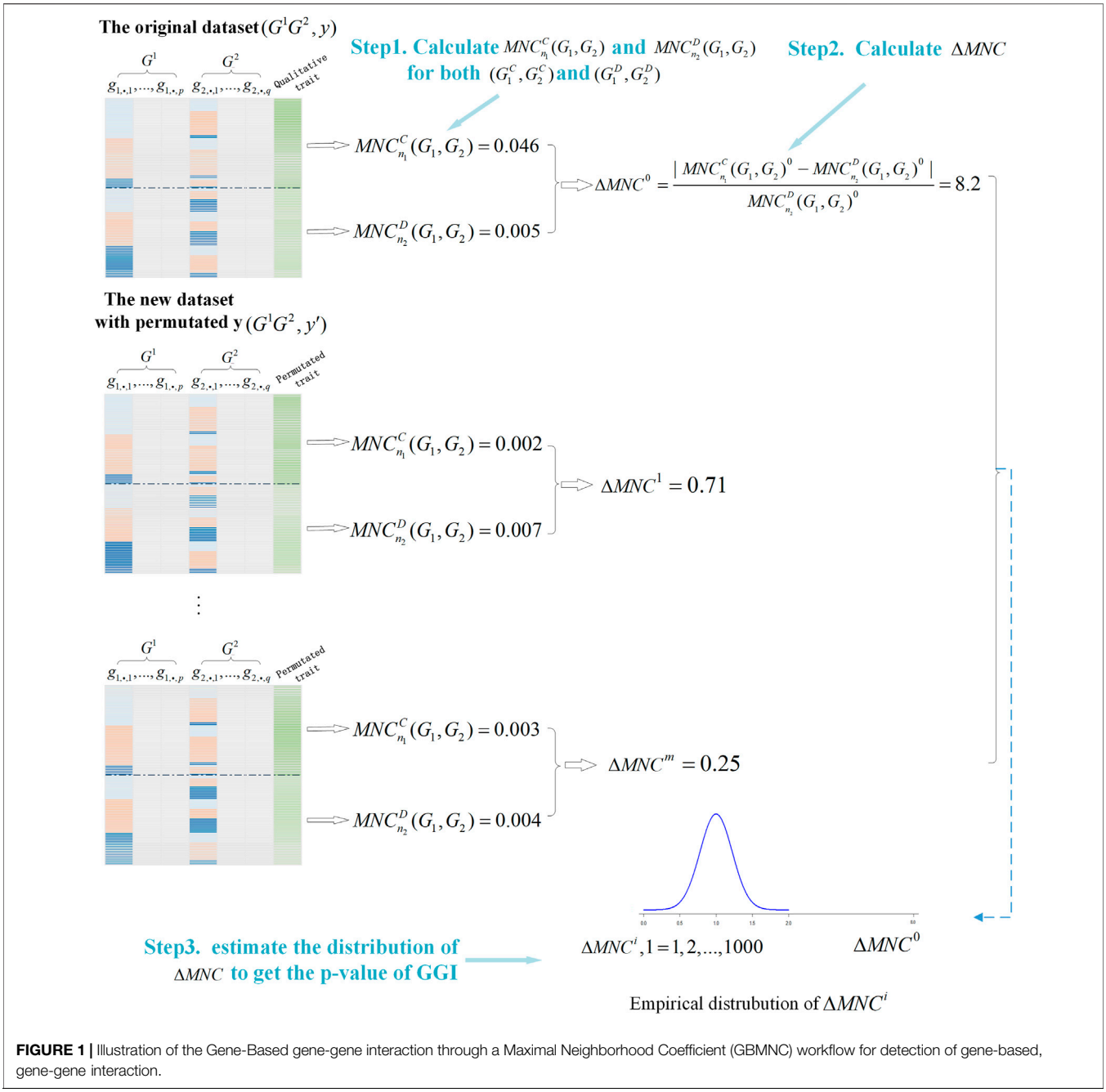


FIGURE 1 | Illustration of the Gene-Based gene-gene interaction through a Maximal Neighborhood Coefficient (GBMNC) workflow for detection of gene-based, gene-gene interaction.

2.2.1 Simulation With GAMETES

The goal of this simulation study was to evaluate the performance of the GBMNC procedure to detect gene-gene interaction. We set all simulated datasets to have 50 SNPs. Among them, two SNPs were functional, and the remaining 48 SNPs were non-functional. The 50 SNPs formed five genes, and each had 10 SNPs. The two functional SNPs were put into the first and second genes. We chose the publicly available tool GAMETES (Urbanowicz et al., 2012) to generate the simulated genotype data. This tool was designed to generate pure and strict epistasis models. Pure and strict epistasis models are the most difficult disease-related patterns to identify. Such associations can only be observed if all n-loci are included in the disease model. This requirement makes these types of models an attractive gold standard for simulation studies of complex multi-locus effects.

Evaluation of Type-I error: The type-I error indicates the ability of a method to reject the null hypothesis when it is true (i.e., the false positive rate). We used GAMETES to generate the custom disease model (**Table 1**) with one causal SNP pair. γ characterizes the baseline odds (i.e., the odds conditional on genotype pair *AABB*). We ran the simulation 100 times with

TABLE 1 | Table of odds for the no effect model without interaction between a pair of SNPs.

	AA	Aa	Aa
BB	γ	γ	γ
Bb	γ	γ	γ
bb	γ	γ	γ

each sample size $n \in \{1k, 2k, 3k, 4k, 5k\}$ and $\gamma = 1$. The significance level α was set to be 0.05.

Evaluation of power of the test: The power of a test indicates the probability that the method rejects the null hypothesis correctly when the alternative hypothesis is true. In this simulation study, we generated 100 data sets for each parameter settings. The power under each parameter setting was expressed by the frequency, and the null hypothesis of the data set was rejected correctly at the significance level of $\alpha = 0.05$.

1) To assess the impact of heritability h, which measured the intensity of correlation between genotype and phenotype, we chose $h \in \{0.01, 0.025, 0.05, 0.1, 0.2\}$ and two different minor allele frequencies MAF $\in \{0.2, 0.4\}$ with population prevalence set to 0.2 and sample size set at 4,000. Under each parameter combination, five models were generated so that we had a total of 100 models that followed Hardy-Weinberg proportions. For a specified genetic constrain combination, the 10 models were sorted roughly by the ascending customized odds ratio (COR) using GAMETES and labeled M1 to M5. COR is a metric of detectability that was calculated directly from the genetic model. The higher it is, the easier it is to detect GGIs. GAMETES generated the penetrance tables for these 100 models in the absence of the main effect. One hundred replicated data sets were generated from each model with balanced cases and controls, which resulted in 5,000 data sets in total in this scenario.
2) To evaluate the influence of sample size, we set heritability to be 0.025, MAF $\in \{0.2, 0.4\}$ and prevalence to be 0.2 with a sample size of 10,000. Then, 100 data sets were generated by random sampling from this large dataset for each of the sample sizes $n \in \{1k, 2k, 3k, 4k, 5k\}$. In this scenario, we had 1,000 datasets in total.

For GBMNC, KCCU, AGGrEGATOr, and GBIGM, if the number of data sets with a significance level less than α is m_1, then the power can be calculated by the following formula:

$$power = \frac{m_1}{100} \tag{7}$$

GBIGM and AGGrEGATOr methods are nonparametric methods, so no parameters need to be specific. We only set the ratio of the trimmed jackknife to 0.05 ($\omega = 0.05$) for KCCU.

2.3 EXPERIMENTS USING RHEUMATOID ARTHRITIS DATA

To evaluate GBMNC's ability to process real GGIs in a qualitative data set, we analyzed the susceptibility of a series of pairs of genes in Rheumatoid Arthritis (RA). RA is a chronic autoimmune disease that causes pannus development and cartilage and bone loss in synovial joints. It leads to progressive bone deterioration and interferes with bone repair. In this work, we used the WTCCC (2007) data set, which includes genotype data from the British population obtained by the Affymetrix GeneGhip 500 k. Our dataset was pre-processed in the following ways:

1) We used pathway hsa05323 from the KEGG pathway database to validate the GGIs in the RA. The WTCCC data set's genotyping coordinates can be found in UCSC hg18/NCBI Build36. This pathway contained 90genes. Many of the genes belonged to the protein combinations MHCII and V-ATPase. Because numerous GGIs happened on their own, we only chose representative genes from each protein combination and then remove the others. Finally, 48genes remained, resulting in a total of $C_{48}^2 = 1128$ pairs of genes to be analyzed.
2) We collected the detailed gene information from the NCBI Build36 annotation file, and for each gene, we inserted a 10 kb buffer region both downstream and upstream of the originally defined gene location. For each gene, all SNPs within the area were chosen.
3) According to the quality control of GWAS, samples that included gender that did not match the chromosome X heterozygote rates were removed. SNPs were also removed if any of the following requirements were met: the missing rate in the sample was $\geq 10\%$, MAF was ≤ 0.05, or the frequency of control violated Hardy-Weinberg equilibrium ($p < 0.0001$). Finally, 385 SNPs remained in 4,966 samples, which included 2,993 control subjects and 1973 case subjects.

3 RESULTS AND DISCUSSION

The experimental environment for all the following results was a workstation with an Intel Xeon CPU E5-2,620 v2 at 2.10GHz, 96 GB of DDR3, and python3.6.

3.1 Simulation Study

3.1.1 Evaluation of Type-I Error

For type-I error, we varied the sample size from 1,000 to 5,000. Except for GBIGM with $n = 1,000$, all methods tested had a type-I error comparable to a significance level $\alpha = 0.05$ (**Table 2**), which implied that these methods controlled for type-I error for various sample sizes quite well.

3.1.2 Evaluation of the Power of GBMNC

Impact of heritability: To evaluate the statistical power of our GBMNC and the other three methods, we used 10 heritability-MAF combinations, with a population prevalence of 0.2, a sample size of 4,000, and heritability that varied from 0.01 to 0.2

TABLE 2 | Type-I error for KCCU, GBIGM, AGGrEGATOr, and GBMNC when varying the sample size from 1,000 to 5,000.

Methods	Sample size				
	1,000	2,000	3,000	4,000	5,000
KCCU	0.02	0.02	0.01	0.05	0.07
GBIGM	0.13	0.06	0.07	0.07	0.07
AGGrEGATOr	0.05	0.06	0.07	0.04	0.02
GBMNC	0.02	0.05	0.07	0.05	0.05

TABLE 3 | The statistical power of simulation studies for GBMNC, AGGrEGATOr, KCCU and GBIGM under 10 heritability-MAF combinations, with $h \in \{0.01, 0.025, 0.05, 0.1, 0.2\}$ and MAF $\in \{0.2, 0.4\}$. Each heritability-MAF combination has five models. Bold font indicates the method that performed best under each model.

MAF	Heritability	Model	M1	M2	M3	M4	M5
		Method					
0.2	0.01	GBMNC	0.13	0.40	0.68	0.72	0.89
		AGGrEGATOr	0.12	0.12	0.89	0.89	1
		KCCU	0.15	0.09	0.29	0.43	0.62
		GBIGM	0.09	0.11	0.13	0.11	0.08
	0.025	GBMNC	0.95	0.75	1	0.96	1
		AGGrEGATOr	1	0.27	1	0.37	1
		KCCU	0.58	0.09	0.74	0.24	0.8
		GBIGM	0.08	0.07	0.11	0.13	0.2
	0.05	GBMNC	0.68	0.83	0.94	1	1
		AGGrEGATOr	0.09	0.59	0.89	1	1
		KCCU	0.13	0.57	0.65	0.84	0.85
		GBIGM	0.18	0.08	0.22	0.17	0.19
	0.1	GBMNC	1	1	1	1	1
		AGGrEGATOr	1	1	1	1	1
		KCCU	0.81	0.93	0.9	0.86	0.91
		GBIGM	0.15	0.14	0.23	0.16	0.16
	0.2	GBMNC	1	1	1	1	1
		AGGrEGATOr	1	1	1	1	1
		KCCU	0.89	0.97	0.94	0.89	0.97
		GBIGM	0.19	0.31	0.18	0.22	0.21
0.4	0.01	GBMNC	0.75	0.66	0.82	0.90	0.96
		AGGrEGATOr	0.71	0.09	0.1	0.94	0.96
		KCCU	0.34	0.05	0.08	0.77	0.29
		GBIGM	0.09	0.08	0.1	0.11	0.07
	0.025	GBMNC	1	0.73	0.85	0.93	0.80
		AGGrEGATOr	0.99	0.56	0.12	0.91	0.26
		KCCU	0.58	0.24	0.08	0.24	0.11
		GBIGM	0.15	0.12	0.14	0.11	0.09
	0.05	GBMNC	1	1	1	0.68	0.86
		AGGrEGATOr	1	0.97	0.91	0.35	0.42
		KCCU	0.86	0.9	0.95	0.41	0.37
		GBIGM	0.11	0.12	0.09	0.08	0.10
	0.1	GBMNC	1	1	1	0.63	1
		AGGrEGATOr	0.98	1	0.96	0.27	1
		KCCU	0.62	1	0.95	0.41	1
		GBIGM	0.12	0.19	0.18	0.26	0.20
	0.2	GBMNC	1	1	1	1	1
		AGGrEGATOr	0.93	1	0.99	1	0.80
		KCCU	0.28	1	0.83	1	0.76
		GBIGM	0.19	0.25	0.31	0.13	0.26

(**Table 3**). The bold in **Table 3** shows the best-performed method in each model under a given heritability-MAF combination. Notice that a larger value indicates better performance. On average, GBMNC was the best performing algorithm in this comparison. It largely outperformed the other methods, but not for all the data sets; it was inferior to AGGrEGATOr for some data sets. However, its performance was remarkably consistent, and it was the top performer for most data sets. AGGrEGATOr achieved the same performance when MAF was 0.2 and heritability was >0.05.

TABLE 4 | Average power for GBMNC, AGGrEGATOr, KCCU, and GBIGM under 10 heritability-MAF combinations, with heritability $\in \{0.01, 0.025, 0.05, 0.1, 0.2\}$ and MAF. $\in \{0.2, 0.4\}$

MAF	Method	GBMNC	AGGrE-GATOr	KCCU	GBIGM
	Heritability				
0.2	0.01	0.564	0.604	0.316	0.104
	0.025	0.932	0.728	0.490	0.118
	0.05	0.890	0.714	0.608	0.168
	0.1	1	1	0.882	0.168
	0.2	1	1	0.932	0.222
0.4	0.01	0.818	0.560	0.306	0.090
	0.025	0.862	0.568	0.250	0.122
	0.05	0.908	0.730	0.698	0.100
	0.1	0.926	0.842	0.796	0.190
	0.2	1	0.944	0.774	0.228

The power of all the methods was significantly affected by heritability (i.e., the effect size of interaction) (**Table 4**). A larger heritability led to better performance for all methods under a specific MAF. When heritability varied from 0.01 to 0.025, GBMNC almost doubled its power for a given sample size of 4,000 with MAF = 0.2. Other methods also show a steady upward trend (**Table 4**). The power also depended on the MAF of the interacting SNPs (e.g., for the cases of $h = 0.01$, the power of GBMNC under model M1-M5 ranged between 0.13–0.89 for MAF = 0.2, but it ranged between 0.66–0.96 for MAF = 0.4 (**Table 3**)). The average power was 0.564 for MAF = 0.2, which was much lower than 0.818 for MAF = 0.4 (**Table 4**).

It is worth noting that under the same combination of habitability and MAF, GBMNC was more stable under models with different COR compared with AGGrEGATOr (**Figure 2**). KCCU detected the interaction of some simulated disease models in our study, and it had a similar performance pattern with AGGrEGATOr. However, AGGrEGATOr was much more powerful in most of the simulated scenarios. GBIGM had little power to detecting pure gene-gene interaction,. This result replicated Emily's (Emily, 2016) result of the simulation.

Impact of sample size: The sample size of the data set had a considerable effect on power. Let the sample size be $n \in \{1k, 2k, 3k, 4k, 5k\}$, $h = 0.025$, and MAF $\in \{0.2, 0.4\}$ (**Table 5**). As the sample size increased, the power of all methods increased almost monotonically under different MAF settings. With all methods, a larger sample size corresponded to better performance.

In conclusion, in simulated studies, our results showed that GBMNC detected gene-gene interaction effectively, in which a

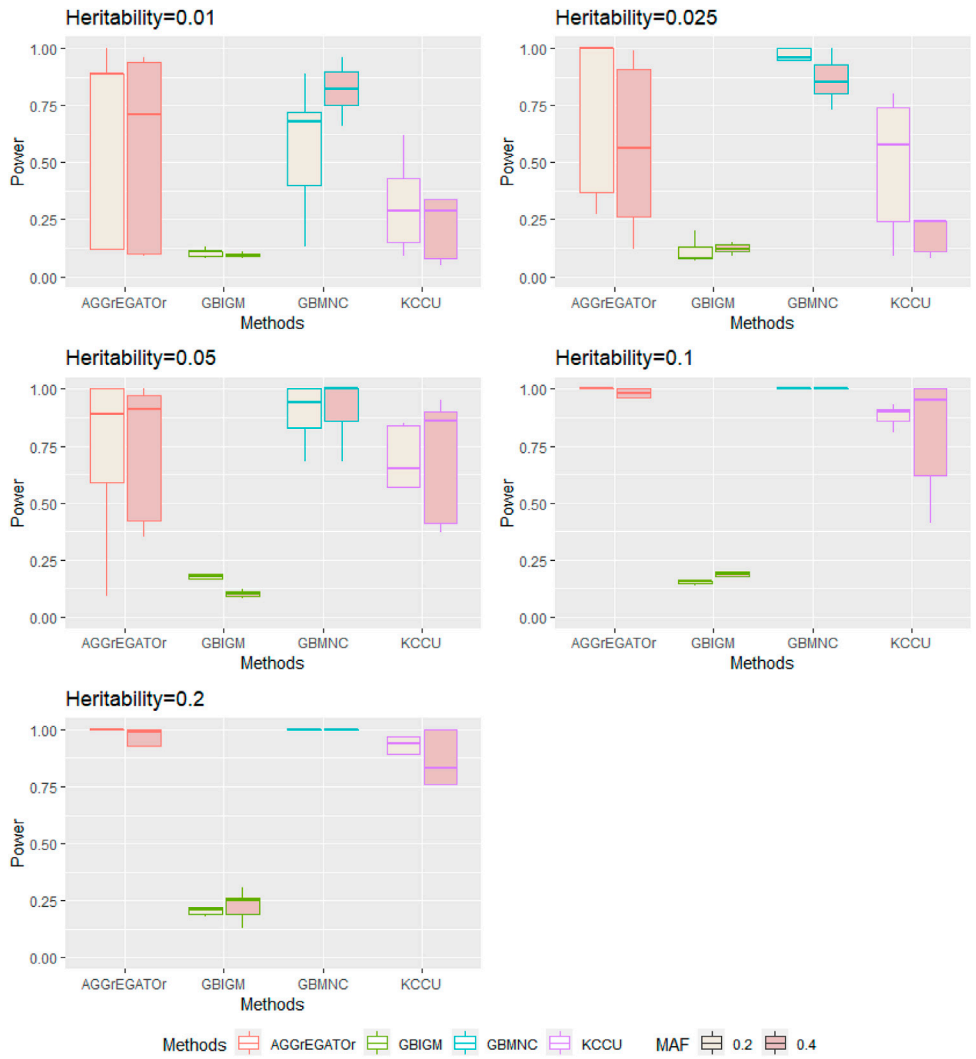


FIGURE 2 | Illustration of the distribution of power of each method in each heritability-MAF combination with $h \in \{0.01, 0.025, 0.05, 0.1, 0.2\}$ and MAF $\in \{0.2, 0.4\}$.

TABLE 5 | The statistical power of simulation studies for GBMNC, AGGrEGATOr, KCCU, and GBIGM under models with $h = 0.025$, MAF $\in \{0.2, 0.4\}$, and sample sizes that varied from $1k$ to $5k$.

MAF	Method / Sample size	GBMNC	AGGrEGATOr	KCCU	GBIGM
0.2	1,000	0.67	0.15	0.11	0.2
	2000	0.83	0.18	0.38	0.16
	3,000	1	0.20	0.55	0.23
	4,000	1	0.31	0.76	0.21
	5,000	1	0.29	0.87	0.12
0.4	1,000	0.68	0.16	0.13	0
	2000	0.97	0.20	0.11	0.04
	3,000	1	0.35	0.2	0.11
	4,000	1	0.54	0.37	0.11
	5,000	1	0.65	0.58	0.05

pair of SNPs was a causal factor by the purely and strictly epistasis model without main effect, which can only be observed if all 2-loci are included in the disease model. Compared with other methods, GBMNC identified a broad range of epistatic signals accurately.

3.2 EXPERIMENTS USING RHEUMATOID ARTHRITIS DATA

RA is a chronic autoimmune disease where HLA genes, TNF family, and TRAF1 are important genetic risk factors in the development. Each unique gene pair of the hsa05323 pathway was evaluated in the RA study, which resulted in $C_{48}^2 = 1128$ total pairs for 48 genes. With a significance level $\alpha = 0.01$ and multiple testing adjustment, for KCCU and GIGBM, we obtained 159 and 134 significant GGIs, respectively. Among them, 30 and 65 had p-values equal to 0; hence we were unable to rank them in the order of significance. AGGrGETOr did not show any significant results. Following Emily (Emily, 2016), and after removing the multiple testing correction, AGGrGETOr exhibited 17 significant GGIs, which we ranked by their

TABLE 6 | The calculated *p*-value for the 20 gene pairs using GBMNC and AGGrEGATOr. *p*-values in bold font indicate that they are significant. The "Chr" column indicates the chromosome number of the human genome where the gene is located.

Gene1	Chr	Gene2	Chr	*p*-value	
				GBMNC	AGGrEGATOr
TGF-β 2	1	CXCL8	4	0.0	1
CTLA4	2	GM-CSF	5	0.0	0.327
CD80	3	HLA-classII	6	0.0	0.37
GM-CSF	5	TRAP	19	0.0	0.01
TLR-4	9	FLT-1	13	0.0	0.069
IL-17	6	TNFSF13B	13	0.0	0.185
CXCL6	4	ICAM1	19	0.0	1
CD28	2	CXCL6	4	0.0	0.512
CTLA4	2	CXCL6	4	0.0	0.849
MMP-3	11	FLT-1	13	0.0	0.089
CD80	3	April	17	0.99	0.0007
CTSK	1	TNFSF13B	13	0.615	0.0008
JUN	1	IL-6	7	0.445	0.0019
CD80	3	CTSL	25	0.0	0.002
CXCL6	4	FLT-1	13	0.297	0.0021
CTLA4	2	FOS	37	0.727	0.0022
FLT-1	13	LFA-1	39	0.815	0.0033
CCL3	17	TRAP	19	0.564	0.0034
IL-18	11	TGF-β 3	14	0.693	0.004
IL-1	2	CXCL12	10	0.081	0.004

p-values. We chose the top 10 gene pairs obtained by GBMNC and by AGGrGETOr to analyze, which comprised approximately 1% of the total interactions (**Table 6**).

We found that some of our findings were supported by prior research (Xiao et al., 2008; Klocke et al., 2016; Cen et al., 2019). For instance, our method detected a significant interaction between IL17 and TNFSF13B. Studies (Xiao et al., 2008) show that both B cells and T cells formed aggregates in the synovium of inflamed joints and mediated the pathogenesis of RA, and B-cell-activating factor (BAFF, also named TNFSF13B, BLys) played a vital role in B-cell survival and maturation. After activation and expansion, $CD4^+$ T cells developed into different T helper cell subsets with different cytokine profiles and distinct effector functions. In addition to Th1 and Th2 cells, Th17 cells were a third T helper cell and produce IL-17. Th17 cells can recruit and activate inflammatory cells and they have been recognized as a primary cause of bone destruction and inflammation in autoimmune diseases. BAFF promoted Th17 cell proliferation and expansion preferentially (Lai Kwan Lam et al., 2008). IL-17 was a key cytokine for BAFF-mediated proinflammatory effects during collagen-induced arthritis pathogenesis. Only one pair of potential interactions between CD80 and CTSL was captured by both methods within the top 10 GGIs. However, there is not yet direct evidence to show the interaction between CD80 and CTSL.

4 CONCLUSION

The study of detecting GGIs is of great importance in understanding the pathogenesis of complex human diseases. In this paper, we proposed a gene-based GGI detection method called GBMNC based on a maximal neighborhood coefficient and a permutation strategy for case-control studies in GWAS. The method not only benefited from the ability of a maximal neighborhood coefficient, which considered the neighborhood structure of each sample and captured a wide range of associations, but also from the robustness of our permutation-based hypothesis testing scheme.

We designed a statistic to capture the different intensities of interaction between two genes in both cases and controls, then transformed the problem of GGI detection into a form of hypothesis testing; our null hypothesis was there was no significant difference in the relationship between the two genes in the disease data and the control data. This hypothesis did not limit the form of interaction between genes, and it enhanced the method's ability to detect different types of interactions. We demonstrated the effectiveness of our method through a simulation study and retrospective analysis of rheumatoid arthritis. Under a large range of settings, GBMNC outperformed previous methods in the power to detect GGIs. The statistical power of our method increased monotonically with the increase in the heritability and the MAF. The method was also stable to sample size based on a test of false positive rates. MNC did not restrict the dimension of two random vectors. Therefore, it is possible to generalize the method for marker-based detection of gene pairs that are identified as interactive. Investigating the mechanism of gene-based interaction at the marker level might point the way for further research. In summary, GBMNC is a helpful addition to the current toolbox of statistical models to elucidate GGIs in case-control studies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.wtccc.org.uk/info/access_to_data_samples.html.

AUTHOR CONTRIBUTIONS

YG: Conceptualization, Methodology, Investigation, Funding acquisition, Writing-Original Draft. HC: Methodology, Formal analysis, Writing-Original Draft. ZY: Software, Formal analysis. ZL: Resources, Data Curation. YW: Formal analysis, Writing-Review and Editing. DD: Conceptualization, Project administration.

FUNDING

The work was supported by the National Natural Science Foundation of China (No. 62002243, No. 31900306), the Post-doctoral Foundation Project of Shenzhen Polytechnic China (6020330004K), the Post-doctoral Foundation Project of Shenzhen Polytechnic China (6020330004K). Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (2021L286).

REFERENCES

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of Published Genome-wide Association Studies, Targeted Arrays and Summary Statistics 2019. *Nucleic Acids Res.* 47 (D1), D1005–D1012. doi:10.1093/nar/gky1120

Cai, L., Wang, L., Fu, X., Xia, C., Zeng, X., and Zou, Q. (2020). 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

Cen, S., Wang, P., Xie, Z., Yang, R., Li, J., Liu, Z., et al. (2019). Autophagy Enhances Mesenchymal Stem Cell-Mediated CD4+ T Cell Migration and Differentiation through CXCL8 and TGF-B1. *Stem Cel Res Ther.* 10 (1), 265. doi:10.1186/s13287-019-1380-0

Chang, Y.-C., Wu, J.-T., Hong, M.-Y., Tung, Y.-A., Hsieh, P.-H., Yee, S. W., et al. (2020). GenEpi: Gene-Based Epistasis Discovery Using Machine Learning. *BMC Bioinformatics* 21 (1), 68. doi:10.1186/s12859-020-3368-2

Chen, L., Li, J., and Chang, M. (2020). Cancer Diagnosis and Disease Gene Identification via Statistical Machine Learning. *Curr. Bioinformatics* 15 (9), 956–962. doi:10.2174/1574893615666200207094947

Cheng, H., Liang, J., Qian, Y., and Hu, Z. (2020). Association Mining Method Based on Neighborhood Perspective. *Sci. Sin.-Inf.* 50 (6), 824–844. doi:10.1360/ssi-2020-0009

Cordell, H. J. (2009). Detecting Gene-Gene Interactions that Underlie Human Diseases. *Nat. Rev. Genet.* 10 (6), 392–404. doi:10.1038/nrg2579

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., et al. (2008). Exploration of Gene-Gene Interaction Effects Using Entropy-Based Methods. *Eur. J. Hum. Genet.* 16 (2), 229–235. doi:10.1038/sj.ejhg.5201921

Emily, M. (2016). AGGrEGATOr: A Gene-Based GEne-Gene interActTiOn Test for Case-Control Association Studies. *Stat. Appl. Genet. Mol. Biol.* 15 (2), 151–171. doi:10.1515/sagmb-2015-0074

Emily, M. (2018). A Survey of Statistical Methods for Gene-Gene Interaction in Case-Control Genome-wide Association Studies. *Journal de la société française de statistique*. 159 (1), 27–67.

Emily, M. (2012). IndOR: a New Statistical Procedure to Test for SNP-SNP Epistasis in Genome-wide Association Studies. *Statist. Med.* 31 (21), 2359–2373. doi:10.1002/sim.5364

Emily, M., Sounac, N., Kroell, F., and Houée-Bigot, M. (2020). Gene-Based Methods to Detect Gene-Gene Interaction in R: The GeneGeneInteR Package. *J. Stat. Softw.* 95 (12), 1–32. doi:10.18637/jss.v095.i12

Fang, G., Wang, W., Paunic, V., Heydari, H., Costanzo, M., Liu, X., et al. (2019). Discovering Genetic Interactions Bridging Pathways in Genome-wide Association Studies. *Nat. Commun.* 10 (1), 4274. doi:10.1038/s41467-019-12131-7

Guo, F., Wang, D., and Wang, L. (2018). Progressive Approach for SNP Calling and Haplotype Assembly Using Single Molecular Sequencing Data. *Bioinformatics* 34 (12), 2012–2018. doi:10.1093/bioinformatics/bty059

Guo, Y., Yan, K., Lv, H., and Liu, B. (2021). PreTP-EL: Prediction of Therapeutic Peptides Based on Ensemble Learning. *Brief. Bioinform.* 22 (6), bbab358. doi:10.1093/bib/bbab358

Guo, Y., Zhong, Z., Yang, C., Hu, J., Jiang, Y., Liang, Z., et al. (2019). Epi-GTBN: an Approach of Epistasis Mining Based on Genetic Tabu Algorithm and Bayesian Network. *BMC Bioinformatics* 20 (1), 444. doi:10.1186/s12859-019-3022-z

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential Etiologic and Functional Implications of Genome-wide Association Loci for Human Diseases and Traits. *Proc. Natl. Acad. Sci.* 106 (23), 9362–9367. doi:10.1073/pnas.0903103106

Hu, Y., Qiu, S., and Cheng, L. (2021). Integration of Multiple-Omics Data to Analyze the Population-specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* 2021, 7036592. doi:10.1155/2021/7036592

Hu, Y., Sun, J. Y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (2021). rs1990622 Variant Associates with Alzheimer's Disease and Regulates TMEM106B Expression in Human Brain Tissues. *BMC Med.* 19 (1), 11. doi:10.1186/s12916-020-01883-5

Hu, Y., Zhang, H., Liu, B., Gao, S., Wang, T., Han, Z., et al. (2020). rs34331204 Regulates TSPAN13 Expression and Contributes to Alzheimer's Disease with Sex Differences. *Brain* 143 (11), e95. doi:10.1093/brain/awaa302

Jiang, L., Xiao, Y., Ding, Y., Tang, J., and Guo, F. (2018). FKL-Spa-LapRLS: an Accurate Method for Identifying Human microRNA-Disease Association. *Bmc Genomics* 19, 911. doi:10.1186/s12864-018-5273-x

Jiang, Q., Jin, S., Jiang, Y., Liao, M., Feng, R., Zhang, L., et al. (2017). Alzheimer's Disease Variants with the Genome-wide Significance Are Significantly Enriched in Immune Pathways and Active in Immune Cells. *Mol. Neurobiol.* 54 (1), 594–600. doi:10.1007/s12035-015-9670-8

Képíró, L., Széll, M., Kovács, L., Keszthelyi, P., Kemény, L., and Gyulai, R. (2021). The Association of HLA-C and ERAP1 Polymorphisms in Early and Late Onset Psoriasis and Psoriatic Arthritis Patients of Hungary. *Postepy Dermatol. Alergol.* 38 (2), 43–51. doi:10.5114/ada.2021.104277

Klocke, K., Sakaguchi, S., Holmdahl, R., and Wing, K. (2016). Induction of Autoimmune Disease by Deletion of CTLA-4 in Mice in Adulthood. *Proc. Natl. Acad. Sci. USA* 113 (17), E2383–E2392. doi:10.1073/pnas.1603892113

Lai Kwan Lam, Q., King Hung Ko, O., Zheng, B.-J., and Lu, L. (2008). Local BAFF Gene Silencing Suppresses Th17-Cell Generation and Ameliorates Autoimmune Arthritis. *Proc. Natl. Acad. Sci.* 105 (39), 14993–14998. doi:10.1073/pnas.0806044105

Larson, N. B., Jenkins, G. D., Larson, M. C., Vierkant, R. A., Sellers, T. A., Phelan, C. M., et al. (2013). Kernel Canonical Correlation Analysis for Assessing Gene-Gene Interactions and Application to Ovarian Cancer. *Eur. J. Hum. Genet.* 22 (1), 126–131. doi:10.1038/ejhg.2013.69

Li, H.-L., Pang, Y.-H., and Liu, B. (2021). BioSeq-BLM: a Platform for Analyzing DNA, RNA and Protein Sequences Based on Biological Language Models. *Nucleic Acids Res.* gkab829. doi:10.1093/nar/gkab829

Li, J., Huang, D., Guo, M., Liu, X., Wang, C., Teng, Z., et al. (2015). A Gene-Based Information Gain Method for Detecting Gene-Gene Interactions in Case-Control Studies. *Eur. J. Hum. Genet.* 23 (11), 1566–1572. doi:10.1038/ejhg.2015.16

Li, M.-X., Gui, H.-S., Kwan, J. S. H., and Sham, P. C. (2011). GATES: a Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *Am. J. Hum. Genet.* 88 (3), 283–293. doi:10.1016/j.ajhg.2011.01.019

Li, P., Guo, M., Wang, C., Liu, X., and Zou, Q. (2015). An Overview of SNP Interactions in Genome-wide Association Studies. *Brief. Funct. Genomics* 14 (2), 143–155. doi:10.1093/bfgp/elu036

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, G., Hu, Y., Han, Z., Jin, S., and Jiang, Q. (2019). Genetic Variant Rs17185536 Regulates SIM1 Gene Expression in Human Brain Hypothalamus. *Proc. Natl. Acad. Sci. USA* 116 (9), 3347–3348. doi:10.1073/pnas.1821550116

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, J., Su, R., Zhang, J., and Wei, L. (2021). Classification and Gene Selection of Triple-Negative Breast Cancer Subtype Embedding Gene Connectivity Matrix in Deep Neural Network. *Brief. Bioinform.* 22, 1477–4054. (Electronic). doi:10.1093/bib/bbaa395

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., et al. (2010). A Versatile Gene-Based Test for Genome-wide Association Studies. *Am. J. Hum. Genet.* 87 (1), 139–145. doi:10.1016/j.ajhg.2010.06.009

Loos, R. J. F. (2020). 15 Years of Genome-wide Association Studies and No Signs of Slowing Down. *Nat. Commun.* 11 (1), 5900. doi:10.1038/s41467-020-19653-5

Luo, J., Meng, Y., Zhai, J., Zhu, Y., Li, Y., and Wu, Y. (2020). Screening of SLE-Susceptible SNPs in One Chinese Family with Systemic Lupus Erythematosus. *Cbio* 15 (7), 778–787. doi:10.2174/1574893615666200120105153

Lyu, P., Hou, J., Yu, H., and Shi, H. (2020). High-density Genetic Linkage Map Construction in Sunflower (Helianthus Annuus L.) Using SNP and SSR Markers. *Curr. Bioinformatics* 15 (8), 889–897. doi:10.2174/1574893615666200324134725

Ma, L., Clark, A. G., and Keinan, A. (2013). Gene-based Testing of Interactions in Association Studies of Quantitative Traits. *Plos Genet.* 9 (2), e1003321. doi:10.1371/journal.pgen.1003321

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the Missing Heritability of Complex Diseases. *Nature* 461 (7265), 747–753. doi:10.1038/nature08494

Moore, J. H., Asselbergs, F. W., and Williams, S. M. (2010). Bioinformatics Challenges for Genome-wide Association Studies. *Bioinformatics* 26 (4), 445–455. doi:10.1093/bioinformatics/btp713

Nobre, R., Ilic, A., Santander-Jimenez, S., and Sousa, L. (2021). Retargeting Tensor Accelerators for Epistasis Detection. *IEEE Trans. Parallel Distrib. Syst.* 32 (9), 2160–2174. doi:10.1109/tpds.2021.3060322

Peng, Q., Zhao, J., and Xue, F. (2010). A Gene-Based Method for Detecting Gene-Gene Co-association in a Case-Control Association Study. *Eur. J. Hum. Genet.* 18 (5), 582–587. doi:10.1038/ejhg.2009.223

Ritchie, M. D., Hahn, L. W., and Moore, J. H. (2003). Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity. *Genet. Epidemiol.* 24 (2), 150–157. doi:10.1002/gepi.10218

Ritchie, M. D., and Van Steen, K. (2018). The Search for Gene-Gene Interactions in Genome-wide Association Studies: Challenges in Abundance of Methods, Practical Considerations, and Biological Interpretation. *Ann. Transl. Med.* 6 (8), 157. doi:10.21037/atm.2018.04.05

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 (3), bbaa144. doi:10.1093/bib/bbaa144

Song, B., Li, F., Liu, Y., and Zeng, X. (2021). Deep Learning Methods for Biomedical Named Entity Recognition: a Survey and Qualitative Comparison. *Brief. Bioinform.* 22 (6), bbab282. doi:10.1093/bib/bbab282

Su, R., Liu, X., Jin, Q., Liu, X., and Wei, L. (2021). Identification of Glioblastoma Molecular Subtype and Prognosis Based on Deep MRI Features. *Knowledge-Based Syst.* 232, 107490. doi:10.1016/j.knosys.2021.107490

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A Deep forest Model to Predict Anti-cancer Drug Response. *Methods* 166, 91–102. doi:10.1016/j.ymeth.2019.02.009

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

Urbanowicz, R. J., Kiralis, J., Sinnott-Armstrong, N. A., Heberling, T., Fisher, J. M., and Moore, J. H. (2012). GAMETES: a Fast, Direct Algorithm for Generating Pure, Strict, Epistatic Models with Random Architectures. *BioData Mining* 5 (1), 16. doi:10.1186/1756-0381-5-16

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., et al. (2010). BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *Am. J. Hum. Genet.* 87 (3), 325–340. doi:10.1016/j.ajhg.2010.07.021

Wang, H. T., Tang, J., 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 (5), bbaa409. doi:10.1093/bib/bbaa409

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., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2019). Exploring Sequence-Based Features for the Improved Prediction of DNA N4-Methylcytosine Sites in Multiple Species. *Bioinformatics* 35 (8), 1326–1333. doi:10.1093/bioinformatics/bty824

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

Xiao, Y., Motomura, S., and Podack, E. R. (2008). APRIL (TNFSF13) Regulates Collagen-Induced Arthritis, IL-17 Production and Th2 Response. *Eur. J. Immunol.* 38 (12), 3450–3458. doi:10.1002/eji.200838640

Young, A. I. (2019). Solving the Missing Heritability Problem. *Plos Genet.* 15 (6), e1008222. doi:10.1371/journal.pgen.1008222

Yuan, Z., Gao, Q., He, Y., Zhang, X., Li, F., Zhao, J., et al. (2012). Detection for Gene-Gene Co-association via Kernel Canonical Correlation Analysis. *BMC Genet.* 13, 83. doi:10.1186/1471-2156-13-83

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and Validation of Disease Genes Using HeteSim Scores. *Ieee/acm Trans. Comput. Biol. Bioinf.* 14 (3), 687–695. doi:10.1109/tcbb.2016.2520947

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a Network-Based Deep Learning Approach to In Silico Drug Repositioning. *Bioinformatics* 35 (24), 5191–5198. doi:10.1093/bioinformatics/btz418

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, T., Hu, Y., Wu, X., Ma, R., Jiang, Q., and Wang, Y. (2016). Identifying Liver Cancer-Related Enhancer SNPs by Integrating GWAS and Histone Modification ChIP-Seq Data. *Biomed. Res. Int.* 2016, 2395341. doi:10.1155/2016/2395341

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. (2019). Meta-path Methods for Prioritizing Candidate Disease miRNAs. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16 (1), 283–291. doi:10.1109/tcbb.2017.2776280

Zhang, Y., and Liu, J. S. (2007). Bayesian Inference of Epistatic Interactions in Case-Control Studies. *Nat. Genet.* 39 (9), 1167–1173. doi:10.1038/ng2110

Zhu, H., Du, X., and Yao, Y. (2020). ConvsPPIS: Identifying Protein-Protein Interaction Sites by an Ensemble Convolutional Neural Network with Feature Graph. *Cbio* 15 (4), 368–378. doi:10.2174/1574893614666191105155713

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.

Copyright © 2021 Guo, Cheng, Yuan, Liang, Wang and Du. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

REVIEW
published: 15 December 2021
doi: 10.3389/fgene.2021.810875

Application of Sparse Representation in Bioinformatics

Shuguang Han[1†], *Ning Wang*[2†], *Yuxin Guo*[1,3], *Furong Tang*[1,4], *Lei Xu*[4], *Ying Ju*[5*] *and Lei Shi*[6*]

[1]*Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China,* [2]*Beidahuang Industry Group General Hospital, Harbin, China,* [3]*School of Mathematics and Statistics, Hainan Normal University, Haikou, China,* [4]*School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China,* [5]*School of Informatics, Xiamen University, Xiamen, China,* [6]*Department of Spine Surgery, Changzheng Hospital, Naval Medical University, Shanghai, China*

Inspired by L1-norm minimization methods, such as basis pursuit, compressed sensing, and Lasso feature selection, in recent years, sparse representation shows up as a novel and potent data processing method and displays powerful superiority. Researchers have not only extended the sparse representation of a signal to image presentation, but also applied the sparsity of vectors to that of matrices. Moreover, sparse representation has been applied to pattern recognition with good results. Because of its multiple advantages, such as insensitivity to noise, strong robustness, less sensitivity to selected features, and no "overfitting" phenomenon, the application of sparse representation in bioinformatics should be studied further. This article reviews the development of sparse representation, and explains its applications in bioinformatics, namely the use of low-rank representation matrices to identify and study cancer molecules, low-rank sparse representations to analyze and process gene expression profiles, and an introduction to related cancers and gene expression profile database.

Keywords: sparse representation, gene expression profile, machine learning, low-rank representation, cancer

OPEN ACCESS

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

Reviewed by:
Liang Cheng,
Harbin Medical University, China
Yongqing Zhang,
Chengdu University of Information Technology, China

***Correspondence:**
Ying Ju
yju@xmu.edu.cn
Lei Shi
slspine@163.com

†These authors have contributed equally to this work

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

Received: *08 November 2021*
Accepted: *01 December 2021*
Published: *15 December 2021*

Citation:
Han S, Wang N, Guo Y, Tang F, Xu L, Ju Y and Shi L (2021) Application of Sparse Representation in Bioinformatics. Front. Genet. 12:810875. doi: 10.3389/fgene.2021.810875

INTRODUCTION

In recent years, inspired by L1-norm minimization methods, such as basis pursuit (Donoho and Huo, 2001), compressed sensing (Candes et al., 2004; Candes and Tao, 2005; Lustig et al., 2007), and Lasso feature selection (Tibshirani, 1996), sparse representation shows up as a novel and potent data processing method. Sparse representation has been applied to pattern recognition, for example, digit recognition, speech recognition, and face recognition, and achieved good results. Hang and Wu (2009) first introduced sparse representation to the analysis of tumor gene expression data. They applied sparse representation to classify two multi-class tumor data, compared them with the classification performance of a support vector machine (SVM), and concluded that sparse representation was superior to SVM. Sparse representation was subsequently adopted for feature selection and the classification of tumor gene expression data. Hang applied it to gene selection and obtained sound classification results (Hang, 2009). Zheng et al. (Gan et al., 2013) proposed a sparse representation classification method based on meta-samples. The method uses singular value decomposition to extract the meta-samples of various training samples, and then uses the meta-samples to linearly represent test samples and categorizes them based on representation coefficients. The test samples compare the classification performance of this method with other classic methods on multiple two-class and multi-class datasets. The experimental results demonstrated that this method is superior to a classic SVM and other methods. These results testify the application potential of sparse representation methods in tumor gene expression data analysis.

The low-rank sparse representation model based on sparse representation has also become a topic of great interest in fields such as machine vision, machine learning, and image processing, and has been applied successfully in video image processing, target recognition, task learning, and recommendation systems (Huang et al., 2017; Yu and Gao, 2019; Liu et al., 2020a; Yu et al., 2020). In low-rank sparse representation theory, a noisy or missing data matrix is decomposed into an accurate data matrix and a singular/sparse data matrix, where the accurate data matrix has low-rank characteristics, and the singular/sparse data matrix contains data noise and singular data (Tang et al., 2020). Wright et al. proposed a classification algorithm based on sparse representation (Wright et al., 2009a) that successfully applies sparse representation theory to face recognition. Meanwhile, researchers have applied the sparsity of vectors to that of matrices, and proposed low-rank matrix recovery theory (Wright et al., 2009b; Emmanuel et al., 2009) and matrix low-rank representation (Liu et al., 2010). Low-rank representation has also received extensive attention from researchers and has become another important data representation method. It has demonstrated great potential. Sparse representation has many advantages, such as insensitivity to noise, strong robustness, insensitivity to selected features, and no "overfitting" phenomenon. Therefore, the application of sparse representation in bioinformatics should be studied further.

In recent years, inspired by discriminant analysis, researchers have combined discriminative ideas with sparse representation or low-rank representation theory to extract discriminative information from samples further to improve recognition performance. Discriminant analysis is a multivariate statistical analysis method that analyzes various characteristic values of sample data, and then discriminates the category of the observed sample. For example, Fisher Linear Discrimination (FLD). The essence of the FLD is to project sample points into a low-dimensional space so that, in the projected space, the distance between sample points of the same category is small and the distance between sample points of varying categories is large.

And because gene expression profile data research plays a vital role in genetic engineering, protein design, new drug development, etc., the use of machine learning methods including deep learning to explore gene expression profile data modeling methods has led to the biological field Wide attention of researchers. At the same time, the innovation of this article are; 1) The low-rank representation (LRR) is modified, and a new type of low-rank representation model is constructed by introducing manifold regularization and class label restriction mechanism, which is used for low-rank scoring of gene features and selecting the optimal gene subset; 2) Introduce the idea of deep learning to the low-rank sparse model, and propose a deep feature representation method for gene expression profile data, and realize the classification and clustering of gene data on this basis; 3) Propose a feature selection mechanism for gene expression profile data based on low-rank graphs; 4) Establish a genetic feature correlation measurement criterion based on low-rank representation coefficients, use this criterion to obtain a new genetic feature selection method, and use Robust Principal Component Analysis (RPCA) and Maximum Interval Criterion (MMC) to build a two-step genetic feature selection method.

DATABASE FOR THE APPLIED RESEARCH OF SPARSE REPRESENTATION

As sparse representation and low-rank representation have been widely applied to the analysis and research of cancer and gene expression profiles in recent years, the databases of cancer and gene expression profiles can be adopted, respectively, for the research and application of sparse representation methods. **Tables 1**, **2** show the specific database description.

APPLICATION OF SPARSE REPRESENTATION IN BIOINFORMATICS

The development of bioinformatics is mainly divided into three stages: gene stage, genomic stage, and post-genomic stage. The first two stages mainly focus on the research of gene sequences (Yu et al., 2019; Cai et al., 2020a; Fu et al., 2020; Wang et al., 2020; Dao et al., 2021a; Dao et al., 2021b; Huang et al., 2021). In the post-genome stage, bioinformatics has entered a new development period, and its research focus has shifted from the study of gene sequences to the study of gene functions (Wang et al., 2013; Dong et al., 2020; Wang et al., 2021a; Lv et al., 2021; Yu et al., 2021). It incorporates all aspects of the process of acquiring, storing, processing, distributing, and explaining biological information, and combines various tools of applied mathematics, computer science, and biology to clarify and understand biological significance in biological data.

Cancer Molecular Study Based on Low-Rank Representation Learning

As a common malignant tumor, cancer is a common fatal disease worldwide because of its complex pathogenic factors, high treatment difficulty, and high risk of recurrence and metastasis. In China, deaths from cancer are always high, and it is a severe threat to the lives and health of Chinese people (Silverberg and Lubera, 1998; Chen et al., 2020). How to prevent and treat cancer effectively has become a topic of widespread concern the world over. With the development of high-throughput sequencing technology, scientists can observe the gene expression of cancer cells at the single-cell level. Feature mining methods for cancer molecules are divided into supervised and unsupervised learning, as shown in **Figure 1**. The supervised method generally includes two steps: 1) First obtain the cancer classification information of the research sample through known prior information or other models. For example, using marker genes, clustering methods, or SNF algorithms. 2) Based on the sample typing information obtained in the previous step, the candidate molecular characteristics are screened out in the training data set, and then these candidate molecular characteristics are classified or survival analysis in the validation data set to determine the final effective molecular

TABLE 1 | Common cancer databases.

Database name	Database introduction
GEO Edgar et al. (2008)	The GEO database stores the records (series, samples, and platforms) provided by the original submitter and the sorted data set, but not all the records provided by the original submitter have been assembled into a selected data set. And the selected data sets form the basis of GEO's advanced data display and analysis functions
TCGA Tomczak et al. (2015)	The Cancer Genome Atlas (TCGA) is a publicly funded project aimed at cataloging and discovering major oncogenic genome changes in order to create a comprehensive "atlas" of cancer genome maps. So far, TCGA researchers have passed large-scale genome sequencing and synthesis Multidimensional analysis analyzed a large cohort of more than 30 human tumors
KEGG Rédei. (2012)	The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledge base for analyzing gene function based on genetic and molecular network systems. KEGG maintains the GENES database and the LIGAND database
COSMIC Forbes et al. (2011)	COSMIC provides comprehensive information about somatic mutations in human cancers. Version v48 (July 2010) describes more than 136,000 coding mutations in nearly 542,000 tumor samples; it aims to collect, manage, organize and present cancer somatic mutations in the world. The information is provided free of charge in a variety of useful ways and can be accessed at http://www.sanger.ac.uk/cosmic
UCSC Cancer Genomics Browser	UCSC Cancer Genomics Browser is a set of web-based tools designed to integrate, visualize and analyze genomic and clinical data. It consists of three main components: hgHeatmap, hgFeatureSorter and hgPathSorter, which can be browsed at https://cancer.cse.ucsc.edu/. And because UCSC Cancer Genomics Browser is an extension of UCSC Genome Browser; therefore, it inherits and integrates the rich human biology and genetics data set of Genome Browser to enhance the interpretability of cancer genomics data
ArrayMapCancer	ArrayMap provides preprocessed tumor genome chip data and CNA maps. In the ArrayMap database, users can search for samples they are interested in, and on this basis, analyze the CNA on the gene or genome fragment of interest

TABLE 2 | Commonly used gene expression profile database.

Name database	The data source	Database introduction
RNA-Seq Atlas	Network-based RNA-Seq gene expression profile and query tool library	This is the first open-access database that provides data mining tools and large-scale RNA-Seq expression profiling. Its application will be multifaceted, because it will help to identify tissue-specific genes and expression profiles, compare gene expression profiles between different tissues, and systems biology methods that link tissue function to changes in gene expression
GEO	The National Center for Biotechnology Information (NCBI) was established	The initial goal was to serve as a public repository for high-throughput gene expression data mainly generated by microarray technology. In addition, the database also includes comparative genome analysis, chromatin immunoprecipitation analysis describing genomic protein interactions, non-coding RNA analysis, SNP genotyping, and genome methylation status analysis
ArrayExpress	Alvis Brazma from EBI et al	It is a functional genomics database under the European Bioinformatics Association (EMBL-EBI), which collects and organizes data from genomics experiments based on microarrays and sequencing to support reproducible research. It is also one of the main knowledge bases for functional genomics experiments based on microarray and high-throughput sequencing. All data is provided in MAGE-TAB format

characteristics. The methods often used in this step mainly include difference hypothesis testing, support vector machine algorithm, random forest and linear discriminant analysis. Another type of unsupervised method does not require the typing information of a given sample set. It is mainly based on model assumptions and related data theories. At the same time, the molecular features and samples in the data are grouped to obtain a molecular set or module, and for the "liveness" value of the sample in the new feature space, commonly used methods include bi-clustering algorithm, matrix decomposition and manifold learning. However, existing unsupervised methods (Chen et al., 2020; Zou et al., 2020) fail to distinguish different feature subspaces. Hence, they may produce errors, or even invalid results, when applied to cancer molecular feature mining. Thus, a low-rank representation learning algorithm (Chen and Yanga, 2014) is presented based on the presumption that the sample subspace exists, and samples in the same subspace can represent each other, while those in different subspaces cannot. The algorithm can accurately identify a "clustered" structure or grouping information of inherent samples in the heterogeneous data. The effectiveness of this method has been widely recognized in image processing, and it also provides new ideas and directions for establishing accurate models for mining cancer molecular characteristics. Therefore, a mathematical model based on low-rank representation can be established by combining multiple scales, including molecules, modules, functional networks, and multi-omics molecular features. This model can be studied from the three aspects described below, and a series of mathematical models that are more in line with the heterogeneous structure of

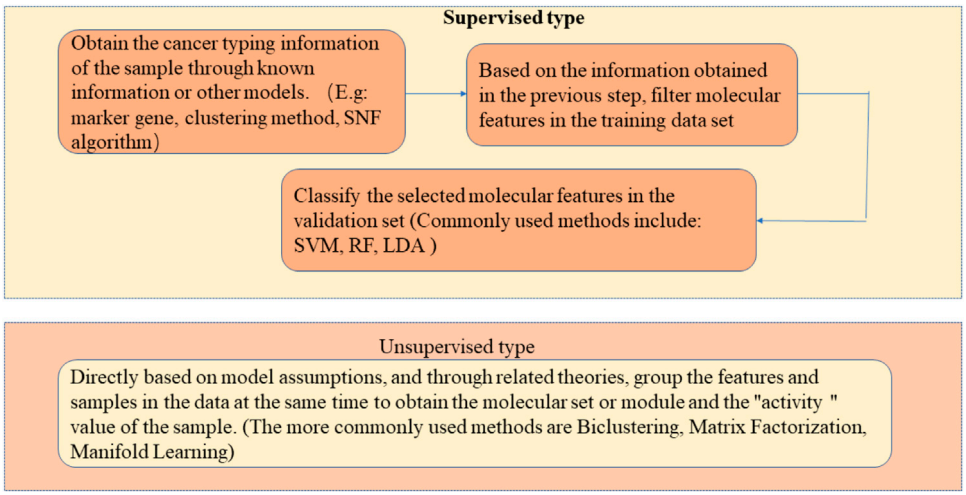


FIGURE 1 | Method for mining cancer molecular features.

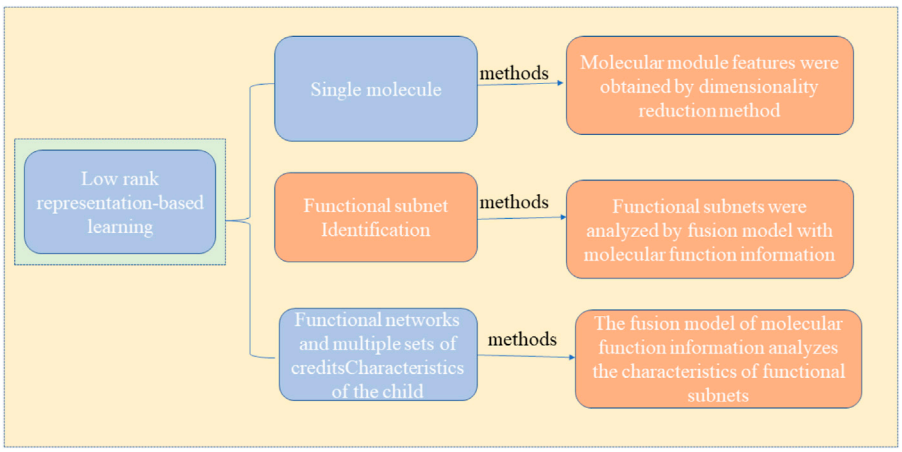


FIGURE 2 | Method for mining cancer molecular features using a low-rank representation matrix.

data and the biological characteristics of the disease are proposed, and a fair evaluation of the validity and practicability of the model is provided using simulated cases and the application of real data, and theoretical modeling and tools for analyzing multi-scale molecular characteristics of cancer are provided. **Figure 2** shows the method of applying a low-rank representation matrix to mine the molecular characteristics of cancer.

1) A dimensionality reduction method is adopted to obtain the characteristics of the molecular module specific to the cancer subtype (Cheng et al., 2018; Tang et al., 2018; Yu et al., 2018; Zhang et al., 2018; Jiang et al., 2019; Su et al., 2019; Liu et al., 2020b; Su et al., 2020). It can address nonlinear sample structure issues that the traditional dimensionality reduction method cannot identify. This is because the dimensionality reduction model fused with low-rank representation learning can process highly heterogeneous data, adaptively capture sample cluster structure and subtype-specific module features, and improve the ability to classify tumor subtypes and obtain reliable molecular modules.
2) The fusion model with molecular function information was used to analyze the characteristics of functional subnets. Makes full use of the advantages of known functional information in biological interpretability (Liu et al., 2019; Cai et al., 2020b), deeply probes into functionally abnormal biological pathways or molecular behaviors, obtains subtype-specific functional subnets, and clarifies the molecular mechanism of cancer from a functional level.
3) A fusion model with molecular function information analyzes the features of functional subnets, makes full use of the biological characteristics of the sample representation relationship consistency of multi-omics data, further explores synergistic or complementary molecular

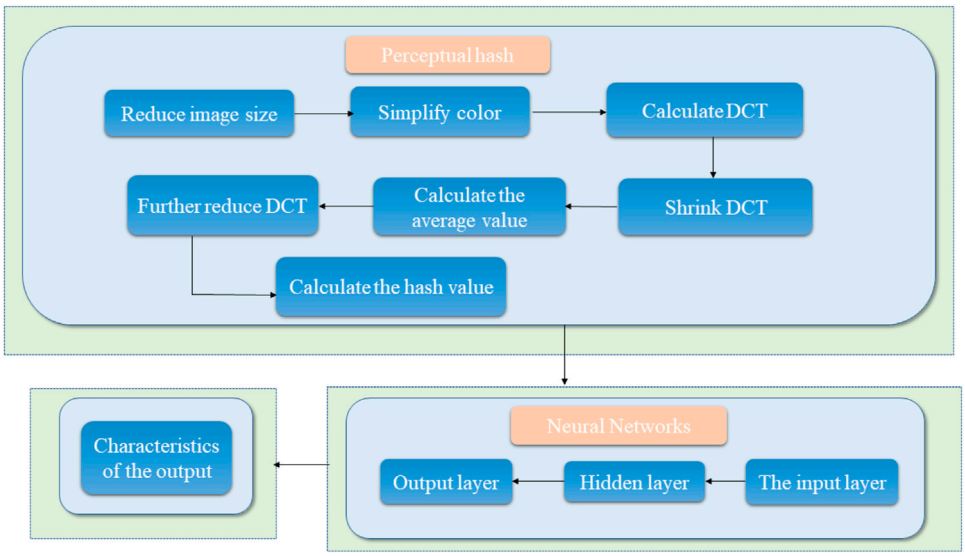


FIGURE 3 | Schematic diagram of the robustness feature acquisition of medical images based on perceptual hashing and a neural network.

characteristic information at the system level, and provides new clue to enable the understanding of the cross-omics pathogenic factors of cancers.

At the same time, medical imaging is also playing an increasingly major role in helping doctors to conduct a precise diagnosis of cancer. Even medical imaging cloud and remote image center can be used for cloud reading, remote consultation, health management, disease diagnosis, image archiving and communication, etc. (Mehto and Mehra, 2016; Ma et al., 2020; Meziane, 2020; Zhang et al., 2021). Therefore, how to protect patients' personal information in medical images, such as CT, MRI, and other medical images, so that this personal information and patients' electronic medical records cannot be leaked has constituted a key issue for the medical industry that needs to be resolved urgently against the background of machine learning cloud computing and big data. Using medical image digital watermarking technology is an effective method to work out this problem (Hong et al, 2016; Vairaprakash and Shenbagavalli, 2017; Shen et al., 2018; Yang et al., 2018; Zhou et al., 2020). Compared with general digital watermarking technology, digital watermarking technology used in medical images, theoretically, should satisfy three characteristics: reliability, availability, and confidentiality.

Common medical image digital watermarking algorithms are divided into three categories: 1) a medical image watermarking algorithm based on non-interest area RONI (Thanki et al., 2017), which diminishes the watermark embedding capacity (Liu et al., 2016b; Gangadhar et al., 2018) and demonstrates poor robustness; 2) reversible digital watermarking; 3) classic conventional digital watermarking algorithms used to process medical images. However, these conventional watermarking algorithms demonstrate poor resistance to geometric attacks; hence, other models that can resist conventional attacks and geometric attacks effectively are necessary. Thus, the design and construction of a new medical image algorithm model based on perceptual hashing technology and neural network technology should be attempted to resolve the contradiction between the robustness and invisibility of medical image digital watermarking. Perceptual hashing mainly resolves the issue of conventional attacks and the neural network mainly resolves geometric attacks. The framework diagram is shown in **Figure 3**. The model process roughly uses the output vector of the hash algorithm as the input vector of the neural network, and finally obtains the output result. Perceptual hashing is a type of hashing algorithm, and its workflow has 7 main steps: 1) Reduce the size, reduce the picture to 8×8 size, a total of 64 pixels; 2) Simplify the color, that is, convert the reduced image to 64-level grayscale; 3) Calculate DCT. DCT is to decompose the frequency of the picture and gather it into a trapezoid shape. Here, a 32×32 DCT transform is used; 4) Reduce the DCT and keep the 8*8 matrix in the upper left corner, showing the lowest frequency in the picture; 5) Calculate the average of all 64 values; 6) To further reduce the DCT, set a 64-bit hash value of 0 or 1 according to the 8×8 DCT matrix, set the value greater than or equal to the average value of DCT to "1", and set the value less than the average value of DCT to "0"; 7) Calculate the hash value. The neural network is a mathematical model or calculation model that imitates the structure and function of a biological neural network. It is calculated by connecting a large number of artificial neurons, mainly including an input layer, a hidden layer and an output layer.

The robustness and invisibility of digital watermark images can be studied from the following perspectives:

1) Regarding anti-conventional attacks, research is based on the extraction of perceptual hashing medical image features in the transform domain. It is used to study the human visual

system, and by combining with perceptual hashing technology, establishes a transform domain perception hash algorithm model, and locates a vector that conforms to the human visual characteristic and is robust against conventional attacks.

2) Regarding anti-geometric attacks, the extraction of medical image features based on perceptual hashing and a neural network is studied. The Osirix DICOM image library and existing medical images are adopted to construct a medical image database that is attacked using nonlinear geometry. Then, the neural network model is designed to train the 2D and 3D medical images after nonlinear geometric attacks, and find the robust feature vectors against nonlinear geometric attacks, which are used as the features of designing robust watermarking algorithms for medical images against geometric attacks.
3) Research on methods for extracting robust perceptual hashing sequences from medical images based on perceptual hashing and neural networks.
4) Regarding research on how to embed large-capacity digital watermarks in medical images, perpetual hashing sequence feature vectors that counter conventional attacks and geometric attacks are used to generate a secret key by combining with the encrypted watermark to complete the embedding and extraction of a large-capacity watermark.

Research on Gene Expression Profile Data Based on Low-Rank Sparse Representation

The emergence of gene expression profile data helps the understanding of the pathological process of cancer cells at the molecular level. Tens of thousands of varying genes in tissue samples can be detected by gene chips, and then the gene chip expression profile data can be analyzed and processed. Thus, tumors are classified so that patients can be treated effectively. However, gene expression profiles are characterized by high dimensionality, large noise, a small number of gene samples, missing data, data redundancy, and an unbalanced distribution of class samples. Thus, advanced data modeling methods must be used to extract the classification characteristics of samples effectively from tens of thousands of gene expression profiles. With the rapid development of artificial intelligence and machine learning in speech and machine vision in recent years, the use of machine learning methods, including deep learning, to explore gene expression profile data modeling methods is destined to be a development trend in the future.

Presently, research on gene expression profiles mainly covers the following: 1) the preprocessing of gene expression profile data, 2) extraction of gene expression profile data features, 3) selection of gene expression profile data features, and 4) clustering and classification research of gene expression profile data. Common gene feature selection methods are categorized into three types: the filter method, wrapper method, and embedded method (Bolón-Canedo et al., 2014). They can also be based on low-rank scoring, low-rank representation coefficient-based gene feature correlation measurement, and a two-step method based on robust principal component analysis (RPCA) (Partridge and Jabri, 2002) and the maximum margin criterion (MMC) for feature selection. RPCA, low-rank representation (Shu et al., 2017), and matrix completion (Cao et al., 2011; Zeng et al., 2017; Liu et al., 2020c; Ran et al., 2020; Zhao et al., 2020) are three main research areas for low-rank sparse theory. As the name implies, sparse representation refers to a linear combination of fewer basic signals to express most or all of the original signal. Among them, these basic signals are called atoms, which are selected from the over-complete dictionary; and the over-complete dictionary is gathered from atoms whose number exceeds the signal dimension. Therefore, it can be seen that any signal has different sparse representations under different atom groups. For example, a $M \times N$ matrix is used to represent the data set X, each row represents a sample, and each column represents an attribute of the sample. Generally speaking, the matrix is dense, that is, most elements are not 0. The meaning of sparse representation is to find a coefficient matrix $A(K \times N)$ and a dictionary matrix $B(M \times K)$, so that $B \times A$ restores X as much as possible, and A is as sparse as possible. A is the sparse representation of X.

Low-rank sparse representation models have been applied in many fields (Cheng et al., 2016; Chen et al., 2017; Zhang et al., 2017; Brbic and Kopriva, 2018; Chen et al., 2018; Xie et al., 2018; Yuanyuan et al., 2018; Zeng et al., 2018; Ding et al., 2019; Shen et al., 2019; Zhang et al., 2019; Li et al., 2020; Wu and Yu, 2021), which demonstrate high superiority, particularly in terms of dimensionality reduction and subspace segmentation. Considering existing analysis methods, introduce a low-rank sparse representation model for gene expression profile data analysis, several new methods for feature selection and feature extraction of gene expression profile data based on low-rank sparse representation models are explored, and they are applied to gene expression profile clustering and classification. As shown in **Figure 4**, this section mainly uses the following process to study gene expression profile data based on low-rank sparse representation analysis. In typical cases, the following three specific research areas are mainly involved when studying gene expression profile data.

1) Estimation of missing points in gene expression profile data.

In recent years, missing point estimation methods have included the following: 1) list deletion method; 2) duplicate value filling; 3) average value substitution method; and 4) the use of statistical methods for estimation, such as K-nearest neighbor (KNN) (Olga et al., 2001), singular value decomposition, and local least squares.

2) Feature selection for gene expression profile data.

Feature selection is a major prerequisite for the classification and clustering of gene expression profile data (Lu and Zhao, 2019; Zou et al., 2020; Qi et al., 2021a; Zulfiqar et al., 2021). Three common gene feature selection methods exist: the filter method, wrapper method, and embedded method. And Low-rank scoring, gene feature correlation measurement based on a low-rank representation coefficient, and a two-step method based on

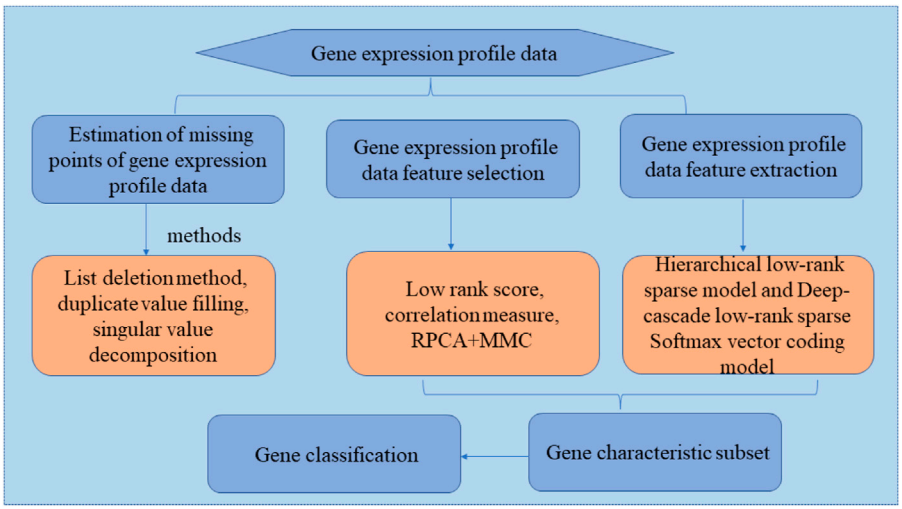


FIGURE 4 | Research procedure for gene database analysis based on low-rank sparse representation.

RPCA and MMC can also be used to select features. To overcome the shortcomings of traditional low-rank representation models, feature selection introduces manifold regularization constraints and class-label information constraints, sets up a manifold regularized low-rank representation model and a class-label constrained low-rank representation model, and solve the low-rank representation coefficient matrix in the two models. On this basis, two different low-rank graphs are set up, the low-rank graphs are used to score each gene feature, and a set of optimal gene feature subsets is selected according to the score.

3) Gene expression profile data feature extraction.

Common feature extraction methods can be divided into linear and nonlinear transformations. Typical linear feature extraction algorithms include sparse principal component analysis (PCA) (Min et al., 2018; Islam et al., 2020), independent component analysis (Moysés et al., 2017), and LDA. Nonlinear transformation methods primarily include neural networks, kernel methods (Qi et al., 2021b), manifold learning (Shen et al., 2017), sparse representation (Min et al., 2017), and matrix factorization methods (Wang et al., 2017; Yang et al., 2017; Yang and Hu, 2017; McCall et al., 2019). With the continuous development of machine learning and data mining, new feature extraction methods continue to arise. For example, PCA, FA, and ICA are three characteristic methods commonly used in gene expression profile data mining.

Gene expression profile data analysis has attracted widespread attention from scholars, and a series of gene expression profile analysis methods have been proposed. Classic methods such as PCA, LDA, KNN, decision-making tree method, ensemble learning, SVM, extreme learning machine, neural network, sparse representation, and gene bi-clustering method based on qualitative/quantitative measurement have been widely applied to the classification and clustering of gene expression profile data. Meanwhile, these technologies can provide techniques and comparisons for low-rank sparse representation methods. The core of the low-rank sparse representation method is low-rank sparse modeling theory. As an effective tool for large-scale data analysis, this theory has made great progress in recent years. Additionally, it has been widely used in subspace segmentation, image processing and recognition, machine vision, system modeling and control, and other large-scale data analysis.

CONCLUSION

Therefore, it has become an inevitable trend to apply low-rank sparse representation models to study them. Low-rank sparse representation models have been applied in multiple fields, particularly in dimensionality reduction and subspace segmentation. For example, in feature extraction, traditional graph-based learning algorithm feature extraction methods are constrained using a graph construction method, and the effectiveness of the extracted feature vectors is reduced. By contrast, low-rank graphs have better local and global data description capabilities. A dimensionality reduction method based on low-rank graphs is a more effective feature extraction method. Moreover, with the advancement of biological sequencing technology, scientists have been able to observe the gene expression of cancer cells at the single-cell level, and discovered that the heterogeneity of cancer tissue far exceeds previous estimates. However, so far, low-rank sparse representation models are rarely used for gene data analysis. Therefore, this article introduces low-rank sparse representation models for gene expression profile data analysis based on existing analysis methods. Discuss new methods for feature selection and feature extraction of gene expression profile data based on low-rank sparse representation model, and use it for gene expression profile clustering and classification.

At the same time, with the advancement of biological sequencing technology, scientists have been able to observe the

gene expression of cancer cells at the single-cell level, and found that the heterogeneity of cancer tissues far exceeds previous estimates. The observation samples of potential strongly heterogeneous data are likely to be in multiple feature subspaces. Each subspace is composed of the same set of molecular features that represent the same cancer class (subtype), and samples from different subspaces belong to different cancer class (subtype). However, many unsupervised methods proposed before cannot distinguish different feature subspaces, so errors or even invalid results may occur when these methods are used for cancer molecular feature mining. After research, it is found that the low-rank representation learning algorithm can accurately identify the inherent sample "cluster" structure or grouping information in heterogeneous data. The algorithm assumes that the sample subspace exists, and samples in the same subspace can characterize each other. Samples in different subspaces cannot characterize each other. Moreover, the effectiveness of this algorithm has been widely recognized in the field of image processing, and it also provides us with new ideas and new directions for establishing accurate models for mining cancer molecular characteristics.

PENDING ISSUES AND PROSPECTS

Gene expression profile data analysis has attracted widespread attention from scholars at home and abroad. Not only have they proposed a series of gene expression profile analysis methods, but they also developed a variety of gene software based on gene public databases, such as EASE network platform, pathway analysis software Gen-MAPP2 and the development of the domestically developed pathway analysis platform KOBAS, the development of these software provides a basis for the subsequent further research on gene expression profiles.

This article mainly uses low-rank sparse modeling theory to analyze experimental data. As one of the effective tools for large-scale data analysis, this theory has been widely used in different aspects in recent years. For example, sparse representation has been applied to the field of pattern recognition and has yielded fruitful results. The low-rank sparse representation model based on sparse representation has also become a research focus in machine vision, machine learning, and image processing, and has been applied successfully in video image processing, target recognition, task learning, bioinformatics (Ding et al., 2020; Hong et al., 2020; Hu et al., 2020; Lu et al., 2020; Hu et al., 2021a; Hu et al., 2021b; Wang et al., 2021b), and recommendation systems (Wei et al., 2014; Wei et al., 2017a; Wei et al., 2017b). However, further attention should be paid to low-rank representation learning. In specific applications, LRR generally uses original data as a dictionary, which requires a sufficient number of observed data samples, and only part of the data in the dictionary can be damaged. In real-world scenarios, the aforementioned assumptions may not be tenable; hence, LatLRR can be considered, and a dictionary can be constructed using observed and unobserved data.

At the same time, sparse representation also has important clinical significance. For example, data released by the National Cancer Center reveal that there are approximately 4.29 million new cancer patients in China every year, which accounts for 20% of new cases globally, and deaths have reached 2.81 million. Approximately 10,000 patients are diagnosed with cancer in China every day, that is, one patient every 7 min. Therefore, the prevention and treatment of cancers are not optimistic. It is expected that the incidence of cancers will continue to rise in the next one or two decades. The high incidence of cancer cases has resulted in severe challenges to domestic economic development and residents' healthy life. How to prevent and treat cancer effectively has become a topic of great interest worldwide. With the advancement of high-throughput technology, biomedicine is rapidly stepping into the era of big data. Omics data represented by gene expression profiles have demonstrated particular leaps. The emergence of gene expression profile data helps people to understand the pathological process of cancer cells at the molecular level. Thousands of genes in tissue samples can be detected by gene chips, and then the tumor can be classified by analyzing and processing the gene chip expression profile data so that patients can be treated effectively. However, because of the characteristics of gene expression profile data, there are still many problems in the research field. With the rapid development of artificial intelligence and machine learning in the field of speech and machine vision, in the next few years, artificial intelligence and machine learning will play an increasingly important role in genetic biology, genomic medicine and precision medicine, especially deep learning. The rapid development has attracted widespread attention from researchers in the biomedical field, so it has become an inevitable trend to use low-rank sparse representation models to study them. An extensive application of sparse representation in bioinformatics helps to address the problem that some unsupervised algorithms cannot distinguish different feature subspaces of cancer molecules. Moreover, it is expected that, in the near future, it can provide technological references for the prevention and treatment of critical illness, and the research and development of new drugs.

However, sparse representation in bioinformatics still has varying degrees of limitations. For example: 1) Constructing a more flexible sparse representation model. In the existing sparse representation model, there is an objective function and a constraint function, the objective function is generally to minimize the energy of the noise under the assumption that the observation signal has a linear model form and contains Gaussian white noise, constraint function generally refers to sparse constraint term. On the one hand, this objective function treats the sparse components equally; on the other hand, it ignores the existence of other goals in different applications, because if you look at it from the standpoint of representation alone, it does not necessarily require the sparsest solution to be unique or the sparsest solution is not the most ideal. Therefore, it is necessary to construct a sparse representation model with multiple targets and variable regular parameters to meet the characteristics and needs of more application problems. 2) When determining the regular parameter λ and the parameter k representing the degree of sparseness for the model, a manual pre-determined method is generally used to assign values to the

two hyperparameters. After determining its value, perform the solution, and then compare the solution result with the target demand. If it does not meet the requirements, then adjust the parameters. This inevitably results in non-adaptability or non-automation of the solution process, and also limits the application of sparse representation methods in some fields that require a high degree of automation. Therefore, it is necessary to study the adaptive solution of sparse representation model, and construct the functional relationship between hyperparameters and observation signals and sparse vectors. 3) At present, the application scope of sparse representation is mainly limited to the field of natural signals. The application prospects in the field of unnatural data signals are still unclear. According to the characteristics of sparse representation in various fields, the application types of sparse representation can be divided into reconstruction based Applications and classification-based applications. Reconstruction-based applications mainly include image denoising, image signal reconstruction, audio signal recovery, compressed sensing, SAR imaging, etc. The common point of this category of applications is that the characteristics of the target signal need to be obtained first, and the sparse vector is constructed using the characteristics. The mathematical model in the sparse representation theory is then used to solve the problem to achieve the effect of reconstructing the original signal within the allowable error range. Classification-based applications mainly include face recognition, target tracking, text detection, blind source separation, etc. Classification-based applications all construct sparse feature vectors by extracting feature information from objects. These feature vectors are strongly distinguishable and can differentiate different types of signals, and then according to the optimization method of sparse representation, determine the distance between the target signal and these feature vectors, and when a certain threshold is met, it is determined to belong to the category to achieve the effect of pattern recognition and classification. Therefore, sparse representation has some limitations in the application of bioinformatics, which requires further research and discussion by scholars.

At the same time, sparse representation provides a powerful means in blind source separation technology, because blind source separation technology is to solve the unknown input and unknown transmission channel and output the known signal processing technology. The sparse representation technology reduces the complexity of the algorithm by separating the estimation process of the mixing matrix and the estimation process of the source signal, and improves the accuracy of the source signal separation. Therefore, sparse representation has become a popular method in the current blind source separation problem.

AUTHOR CONTRIBUTIONS

Conceptualization, LS and YJ; data collection or analysis, YG and LX; validation, YG and FT; writing—original draft preparation, SH and NW; writing—review and editing, NW and SH. All authors have read and agreed to the published version of the manuscript.

FUNDING

The work was supported by the National Key R and D Program of China (No. 2020YFB2104400), the Special Science Foundation of Quzhou (2021D004), and the research foundation of Shenzhen Polytechnic (No. 6021310016K).

ACKNOWLEDGMENTS

We thank Maxine Garcia, PhD, from Liwen Bianji (Edanz) (www.liwenbianji.cn/) for editing the English text of a draft of this manuscript. As well as thanks to the guidance of the tutor and the joint efforts of other authors, the success of this article is the result of everyone's joint efforts.

REFERENCES

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., and José, B. (2014). A Review of Microarray Datasets and Applied Feature Selection Methods. *Inf. Sci. Int. J.* 282, 111–135. doi:10.1016/j.ins.2014.05.042

Brbic, M., and Kopriva, I. (2018). Multi-view Low-Rank Sparse Subspace Clustering. *Pattern Recognition J. Pattern Recognition Soc.* 73, 247–258. doi:10.1016/j.patcog.2017.08.024

Cai, L., Ren, X., Fu, X., Peng, L., Gao, M., Zeng, X., et al. (2020). Interpretable Sequence-Based Enhancers and Their Strength Predictor. *Bioinformatics* 37 (8), 1060–1067. doi:10.1093/bioinformatics/btaa914

Cai, L., Wang, L., Fu, X., Xia, C., Zeng, X., and Zou, Q. (2020). 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

Candes, E. J., and Tao, T. (2005). Decoding by Linear Programming. *IEEE Trans. Inform. Theor.* 51 (12), 4203–4215. doi:10.1109/tit.2005.858979

Candes, E., Romberg, J., and Tao, T. (2004). *Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information*.

Cao, F., Cai, M. M., and Tan, Y. (2011). *Image Interpolation via Low-Rank Matrix Completion and Recovery, International Workshop on Java Technologies for Real-Time & Embedded Systems*.

Chen, B., Yang, Z., and Yang, Z. (2018). An Algorithm for Low-Rank Matrix Factorization and its Applications. *Neurocomputing* 275, 1012–1020. doi:10.1016/j.neucom.2017.09.052

Chen, J., Mao, H., Sang, Y., and Yi, Z. (2017). Subspace Clustering Using a Symmetric Low-Rank Representation. *Knowledge-Based Syst.* 127, 46–57. doi:10.1016/j.knosys.2017.02.031

Chen, J., and Yanga, J. (2014). Robust Subspace Segmentation via Low-Rank Representation. *IEEE Trans. Cybernetics* 44 (8), 1432.

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., 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

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., 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

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. Bioinformatics* 22 (4), bbaa356. doi:10.1093/bib/bbab047

Dao, F. Y., Lv, H., Zhang, D., Zhang, Z. M., Liu, L., and Lin, H. (2021). DeepYY1: a Deep Learning Approach to Identify YY1-Mediated Chromatin Loops. *Brief. Bioinformatics* 22 (4). doi:10.1093/bib/bbaa356

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

Dong, L., Wang, J., and Wang, G. (2020). BYASE: a Python Library for Estimating Gene and Isoform Level Allele-specific Expression. *Bioinformatics* 36 (19), 4955–4956. doi:10.1093/bioinformatics/btaa636

Donoho, D. L., and Huo, X. (2001). Uncertainty Principles and Ideal Atomic Decomposition. *IEEE Trans. Inform. Theor.* 47 (7), 2845–2862. doi:10.1109/18.959265

Edgar, R., Domrachev, M., and Lash, A. E. (2008). *Gene Expression Omnibus*. Springer Netherlands.

Emmanuel, J., Cande`, E., and Xiaodong, L. I. (2009). *Robust Principal Component Analysis?*

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., et al. (2011). COSMIC: Mining Complete Cancer Genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* 39, D945–D950. doi:10.1093/nar/gkq929

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

Gan, B., Zheng, C. H., and Liu, J. X. (2013). "Metasample-Based Robust Sparse Representation for Tumor Classification," in International Conference on Biomedical Engineering(ICBE).

Gangadhar, Y., Giridhar Akula, V. S., and Reddy, P. C. (2018). An Evolutionary Programming Approach for Securing Medical Images Using Watermarking Scheme in Invariant Discrete Wavelet Transformation. *Biomed. Signal Process. Control.* 43, 31–40. doi:10.1016/j.bspc.2018.02.007

Hang, X., and Wu, F. X. (2009). Sparse Representation for Classification of Tumors Using Gene Expression Data. *J. Biomed. Biotechnol.* 2009 (10), 403689. doi:10.1155/2009/403689

Hang, X. (2009). "Multiclass Gene Selection on Microarray Data Using L1-Norm Least Square Regression," in International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, 52–55. doi:10.1109/IJCBS.2009.76

Hong, L., Xiao, Di., Zhang, R., Zhang, Y., and Bai, S. (2016). Robust and Hierarchical Watermarking of Encrypted Images Based on Compressive Sensing. *Signal. Process. Image Commun. A Publ. Eur. Assoc. Signal Process.* 45, 41–51. doi:10.1016/j.image.2016.04.002

Hong, Q., Yan, R., Wang, C., and Sun, J. (2020). Memristive Circuit Implementation of Biological Nonassociative Learning Mechanism and its Applications. *IEEE Trans. Biomed. Circuits Syst.* 14 (5), 1036–1050. doi:10.1109/tbcas.2020.3018777

Hu, Y., Qiu, S., and Cheng, L. (2021). Integration of Multiple-Omics Data to Analyze the Population-specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* 2021, 7036592. doi:10.1155/2021/7036592

Hu, Y., Sun, J. Y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (2021). rs1990622 Variant Associates with Alzheimer's Disease and Regulates TMEM106B Expression in Human Brain Tissues. *BMC Med.* 19 (1), 11. doi:10.1186/s12916-020-01883-5

Hu, Y., Zhang, H., Liu, B., Gao, S., Wang, T., Han, Z., et al. (2020). rs34331204 Regulates TSPAN13 Expression and Contributes to Alzheimer's Disease with Sex Differences. *Brain* 143 (11), e95. doi:10.1093/brain/awaa302

Huang, L., Li, X., Guo, P., Yao, Y., Liao, B., Zhang, W., et al. (2017). Matrix Completion with Side Information and its Applications in Predicting the Antigenicity of Influenza Viruses. *Bioinformatics* 33 (20), 3195–3201. doi:10.1093/bioinformatics/btx390

Huang, S., He, X., Wang, G., and Bao, E. (2021). AlignGraph2: Similar Genome-Assisted Reassembly Pipeline for PacBio Long Reads. *Brief Bioinform* 22 (5), bbab022. doi:10.1093/bib/bbab022

Islam, M. A., Kundu, S., and Hassan, R. (2020). Gene Therapy Approaches in an Autoimmune Demyelinating Disease: Multiple Sclerosis. *Cgt* 19 (6), 376–385. doi:10.2174/1566523220666200306092556

Jiang, L., Xiao, Y., Ding, Y., and Tang, J. (2019). Discovering Cancer Subtypes via an Accurate Fusion Strategy on Multiple Profile Data. *Front. Genet.* 10, 20. doi:10.3389/fgene.2019.00020

Li, J., Chang, M., Gao, Q., Song, X., and Gao, Z. (2020). Lung Cancer Classification and Gene Selection by Combining Affinity Propagation Clustering and Sparse Group Lasso. *Cbio* 15 (7), 703–712. doi:10.2174/1574893614666191017103557

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Mol. Ther. - Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. (2010). *Robust Recovery of Subspace Structures by Low-Rank Representation*.

Liu, L., Li, Q.-Z., Jin, W., Lv, H., and Lin, H. (2019). Revealing Gene Function and Transcription Relationship by Reconstructing Gene-Level Chromatin Interaction. *Comput. Struct. Biotechnol. J.* 17, 195–205. doi:10.1016/j.csbj.2019.01.011

Liu, Y., Yu, Z., Chen, C., Han, Y., and Yu, B. (2020). Prediction of Protein Crotonylation Sites through LightGBM Classifier Based on SMOTE and Elastic Net. *Anal. Biochem.* 609, 113903. doi:10.1016/j.ab.2020.113903

Liu, Y., Qu, X., and Xin, G. (2016). A ROI-Based Reversible Data Hiding Scheme in Encrypted Medical Images. *J. Vis. Commun. Image Representation* 39, 51–57. doi:10.1016/j.jvcir.2016.05.008

Liu, Y., Wen, Z., and Li, M. (2020). The Power of Matrix Factorization: Methods for Deconvoluting Genetic Heterogeneous Data at Expression Level. *Curr. Bioinformatics* 15 (8), 841–853.

Lu, X.-X., and Zhao, S.-Z. (2019). Gene-based Therapeutic Tools in the Treatment of Cornea Disease. *Cgt* 19 (1), 7–19. doi:10.2174/1566523219666181213120634

Lu, X., Wang, X., Ding, L., Li, J., Gao, Y., and He, K. (2020). frDriver: A Functional Region Driver Identification for Protein Sequence. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 18 (5), 1773–1783. doi:10.1109/TCBB.2020.3020096

Lustig, M., Donoho, D., and Pauly, J. M. (2007). Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging. *Magn. Reson. Med.* 58 (6), 1182–1195. doi:10.1002/mrm.21391

Lv, H., Dao, F. Y., Zulfiqar, H., Su, W., Ding, H., Liu, L., et al. (2021). A Sequence-Based Deep Learning Approach to Predict CTCF-Mediated Chromatin Loop. *Brief. Bioinformatics* 22 (5), bbab031. doi:10.1093/bib/bbab031

Ma, X., Xi, B., Zhang, Y., Zhu, L., Sui, X., Tian, G., et al. (2020). A Machine Learning-Based Diagnosis of Thyroid Cancer Using Thyroid Nodules Ultrasound Images. *Cbio* 15 (4), 349–358. doi:10.2174/1574893614666191017091959

McCall, A. L., Stankov, S. G., Cowen, G., Cloutier, D., Zhang, Z., Yang, L., et al. (2019). Reduction of Autophagic Accumulation in Pompe Disease Mouse Model Following Gene Therapy. *Cgt* 19 (3), 197–207. doi:10.2174/1566523219666190621113807

Mehto, A., and Mehra, N. (2016). Adaptive Lossless Medical Image Watermarking Algorithm Based on DCT & DWT. *Proced. Comp. Sci.* 78, 88–94. doi:10.1016/j.procs.2016.02.015

Meziane, B. (2020). A Self-Sustained Oscillator to the Lorenz-Haken Dynamics. *Physica Scripta* 95 (5). doi:10.1088/1402-4896/ab6e4c

Min, C., He, X., Shao, B. D., and Ying, W. D. (2017). A Novel Gene Selection Method Based on Sparse Representation and Max-Relevance and Min-Redundancy. *Comb. Chem. High Throughput Screen.* 20 (999). doi:10.2174/1386207320666170126114051

Min, W., Liu, J., and Zhang, S. (2018). Edge-group Sparse PCA for Network-Guided High Dimensional Data Analysis. *Bioinformatics* 34 (20), 3479–3487. doi:10.1093/bioinformatics/bty362

Moysés, N., Sff, E., Thelma, S., Campana, N., Maciel, F., Azevedo, B., et al. (2017). Independent Component Analysis (ICA) Based-Clustering of Temporal RNA-Seq Data. *Plos One* 12 (7), e0181195. doi:10.1371/journal.pone.0181195

Olga, T., Michael, C., Gavin, S., Pat, B., Trevor, H., Robert, T., et al. (2001). Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics* 17 (6), 520–525. doi:10.1093/bioinformatics/17.6.520

Partridge, M., and Jabri, M. (2002). , 1, 289–298. doi:10.1109/NNSP.2000.889420Robust Principal Component Analysis*Neural Networks Signal. Process. X, IEEE Signal. Process. Soc. Workshop*

Qi, C., Wang, C., Zhao, L., Zhu, Z., Wang, P., Zhang, S., et al. (2021). SCovid: Single-Cell Atlases for Exposing Molecular Characteristics of COVID-19 across 10 Human Tissues. *Nucleic Acids Res.* doi:10.1093/nar/gkab881

Qi, R., Wu, J., Guo, F., Xu, L., and Zou, Q. (2021). A Spectral Clustering with Self-Weighted Multiple Kernel Learning Method for Single-Cell RNA-Seq Data. *Brief Bioinform* 22 (4), bbaa216. doi:10.1093/bib/bbaa216

Ran, W., Chen, X., Wang, B., Yang, P., Li, Y., Xiao, Y., et al. (2020). Whole-exome Sequencing of Tumor-Only Samples Reveals the Association between Somatic Alterations and Clinical Features in Pancreatic Cancer. *Curr. Bioinformatics* 15 (10), 1160–1167. doi:10.2174/1574893615999200626190346

Rédei, G. (2012). *Kyoto Encyclopedia of Genes and Genomes*.

Shen, C., Ding, Y., Tang, J., Xu, X., and Guo, F. (2017). An Ameliorated Prediction of Drug-Target Interactions Based on Multi-Scale Discrete Wavelet Transform and Network Features. *Ijms* 18 (8), 1781. doi:10.3390/ijms18081781

Shen, M., Ma, B., Zhu, L., Mijumbi, R., Du, X., and Hu, J. (2018). Cloud-Based Approximate Constrained Shortest Distance Queries over Encrypted Graphs with Privacy Protection. *IEEE Trans. Inf. Forensics Security* 13 (4), 940–953. doi:10.1109/TIFS.2017.2774451

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

Shu, Z., Fan, H., Huang, P., Wu, D., Ye, F., and Wu, X. (2017). Multiple Laplacian Graph Regularised Low-rank Representation with Application to Image Representation. *Iet Image Process.* 11 (6), 370–378. doi:10.1049/iet-ipr.2016.0391

Silverberg, E., and Lubera, J. A. (1998). Cancer Statistics, 1989. *Ca Cancer J. Clin.* 39 (1), 3–20. doi:10.3322/canjclin.39.1.3

Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2020). Empirical Comparison and Analysis of Web-Based Cell-Penetrating Peptide Prediction Tools. *Brief. Bioinformatics* 21 (2), 408–420. doi:10.1093/bib/bby124

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A Deep forest Model to Predict Anti-cancer Drug Response. *Methods* 166, 91–102. doi:10.1016/j.ymeth.2019.02.009

Tang, H., Zhao, Y.-W., Zou, P., Zhang, C.-M., Chen, R., Huang, P., et al. (2018). HBPred: a Tool to Identify Growth Hormone-Binding Proteins. *Int. J. Biol. Sci.* 14 (8), 957–964. doi:10.7150/ijbs.24174

Tang, X., Cai, L., Meng, Y., Xu, J., Lu, C., and Yang, J. (2020). Indicator Regularized Non-negative Matrix Factorization Method-Based Drug Repurposing for COVID-19. *Front. Immunol.* 11, 603615. doi:10.3389/fimmu.2020.603615

Thanki, R., Borra, S., Dwivedi, V., and Borisagar, K. (2017). A RONI Based Visible Watermarking Approach for Medical Image Authentication. *J. Med. Syst.* 41 (9), 143. doi:10.1007/s10916-017-0795-3

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* 58 (1). doi:10.1111/j.2517-6161.1996.tb02080.x

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemp. Oncol. (Pozn)* 19 (1A), A68–A77. doi:10.5114/wo.2014.47136

Vairaprakash, S., and Shenbagavalli, A., A Discrete Rajan Transform-Based Robustness Improvement Encrypted Watermark Scheme Backed by Support Vector Machine ☆. *Comput. Electr. Eng.* 2017, 70: 826–843. doi:10.1016/j.compeleceng.2017.12.029

Wang, D., Zhang, Z., Jiang, Y., Mao, Z., Wang, D., Lin, H., et al. (2021). DM3Loc: Multi-Label mRNA Subcellular Localization Prediction and Analysis Based on Multi-Head Self-Attention Mechanism. *Nucleic Acids Res.* 49 (8), e46. doi:10.1093/nar/gkab016

Wang, G., Qi, K., Zhao, Y., Li, Y., Juan, L., Teng, M., et al. (2013). Identification of Regulatory Regions of Bidirectional Genes in Cervical Cancer. *BMC Med. Genomics* 6 (Suppl. 1), S5. doi:10.1186/1755-8794-6-S1-S5

Wang, H., Tang, J., 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. Bioinformatics* 22 (5), bbaa409. doi:10.1093/bib/bbaa409

Wang, J., Chen, S., Dong, L., and Wang, G. (2020). CHTKC: a Robust and Efficient K-Mer Counting Algorithm Based on a Lock-free Chaining Hash Table. *Brief Bioinform* 22.

Wang, J., Liu, J. X., Zheng, C. H., Wang, Y. X., Kong, X. Z., and Weng, C. G. (2017). A Mixed-Norm Laplacian Regularized Low-Rank Representation Method for Tumor Samples Clustering. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 16 (1), 172–182. doi:10.1109/TCBB.2017.2769647

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., 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

Wright, J., Ganesh, A., Rao, S., and Ma, Y. (2009). *Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices*. IEEE.

Wright, J., Yang, A., Ganesh, A., and Sastry, S. (2009). Robust Face Recognition via Sparse Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (2), 210–227. doi:10.1109/tpami.2008.79

Wu, X., and Yu, L. (2021). EPSOL: Sequence-Based Protein Solubility Prediction Using Multidimensional Embedding. *Bioinformatics* 37 (23), 4314–4350. doi:10.1093/bioinformatics/btab463

Xie, L., Yin, M., Yin, X., Liu, Y., and Yin, G. (2018). Low-Rank Sparse Preserving Projections for Dimensionality Reduction. *IEEE Trans. Image Process.* 27, 5261–5274. doi:10.1109/TIP.2018.2855426

Yang, G., and Hu, Z. (2017). Gene Feature Extraction Based on Nonnegative Dual Graph Regularized Latent Low-Rank Representation. *Biomed. Res. Int.* 2017, 1–8. doi:10.1155/2017/1096028

Yang, H., Yin, J., and Jiang, M. (2018). Perceptual Image Hashing Using Latent Low-Rank Representation and Uniform LBP. *Appl. Sci.* 8 (2), 317. doi:10.3390/app8020317

Yang, J., Hagen, J., Guntur, K. V., Allette, K., Schuyler, S., Ranjan, J., et al. (2017). A Next Generation Sequencing Based Approach to Identify Extracellular Vesicle Mediated mRNA Transfers between Cells. *BMC Genomics* 18 (1), 987. doi:10.1186/s12864-017-4359-1

Yu, L., Yao, S., Gao, L., and Zha, Y. (2019). Conserved Disease Modules Extracted from Multilayer Heterogeneous Disease and Gene Networks for Understanding Disease Mechanisms and Predicting Disease Treatments. *Front. Genet.* 9, 745. doi:10.3389/fgene.2018.00745

Yu, L., and Gao, L. (2019). Human Pathway-Based Disease Network. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16 (4), 1240–1249. doi:10.1109/tcbb.2017.2774802

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, 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., Zhao, J., and Gao, L. (2018). Predicting Potential Drugs for Breast Cancer Based on miRNA and Tissue Specificity. *Int. J. Biol. Sci.* 14 (8), 971–982. doi:10.7150/ijbs.23350

Yuanyuan, C., Lei, Z., and Zhang, Yi. (2018). Subspace Clustering Using a Low-Rank Constrained Autoencoder. *Inf. Sci. Int. J.* 424, 27–38. doi:10.1016/j.ins.2017.09.047

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and Validation of Disease Genes Using HeteSim Scores. *Ieee/acm Trans. Comput. Biol. Bioinf.* 14 (3), 687–695. doi:10.1109/tcbb.2016.2520947

Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of Potential Disease-Associated microRNAs Using Structural Perturbation Method. *Bioinformatics* 34 (14), 2425–2432. doi:10.1093/bioinformatics/bty112

Zhang, S., Wang, Y., Gu, Y., Zhu, J., Ci, C., Guo, Z., et al. (2018). Specific Breast Cancer Prognosis-subtype Distinctions Based onDNAmethylation Patterns. *Mol. Oncol.* 12 (7), 1047–1060. doi:10.1002/1878-0261.12309

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. (2019). Meta-Path Methods for Prioritizing Candidate Disease miRNAs. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16 (1), 283–291. doi:10.1109/tcbb.2017.2776280

Zhang, Y., Xiang, M., and Yang, B. (2017). Low-rank Preserving Embedding. *Pattern Recognition* 70, 112–125. doi:10.1016/j.patcog.2017.05.003

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 (16), 4466–4472. doi:10.1093/bioinformatics/btaa428

Zhou, X., Li, Z., Xie, H., Feng, T., Lu, Y., Wang, C., et al. (2020). Leukocyte Image Segmentation Based on Adaptive Histogram Thresholding and Contour Detection. *Cbio* 15 (3), 187–195. doi:10.2174/1574893614666190723115832

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2020). Sequence Clustering in Bioinformatics: an Empirical Study. *Brief. Bioinform.* 21 (1), 1–10.

Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Sun, Z.-J., Dao, F.-Y., Yu, X.-L., et al. (2021). Identification of Cyclin Protein Using Gradient Boost Decision Tree Algorithm. *Comput. Struct. Biotechnol. J.* 19, 4123–4131. doi:10.1016/j.csbj.2021.07.013

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.

The handling editor declared a past co-authorship with one of the authors LX.

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.

Copyright © 2021 Han, Wang, Guo, Tang, Xu, Ju and Shi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 December 2021
doi: 10.3389/fgene.2021.798107

Using Machine Learning Approaches to Predict Target Gene Expression in Rice T-DNA Insertional Mutants

*Ching-Hsuan Chien[1†], Lan-Ying Huang[1†], Shuen-Fang Lo[2], Liang-Jwu Chen[3,4], Chi-Chou Liao[3], Jia-Jyun Chen[5] and Yen-Wei Chu[1,2,3,5,6,7,8]**

[1]Ph.D. Program in Medical Biotechnology, National Chung Hsing University, Taichung, Taiwan, [2]Biotechnology Center, National Chung Hsing University, Taichung, Taiwan, [3]Institute of Molecular Biology, National Chung Hsing University, Taichung, Taiwan, [4]Advanced Plant Biotechnology Center National Chung Hsing University, Taichung, Taiwan, [5]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung, Taiwan, [6]Agricultural Biotechnology Center, National Chung Hsing University, Taichung, Taiwan, [7]Ph.D. Program in Translational Medicine, National Chung Hsing University, Taichung, Taiwan, [8]Rong Hsing Research Center for Translational Medicine, National Chung Hsing University, Taichung, Taiwan

OPEN ACCESS

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

Reviewed by:
Zhibin Lv,
Sichuan University, China
Jinyan Li,
University of Technology Sydney, Australia

***Correspondence:**
Yen-Wei Chu
ywchu@nchu.edu.tw

[†]These authors have contributed equally to this work and share first authorship

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

Received: 19 October 2021
Accepted: 15 November 2021
Published: 17 December 2021

Citation:
Chien C-H, Huang L-Y, Lo S-F, Chen L-J, Liao C-C, Chen J-J and Chu Y-W (2021) Using Machine Learning Approaches to Predict Target Gene Expression in Rice T-DNA Insertional Mutants. Front. Genet. 12:798107. doi: 10.3389/fgene.2021.798107

To change the expression of the flanking genes by inserting T-DNA into the genome is commonly used in rice functional gene research. However, whether the expression of a gene of interest is enhanced must be validated experimentally. Consequently, to improve the efficiency of screening activated genes, we established a model to predict gene expression in T-DNA mutants through machine learning methods. We gathered experimental datasets consisting of gene expression data in T-DNA mutants and captured the PROMOTER and MIDDLE sequences for encoding. In first-layer models, support vector machine (SVM) models were constructed with nine features consisting of information about biological function and local and global sequences. Feature encoding based on the PROMOTER sequence was weighted by logistic regression. The second-layer models integrated 16 first-layer models with minimum redundancy maximum relevance (mRMR) feature selection and the LADTree algorithm, which were selected from nine feature selection methods and 65 classified methods, respectively. The accuracy of the final two-layer machine learning model, referred to as TIMgo, was 99.3% based on fivefold cross-validation, and 85.6% based on independent testing. We discovered that the information within the local sequence had a greater contribution than the global sequence with respect to classification. TIMgo had a good predictive ability for target genes within 20 kb from the 35S enhancer. Based on the analysis of significant sequences, the G-box regulatory sequence may also play an important role in the activation mechanism of the 35S enhancer.

Keywords: rice, CaMV 35S enhancer, T-DNA activation tagging, gene expression, machine learning

1 INTRODUCTION

Rice is one of the most important models of monocotyledon plants for the analysis of plant gene function. Rice is one of three major food crops throughout the world, and it is the staple food of more than half of the world's population. Rice production has doubled in the past 30 years, although the supply of rice is expected to gradually become insufficient with the rapid increase in the world population, climate change, and a shortage of water (Ray et al., 2013). It will not be easy to increase

food production to the necessary levels. In 2004, the International Rice Genome Sequencing Project (IRGSP) completed the sequencing of the rice genome (IRGSP, 2005). The ultimate goal of genome analysis is to realize the structure and function of each gene within an organism. To further confirm the function of and metabolic pathways related to each gene in rice, scientists have focused their efforts on analyzing the rice genome and are committed to promoting rice genome annotation to move rice research into the post-genome era.

T-DNA insertion activation-tagging technology is widely used in the analysis of the function of rice genes (Jeong et al., 2002; Yang et al., 2013). This method results in the construction of four tandem cauliflower mosaic virus (CaMV) 35S enhancers on a T-DNA plasmid; when this T-DNA is inserted into the rice genome, it activates genes that flank the T-DNA insertion site (Hsing et al., 2007). The CaMV 35S enhancer can activate gene expression in dicots and monocots and is widely used in T-DNA transformation. Gene expression gradually increases with the number of 35S enhancers on T-DNA, which led to the incorporation of four tandem repeat CaMV 35S enhancers for enhanced gene expression with this approach (Odell et al., 1985; Fang et al., 1989; Kardailsky et al., 1999; Weigel et al., 2000; Huang et al., 2001; Ichikawa et al., 2003). Agrobacterium-mediated T-DNA transformation tends to insert one copy of T-DNA, an average of 1.4 loci of T-DNA inserts in transgenic plants (Jeon et al., 2000), reducing the complexity of rice gene research. T-DNA inserted into the rice genome with a 35S enhancer resulted in two states:

(1) Gene knockdown: when T-DNA is inserted into the coding region or promoter of a gene, it is likely to destroy the structure of the gene, resulting in reduced function or loss of function of the gene.

(2) Activation tagging: T-DNA might enhance the activity of genes that flank the T-DNA insertion site through the effect of the 35S enhancers.

Thus, we can make use of T-DNA insertion activation tagging to study the association between genetic function and morphological traits (Hsing et al., 2007). However, there has been no basis for determining whether a target gene is activated by the enhancer prior to experimental analyses. There has even been a study indicating that the enhancer can activate genes that are millions of base pairs away from the enhancer (Li et al., 2012). Not all of the genes that flank the T-DNA insertion site are expected to be activated by the 35S enhancer. In some T-DNA mutants, the 35S enhancer does not activate the closer gene but instead activates a gene that is farther away from the 35S enhancer (Ren et al., 2004). Researchers thus cannot rely on the distance between the enhancer and a particular gene to judge whether that gene would be activated. They must instead determine the activated genes experimentally to explore the related genetic function and morphological traits. Therefore, it is a time-consuming and laborious process to check for the expression of a target gene.

Our team had developed a website platform, EAT-Rice (Liao et al., 2019), for predicting the expression status of rice genes that flank the T-DNA insertion site in activating mutants. In this study, we used a machine learning approach to predict target gene expression in rice T-DNA insertion mutants and improved the efficiency of finding activated target genes. The system of EAT-Rice applied the distance factor from T-DNA insertion site to gene loci to weight feature encoding and used two kinds of algorithms to build a two-layer model of machine learning. Based on EAT-Rice with a modified sequence capturing method, system architecture, and other additional features, we built a more comprehensive system for target gene expression prediction in T-DNA insertion mutants.

The datasets used in this study were experimentally validated. We first characterized genes based on their activation by the 35S enhancer; these genes were divided into activated genes and nonactivated genes. The system we built refers to EAT-Rice. We captured the DNA sequence of the promoter and the central region of each activated gene from the start codon of the target gene to the 35S enhancer and used nine features—CpG islands (CGIs), Motif, Kmer, reverse complementary kmer (RevKmer), DNP, TNP, DACC, TACC, and PseKNC—for encoding. Moreover, we carried out a logistic regression to weight the features of the first-layer model, depending on the probability of gene activation and the distance from the enhancer to the gene start codon. We then used LIBSVM (Chang and Lin, 2011) and LADTree (Boros et al., 2011) algorithms to build a two-layer model of machine learning. In the second layer, we used the minimum redundancy maximum relevance (mRMR) (Peng et al., 2005) method and incremental feature selection to determine the most relevant features. This system is referred to as TIMgo.

The TIMgo performance was 99.3% based on fivefold cross-validation and 85.6% based on independent testing. TIMgo had >80% accuracy for target genes within 20 kb from the 35S enhancer, but genes that were >20 kb away were still predicted with >60% accuracy. We also discovered that the value of the *k* parameter for Kmer, RevKmer, and PseKNC encoding within the PROMOTER sequences was higher than that of MIDDLE sequences. This suggested that for the analysis of longer sequences, a greater number of features was needed to improve the prediction performance. Finally, the G-box cis-element has an important function in gene activation by the 35S enhancer based on the motif analysis, and among the G-box-associated binding proteins, most are bZIP (basic region/leucine zipper) transcription factors.

2 MATERIALS AND METHODS

2.1 Sources for T-DNA Mutant Data and Datasets

The experimental data were collected from 11 rice T-DNA mutants from Liang-Jwu Chen's laboratory at NCHU and 316 mutants from Su-May Yu's research team at Academia Sinica. These data consisted of the T-DNA insertion point and expression status of flanking genes [as detected by RT-PCR (Ohan and Heikkila, 1993)]. The expression status of each gene was characterized based on the following four categories: activated gene (Ac), no significant effect (NE), non-detectable (ND), and knockout (Ko). The data distribution for the expression status of these genes is shown in **Table 1**.

TABLE 1 | Data distribution of flanking analyzed genes in rice T-DNA mutants.

Data source	Number of mutant lines	Gene expression status				Validated genes[a]
		Ac	NE	ND	Ko	
NCHU[b]	11	26	22	17	0	65
Academia Sinica[c]	316	262	143	13	2	420
Total	327	288	165	30	2	485

Ac, activated gene; NE, nonactivated gene; ND, non-detectable gene; Ko, knockout gene.
[a]*Validated genes indicate the target genes that were detected by RT-PCR.*
[b]*NCHU, experimental data were collected from Liang-Jwu Chen's laboratory.*
[c]*Academia Sinica, experimental data were collected by Su-May Yu's research team.*

TABLE 2 | Data distribution of the training dataset and independent-testing dataset.

Data sources	Training dataset (D300)		Testing dataset (D153)	
	Ac	NAc	Ac	NAc
NCHU	20	20	6	2
Academia Sinica	130	130	132	13
Total	150	150	138	15

To maintain dataset quality and consistency, we removed the 30 ND genes from the dataset. The collected data included two Ko genes, in which the T-DNA insertion point was located inside the gene, thus disrupting the gene structure and most likely leading to a loss of function. Because Ko genes were not a focus of this study, we removed them from the dataset. We defined NE genes as nonactivated (NAc) genes to differentiate them from the Ac genes. Ultimately, data for 453 genes were collected in this study.

A training set was used to determine the performance of the subsequent system. As the ratio of positive data (Ac genes) to negative data (NAc genes) affects the performance of machine learning (Akbani et al., 2004), this study used EAT-Rice with a 1:1 ratio to carry out the selection of the training dataset. We used data from 300 genes in the training dataset, which was referred to as D300. Data from the remaining 153 genes were used for independent testing to evaluate system accuracy; this dataset was referred to as D153 (**Table 2**).

2.2 Target Gene Sequence Retrieval

The analyzed genes provided from Liang-Jwu Chen's laboratory and Su-May Yu's team were annotated according to the Rice Genome Automated Annotation System (RiceGAAS) (Sakata et al., 2002) and the MSU Rice Genome Annotation Project (TIGR) (Yuan et al., 2003; Ouyang et al., 2007) rice gene annotation database. We hypothesized that we could predict the expression status of a target gene by analyzing the sequence of Ac and NAc genes. Thus, with reference to the EAT-Rice construction process and the enhancer-related hypothesis mechanisms (Singer et al., 2010; Singer et al., 2011), we extracted nucleotide sequences for each gene from two regions: (1) a 1,500-bp region upstream relative to the translation start site (TLS), referred to as the PROMOTER region, and (2) a central region of 300 bp centered between the TLS of the target gene and the 35S enhancer, referred to as the MIDDLE region (**Supplementary Figure S1**).

2.3 Feature Encoding

In this study, we encoded information about nine features of the sequences: five sequence information codes and four biological functional codes. The sequence codes consisted of two local sequence codes, two global sequence codes, and a code to reflect both the local and global sequence information simultaneously. The local sequence characteristics consisted of Kmer and RevKmer values, which were coded by the DNA composition; such characteristics have been successfully applied toward human gene regulatory sequence prediction (Noble et al., 2005; Gupta et al., 2008) and enhancer identification (Lee et al., 2011), among others. The two global sequence codes, dinucleotide-based auto-cross covariance (DACC) and trinucleotide-based auto-cross covariance (TACC), were coded by calculating the sequence autocorrelation as global sequence characteristics; this type of feature has been used to predict sequence-based protein–protein interactions (Guo et al., 2008). Another coding method, PseKNC, has been used to identify promoters in prokaryotes (Lin et al., 2014) and incorporates the information of contiguous local sequence order and the global sequence order into the feature vector. The biological characteristics included the presence of CGIs, regulatory cis-elements (Motif), and conformational and physicochemical properties of dinucleotide and trinucleotide sequences (DNP and TNP, respectively). Each of these features is described in more detail below.

2.3.1 CGIs

DNA methylation on CGIs reduces or silences gene expression based on enhancer–promoter interactions (Antequera et al., 1990; Volpe et al., 2002). For this analysis, we used the EMBOSS Newcpgreport tools of EMBL-EBI to predict CGIs and encoded their corresponding number, length, distance from the TLS, CpG ratio, and OE (observed/expected) value, resulting in the feature CGIs (**Supplementary Equations S1–S5**).

2.3.2 Regulatory *Cis*-Elements (Motif)

Considering that the rice transcription factor binding sites (TFBSs) that have been confirmed may not be comprehensive enough yet, we therefore incorporated other proven plant TFBSs. Data for 2,087 motifs were collected from PLACE (Higo et al., 1999) and the RegSite database (http://linux1.softberry.com/berry.phtml?topic=regsitelist). The tool Find Individual Motif Occurrences (FIMO) (Grant et al., 2011) in the MEME suite was used to scan for regulatory sequences in the PROMOTER region, and the scanning results were encoded by FIMO (Beer and

Tavazoie, 2004; Yuan et al., 2007). These types of feature encoding are referred to as follows.

$$\text{Motif_Number}_{(i)} = \begin{cases} j, & j \in \text{N} \\ 0, & \text{otherwise} \end{cases}, \quad \text{i} \in \{1, 2, \cdots, 2087\} \tag{1}$$

$$\text{Motif_Conserve}_{(i)} = \frac{M_i \text{ alignment score in promoter}}{\text{Motif_Number}\,(i)} \tag{2}$$

$$\text{Motif_Orientation}_{(i)} = \frac{pos \text{ in Motif_Number}_{(i)}}{\text{Motif_Number}_{(i)}} \tag{3}$$

$$\text{Motif_Dis}_{(i)} = \frac{\left|\text{geneTLS} - \text{Motif location site}\right|}{\text{Motif_Number}_{(i)}} \tag{4}$$

The number of regulatory elements was coded by the number (j) of motifs found in the PROMOTER (**Equation 1**). The conservation score was calculated by FIMO; we used the value from the summed motif conserved scores divided by the number of motifs in the sequence (**Equation 2**). As motifs can be located on both the DNA coding strand (codons) and the template strand (anticodons), the orientation characteristic was calculated to determine the proportion of motifs on the coding strand. We thus used the number of motifs on the coding strand (i.e., positive motifs, pos) as the numerator, and the denominator is the number of all motifs (**Equation 3**). The distance characteristic was determined based on the distance (in base pairs) from each motif to the TLS, which was summed for all motif sites within a given sequence, divided by the number of motifs (**Equation 4**). In these equations, i indicates the kinds of motifs, M_i indicates a specific motif, and *geneTLS* refers to the translation start site of a target gene.

2.3.3 Kmer and RevKmer

Kmer refers to the local sequence information and indicates a subsequence containing k neighboring nucleic acids in a DNA sequence. Using a coding strand as the template, the Kmer feature will scan for the number of occurrences of the nucleic acid subsequence in the template. For example, when k is 2, the subsequence composition of a Kmer will be called a 2-mer, which contains 16 subsequences (based on the four nucleotides G, A, T, and C). In the case of the dinucleotide AA, if this subsequence appeared twice in the DNA template, it would be encoded as 2; if it was not present in the template, it would be encoded as 0. In eukaryotes, the average length of TFBSs is 10 bp (Stewart et al., 2012), which suggests that the number of k neighboring nucleic acids in this study could be increased. We encoded the sequence with 3- to 6-mer, 3- to 7-mer, 3- to 8-mer, and 3- to 9-mer, which produced 5,440, 21,824, 87,360, and 349,504 different nucleotide compositions, respectively. The Kmer encoding was carried out based on the number of occurrences in the template sequence (**Supplementary Equation S6**).

RevKmer is a variant of kmer, in which the kmers are not expected to be strand specific, so reverse complements are collapsed into a single value. In this study, the RevKmer feature was encoded in the same manner as Kmer and produced 2,760, 10,952, 43,848, and 174,920 nucleotide compositions for the 3- to 6-mer, 3- to 7-mer, 3- to 8-mer, and 3- to 9-mer, respectively. RevKmer encoding was carried out according to the number of occurrences in the template sequence (**Supplementary Equation S7**).

2.3.4 Nucleotide Conformational and Physicochemical Properties (DNP and TNP)

The nucleotide conformation and physicochemical properties of dinucleotides and trinucleotides were also encoded. DiProDB provides information about 125 properties of dinucleotides, and these 125 properties were integrated into 15 characteristics through a statistical principal components analysis (PCA) method (Friedel et al., 2009). The value of each property is based on the dinucleotide as a unit, and each property has 16 values corresponding to all possible dinucleotide combinations. We used the property of the dinucleotide to produce a training model with 240 dimensions; this feature is referred to as the DNP (dinucleotide conformation and physicochemical properties) (**Supplementary Equation S8**). PseKNC-General (the general form of pseudo k-tuple nucleotide composition) is a tool that provides the conformation and physicochemical properties of oligonucleotides (Chen et al., 2015). In this study, 12 trinucleotide properties were used for coding. There were 64 combinations of trinucleotides, which generated a training model with 768 dimensions based on the 12 trinucleotide properties; this feature is referred to as the TNP (trinucleotide conformation and physicochemical properties) (**Supplementary Equation S9**).

2.3.5 Autocorrelation (DACC and TACC)

Pse-in-One provides a pseudo-component mode reflecting the correlation between two dinucleotides or trinucleotides within a DNA sequence via their physicochemical properties (Liu et al., 2015). In this study, we used dinucleotide-based auto-cross covariance (DACC) and trinucleotide-based auto-cross covariance (TACC) as provided by Pse-in-One for encoding (**Supplementary Equations S10–S12**).

In this study, DACC was based on the 15 properties from DiProDB, and the lag value was 4, generating a training model with 900 dimensions. TACC used the 12 Pse-in-One built-in properties, and the lag value was 4; it generated a training model with 576 dimensions.

2.3.6 Pseudo k-Tuple Nucleotide Composition

Pseudo k-tuple nucleotide composition (PseKNC) is one of the encoding modes supplied by Pse-in-One. It incorporates both the contiguous local sequence order information (like Kmer and RevKmer) and the global sequence order information (like DACC and TACC) into the feature vector of the DNA sequence.

$$D = R_1R_2R_3R_4R_5R_6 \cdots R_L \tag{5}$$

$$\text{PseKNC}_{(u)} = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j}, u \in \left\{1, 2, \cdots 4^k\right\} \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j}, u \in \left\{4^{k+1}, \left(4^{k+1}+1\right), \cdots, \left(4^{k+1}+\lambda\right)\right\} \end{cases} \tag{6}$$

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\left\{\frac{1}{\mu}\sum_{v=1}^{u}\left[P_v\left(R_iR_{i+1}\right) - P_v\left(R_{i+j}R_{i+j+1}\right)\right]^2\right\}, j \in \{1, 2, \cdots, \lambda\}, \lambda < L \tag{7}$$

For a DNA sequence D with L nucleic acid residues, R_1 represents the nucleic acid residue at the sequence position 1, R_2 the nucleic acid residue at position 2, and so on (**Equation 5**). PseKNC will calculate the occurrence frequency (f) of dinucleotides in the DNA sequence and the correlation between two oligonucleotides that are 1 to λ nucleotides apart from each other. In **Equation 6**, f_u is the occurrence frequency of dinucleotides in the DNA sequence, which is normalized to $\sum_{i=1}^{4^k} f_i = 1$; w is the weight factor; θ_j represents the correlation factor that reflects the sequence-order correlation between all two dinucleotides that are j nucleotides away from each other along a DNA sequence; μ is the number of physicochemical indices; $P_v(R_iR_{i+1})$ represents the numerical value of the dinucleotide located at the ith position (R_iR_{i+1}) of the vth ($v = 1, 2, \ldots, \mu$) physicochemical property (**Equation 7**). The feature number of PseKNC will be λ multiplied by 4 to the power k. In this study, the PseKNC feature was determined with a λ value of 4, w is 0.2, and k is from 2 to 6.

2.4 Significant Sequence Fragment Analysis

Because there are numerous features in this first-layer model, the complexity of the model is relatively high. To reduce the interference of excessive noise, we used independent two-sample t-tests (implemented in R) to select features from the high-dimension models. We used the occurrence of specific oligonucleotides in the Ac and NAc groups to generate the t-test (**Supplementary Figure S2**) and retained the oligonucleotides with $p < 0.05$ to encode these significant fragments.

2.5 Model Evaluation and Cross-Validation

We used a five-fold cross-validation method and independent-testing data to evaluate the predictive performance of the model. Our evaluation methods included accuracy (Acc), sensitivity (Sn), specificity (Sp), and Matthews correlation coefficient (MCC). Acc is used to estimate the prediction accuracy of the global prediction capability, with values closer to 100% indicating better overall predictive performance of a model (**Equation 8**). Sn and Sp evaluate the accuracy of the prediction of positive and negative data, respectively (**Equations 9** and **10**). When the number of positive and negative data differs, Acc is not a good evaluation indicator. MCC is, however, suitable for assessing a dataset in which there is an imbalance between positive data and negative data (**Equation 11**). When the MCC score is closer to 1, the prediction capability is better; a score closer to −1 indicates a worse prediction capability.

$$\mathrm{Acc} = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$\mathrm{Sn} = \frac{TP}{TP + FN} \tag{9}$$

$$\mathrm{Sp} = \frac{TN}{TN + FP} \tag{10}$$

$$\mathrm{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \tag{11}$$

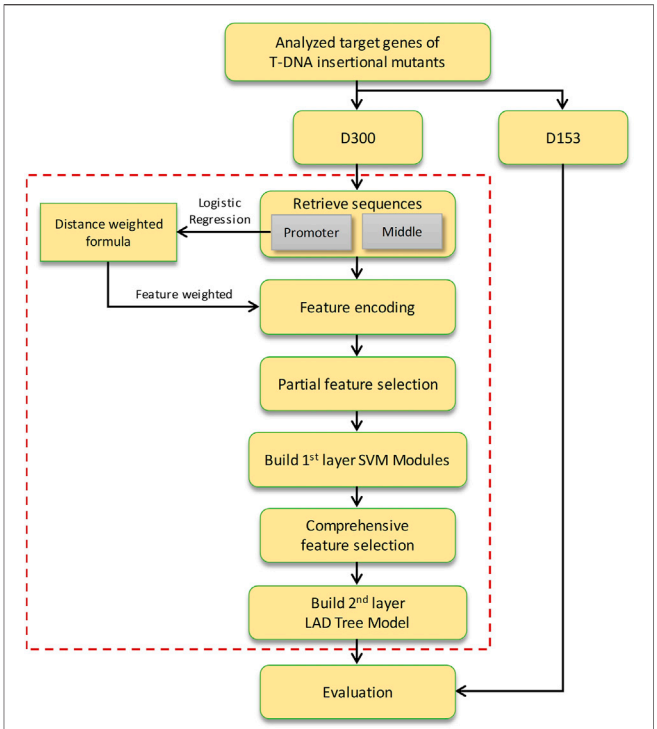


FIGURE 1 | Flow chart of the TIMgo predictive system.

2.6 Framework of TIMgo

TIMgo is a two-layer machine learning model constructed for predicting the effect of a 35S enhancer on the expression of the target gene (**Figure 1**). The D453 was divided into a training dataset (D300) and independent testing data (D153). The DNA sequences of PROMOTER and MIDDLE were retrieved for analysis between NAc and Ac genes. In the first-layer module, the support vector machine (SVM) models were constructed within nine feature-encoding methods. And the significant sequences were analyzed by Student's t-test, and a model of logistic regression was used to assist in training, which is based on the relationship between distance from the 35S enhancer to the target gene and states of gene expression. The features encoded from the PROMOTER region were weighted by a logical regression model for probability of gene activation. Then, we adopted feature selection by the LIBSVM built-in tool in the partial SVM models. The prediction results of the first-layer module were integrated into the second-layer model, and mRMR (Peng et al., 2005) was used for feature selection and building the LAD tree model. Finally, we evaluated the prediction efficacy of TIMgo with the D153 independent-testing dataset.

3 RESULTS

3.1 Correlation Between Gene Activation and Distance From the 35S Enhancer to the TLS

The distance between the enhancer and a target gene cannot be directly used to determine whether the target gene will be

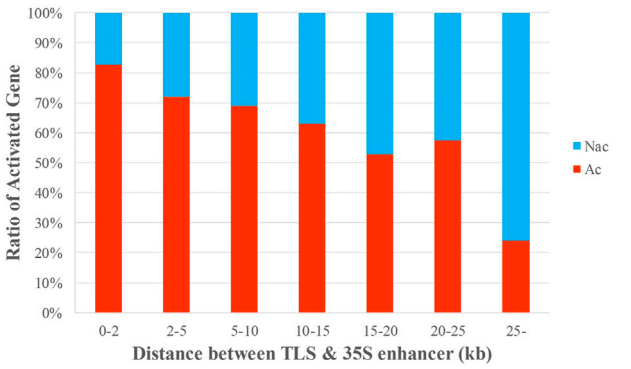


FIGURE 2 | Correlation between distance and gene activation. The data were sorted by the distance between the 35S enhancer and the TLS, and the ratio of Ac to NAc genes in each group was calculated. The *x*-axis is the distance from the 35S enhancer to the TLS of a target gene; the *y*-axis is the proportion of Ac and NAc genes in each group.

activated, although it does have some relevance for determining gene activation (Vandergeest and Hall, 1997; Jagannath et al., 2001). A target gene is more likely to be activated if it is closer to the enhancer (Marenduzzo et al., 2007). We characterized each of the 453 genes in the entire dataset (D453) based on the distance from the CaMV 35S enhancer on the inserted T-DNA to TLS and calculated the ratio of Ac genes and NAc genes. We found a negative correlation between this distance and gene activation. Genes closer to the 35S enhancer had a greater probability of activation ($p < 0.001$) (**Supplementary Figure S3**). The results are the same as those indicated in a previous study (Liao et al., 2019) (**Figure 2**, **Supplementary Table S1**).

Among the D453 dataset, there were 94 sets of duplicated data which consist of multiple genes, and the PROMOTER sequences corresponding to these genes were identical. Each of the experimental data in this study represented the effect of a single insertion event on its target gene. In the experimental data collected in this study, when the same gene was detected for multiple T-DNA insertion events, the PROMOTER sequences from those genes were identical. For different T-DNA insert events, the 35S enhancer may result in different states of expression for the same target gene, which will lead to contradictory results while building the machine learning model. To distinguish between these PROMOTER sequences, we used logistic regression to build a regression model of the distance coefficient and the target gene activation probability (**Supplementary Equation S13**). In this study, the values calculated by logistic regression were used to weight the promoter sequence feature, so that the same sequence could be distinguished when quantified based on numerical values.

3.2 Comparison of Kmer and RevKmer Combined With Motif

In the Kmer and RevKmer feature models, a *t*-test was used to calculate the number of occurrences of specific sequence fragments in Ac and NAc genes, respectively, from sequence lengths (k) of three to nine nucleotides. The specific sequence fragments with $p < 0.05$ were then used for encoding. These fragments were combined as 3–6, 3–7, 3–8, and, 3–9 combinations for Kmer and RevKmer. The Motif feature was used to carry out a similar analysis. The Kmer and RevKmer features associated with the PROMOTER region were combined with the Motif feature (**Supplementary Table S2**). The features from Kmer, RevKmer, Kmer + Motif, and RevKmer + Motif were used to build SVM models, and the best model was selected for the second-layer model integration (**Supplementary Table S3**).

Before combining Motif with Kmer or RevKmer, the Acc scores of the SVM models of Kmer and RevKmer were 55%–85%, whereas the Acc scores of the Motif models were 52%–75%. After combining Motif with Kmer or RevKmer, the Acc scores were 78%–86%, and the Acc consistently increased with the k value for Kmer and RevKmer (**Table 3**).

3.3 First-Layer Model Evaluation

In the first-layer models, nine feature coding methods and two types of sequences were used to construct 16 feature models (**Supplementary Table S4**). The prediction ability of each feature model was evaluated with fivefold cross-validation and independent testing with the D153 data (**Table 4**). For the Pse-in-One feature encoding, one gene sequence from the training dataset (D300) did not conform to the encoding requirements. Therefore, in the DACC, TACC, and PseKNC models, this information was removed from the training data, and the training dataset consisting of the remaining 299 genes was referred to as D299. The PseKNC models used k values of 2–7, and eight models each were established for the PROMOTER and MIDDLE sequences. A PseKNC model with $k = 6$ that was selected among the PROMOTER models had an Acc of 75.3% with fivefold cross-validation. The PseKNC model with $k = 2$ that was selected among the MIDDLE models had an Acc of 59.5% (**Supplementary Table S5**).

In the evaluation results of the first-layer feature models (**Table 4**), the Kmer, RevKmer, Kmer + Motif, and RevKmer + Motif had the best predictive performance based on the Kmer feature provided. Their Acc values were 79.0%–88.3% with fivefold cross-validation. With independent testing, their Acc values were 80.4%–84.3%, with the exception of RevKmer, which had 67.3%. The PseKNC model built using the PROMOTER sequence was slightly inferior to the model built using Kmer-related features. The Acc and MCC values for PseKNC were 75.3% and 52.9% with cross-validation, respectively, and 56.2% for Acc and 16.5% for MCC with independent testing. The DACC, TACC, DNP, CGIs, and TNP constructed by the PROMOTER sequence and the PseKNC constructed by the MIDDLE sequence had lower predictive performance, with Acc values of 58.2%–69.9% and MCC values of 16.4%–39.8%. Among these 16 models, CGIs and TNP constructed using the MIDDLE sequence were the least accurate in cross-validation, with an Acc of ~47%. Their Acc values for independent testing were 11.8% and 62.1%, respectively. In terms of overall predictive performance, the PROMOTER sequence is thus more important than the MIDDLE sequence, and Kmer, RevKmer, Kmer + Motif, and RevKmer + Motif features have the highest correlation with the activation of genes.

TABLE 3 | Data distribution of the training dataset and independent-testing dataset.

Feature	k^a	Without motif					With motif				
		Sp (%)	Sn (%)	Acc (%)	MCC (%)	AUC (%)	Sp (%)	Sn (%)	Acc (%)	MCC (%)	AUC (%)
Kmer	6	72.7	66.0	69.3	38.8	79.0	79.3	77.3	78.3	56.7	88.1
	7	86.7	73.3	80.0	60.5	89.1	83.3	78.7	81.0	62.1	89.7
	8	75.3	35.3	55.3	11.6	65.3	83.3	84.7	84.0	68.0	93.6
	9	84.7	85.3	85.0	70.0	93.2	86.7	85.3	86.0	72.0	93.7
RevKmer	6	71.3	60.7	66.0	32.2	72.7	78.0	77.3	77.7	55.3	85.7
	7	84.7	76.0	80.3	60.9	87.9	79.3	77.3	78.3	56.7	88.1
	8	77.3	32.7	55.0	11.2	64.9	84.0	80.0	82.0	64.1	91.5
	9	74.7	88.0	81.3	63.2	90.6	84.0	84.7	84.3	68.7	92.9

[a]k *refers to the maximum* k *value used in Kmer and RevKmer, with a range of 3-*k *nucleotides in length for each analysis.*

TABLE 4 | Performance of the first-layer features with the SVM models.

Feature encoding	Sequence	Cross-validation					Independent testing				
		Sp (%)	Sn (%)	Acc (%)	MCC (%)	AUC (%)	Sp (%)	Sn (%)	Acc (%)	MCC (%)	AUC (%)
CGIs	PROMOTER	71.3	48.7	60.0	20.5	58.5	53.3	40.6	41.8	–3.7	48.2
	MIDDLE	77.3	18.0	47.7	–5.8	47.2	100.0	2.2	11.8	4.7	65.0
DNP	PROMOTER	56.0	64.7	60.3	20.7	64.3	26.7	71.7	67.3	–1.1	45.1
	MIDDLE	59.3	62.0	60.7	21.3	60.0	60.0	53.6	54.3	8.1	48.7
TNP	PROMOTER	56.0	61.3	58.7	17.4	62.2	53.3	68.1	66.7	13.5	57.4
	MIDDLE	64.7	30.0	47.3	–5.7	47.4	26.7	65.9	62.1	–4.7	45.0
Kmer + Motif	PROMOTER	86.7	85.3	86.0	72.0	93.7	73.3	85.5	84.3	43.5	79.1
RevKmer + Motif	PROMOTER	84.0	84.7	84.3	68.7	92.9	73.3	81.2	80.4	37.8	83.6
Kmer	MIDDLE	92.0	84.7	88.3	76.9	94.2	66.7	86.2	84.3	40.1	86.4
RevKmer	MIDDLE	85.3	72.7	79.0	58.5	88.2	53.3	68.8	67.3	14.0	66.5
DACC	PROMOTER	67.1	72.7	69.9	39.8	78.6	46.7	59.4	58.2	3.7	54.6
	MIDDLE	76.5	58.0	67.2	35.1	74.1	53.3	49.3	49.7	1.6	47.5
TACC	PROMOTER	60.4	58.0	59.2	18.4	60.3	13.3	63.0	58.2	–14.8	41.6
	MIDDLE	59.7	56.7	58.2	16.4	57.8	46.7	45.7	45.8	–4.6	45.1
PseKNC	PROMOTER	89.9	60.7	75.3	52.9	84.5	73.3	54.3	56.2	16.5	59.1
	MIDDLE	56.4	52.7	59.5	19.1	61.7	66.7	58.0	58.8	14.7	54.5

3.4 Comprehensive Feature Selection in the Second-Layer Model

The second-layer model integrated the prediction results from the 16 feature models in the first layer and obtained the ultimate prediction result by machine learning. The features used in the second-layer model of this study included predictive results and positive and negative predictive confidence scores, generating 48 features. We used incremental feature selection and an SVM model with cross-validation to carry out comprehensive feature selection among these 48 features to pick out the best feature combinations with nine feature selection methods. The top 33 features of the mRMR (Peng et al., 2005) were selected as the best feature combination with the highest Acc and the fewest features (**Figure 3**, **Supplementary Table S6**). Among the 33 selected features, we knew that the encoding contributed for classification is Kmer related, DACC was better than PseKNC and TACC, and CGIs, TNP, and DNP are worse.

3.5 Second-Layer Model Evaluation

We assessed the best-suited machine learning algorithm for the second-layer model through the WEKA (Holmes et al., 1994) analysis platform. In this study, we used the 65 algorithms provided by WEKA to establish the model separately and evaluated the effectiveness of these models with cross-validation (**Supplementary Table S7**). In this experiment, the LADTree algorithm was used to construct the second-layer integration model according to the above conditions. The Acc was 99.3%, MCC was 98.7%, and Sn and Sp were 99.3%. In independent testing, the model Acc reached 85.6%, MCC was 35.3%, Sn was 89.1%, and Sp was 53.3%. Among the testing data, there were only 15 negative data, such that each predictive result with these data would lead to a substantial impact on the overall predictive effectiveness assessment. Among these models built with multiple algorithms, Sp values ranged from 46.7% to 73.3%, which corresponded to a difference of only six correctly predicted negative data.

3.6 Correlation Between Predictive Accuracy and Distance From the 35S Enhancer to TLS

To analyze the relationship between distance and TIMgo prediction accuracy, the training dataset and independent-testing dataset were grouped according to the distance between the TLS and 35S enhancer (**Figure 4**). In cross-validation, Acc was 99.3%, and predictions for

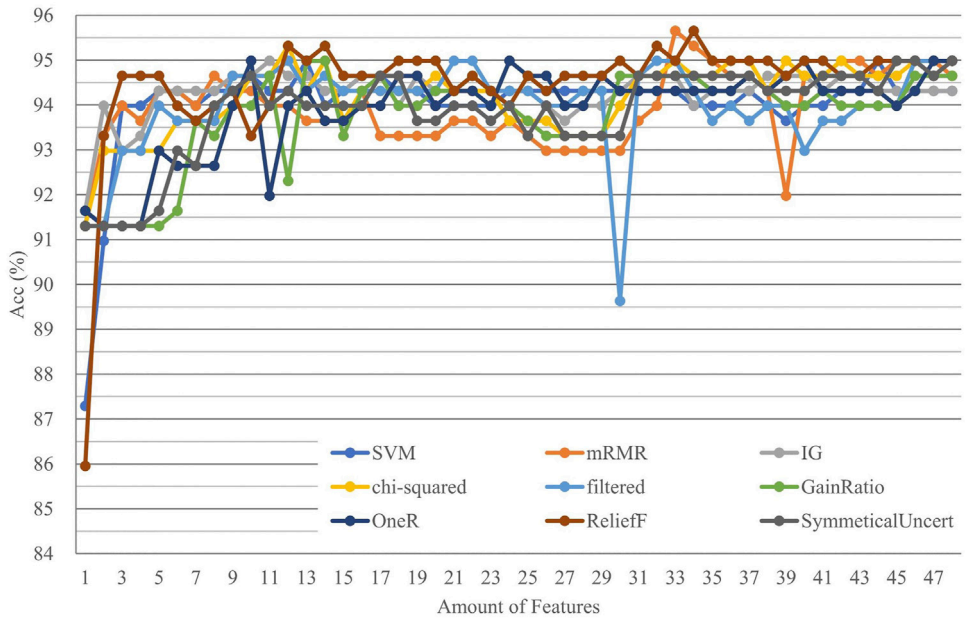


FIGURE 3 | Accuracy trend in the second-layer feature selection.

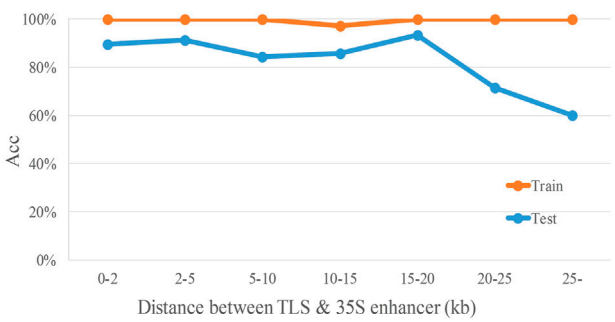


FIGURE 4 | Accuracy trend of TIMgo for cross-validation and independent testing of data within different distances. Train represents the Acc from fivefold cross-validation with D299. Test represents the Acc from independent testing with D153. The *x*-axis indicates each distance interval, and the *y*-axis indicates the predictive accuracy.

only two genes were incorrect (**Table 5**); these two genes were 10–15 kb away from the 35S enhancer. In independent testing, the prediction accuracy for genes within 20 kb from the 35S enhancer was >84%. For genes located >20 kb from the 35S enhancer, the prediction accuracy decreased with increasing distance but still was >60% (**Table 6**).

4 DISCUSSION

4.1 Comparison of the Framework Between TIMgo and EAT-Rice

In a previous study, the PROMOTER region for most genes was defined as the upstream region from the transcription start site (TSS) (Chang et al., 2008). For the EAT-Rice analysis, however, as the collected gene data had information about only the TLS, the PROMOTER region, including the upstream sequence of the TSS, was based on a 1,000-bp region upstream of the TLS. The upstream sequence of the TSS contains the 5′ untranslated region of the mRNA, and sequences downstream of the TSS may also be involved with transcription factor regulation of gene expression (Heyndrickx et al., 2014). Given an average length of 500 bp for 5′ untranslated regions in rice and the 1,000 bp upstream of the TSS as the condition, we used the 1,500-bp sequence upstream of the TLS as the PROMOTER region in this study.

For our prediction models, we retained the EAT-Rice CGIs and DNP (dinucleotide conformation and physicochemical properties encoding) and increased the TNP coding with the DNP coding concept. We also used the Pse-in-One tool to generate codes for DACC, TACC, and PseKNC. Given the strand specificity of Kmer, we added RevKmer coding, and the Motif coding of the PROMOTER region was combined with Kmer and with RevKmer. The ranges of overall predictive accuracy for Kmer + Motif and RevKmer + Motif models were small, which indicated that Motif was complementary with Kmer and RevKmer, and the combination of these two features could improve the classification ability. Predictive accuracy increased with the length of *k* for both Kmer and RevKmer, because that Motif feature consisted of experimentally validated regulatory sequences, but the number of proven regulatory sequences in plants is limited, whereas Kmer and RevKmer considered all the sequence combinations that provided higher data integrity than Motif, so using longer Kmer and RevKmer should lead to better prediction

TABLE 5 | Performance of the LADTree model in the second-layer.

	TP	FP	TN	FN	Sn (%)	Sp (%)	Acc (%)	MCC (%)
Cross-validation	149	1	148	1	99.3	99.3	99.3	98.7
Independent testing	123	7	8	15	89.1	53.3	85.6	35.3

TABLE 6 | Predictive accuracy of TIMgo for different distance groups.

Distance from the 35S enhancer (kb)							
Dataset	**0–2**	**2–5**	**5–10**	**10–15**	**15–20**	**20–25**	**>25**
Training set	100.0%	100.0%	100.0%	97.0%	100.0%	100.0%	100.0%
Testing set	89.0%	91.0%	84.0%	86.0%	93.0%	71.0%	60.0%

performance. Although Kmer and RevKmer had higher data integrity than Motif, the complexity of the Kmer and RevKmer data increased exponentially with the increase in sequence length, resulting in processing time that was too lengthy. Therefore, we used Kmer (and RevKmer) with limited *k* length and retained Motif with longer sequences, to preserve important regulatory sequence data and reduce the computational complexity significantly.

4.2 Specific Regulatory Sequences Within Genes Activated by the 35S Enhancer

To find out whether a specific regulatory sequence was related to gene activation in the T-DNA insertion mutants, we analyzed the 2,087 motifs with a *t*-test. We found that there were 181 regulatory sequences that had significant difference in their occurrence frequency between Ac and NAc genes. Among these 181 regulatory sequences, 20 were G-box and G-box-related sequences. The G-box contains a core region, CACGTG, and flanking sequences that are composed of other nucleotides. The G-box-binding protein has different binding preferences and affinities according to the different flanking sequences in the G-box. bZIP (basic region/leucine zipper) transcription factors account for the majority of G-box-binding proteins. Transcription regulation in plants is often affected by G-box sequences, such as stress hormones (e.g., abscisic acid), seed germination, protein storage, and light response (Marcotte et al., 1989; Donald and Cashmore, 1990; Mason et al., 1993). Thus, the G-box may have important biological significance in the regulation of gene expression by the 35S enhancer and may affect whether the 35S enhancer will activate a target gene in rice.

4.3 Correlation Between Length of Sequence and Nucleotide Length Parameter

In the feature coding of TIMgo, the coding of Kmer, RevKmer, and PseKNC can be adjusted based on the nucleotide length parameter (*k*). We needed to find a suitable nucleotide length parameter for encoding. For these three kinds of coding, the *k* value selected for the PROMOTER region was greater than that for the MIDDLE region. A higher value for *k* results in a higher number of features being generated, which requires more features to be improved to increase the predictive accuracy of the PROMOTER region, relative to the MIDDLE region. Thus, an excessive number of features would reduce the predictive performance of the model. From the optimal *k* value for the MIDDLE sequence, we could see that a higher number of features did not necessarily make the classification better. By comparing the optimal *k* value selected for the PROMOTER and MIDDLE regions, we note that a longer sequence does seem to require more features to make the classification better. Moreover, among the local, global, and local + global sequence characteristics used to build the TIMgo, the local sequences had a greater contribution with respect to identifying activation of the target genes (**Table 4**).

4.4 Performance Comparison of TIMgo and EAT-Rice

To confirm that the model constructed by the framework of TIMgo is superior to that of EAT-Rice, the training dataset and testing dataset used to develop EAT-Rice were used to build models in the TIMgo framework and to evaluate TIMgo by comparing their predictive performance. The training dataset used with EAT-Rice had data for 280 validated genes, and these 280 data points were separated into two subsets (subset1 and subset2) with 180 validated genes (Liao, et al., 2019). The independent-testing dataset used with EAT-Rice had 48 validated genes. Two training datasets (subset1 and subset2) were used to build training models within the framework of TIMgo, and the predictive efficacy of EAT-Rice and TIMgo was evaluated with an independent-testing dataset consisting of an additional 48 validated genes (**Table 7**). With the use of subset1 as the training dataset and of the EAT-Rice system to establish the model, the Acc in the independent testing was 72.9%, the Acc for TIMgo was 79.2%, and the Sp value of TIMgo was 12.8% higher than that of EAT-Rice. With subset2 as the training dataset, the Acc with independent testing was 77.1% for EAT-Rice and 77.6% for TIMgo. In the case of using the same training dataset and testing dataset, the accuracy of the TIMgo framework is better than that of EAT-Rice.

TABLE 7 | Comparison of TIMgo and EAT-Rice with independent-testing evaluation.

System	Subset1				Subset2			
	Sp (%)	Sn (%)	Acc (%)	AUC (%)	Sp (%)	Sn (%)	Acc (%)	AUC (%)
EAT-Rice	59.1	84.6	72.9	79.4	59.1	92.3	77.1	83.2
TIMgo	72.7	84.6	79.2	87.4	78.3	76.7	77.6	84.4

5 CONCLUSION

In this study, we analyzed the DNA sequence and constructed a two-layer model system using the machine learning method to predict whether the 35S enhancer would affect the expression of a target gene in T-DNA insertion mutants. The first layer of the system was built with the PROMOTER and MIDDLE sequences and was encoded using nine features. We analyzed significant sequence fragments in Motif, Kmer, and RevKmer and weighted the PROMOTER based on a logistic regression analysis of the distance between the 35S enhancer and the TLS of each gene. Some of the first-layer SVM models were built with LIBSVM feature selection. The second-layer model used the mRMR feature selection tool to select the predicted values from the 16 models in the first layer, and these were integrated with the LADTree algorithm as the second-layer model. The predictive performance of TIMgo had Acc of 99.3% and 85.6% with cross-validation and with independent testing, respectively. TIMgo can more accurately predict the activation of genes located within 20 kb of the 35S enhancer. We analyzed the 2,087 motifs and found that there was a significant difference in the frequency of G-box sequences between Ac and NAc genes, suggesting that the G-box may play an important role in the activation mechanism of 35S enhancer genes. Our model has improved the predictive ability of determining target gene activation based on the CaMV 35S enhancer in rice T-DNA insertion mutants.

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

C-HC and L-YH contributed to data collection, design of experimental processes, and system architecture. J-JC drafted the manuscript. S-FL and L-JC supported the experimental data and data interpretation. C-CL and Y-WC conceived of the study goal, supervised the study, and provided advice with respect to the study direction. All authors read and approved the manuscript.

FUNDING

This research was supported by (1) the Ministry of Science and Technology, Taiwan, under grant number 110-2221-E-005-062-MY3; (2) the Ministry of Science and Technology, Taiwan, under grant number 110-2321-B-005-005; and (3) National Chung Hsing University and Changhua Christian Hospital: NCHU-CCH 11006.

ACKNOWLEDGMENTS

This work was supported in part by the Advanced Plant Biotechnology Center from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.798107/full#supplementary-material

REFERENCES

Akbani, R., Kwek, S., and Japkowicz, N. (2004). Applying Support Vector Machines to Imbalanced Datasets. *Machine Learn. Ecml 2004, Proc.* 3201, 39–50. doi:10.1007/978-3-540-30115-8_7

Antequera, F., Boyes, J., and Bird, A. (1990). High Levels of De Novo Methylation and Altered Chromatin Structure at CpG Islands in Cell Lines. *Cell* 62, 503–514. doi:10.1016/0092-8674(90)90015-7

Beer, M. A., and Tavazoie, S. (2004). Predicting Gene Expression from Sequence. *Cell* 117, 185–198. doi:10.1016/s0092-8674(04)00304-6

Boros, E., Crama, Y., Hammer, P. L., Ibaraki, T., Kogan, A., and Makino, K. (2011). Logical Analysis of Data: Classification with Justification. *Ann. Oper. Res.* 188, 33–61. doi:10.1007/s10479-011-0916-1

Chang, C. C., and Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. *Acm Trans. Intell. Syst. Techn.* 2, 1–27. doi:10.1145/1961189.1961199

Chang, W.-C., Lee, T.-Y., Huang, H.-D., Huang, H.-Y., and Pan, R.-L. (2008). PlantPAN: Plant Promoter Analysis Navigator, for Identifying Combinatorial Cis-Regulatory Elements with Distance Constraint in Plant Gene Groups. *BMC Genomics* 9, 561. doi:10.1186/1471-2164-9-561

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.-C. (2015). PseKNC-General: a Cross-Platform Package for Generating Various Modes of Pseudo Nucleotide Compositions. *Bioinformatics* 31, 119–120. doi:10.1093/bioinformatics/btu602

Donald, R. G., and Cashmore, A. R. (1990). Mutation of Either G Box or I Box Sequences Profoundly Affects Expression from the Arabidopsis rbcS-1A Promoter. *EMBO J.* 9, 1717–1726. doi:10.1002/j.1460-2075.1990.tb08295.x

Fang, R. X., Nagy, F., Sivasubramaniam, S., and Chua, N. H. (1989). Multiple Cis Regulatory Elements for Maximal Expression of the Cauliflower Mosaic Virus 35S Promoter in Transgenic Plants. *Plant Cell* 1, 141–150. doi:10.1105/tpc.1.1.141

Friedel, M., Nikolajewa, S., Sühnel, J., and Wilhelm, T. (2009). DiProDB: a Database for Dinucleotide Properties. *Nucleic Acids Res.* 37, D37–D40. doi:10.1093/nar/gkn597

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: Scanning for Occurrences of a Given Motif. *Bioinformatics* 27, 1017–1018. doi:10.1093/bioinformatics/btr064

Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using Support Vector Machine Combined with Auto Covariance to Predict Protein-Protein Interactions from Protein Sequences. *Nucleic Acids Res.* 36, 3025–3030. doi:10.1093/nar/gkn159

Gupta, S., Dennis, J., Thurman, R. E., Kingston, R., Stamatoyannopoulos, J. A., and Noble, W. S. (2008). Predicting Human Nucleosome Occupancy from Primary Sequence. *Plos Comput. Biol.* 4, e1000134. doi:10.1371/journal.pcbi.1000134

Heyndrickx, K. S., de Velde, J. V., Wang, C., Weigel, D., and Vandepoele, K. (2014). A Functional and Evolutionary Perspective on Transcription Factor Binding in Arabidopsis thaliana. *Plant Cell* 26, 3894–3910. doi:10.1105/tpc.114.130591

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant Cis-Acting Regulatory DNA Elements (PLACE) Database: 1999. *Nucleic Acids Res.* 27, 297–300. doi:10.1093/nar/27.1.297

Holmes, G., Donkin, A., and Witten, I. H. (1994). "Weka: A Machine Learning Workbench," in Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference, Brisbane, QLD, Australia, 29 Nov.-2 Dec. 1994 (IEEE), 357–361.

Hsing, Y.-I., Chern, C.-G., Fan, M.-J., Lu, P.-C., Chen, K.-T., Lo, S.-F., et al. (2007). A rice Gene Activation/knockout Mutant Resource for High Throughput Functional Genomics. *Plant Mol. Biol.* 63, 351–364. doi:10.1007/s11103-006-9093-z

Huang, S., Cerny, R. E., Bhat, D. S., and Brown, S. M. (2001). Cloning of an Arabidopsis Patatin-like Gene, STURDY, by Activation T-DNA Tagging. *Plant Physiol.* 125, 573–584. doi:10.1104/pp.125.2.573

Ichikawa, T., Nakazawa, M., Kawashima, M., Muto, S., Gohda, K., Suzuki, K., et al. (2003). Sequence Database of 1172 T-DNA Insertion Sites in Arabidopsis Activation-Tagging Lines that Showed Phenotypes in T1 Generation. *Plant J.* 36, 421–429. doi:10.1046/j.1365-313x.2003.01876.x

IRGSP (2005). The Map-Based Sequence of the rice Genome. *Nature* 436, 793–800. doi:10.1038/nature03895

Jagannath, A., Bandyopadhyay, P., Arumugam, N., Gupta, V., Burma, P. K., and Pental, D. (2001). The Use of a Spacer DNA Fragment Insulates the Tissue-specific Expression of a Cytotoxic Gene (Barnase) and Allows High-Frequency Generation of Transgenic Male Sterile Lines in Brassica Juncea L. *Mol. Breed.* 8, 11–23. doi:10.1023/a:1011916216191

Jeon, J.-S., Lee, S., Jung, K.-H., Jun, S.-H., Jeong, D.-H., Lee, J., et al. (2000). T-DNA Insertional Mutagenesis for Functional Genomics in rice. *Plant J.* 22, 561–570. doi:10.1046/j.1365-313x.2000.00767.x

Jeong, D.-H., An, S., Kang, H.-G., Moon, S., Han, J.-J., Park, S., et al. (2002). T-DNA Insertional Mutagenesis for Activation Tagging in rice. *Plant Physiol.* 130, 1636–1644. doi:10.1104/pp.014357

Kardailsky, I., Shukla, V. K., Ahn, J. H., Dagenais, N., Christensen, S. K., Nguyen, J. T., et al. (1999). Activation Tagging of the floral Inducer FT. *Science* 286, 1962–1965. doi:10.1126/science.286.5446.1962

Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative Prediction of Mammalian Enhancers from DNA Sequence. *Genome Res.* 21, 2167–2180. doi:10.1101/gr.121905.111

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., et al. (2012). Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* 148, 84–98. doi:10.1016/j.cell.2011.12.014

Liao, C.-C., Chen, L.-J., Lo, S.-F., Chen, C.-W., and Chu, Y.-W. (2019). EAT-Rice: A Predictive Model for Flanking Gene Expression of T-DNA Insertion Activation-Tagged rice Mutants by Machine Learning Approaches. *Plos Comput. Biol.* 15, e1006942. doi:10.1371/journal.pcbi.1006942

Lin, H., Deng, E.-Z., Ding, H., Chen, W., and Chou, K.-C. (2014). iPro54-PseKNC: a Sequence-Based Predictor for Identifying Sigma-54 Promoters in Prokaryote with Pseudo K-Tuple Nucleotide Composition. *Nucleic Acids Res.* 42, 12961–12972. doi:10.1093/nar/gku1019

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a Web Server for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Nucleic Acids Res.* 43, W65–W71. doi:10.1093/nar/gkv458

Marcotte, W. R., Jr., Russell, S. H., and Quatrano, R. S. (1989). Abscisic Acid-Responsive Sequences from the Em Gene of Wheat. *Plant Cell* 1, 969–976. doi:10.1105/tpc.1.10.969

Marenduzzo, D., Faro-Trindade, I., and Cook, P. R. (2007). What Are the Molecular Ties that Maintain Genomic Loops? *Trends Genet.* 23, 126–133. doi:10.1016/j.tig.2007.01.007

Mason, H. S., Dewald, D. B., and Mullet, J. E. (1993). Identification of a Methyl Jasmonate-Responsive Domain in the Soybean vspB Promoter. *Plant Cell* 5, 241–251. doi:10.1105/tpc.5.3.241

Noble, W. S., Kuehn, S., Thurman, R., Yu, M., and Stamatoyannopoulos, J. (2005). Predicting the *In Vivo* Signature of Human Gene Regulatory Sequences. *Bioinformatics* 21 (Suppl. 1), i338–43. doi:10.1093/bioinformatics/bti1047

Odell, J. T., Nagy, F., and Chua, N.-H. (1985). Identification of DNA Sequences Required for Activity of the Cauliflower Mosaic Virus 35S Promoter. *Nature* 313, 810–812. doi:10.1038/313810a0

Ohan, N. W., and Heikkila, J. J. (1993). Reverse Transcription-Polymerase Chain Reaction: an Overview of the Technique and its Applications. *Biotechnol. Adv.* 11, 13–29. doi:10.1016/0734-9750(93)90408-f

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., et al. (2007). The TIGR Rice Genome Annotation Resource: Improvements and New Features. *Nucleic Acids Res.* 35, D883–D887. doi:10.1093/nar/gkl976

Peng, H., Fuhui Long, F., and Ding, C. (2005). Feature Selection Based on Mutual Information Criteria of max-dependency, max-relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Machine Intell.* 27, 1226–1238. doi:10.1109/tpami.2005.159

Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield Trends Are Insufficient to Double Global Crop Production by 2050. *PLoS One* 8, e66428. doi:10.1371/journal.pone.0066428

Ren, S., Johnston, J. S., Shippen, D. E., and Mcknight, T. D. (2004). TELOMERASE ACTIVATOR1 Induces Telomerase Activity and Potentiates Responses to Auxin in Arabidopsis. *Plant Cell* 16, 2910–2922. doi:10.1105/tpc.104.025072

Sakata, K., Nagamura, Y., Numa, H., Antonio, B. A., Nagasaki, H., Idonuma, A., et al. (2002). RiceGAAS: an Automated Annotation System and Database for rice Genome Sequence. *Nucleic Acids Res.* 30, 98–102. doi:10.1093/nar/30.1.98

Singer, S. D., Cox, K. D., and Liu, Z. (2010). Both the Constitutive Cauliflower Mosaic Virus 35S and Tissue-specific AGAMOUS Enhancers Activate Transcription Autonomously in Arabidopsis thaliana. *Plant Mol. Biol.* 74, 293–305. doi:10.1007/s11103-010-9673-9

Singer, S. D., Cox, K. D., and Liu, Z. (2011). Enhancer-promoter Interference and its Prevention in Transgenic Plants. *Plant Cel Rep* 30, 723–731. doi:10.1007/s00299-010-0977-7

Stewart, A. J., Hannenhalli, S., and Plotkin, J. B. (2012). Why Transcription Factor Binding Sites Are Ten Nucleotides Long. *Genetics* 192, 973–985. doi:10.1534/genetics.112.143370

van der Geest, A. H. M., and Hall, T. C. (1997). The Beta-Phaseolin 5' Matrix Attachment Region Acts as an Enhancer Facilitator. *Plant Mol. Biol.* 33, 553–557. doi:10.1023/a:1005765525436

Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I. S., and Martienssen, R. A. (2002). Regulation of Heterochromatic Silencing and Histone H3 Lysine-9 Methylation by RNAi. *Science* 297, 1833–1837. doi:10.1126/science.1074973

Weigel, D., Ahn, J. H., Bla´zquez, M. A., Borevitz, J. O., Christensen, S. K., Fankhauser, C., et al. (2000). Activation Tagging in Arabidopsis. *Plant Physiol.* 122, 1003–1014. doi:10.1104/pp.122.4.1003

Yang, Y., Li, Y., and Wu, C. (2013). Genomic Resources for Functional Analyses of the rice Genome. *Curr. Opin. Plant Biol.* 16, 157–163. doi:10.1016/j.pbi.2013.03.010

Yuan, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R., et al. (2003). The TIGR rice Genome Annotation Resource: Annotating the rice Genome and Creating Resources for Plant Biologists. *Nucleic Acids Res.* 31, 229–233. doi:10.1093/nar/gkg059

Yuan, Y., Guo, L., Shen, L., and Liu, J. S. (2007). Predicting Gene Expression from Sequence: a Reexamination. *Plos Comput. Biol.* 3, e243. doi:10.1371/journal.pcbi.0030243

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.

Copyright © 2021 Chien, Huang, Lo, Chen, Liao, Chen and Chu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 December 2021
doi: 10.3389/fgene.2021.809001

SNAREs-SAP: SNARE Proteins Identification With PSSM Profiles

Zixiao Zhang[1], Yue Gong[1], Bo Gao[2], Hongfei Li[1], Wentao Gao[1], Yuming Zhao[1]* and Benzhi Dong[1]**

[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2]Department of Radiology, The Second Affiliated Hospital, Harbin Medical University, Harbin, China

Soluble N-ethylmaleimide sensitive factor activating protein receptor (SNARE) proteins are a large family of transmembrane proteins located in organelles and vesicles. The important roles of SNARE proteins include initiating the vesicle fusion process and activating and fusing proteins as they undergo exocytosis activity, and SNARE proteins are also vital for the transport regulation of membrane proteins and non-regulatory vesicles. Therefore, there is great significance in establishing a method to efficiently identify SNARE proteins. However, the identification accuracy of the existing methods such as SNARE CNN is not satisfied. In our study, we developed a method based on a support vector machine (SVM) that can effectively recognize SNARE proteins. We used the position-specific scoring matrix (PSSM) method to extract features of SNARE protein sequences, used the support vector machine recursive elimination correlation bias reduction (SVM-RFE-CBR) algorithm to rank the importance of features, and then screened out the optimal subset of feature data based on the sorted results. We input the feature data into the model when building the model, used 10-fold crossing validation for training, and tested model performance by using an independent dataset. In independent tests, the ability of our method to identify SNARE proteins achieved a sensitivity of 68%, specificity of 94%, accuracy of 92%, area under the curve (AUC) of 84%, and Matthew's correlation coefficient (MCC) of 0.48. The results of the experiment show that the common evaluation indicators of our method are excellent, indicating that our method performs better than other existing classification methods in identifying SNARE proteins.

Keywords: SNARE proteins, position-specific scoring matrix, machine learning, support vector machine, SVM-RFE-CBR

OPEN ACCESS

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

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Liran Juan,
Harbin Institute of Technology, China

****Correspondence:***
Zixiao Zhang
zixiao_zhang@nefu.edu.cn
Benzhi Dong
nefudbz@nefu.edu.cn
Yuming Zhao
zym@nefu.edu.cn

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

Received: *04 November 2021*
Accepted: *15 November 2021*
Published: *20 December 2021*

Citation:
Zhang Z, Gong Y, Gao B, Li H, Gao W, Zhao Y and Dong B (2021) SNAREs-SAP: SNARE Proteins Identification With PSSM Profiles. Front. Genet. 12:809001. doi: 10.3389/fgene.2021.809001

1 INTRODUCTION

N-ethylmaleimide sensitive factor (NSF) (Whiteheart et al., 2001) protein and soluble NSF attachment proteins (SNAPS) (Whiteheart et al., 1993) are two essential factors for protein transport between membranes (Hohl et al., 1998) (Hanson et al., 1997). They were first discovered as essential proteins for protein transport from donor to receptor subcellular structures during the processes of Golgi modification and secretion. The discovery of these two proteins led to the discovery of multiple receptor proteins on transport vesicles and plasma membranes and snap receptors, which are collectively called soluble N-ethylmaleimide-sensitive factor activating protein receptor (SNARE) proteins (Ungar and Hughson, 2003; Zhao et al., 2019). According to the SNARE theory, exocytosis and secretory processes are completed by precise coordination between SNARE proteins. The specificity of membrane fusion is based on the specific

binding of SNARE protein members. At the molecular level, when the transport vesicle is close to the target membrane, syntaxin1A/B on the target membrane receives a signal to recognize, approach and combine with SNAP25, which is also located on the target membrane. At the same time, VAMP2 (q-snare) on the transport vesicle also recognizes (Kweon et al., 2003), draws close to and binds to form a 7S R-Q-SNARE complex, which guides the attachment and fusion of the transport vesicle and the target membrane, leading to the secretion of substances in the transport vesicle into the new subcellular structure or out of the cell through exocytosis, completing the intracellular transport and extracellular exocytosis and secretion processes.

The binding sites of SNARE proteins are specific, which is the reason for the specificity and precision of exocytosis and secretion in different organisms and organs (Fasshauer et al., 1998; Yin et al., 2021). SNARE theory convincingly explains the key role of synapses in the process of nerve impulse transmission at the molecular level (Chen and Scheller, 2001). Its new insights in the fields of molecular neurobiology and endocrinology have made research on SNARE proteins a hot spot in the basic life sciences worldwide. Such findings greatly enrich understanding of the regulation of intracellular information transmission, substance transport and exocytosis and secretion at the molecular level and improve knowledge of the interaction between proteins and the plasma membrane (Liu et al., 2019a; Wang et al., 2020a; Xu et al., 2021).

Due to the important roles of SNARE proteins in cell biology, research on SNARE proteins is also developing, and a variety of technologies are used to study SNARE proteins (Wang et al., 2020b; Yin et al., 2020), including the establishment of a SNARE protein database, the retrieval and classification of SNARE proteins, bioinformatics technology that was used to predict the role of SNARE proteins, and construction of a neural network model to recognize SNARE proteins.

With the development of computational biology, the application of machine learning to bioinformatics continues to be deep and widespread (Jiang et al., 2013; Tao et al., 2020; Zhao et al., 2021). Machine learning is complex and cross disciplinary across multiple fields (Cheng, 2020). Machine learning obtains new knowledge through learning from pre-existing knowledge and can continuously advance itself based on large quantities of this pre-existing knowledge and skills. Research on machine learning includes the study of computer algorithms, using data and previous techniques to improve the performance of computer algorithms. Machine learning also has significant implications for the development of artificial intelligence, through which computers continuously progress along a path of constant intelligence. A typical way to predict proteins is to transform each protein sequence into a numerical eigenvector used to represent the protein sequence, training a classification model based on the eigenvectors of the training samples and the labels. After feature construction, the classifiers that make predictions about proteins include covariant discriminant (CD) (Chou, 2000), support vector machine (SVM) (Hua and Sun, 2001), K-nearest neighbor (KNN) (Shen and Chou, 2006), deep learning and ensemble classifiers (Shen and Chou, 2006).

In this study, based on SVM classifier (Liu et al., 2010), we constructed a model to recognize SNARE proteins. We use position-specific scoring matrix (PSSM) profiles of protein sequences to extract features (Kumar et al., 2008), process the feature data by the min-max normalization method, build a model based on SVM, train the model with 10-fold cross validation and measure the performance of the model on an independent dataset.

2 MATERIALS AND METHODS

We developed a method to recognize SNARE proteins based on PSSM (Chou and Shen, 2007; Liu et al., 2019b; Hong et al., 2020a; Hong et al., 2020b) profiles and SVM. Method steps include data collection, data processing, feature extraction, feature selection, model training, and model performance evaluation. The overall flow of our designed method is summarized in **Figure 1**, and each section in the figure is described in detail in the following sections. We carried out experiments through the above process, constantly adjusted in our experiment, and finally constructed an excellent method to identify SNARE proteins. The following is a detailed description of the method.

2.1 Feature Extraction

It is very important to select good feature information for protein recognition (Zuo et al., 2017; Zheng et al., 2019; Tang et al., 2020a; Guo et al., 2020; Zhang et al., 2021). We chose the method based on PSSM profiles to extract the feature information of protein sequence data. We use the National Center for Biotechnology Information basic local alignment search tool (NCBI-BLAST) and select a non-redundant (NR) protein sequence database as a comparison dataset. We use the prepared SNARE protein FASTA sequence files to generate PSSM profiles. Each amino acid of the original sequence in the PSSM profiles consists of a vector of 20 values. Then, we transform the original PSSM files into PSSM profiles with 400 dimensions. Finally, 400-dimensional data are extracted as the feature data of each protein sequence for the experiment.

2.2 Data Processing

The feature data in the datasets are seriously unbalanced, especially the ratio of positive samples to negative samples in the independent dataset, which varies tremendously. The model would exhibit the problem of poor generalization, and the applicability would be low, so it is unable to effectively identify SNARE proteins. Therefore, we need to choose the appropriate method to deal with the data. In this study, the data processing methods we chose included Z-score standardization, min-max normalization and L2 regularization.

Normalization: Data can be changed to [0, one] ranges using the normalization method. Normalization, as an effective way to simplify calculation and scale down data values, can change the absolute values of data in the dataset into a relationship of some relative value. After normalization, the data can be calculated conveniently and quickly. This is mainly for the convenience of data processing, mapping the data to the range of 0–1, which will be convenient and fast to use. The method is defined as:

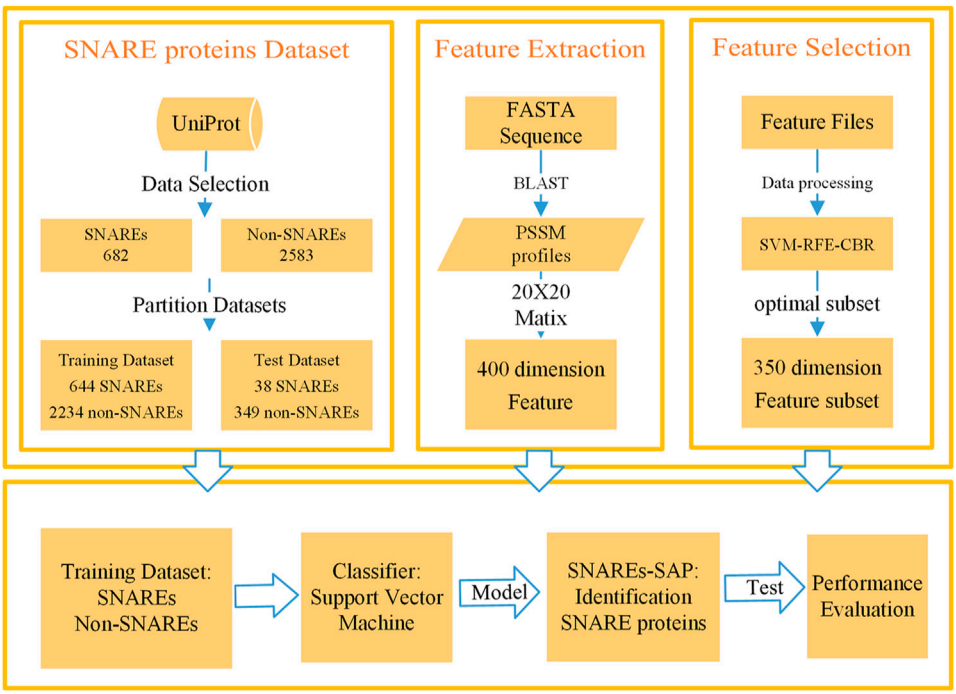


FIGURE 1 | Flow chart of SNARE proteins recognition based on PSSM profiles matrix and SVM.

$$x^{*} = \frac{x - \min}{\max - \min}, \# \quad (1)$$

The distribution of original data can be changed by normalization, and then the weights of each feature dimension can be balanced by varying the feature dimension, such as converting the distribution of data from planar to circular. Normalization can remove the influence of dimensionality on the experimental results by reducing the difference in dimensionality. After normalization, the data of different variables can be compared. Although the maximum and minimum values of the resulting data in the normalization process are affected by outliers in the dataset, and the resulting data are less robust, normalization does improve the accuracy of iterations in the operational data process as well as the efficiency of data convergence.

2.3 Feature Selection

Feature selection refers to sorting features by suitable techniques and algorithms and filtering out the better characterized subset of features based on the sorted results; this is a common technique in bioinformatics (Cheng et al., 2018; Zhu et al., 2019; Zhao et al., 2020a; Zhao et al., 2020b; Shao and Liu, 2021; Yu et al., 2021). After feature selection, the optimal feature subset selected from existing features is used to build the model, which can improve the performance of the model. Feature selection is a very important part of building models for pattern recognition and is a high priority in data processing (Wei et al., 2018; Xue et al., 2018; Li et al., 2020a; Yang et al., 2020a; Su et al., 2020; Wei et al., 2020; Yu et al., 2020; Zhang et al., 2020; Zheng et al., 2020; Wang et al., 2021a; Shang et al., 2021; Shao et al., 2021). Selecting the effective features from the original feature dataset and removing the redundant features can reduce the dimensionality of the feature data, and using more effective feature data can improve the performance of the model. Our original feature is based on PSSM to extract 400 dimensional features. In these original feature spaces, there will be irrelevant, noisy, and redundant features. Suitable feature selection methods with excellent performance are required for accurate screening of redundant features. In our experiment, we finally chose the SVM-RFE-CBR (Yan and Zhang, 2015) algorithm to screen features after comparing multiple feature selection methods. The algorithm ranks the importance of features and selects the optimal subset of features based on the sorted results.

SVM-RFE-CBR is an improved algorithm based on support vector machine recursive feature elimination (SVM-RFE), which introduces the strategy of correlation deviation reduction (CBR) into the process of feature elimination. SVM-RFE estimates feature importance based on the coefficient of the SVM model, and it is a powerful feature selection algorithm. There are linear and nonlinear versions. The SVM-RFE-CBR method adds the correlation reduction strategy (CBR) to the SVM-RFE algorithm to reduce the potential deviation of the algorithm, and the result of feature selection is improved by the integrated CBR strategy. SVM-RFE uses the sequential backward selection algorithm in SVM, which is based on the principle of maximum interval. During the model training process, SVM-RFE sort features based on the score of every feature, deletes the feature with the lowest score, puts the remaining feature data into the next round of training of the model, and finally outputs the feature sort result to

a table. The optimal feature subset can be selected according to the results of sorting. SVM is an excellent machine learning classification algorithm. The feature sort result derived from the SVM model has better performance, and it is also more convenient for subsequent experiments.

2.4 Support Vector Machine

SVM is currently a commonly used classifier in machine learning that classifies data by supervised learning (Cheng et al., 2019a; Cheng et al., 2019b). SVM is commonly used in data dichotomization. In addition, SVM can classify nonlinearly by using the kernel function (Ding et al., 2020a; Liu et al., 2020a; Yang et al., 2020b). SVM was developed from the generalized portrait algorithm in pattern recognition. The basic idea of SVM is to construct a model that separates the dataset accurately according to the geometric interval of the hyperplane with the maximum separation of samples. SVM can map the features of a dataset to points in space and draw a line to distinguish these points effectively. SVM uses a hinge loss function to computationally predict the presence of empirical risk, and a regularization term is added to ensure its robustness and correct rate. The process of SVM: Suppose the training set is $\{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in R^D$,$y_i \in \{+1, -1\}$, x_i is the *i*th sample, N is the sample size, and D is the number of sample features. SVM finding the optimal classification hyperplane.$\omega \cdot x + b = 0$ The optimization problems that SVM needs to solve are:

$$\begin{aligned} &\min \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}\varepsilon_i.\# \\ &\text{s.t. } y_i(\omega \cdot x_i + b) \geq 1 - \varepsilon_i,\ i = 1, 2, \cdots, N \\ &\varepsilon_i \geq 0,\ i = 1, 2, \cdots, N \end{aligned} \tag{2}$$

Transforming the original problem into the dual problem:

$$\min \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i . \tag{3}$$

$$\text{s.t. } \sum_{i=1}^{N} y_i\alpha_i = 0 .\# \tag{4}$$

$0 \leq \alpha_i \leq C,\ i = 1, 2, \cdots, N$ α_i is a Lagrangian

Finally, the solution of ω is:

$$\omega = \sum_{i=1}^{N}\alpha_i y_i x_i.\# \tag{5}$$

When we use SVM to solve nonlinear problems, we need to choose the appropriate kernel function (Yang et al., 2021a) (Ding et al., 2020b) and then map the data to the high-dimensional space to solve the linearly inseparable problem of the data in the original space.

In the experiment, the Python version of a library for support vector machine (LIBSVM) was selected to build an SVM model and identify SNARE proteins. The selection of different kernel functions using LIBSVM as well as the settings of kernel parameters are described as follows: The kernel function (Ding et al., 2020c) of SVM includes the linear kernel function (LKF), polynomial kernel function (PKF), radial basis function (RBF), and sigmoid kernel function (SKF). Formulas corresponding to four kernel functions are as follows:

Linear kernel function defined as:

$$K(x_i, x_j) = x_i^T x_j.\# \tag{6}$$

Polynomial kernel function:

$$K(x_i, x_j) = (\nu x_i^T x_j + r)^d,\ \nu > 0.\# \tag{7}$$

Radial basis functions:

$$K(x_i, x_j) = \exp(-\nu||x_i - x_j||^2),\ \nu > 0.\# \tag{8}$$

Sigmoid kernel function:

$$K(x_i, x_j) = \tanh(\nu x_i^T + r).\# \tag{9}$$

ν, r, and d in formulas are parameters of kernel function.

Parameters are different in different kernel functions. ν in the formula represents the parameter gamma in the kernel function, the default of which is 1/K (K is the number of classes), and g is used to set it in the LIBSVM.

r in the formula represents the parameter r in the kernel function, the default of which is 0, and r is used to set it in the LIBSVM. d in the formula represents the parameter d in the kernel function; it is used to set the highest number of times in the polynomial kernel function, and its default value is 3.

SVM is a very powerful model that allows the decision boundary to be very complex and performs well on both low-dimensional data and high-dimensional data. SVM has been widely used in bioinformatics, binding protein prediction, protein methylation site prediction and so on. We use the LIBSVM of Scikit-learn library integration in Python to train and build the model. In our experimental process, we optimize the parameters according to the results and finally build the model with the best performance.

3 RESULTS AND DISCUSSION

3.1 Dataset

Our research is devoted to constructing a method to recognize SNARE proteins. To establish a model to effectively distinguish SNARE proteins and non-SNARE proteins, we collected a SNARE protein dataset and a non-SNARE protein dataset for our prediction model. The dataset we use has been used by Le, N.Q.K. and V.-N. Nguyen (Le and Nguyen, 2019) previously. The data come from the UniProt database, which is the most informative and resource-free protein database. We collect all SNARE proteins from the UniProt database according to the keyword SNARE. To avoid the homology of the SNARE protein sequence data that we collect, we use BLAST to address the redundancy of the SNARE protein sequence and eliminate the redundant sequence. Finally, 682 SNARE protein sequences are obtained as a positive sample dataset. At the same time, we select vesicular transport proteins as negative samples to establish a non-SNARE protein dataset. We divide the two datasets into a cross-validation dataset and an independent test dataset, and the size and details of the datasets are summarized in **Table 1**.

TABLE 1 | Summary of SNARE protein and non-SNARE protein datasets.

Dataset	SNARE	Non-SANRE	Total
Original dataset	682	2,583	3,265
Train dataset	644	2,234	2,878
Test dataset	38	349	387

Table 1 shows that SNARE proteins and non-SNARE proteins correspond to two datasets: a training dataset and an independent test dataset, both of which include positive samples and negative samples. We use the cross-validation method to train the model with the training dataset, evaluate the performance of the model developed in this study, and optimize the model by adjusting the parameters according to the results of the training dataset. The independent test dataset is used to test and measure the predictive ability of the prediction model we developed.

3.2 Performance Measurements

Our research aims to establish a model to predict whether an amino acid sequence is a SNARE protein. Therefore, we need to use universally acknowledged evaluation indicators to measure the performance of the model. When training the model, we choose 10-fold cross validation as the training model after various considerations and take the average value of the crossing validation results as the result of model training. We optimize the parameters of SVM, select the best parameters to build the model, and evaluate the performance of the model through an independent test dataset to avoid systematic deviation in the process of cross validation. This study adopts some standard evaluation indicators that are widely used in bioinformatics research (Shen et al., 2019a; Shen et al., 2019b; Ao et al., 2020; Li et al., 2020b; Liu et al., 2020b; Tang et al., 2020b; Yin et al., 2020; Chen et al., 2021). The standard evaluation indicators include sensitivity (Sn), specificity (Sp), accuracy (Acc), area under the curve (AUC), Mathew's correlation coefficient (MCC), and F-score (Zhai et al., 2020; Wang et al., 2021b; Yang et al., 2021b). The calculation formulas are as follows (TP means true positive values, FP means false positive values, TN means true negative values, FN means false negative values):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},\ 0 \le Sn \le 1.\# \tag{10}$$

$$\text{Speccificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},\ 0 \le Sp \le 1.\# \tag{11}$$

$$\text{Accurarcy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}},\ 0 \le \text{Acc} \le 1.\# \tag{12}$$

$$\text{MCC} = \frac{\text{TP*TN} - \text{FP*FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TN} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})}},\ 0 \le \text{MCC} \le 1.\# \tag{13}$$

$$\text{F} - \text{score} = \frac{2\text{*TP}}{2\text{TP} + \text{FN} + \text{FP}},\ 0 \le \text{F} - \text{score} \le 1.\# \tag{14}$$

In machine learning research, receiver operating characteristic (ROC) curves are usually used to test the prediction performance of the model. AUC is a floating-point number from 0 to one of ROC. The AUC value can reflect the quality of the model. The greater the value, the better the performance of the model. ROC curves and AUCs are commonly used to compare the performance of different models as machine learning performance indicators, which is very reliable. MCC is often used to measure imbalanced data sets, which is one of the most important indicators to measure the performance of two kinds of classification in machine learning. We use Python's processing library to process data.

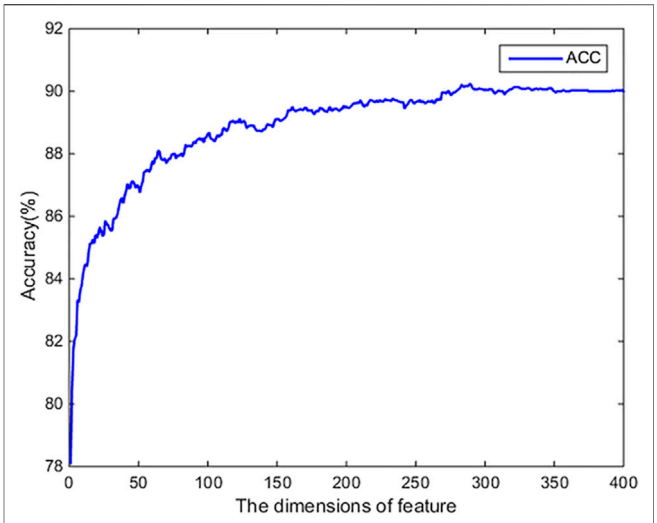


FIGURE 2 | The results of dimension reduction by using SVM-RFE-CBR algorithm.

TABLE 2 | Comparison of prediction results between SVM-RFE-CBR dimension reduction and original dimension.

Feature-dimension	Sn	Sp	Acc	AUC	MCC	F-score
350	**0.68**	**0.94**	**0.92**	**0.84**	**0.48**	**0.5**
400	0.68	0.94	0.91	0.83	0.48	0.5

Comparison of prediction results between SVM-RFE-CBR dimension reduction and original dimension. The bold values mean maximum value in the column.

3.3 Performance Comparison With Different Feature Dimensions

We use the SVM-RFE-CBR algorithm to evaluate the original 400-dimensional feature data. We use MATLAB to implement the SVM-REF-CBR algorithm to sort the features. When sorting features, a performance comparison will be given. The evaluation results are shown in **Figure 2**. From **Figure 2**, it can be found that the ACC achieved highest value, when the top 350-dimensional feature is used in the experiment. Therefore, we choose 350-dimensional feature data for the experiment.

We use the optimal 350-dimensional feature dataset after sorting for the experiment. First, 350-dimensional feature data are selected from the original feature training dataset and test dataset files according to the index obtained by the SVM-RFE-CBR algorithm. Then, the training dataset is 10-fold cross

TABLE 3 | The result of performance compares between SVM and other classification method.

	Sn	Sp	Acc	MCC
KNN	**0.870**	0.906	0.898	**0.73**
Random Forest	0.620	0.962	0.900	0.70
Naïve Bayes	0.853	0.595	0.624	0.28
SVM	0.650	**0.970**	**0.900**	0.70

The result of performances compares between SVM and other classification method. The bold values mean maximum value in the column.

validated, and the model is optimized. After many experiments, the optimal parameters of SVM are obtained. When we choose the radial basis function, penalty coefficient (C) = "11", gamma = "0.1", the model achieves the optimal performance. At the same time, we also use the original 400-dimensional feature data for the experiment and choose the optimal parameterization in the experiment. The comparison of experimental results in different dimensions is shown in **Table 2**.

The experimental results show that both Acc and MCC are improved after feature dimensionality reduction, which eliminates the redundant part of the original feature and improves the performance of the model.

3.4 Comparison of Different Classifier Performance on Dataset

With the development of computers, machine learning has been widely used in bioinformatics (Tang et al., 2019; Wang et al., 2020c; Fu et al., 2020; Cai et al., 2021; Wang et al., 2021c; Jin et al., 2021), and there are many classification models, including the linear classifier, SVM, naive byes, K-nearest neighbor (KNN), decision tree (DT), and ensemble model (random forest/GDBT, etc.). To obtain the most effective classifier method to identify SNARE proteins, we use various machine learning classifiers to construct a model of SNARE protein recognition, including random forest, KNN and naive Bayes.

We compare the experimental results of multiple machine learning classifier training models with the performance measurement results. The performance result of different classifier shown in **Table 3**.

As we can observe from **Table 3**, the results of SVM on training dataset are better than another classifier.

In particular, Sp = 0.970, Acc = 0.900. SVM shows higher performance. Meanwhile, we compare the ROC curves of different classifier method. The result shown in **Figure 3**. As we can observe from **Figure 3**, The ROC curve of SVM is obviously better than the other three classifiers.

3.5 Comparison of Different SNARE Protein Identification Methods

We compare the experimental results of SNARE CNN with the performance measurement results of our research method. The independent test results of using different methods to identify SNARE proteins are shown in **Figure 4**. **Figure 4A** shows the result of performance compares between our classification

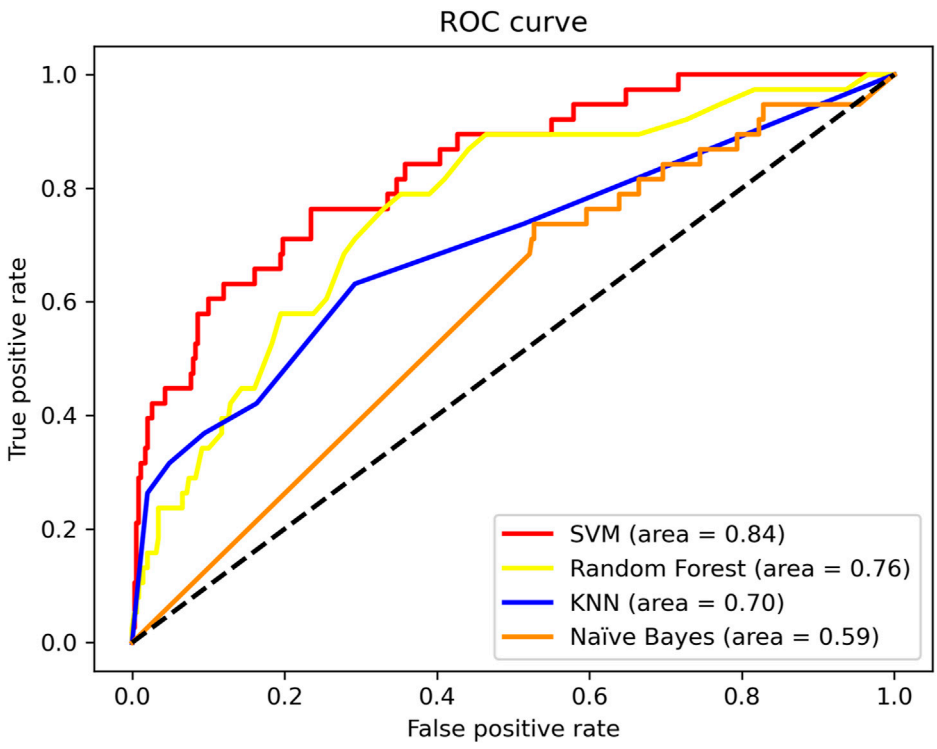


FIGURE 3 | ROC curves of different classifier methods.

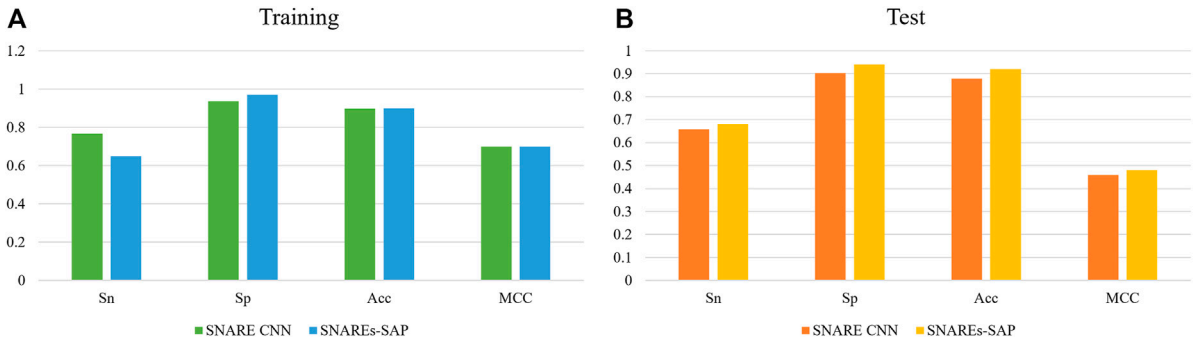


FIGURE 4 | (A) The result of performance compares between our classification method and other classification method on training datasets **(B)** The result of performance compares between our classification method and other classification method on test datasets.

method and other classification method on training datasets. **Figure 4B** shows the result of performance compares between our classification method and other classification method on test datasets.

The results show that our method gives good results in both training and independent test datasets. To compare the performance measurements of our method for identifying SNARE proteins with other methods more accurately, we compare the results of different methods on independent test datasets.As we can observe from **Figure 4B**, the independent test results of our method are better than SANRE CNN. Sn = 0.68, Sp = 0.940, Acc = 0.92 and MCC = 0.48, and all these indicators reach the highest values using our method. As shown above, our method shows higher performance. These results clearly demonstrate the superiority of our method over the existing methods, especially when using an independent dataset test. This means that our method can better recognize SNARE proteins.

4 DISCUSSION

Because of the importance of SNARE proteins and the vital significance of SNARE proteins in vesicular transport, there is an urgent need for classification methods to identify SNARE proteins. Extracting meaningful features and selecting an appropriate machine learning algorithm can greatly increase the model performance of protein prediction. We propose a method based on PSSM profiles to extract features and SVM to construct a model to identify SNARE proteins. We normalize the feature data and use the SVM-RFE-CBR algorithm to reduce the dimensions of feature. Then, we use a 10-fold crossing validation training model and use an independent dataset to test the performance of the model (Li et al., 2017; Li et al., 2020c). The accuracy, specificity, sensitivity, AUC, MCC and other performance indicators of our method have excellent experimental results. All results show that our model has better performance than other machine learning methods and advanced neural networks. Our method can effectively identify SNARE proteins. Taken together, the method proposed in our study is of great significance for the study of SNARE proteins and may also contribute to the prediction of protein function. Future works may include investigation of more kinds of proteins.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories: https://github.com/First-Leaner/Identify-proteins. The names of the repository/repositories and accession number(s) can also be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

ZZ and BD conceived and designed the project. ZZ, HL, and YG conducted the experiments and analyzed the data. ZZ and BG wrote the manuscript. BD, WG, and YZ revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Natural Science Foundation of China (No. 62172129).

REFERENCES

Ao, C., Zhou, W., Gao, L., Dong, B., and Yu, L. (2020). Prediction of Antioxidant Proteins Using Hybrid Feature Representation Method and Random forest. *Genomics* 112 (6), 4666–4674. doi:10.1016/j.ygeno.2020.08.016

Cai, L., Ren, X., Fu, X., Peng, L., Gao, M., and Zeng, X. (2021). iEnhancer-XG: Interpretable Sequence-Based Enhancers and Their Strength Predictor. *Bioinformatics* 37, 1060–1067. doi:10.1093/bioinformatics/btaa914

Chen, Y. A., and Scheller, R. H. (2001). SNARE-Mediated Membrane Fusion. *Nat. Rev. Mol. Cel. Biol.* 2 (2), 98–106. doi:10.1038/35052017

Chen, Y., Ma, T., Yang, X., Wang, J., Song, B., and Zeng, X. (2021). MUFFIN: Multi-Scale Feature Fusion for Drug-Drug Interaction Prediction. *Bioinformatics* 37, 2651–2658. doi:10.1093/bioinformatics/btab169

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., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019). MetSigDis: a Manually Curated Resource for the Metabolic Signatures of Diseases. *Brief Bioinform.* 20 (1), 203–209. doi:10.1093/bib/bbx103

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

Cheng, L. (2020). Omics Data and Artificial Intelligence: New Challenges for Gene Therapy. *Cgt* 20 (1), 1. doi:10.2174/1566523220012006041500 41

Chou, K.-C., and Shen, H.-B. (2007). MemType-2L: A Web Server for Predicting Membrane Proteins and Their Types by Incorporating Evolution Information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* 360 (2), 339–345. doi:10.1016/j.bbrc.2007.06.027

Chou, K.-C. (2000). Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochem. Biophys. Res. Commun.* 278 (2), 477–483. doi:10.1006/bbrc.2000.3815

Ding, Y., Tang, J., and Guo, F. (2020). Identification of Drug-Target Interactions via Fuzzy Bipartite Local Model. *Neural Comput. Applic* 32 (14), 10303–10319. doi:10.1007/s00521-019-04569-z

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

Fasshauer, D., Sutton, R. B., Brunger, A. T., and Jahn, R. (1998). Conserved Structural Features of the Synaptic Fusion Complex: SNARE Proteins Reclassified as Q- and R-SNAREs. *Proc. Natl. Acad. Sci.* 95 (26), 15781–15786. doi:10.1073/pnas.95.26.15781

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

Hanson, P. I., Heuser, J. E., and Jahn, R. (1997). Neurotransmitter Release - Four Years of SNARE Complexes. *Curr. Opin. Neurobiol.* 7 (3), 310–315. doi:10.1016/s0959-4388(97)80057-8

Hohl, T. M., Parlati, F., Wimmer, C., Rothman, J. E., Söllner, T. H., and Engelhardt, H. (1998). Arrangement of Subunits in 20 S Particles Consisting of NSF, SNAPs, and SNARE Complexes. *Mol. Cel.* 2 (5), 539–548. doi:10.1016/s1097-2765(00)80153-7

Hong, J., Luo, Y., Zhang, Y., Ying, J., Xue, W., Xie, T., et al. (2020). Protein Functional Annotation of Simultaneously Improved Stability, Accuracy and False Discovery Rate Achieved by a Sequence-Based Deep Learning. *Brief Bioinform.* 21 (4), 1437–1447. doi:10.1093/bib/bbz081

Hong, J., Luo, Y., Mou, M., Fu, J., Zhang, Y., Xue, W., et al. (2020). Convolutional Neural Network-Based Annotation of Bacterial Type IV Secretion System Effectors with Enhanced Accuracy and Reduced False Discovery. *Brief Bioinform.* 21 (5), 1825–1836. doi:10.1093/bib/bbz120

Hua, S., and Sun, Z. (2001). Support Vector Machine Approach for Protein Subcellular Localization Prediction. *Bioinformatics* 17 (8), 721–728. doi:10.1093/bioinformatics/17.8.721

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

Kumar, M., Gromiha, M. M., and Raghava, G. P. S. (2008). Prediction of RNA Binding Sites in a Protein Using SVM and PSSM Profile. *Proteins* 71 (1), 189–194. doi:10.1002/prot.21677

Kweon, D.-H., Kim, C. S., and Shin, Y.-K. (2003). Regulation of Neuronal SNARE Assembly by the Membrane. *Nat. Struct. Mol. Biol.* 10 (6), 440–447. doi:10.1038/nsb928

Le, N. Q. K., and Nguyen, V.-N. (2019). SNARE-CNN: a 2D Convolutional Neural Network Architecture to Identify SNARE Proteins from High-Throughput Sequencing Data. *PeerJ Comp. Sci.* 5, e177. doi:10.7717/peerj-cs.177

Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., et al. (2017). NOREVA: Normalization and Evaluation of MS-based Metabolomics Data. *Nucleic Acids Res.* 45 (W1), W162–W170. doi:10.1093/nar/gkx449

Li, F., Zhou, Y., Zhang, X., Tang, J., Yang, Q., Zhang, Y., et al. (2020). SSizer: Determining the Sample Sufficiency for Comparative Biological Study. *J. Mol. Biol.* 432 (11), 3411–3421. doi:10.1016/j.jmb.2020.01.027

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.* 22 (3), bbaa159. doi:10.1093/bib/bbaa159

Li, J., Pu, Y., Tang, J., Zou, Q., and Guo, F. (2020). DeepAVP: a Dual-Channel Deep Neural Network for Identifying Variable-Length Antiviral Peptides. *IEEE J. Biomed. Health Inform.* 24 (10), 3012–3019. doi:10.1109/jbhi.2020.2977091

Liu, T., Zheng, X., and Wang, J. (2010). Prediction of Protein Structural Class for Low-Similarity Sequences Using Support Vector Machine and PSI-BLAST Profile. *Biochimie* 92 (10), 1330–1334. doi:10.1016/j.biochi.2010.06.013

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, 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., Li, C.-C., and Yan, K. (2020). DeepSVM-fold: Protein Fold Recognition by Combining Support Vector Machines and Pairwise Sequence Similarity Scores Generated by Deep Learning Networks. *Brief. Bioinform.* 21 (5), 1733–1741. doi:10.1093/bib/bbz098

Liu, B., Zhu, Y., and Yan, K. (2020). Fold-LTR-TCP: Protein Fold Recognition Based on Triadic Closure Principle. *Brief. Bioinform.* 21 (6), 2185–2193. doi:10.1093/bib/bbz139

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 (3), bbaa144. doi:10.1093/bib/bbaa144

Shen, H.-B., and Chou, K.-C. (2006). Ensemble Classifier for Protein Fold Pattern Recognition. *Bioinformatics* 22 (14), 1717–1722. doi:10.1093/bioinformatics/btl170

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

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. Bioinform.* 21 (5), 1628–1640. doi:10.1093/bib/bbz106

Su, R., Liu, X., and Wei, L. (2020). MinE-RFE: Determine the Optimal Subset from RFE by Minimizing the Subset-Accuracy-Defined Energy. *Brief. Bioinform.* 21 (2), 687–698. doi:10.1093/bib/bbz021

Tang, J., Fu, J., Wang, Y., Luo, Y., Yang, Q., Li, B., et al. (2019). Simultaneous Improvement in the Precision, Accuracy, and Robustness of Label-free Proteome Quantification by Optimizing Data Manipulation Chains*. *Mol. Cell Proteomics* 18 (8), 1683–1699. doi:10.1074/mcp.ra118.001169

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

Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., Yang, Q., et al. (2020). ANPELA: Analysis and Performance Assessment of the Label-free Quantification Workflow for Metaproteomic Studies. *Brief Bioinform.* 21 (2), 621–636. doi:10.1093/bib/bby127

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

Ungar, D., and Hughson, F. M. (2003). SNARE Protein Structure and Function. *Annu. Rev. Cel Dev. Biol.* 19 (1), 493–517. doi:10.1146/annurev.cellbio.19.110701.155609

Wang, Z., Liu, D., Xu, B., Tian, R., and Zuo, Y. (2020). Modular Arrangements of Sequence Motifs Determine the Functional Diversity of KDM Proteins. *Brief Bioinform.* 22 (3), bbaa215. doi:10.1093/bib/bbaa215

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, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., et al. (2020). Therapeutic Target Database 2020: Enriched Resource for Facilitating Research and Early Development of Targeted Therapeutics. *Nucleic Acids Res.* 48 (D1), D1031–D1041. doi:10.1093/nar/gkz981

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 (5), bbaa409. doi:10.1093/bib/bbaa409

Wang, X., Yang, Y., Liu, J., and Wang, G. (2021). The Stacking Strategy-Based Hybrid Framework for Identifying Non-coding RNAs. *Brief Bioinform.* 22 (5), bbab023. doi:10.1093/bib/bbab023

Wang, D., Zhang, Z., Jiang, Y., Mao, Z., Wang, D., Lin, H., et al. (2021). DM3Loc: Multi-Label mRNA Subcellular Localization Prediction and Analysis Based on Multi-Head Self-Attention Mechanism. *Nucleic Acids Res.* 49 (8), e46. doi:10.1093/nar/gkab016

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., Hu, J., Li, F., Song, J., Su, R., Zou, Q., et al. (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

Whiteheart, S. W., Griff, I. C., Brunner, M., Clary, D. O., Mayer, T., Buhrow, S. A., et al. (1993). SNAP Family of NSF Attachment Proteins Includes a Brain-specific Isoform. *Nature* 362 (6418), 353–355. doi:10.1038/362353a0

Whiteheart, S. W., Schraw, T., and Matveeva, E. A. (2001). N-ethylmaleimide Sensitive Factor (NSF) Structure and Function. *Int. Rev. Cytol.* 207, 71–112. doi:10.1016/s0074-7696(01)07003-6

Xu, B., Liu, D., Wang, Z., Tian, R., and Zuo, Y. (2021). Multi-substrate Selectivity Based on Key Loops and Non-homologous Domains: New Insight into ALKBH Family. *Cell. Mol. Life Sci.* 78 (1), 129–141. doi:10.1007/s00018-020-03594-9

Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., et al. (2018). What Contributes to Serotonin-Norepinephrine Reuptake Inhibitors' Dual-Targeting Mechanism? the Key Role of Transmembrane Domain 6 in Human Serotonin and Norepinephrine Transporters Revealed by Molecular Dynamics Simulation. *ACS Chem. Neurosci.* 9 (5), 1128–1140. doi:10.1021/acschemneuro.7b00490

Yan, K., and Zhang, D. (2015). Feature Selection and Analysis on Correlated Gas Sensor Data with Recursive Feature Elimination. *Sens. Actuators B: Chem.* 212, 353–363. doi:10.1016/j.snb.2015.02.025

Yang, Q., Li, B., Tang, J., Cui, X., Wang, Y., Li, X., et al. (2020). Consistent Gene Signature of Schizophrenia Identified by a Novel Feature Selection Strategy from Comprehensive Sets of Transcriptomic Data. *Brief Bioinform.* 21 (3), 1058–1068. doi:10.1093/bib/bbz049

Yang, Q., Wang, Y., Zhang, Y., Li, F., Xia, W., Zhou, Y., et al. (2020). NOREVA: Enhanced Normalization and Evaluation of Time-Course and Multi-Class Metabolomic Data. *Nucleic Acids Res.* 48 (W1), W436–W448. doi:10.1093/nar/gkaa258

Yang, C., Ding, Y., Meng, Q., Tang, J., and Guo, F. (2021a). Granular Multiple Kernel Learning for Identifying RNA-Binding Protein Residues via Integrating Sequence and Structure Information. *Neural Comput. Appl.* 33, 11387–11399. 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

Yin, J., Sun, W., Li, F., Hong, J., Li, X., Zhou, Y., et al. (2020). VARIDT 1.0: Variability of Drug Transporter Database. *Nucleic Acids Res.* 48 (D1), D1042–D1050. doi:10.1093/nar/gkz779

Yin, J., Li, F., Zhou, Y., Mou, M., Lu, Y., Chen, K., et al. (2021). INTEDE: Interactome of Drug-Metabolizing Enzymes. *Nucleic Acids Res.* 49 (D1), D1233–D1243. doi:10.1093/nar/gkaa755

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, 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

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, 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.* PP, 1. doi:10.1109/TCBB.2020.2968419

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

Zhao, J., Sun, T., Wu, S., and Liu, Y. (2019). High Mobility Group Box 1: An Immune-Regulatory Protein. *Cgt* 19 (2), 100–109. doi:10.2174/1566523219666190621111604

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36 (16), 4466–4472. 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 Bioinform.* 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., Zuo, Y., et al. (2020). RaacLogo: a New Sequence Logo Generator by Using Reduced Amino Acid Clusters. *Brief Bioinform.* 22 (3), 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

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.

Copyright © 2021 Zhang, Gong, Gao, Li, Gao, Zhao and Dong. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

METHODS
published: 03 January 2022
doi: 10.3389/fgene.2021.808856

VTP-Identifier: Vesicular Transport Proteins Identification Based on PSSM Profiles and XGBoost

Yue Gong[1], Benzhi Dong[1], Zixiao Zhang[1], Yixiao Zhai[1], Bo Gao[2], Tianjiao Zhang[1] and Jingyu Zhang[3]**

[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2]Department of Radiology, The Second Affiliated Hospital, Harbin Medical University, Harbin, China, [3]Department of Neurology, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China

Vesicular transport proteins are related to many human diseases, and they threaten human health when they undergo pathological changes. Protein function prediction has been one of the most in-depth topics in bioinformatics. In this work, we developed a useful tool to identify vesicular transport proteins. Our strategy is to extract transition probability composition, autocovariance transformation and other information from the position-specific scoring matrix as feature vectors. EditedNearesNeighbours (ENN) is used to address the imbalance of the data set, and the Max-Relevance-Max-Distance (MRMD) algorithm is adopted to reduce the dimension of the feature vector. We used 5-fold cross-validation and independent test sets to evaluate our model. On the test set, VTP-Identifier presented a higher performance compared with GRU. The accuracy, Matthew's correlation coefficient (MCC) and area under the ROC curve (AUC) were 83.6%, 0.531 and 0.873, respectively.

Keywords: protein function prediction, vesicular transport proteins, machine learning, XGBoost, position-specific scoring matrix

OPEN ACCESS

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

Reviewed by:
Lei Deng,
Central South University, China
Liang Yu,
Xidian University, China

****Correspondence:***
Tianjiao Zhangmailto tianjiaozhang@nefu.edu.cn
Jingyu Zhangmailto zhangjingyuyx@163.com

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

Received: *04 November 2021*
Accepted: *29 November 2021*
Published: *03 January 2022*

Citation:
Gong Y, Dong B, Zhang Z, Zhai Y, Gao B, Zhang T and Zhang J (2022) VTP-Identifier: Vesicular Transport Proteins Identification Based on PSSM Profiles and XGBoost. Front. Genet. 12:808856. doi: 10.3389/fgene.2021.808856

1 INTRODUCTION

Researchers have paid more attention to vesicular transport proteins in recent years. Vesicular transport is that macromolecular substances or granular substances cannot pass through the cell membrane, but transport across the cell membrane in another special way, that is, substances are wrapped by the membrane, formed vesicular, fused with the membrane or broken in the process of transport in and out of the cell. Vesicular transport proteins are contained in the cell membrane, which can promote the activity of dominant molecules on the vesicle membrane. When macromolecules and particles cannot cross the cell membrane, vesicular transport proteins take on the task of transporting them. To date, many studies have confirmed that abnormal vesicular transport proteins may cause a variety of human diseases (Zhang et al., 2019; Zeng et al., 2020a), such as Hermansky-Pudlaksyndrome and chylomron retention disease (Cláudio et al., 2001; Suzuki et al., 2006). As the relationship between vesicular transport proteins and related diseases is gradually becoming clear, it is particularly important to deepen the study of vesicular transport proteins.

In view of the importance of vesicular transport proteins in eukaryotic cells, researchers in the area of cell biology have been committed to developing experimental techniques that can identify vesicular transport proteins and have achieved excellent results, such as morpholino knockdown (Hager et al., 2010) and dissection (Orci et al., 1989). These techniques can accurately identify vesicular transport proteins, but these technologies are often not very efficient and are expensive, so it

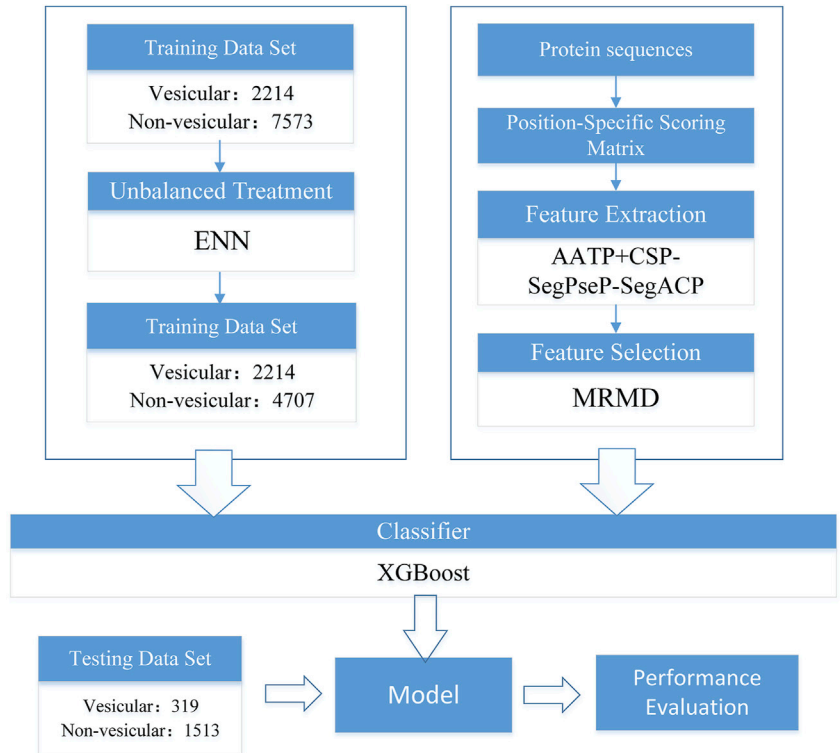


FIGURE 1 | Training flow chart of the prediction model of vesicular transport proteins.

is particularly necessary to find a time-saving and high accuracy method to identify vesicular transport proteins.

In recent years, protein function prediction has been a hot topic in the field of computational biology (Ding et al., 2020a; Fu et al., 2020; Guo et al., 2020; Tao et al., 2020; Wang et al., 2020; Zhai et al., 2020; Cai et al., 2021; Li et al., 2021; Yang, 2021). With the continuous enrichment of protein data, the technology of applying machine learning and data mining to protein function prediction is gradually maturing (Liu et al., 2019; Ding et al., 2020b; Ding et al., 2020c; Liu et al., 2020; Zhao et al., 2021). For example, some researchers used machine learning technology and created high accuracy models by sequence analysis (Chou, 2009; Cui et al., 2019; Jin et al., 2021; Shao et al., 2021), position-specific scoring matrix (PSSM) (Jones, 1999), and to determine various physicochemical and biochemical properties of amino acids (Kawashima and Kanehisa, 2000; Zhang et al., 2021; Zulfiqar et al., 2021). The above studies have shown that the use of computer technology in protein identification is reliable. Deep learning has attracted much attention, and researchers have been trying to create new deep neural networks to solve protein-related problems, such as the prediction of DNA-binding proteins (Qu et al., 2017), human protein subcellular localization (Wei et al., 2018a) and SNARE-CNN (Le and Nguyen, 2019). An increasing number of models and algorithms that can accurately identify proteins have been developed. Therefore, we adopted a machine learning method to obtain a model that can identify vesicular transport proteins.

In the previous study of Nguyen Quoc Khanh Le (Le et al., 2019), the strategy that includes gated recurrent units and PSSM was adopted, and the accuracy and Matthew's correlation coefficient (MCC) of the final model reached 82.3% and 0.52 in the cross-validation set and 85.8% and 0.44 in the independent test data set, which is an excellent result. Deep learning can often achieve high accuracy, but this method will be time-consuming due to training and has a high requirement for computer equipment. Taking PSSM as input to the model for training will also increase the training time, so we hope to find a more efficient and more accurate method to identify vesicular transport proteins.

The method used in this paper extracts information such as transition probability composition, autocovariance transformation and other information from PSSM as a feature vector. We adopted undersampling, oversampling and combined sampling methods to reduce the imbalance of the data set. The Max-Relevance-Max-Distance algorithm (Zou et al., 2016) was used to sort features and reduce the number of features. In this work, we selected XGBoost as the classifier and evaluated our model with 5-fold cross-validation. Finally, we obtained a better model than a previous study, which had high efficiency and accurate identification of vesicular transport proteins.

2 MATERIALS AND METHODS

The flowchart of our work is shown in **Figure 1**, and each section in the figure is described in detail in the following sections.

TABLE 1 | Statistics of the dataset in this work.

	Total	Train	Test
Vesicular	2,533	2,214	319
Non-vesicular	9,086	7,573	1,513

2.1 Benchmark Dataset

In this work, we used the dataset provided in Nguyen Quoc Khanh Le's study (Le et al., 2019) as the benchmark dataset. The numbers of vesicular transport proteins and non-vesicular transport proteins were 2,533 and 9,086, respectively, and we took vesicular transport proteins as positive samples and non-vesicular transport proteins as negative samples. We divided the data set into a training set and a testing set, and the details are shown in **Table 1**.

2.2 Unbalanced Datasets Treatment

We used seven methods from an unbalanced-learning library (Lemaître et al., 2017) to address the imbalance in the dataset. The methods used for undersampling were RandomUnder, ClusterCentroids, NearMiss and EditedNearesNeighbours (ENN). The method used for over-sampling was SMOTE, a total of 5,300 positive sample data have been generated. We used default parameters for these methods. For the cleaning undersampling techniques, ENN adjusted the ratio of positive and negative samples to 1:2. The other four methods changed the number of positive and negative samples to equal. The methods used from the combined methods were SMOTEENN and SMOTETomek. We adjusted the parameters of these two methods and adjusted the proportion of positive and negative samples to 1:1.

As the SMOTE, SMOTEENN and SMOTETomek methods will generate new samples, the results of 5-fold cross-validation processed by these methods are not accurate, so special cross-validation should be performed when using these three methods. K-fold cross-validation divides the training dataset into k subdatasets; k-1 subdatasets are used to train the model, and the rest are used for validation. Our method uses SMOTE and other unbalanced data processing methods to train the k-1 subdataset and then uses the validation set to evaluate the model.

2.3 Feature Extraction

To date, a strategy that includes deep learning and PSSM profiles has been frequently adopted to realize the identification of unknown proteins and has achieved excellent results. However, the strategy is slightly inefficient, so in this work, we used other machine learning models and adopted RPSSM (Ding et al., 2014), CSP-SegPseP-SegACP (Liang et al., 2015), AATP (Zhang et al., 2012), DWT (Wang et al., 2017; Wang, 2019) and SOMA (Liang and Zhang, 2017) to extract features from the PSSM matrix and make a comparison. Among them, AATP and CSP-SegPseP-SegACP have the highest MCC and AUC, so they are selected as feature extraction methods.

2.3.1 Position-Specific Scoring Matrix

PSSM can reveal the evolutionary information of proteins (Jones, 1999). PSSM was mainly used to predict protein secondary structure, now it has been widely used in the field of bioinformatics. Previous studies have shown that it is reliable to extract PSSM from protein sequences, and the evolutionary information in PSSM has more research value than the sequence itself (Kim and Park, 2004).

According to the definition of PSSM, we described PSSM by the following formula:

$$P_{PSSM} = \begin{pmatrix} P_{1,1} & P_{1,2} & L & P_{1,j} & L & P_{1,20} \\ P_{i,1} & P_{i,2} & L & P_{i,j} & L & P_{i,20} \\ P_{L,1} & P_{L,2} & L & P_{L,j} & L & P_{L,20} \end{pmatrix}$$

where $P_{i,j}$ represents the score of the ith amino acid residue of the protein sequence that mutates into amino acid type j during evolution and L shows the length of the sequence. In this work, we used PSI-BLAST to compare the sequence with NCBI's nonredundant (NR) database to obtain PSSM. Now, many methods of extracting features from PSSM have been derived. The methods used in this paper are introduced in the following chapters.

2.3.2 AATP

AATP can be extracted from PSSM, which consists of two feature vectors: amino acid composition (AAC) and transition probability composition (TPC). AAC can be described by the following:

$$AAC = \left(x_1, x_1, \ldots, x_{20}\right)^T$$

$$x_j = \left(\frac{1}{L}\right)\sum_{i=1}^{L} P_{i,j},\ j = 1, 2, 3, \ldots, 20$$

where x_j represents the probability that the amino acid residues change into J-type amino acids during evolution.

TPC is a feature vector of 400 dimensions that is extracted from the transition probability matrix (TPM) by:

$$TPC = \left(X_{1,1}, \ldots, X_{1,20}, \ldots, X_{i,1}, \ldots X_{i,20}, \ldots, X_{20,1}, \ldots, X_{20,20}\right)$$

where

$$X = \left(\sum_{k=1}^{L-1} P_{k,i} \times P_{k+1,j}\right) / \left(\sum_{j=1}^{20}\sum_{k=1}^{L-1} P_{k+1,j} \times P_{k,i}\right),\ 1 \le i, j \le 20$$

The new feature vector AATP can be obtained by integrating AAC and TPC, and each protein sequence can extract 20 + 400 = 420 features.

2.3.3 CSP-SegPseP-SegACP

CSP-SegPseP-SegACP consists of the following three parts: Pseudo-position-specific scoring matrix (PsePSSM), Autocovariance Transformation and Consensus Sequence Based on PSSM.

2.3.3.1 PsePSSM

In this step, PSSM is processed twice. For the first time, PSSM was divided into two equal length segments L_1, L_2 by using a similar procedure in (Yang and Chen, 2011). Then, two segments were used to calculate segments. The equations are as follows:

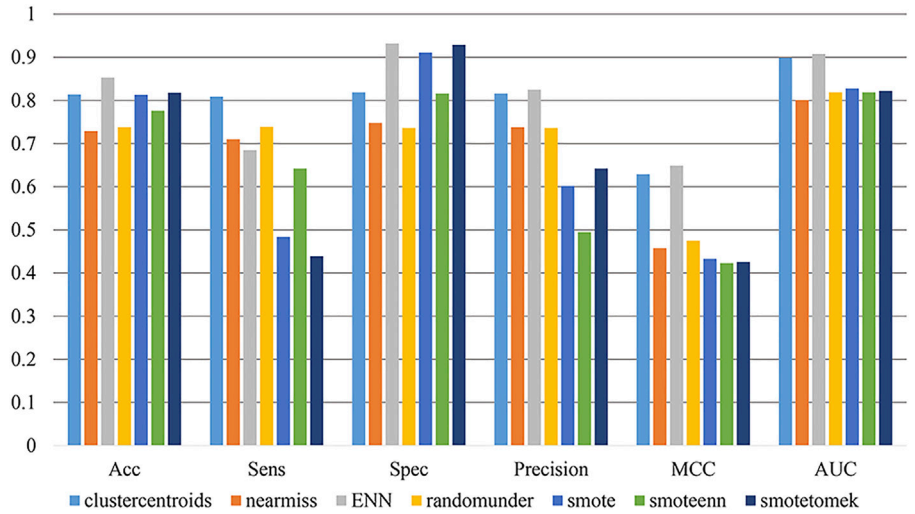


FIGURE 2 | The values of the different unbalanced data processing methods on the training set.

$$\alpha_j^\lambda = \begin{cases} \frac{1}{L_1}\sum_{i=1}^{L_1} P_{i,j}, j = 1,2,\ldots,20, \lambda = 0, \\ \frac{1}{L_1-\lambda}\sum_{i=1}^{L_1-\lambda}\left(P_{i,j}-P_{i+\lambda,j}\right)^2, j = 1,2,\ldots,20, \lambda = 1,2,3,4, \end{cases}$$

$$\beta_j^\lambda = \begin{cases} \frac{1}{L-L_1}\sum_{i=L_1+1}^{L} P_{i,j}, j = 1,2,\ldots,20, \lambda = 0, \\ \frac{1}{L-L_1-\lambda}\sum_{i=L_1+1}^{L-\lambda}\left(P_{i,j}-P_{i+\lambda,j}\right)^2, j = 1,2,\ldots,20, \lambda = 1,2,3,4, \end{cases}$$

where α_j^λ and β_j^λ represent the correlation between amino acids and λ is the contiguous distance of α_j^λ and β_j^λ along the protein sequence of each fragment. The value range of λ is affected by the number of PSSM segments and the length of the shortest series, so λ can be taken to be 0, 1, 2, 3 and 4. Through the above calculation, we can obtain a 200-dimensional feature vector.

Next, the PSSM is divided into three segments L_1, L_2 and L_3; here, λ can be token to 0, 1 and 2. The equations are as follows:

$$\theta_j^\lambda = \begin{cases} \frac{1}{L_1}\sum_{i=1}^{L_1} P_{i,j}, j = 1,2,\ldots,20, \lambda = 0, \\ \frac{1}{L_1-\lambda}\sum_{i=1}^{L_1-\lambda}\sum_{i=1}^{L_1-\lambda}\left(P_{i,j}-P_{i+\lambda,j}\right)^2, j = 1,2,\ldots,20, \lambda = 1,2 \end{cases}$$

$$\mu_j^\lambda = \begin{cases} \frac{1}{L_1}\sum_{i=L_1+1}^{2L_1} P_{i,j}, j = 1,2,\ldots,20, \lambda = 0, \\ \frac{1}{L_1-\lambda}\sum_{i=L_1+1}^{2L_1-\lambda}\left(P_{i,j}-P_{i+\lambda,j}\right)^2, j = 1,2,\ldots,20, \lambda = 1,2 \end{cases}$$

$$\nu_j^\lambda = \begin{cases} \frac{1}{L-2L_1}\sum_{i=L_1+1}^{L_1} P_{i,j}, j = 1,2,\ldots,20, \lambda = 0, \\ \frac{1}{L-2L_1-\lambda}\sum_{i=2L_1+1}^{L-\lambda}\left(P_{i,j}-P_{i+\lambda,j}\right)^2, j = 1,2,\ldots,20, \lambda = 1,2 \end{cases}$$

This time, 180-dimensional feature vectors are obtained. Combined with the results of the previous stage, a 380-dimensional feature vector can be extracted from PSSM.

2.3.3.2 Autocovariance Transformation

In this step, the information contained in the sequence is further extracted by calculating the autocovariance transformation. Similar to the previous step, the PSSM is divided into two segments and three segments, and then the ACT-PSSM feature vector is obtained by the following equations when divided into two segments:

$$AC1_j^{lg} = \frac{1}{L_1-l_g}\sum_{i=1}^{L_1-l_g}\left(P_{i,j}-\alpha_j^0\right)\left(P_{i+l_g,j}-\alpha_j^0\right), j = 1,2,\ldots,20, l_g = 1,2,3,4$$

$$AC2_j^{lg} = \frac{1}{L-L_1-l_g}\sum_{i=L_1+1}^{L-l_g}\left(P_{i,j}-\beta_j^0\right)\left(P_{i+l_g,j}-\beta_j^0\right), j = 1,2,\ldots,20, l_g = 1,2,3,4$$

$$AC1_j^{lg} = \frac{1}{L_1-l_g}\sum_{i=1}^{L_1-l_g}\left(P_{i,j}-\theta_j^0\right)\left(P_{i+l_g,j}-\theta_j^0\right), j = 1,2,\ldots,20, l_g = 1,2$$

$$AC2_j^{lg} = \frac{1}{L_1-l_g}\sum_{i=L_1+1}^{2L_1-l_g}\left(P_{i,j}-\mu_j^0\right)\left(P_{i+l_g,j}-\mu_j^0\right), j = 1,2,\ldots,20, l_g = 1,2$$

$$AC3_j^{lg} = \frac{1}{L - 2L_1 - l_g} \sum_{i=2L_1+1}^{L-l_g} \left(P_{i,j} - \nu_j^0\right)\left(P_{i+l_g,j} - \nu_j^0\right), j = 1, 2, \ldots, 20, l_g = 1, 2$$

where l_g represents the differences between amino acid residues. Finally, the 280-dimensional ACT-PSSM feature vector can be obtained by the above equations.

2.3.3.3 Consensus Sequence Based on PSSM

This step adopts the method in (Patthy, 1987) and generates a consensus sequence as follows:

$$X(i) = arg\,max\left\{P_{i,j}: 1 \le j \le 20\right\}, 1 \le i \le L$$

Next, we compute CSAAC, which shows 20 amino acid composition features of the consensus sequence, and CSCM, which represents 20 composition moment features for CS. Through the combination of the above two feature vectors, we obtain a 40-dimensional feature vector based on CS.

The 700-dimensional CSP-SegPseP-SegACP feature vector is obtained by fusing the features obtained from the above three steps.

2.4 Feature Selection

In this section, we adopted Max-Relevance-Max-Distance algorithm (MRMD) (Zou et al., 2016) to reduce the dimension of the feature vector, MRMD uses the Pearson correlation coefficient to balance the correlation between the subfeature set and the target class and uses various distance functions to obtain the redundancy of each subfeature set. The subfeature set selected by MRMD has low redundancy and strong correlation with the target class.

TABLE 2 | Evaluation of model performance after processing unbalanced data by ENN.

	Acc	Sens	Spec	Precision	MCC	AUC
ENN	0.85	0.701	0.919	0.811	0.659	0.908

2.5 Classification

We compared the performance of four different popular classification methods which are the RF, SVM, KNN and XGBoost to identified VTP. Due to six performance evaluations on the training set, we chose XGBoost as our classification method.

XGBoost (Chen and Guestrin, 2016) is a machine learning method with an excellent classification effect and high efficiency that has been widely used in recent years(Long et al., 2021; Yang et al., 2021). It stands out from many of the challenges of machine learning and data mining. In this paper, XGBoost performed very well, and it still obtained good results under the premise of high training efficiency.

3 RESULTS

3.1 Assessment of Predictive Ability

In this work, our goal was to obtain a model to predict whether the unknown type of protein sequence belongs to vesicular transport proteins, so we took vesicular transport proteins in the data set as positive samples and non-vehicular transport proteins as negative samples. In each section of our work, to evaluate our model, we used 5-fold cross-validation several times and calculated the average value as the final result. After obtaining the results of

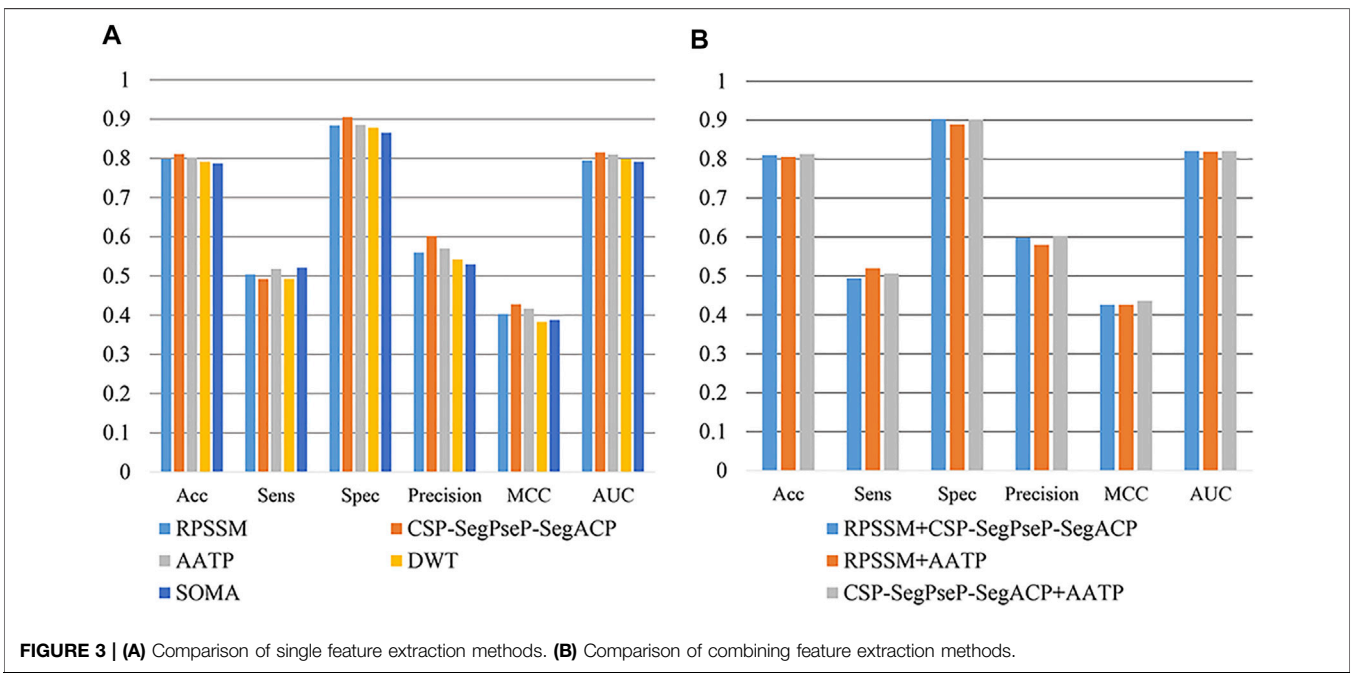


FIGURE 3 | (A) Comparison of single feature extraction methods. **(B)** Comparison of combining feature extraction methods.

TABLE 3 | The results of using different sorting methods in MRMD on the training set.

	Dimension	Acc	Sens	Spec	Precision	MCC	AUC
Hits_a	681	0.852	0.711	0.919	0.805	0.653	0.907
TrustRank	992	0.857	0.709	0.927	0.818	0.658	0.907
PageRank	898	0.855	0.712	0.922	0.81	0.658	0.907
LeadeRank	738	0.854	0.712	0.921	0.809	0.656	0.908
Hits_h	791	0.855	0.713	0.921	0.813	0.66	0.908

TABLE 4 | Comparison of six performance evaluations on the training set.

	Acc	Sens	Spec	Precision	MCC	AUC
RF	0.823	0.582	0.936	0.81	0.574	0.886
SVM	0.843	0.72	0.9	0.773	0.633	0.896
KNN	0.822	0.732	0.865	0.72	0.595	0.879
XGBoost	0.855	0.713	0.921	0.813	0.66	0.908

cross-validation, we used a test data set to test our model and make adjustments.

To evaluate our model comprehensively, we used several methods, including accuracy (ACC), sensitivity (Sens), specificity (Spec), precision, Matthew's correlation coefficient (MCC) and area under the ROC curve (AUC) (Jiang et al., 2013; Wei et al., 2014; Wei et al., 2017; Wei et al., 2018b; Su et al., 2019; Zeng et al., 2020b; Hong et al., 2020; Su et al., 2020; Tang et al., 2020; Dao, 2021; Shao and Liu, 2021; Wang, 2021). These methods are defined as follows:

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

$$Sens = \frac{TP}{TP + FN}$$

$$Spec = \frac{TN}{FP + TN}$$

$$\text{Pr}ecision = \frac{TP}{TP + FP}$$

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

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

where TP, FP, TN and FN represent true positives, false positives, true negatives, and false negatives, respectively.

3.2 Comparison of the Different Unbalanced Data Processing Methods

In the previous section, we selected CSP-SegPseP-SegACP and AATP as the feature extraction methods. Next, we compared the effects of different unbalanced processing methods on the model.

When the samples are in an unbalanced state, the model trained by machine learning tends to be more inclined to a large number of samples. (Fdez-Glez et al., 2018).

We used seven methods from an unbalanced-learning library to address the imbalance in the data set. The methods are RandomUnder, ClusterCentroids, NearMiss, EditedNearesNeighbours (ENN), SMOTE, SMOTEENN and SMOTETomek. The RandomUnder, ClusterCentroids and NearMiss adjusted the number of positive and negative samples to 2214:2214. The ENN adjusted the number of positive and negative samples to 2214:4707. The SMOTE adjusted the number of positive and negative samples to 7573: 7573. The SMOTEENN and SMOTETomek adjusted the number of positive and negative samples to 5000:5000.

In this part of our work, we set the XGBoost parameter scale_pos_weight = default to avoid XGBoost training being more biased towards positive samples. Other parameters of XGBoost are set as follows: learning_rate = 0.1, n_estimators = 1,000, max_depth = 8, min_child_weight = 1, gamma = 0, subsample = 0.8, colsample_bytree = 0.8, objective = "binary: logistic", nthread = 20. We found that the ENN method is the best, and its ACC, MCC, AUC and so on are significantly higher than those of the other methods. Therefore, ENN was selected as the final unbalanced data processing method. The result on the training set after using different imbalance processing algorithms is shown in **Figure 2**.

When dealing with unbalanced data, we cannot precisely control the proportion of positive and negative samples when using ENN. The dataset was still slightly unbalanced, so we continued to adjust the parameter scale_pos_weight of XGBoost, which makes the classifier tend to have small samples in the training process. Finally, we set scale_pos_weight = 0.6. The performance of the model is shown in **Table 2**.

3.3 Comparison of the Different Feature Extraction Methods

In previous studies, the training model using PSSM as input can effectively predict vesicular transport proteins, which indicates that PSSM has important information to identify vesicular transport proteins. In this paper, the methods of extracting features from PSSM were used to further extract the key information in PSSM and to improve the efficiency of the training model.

In this section, RPSSM, AATP, CSP-SegPseP-SegACP, SOMA, and DWT were used to extract features from PSSM. In addition to different feature extraction methods, other experimental conditions are completely consistent. We adopted XGBoost as the classifier, set the scale_pos_weight = 0.1 for the temporary method for dealing with unbalanced data sets, and used cross validation to evaluate our model. The result is shown in **Figure 3A**. By comparison, we found that RPSSM, CSP-SegPseP-SegACP and AATP performed better.

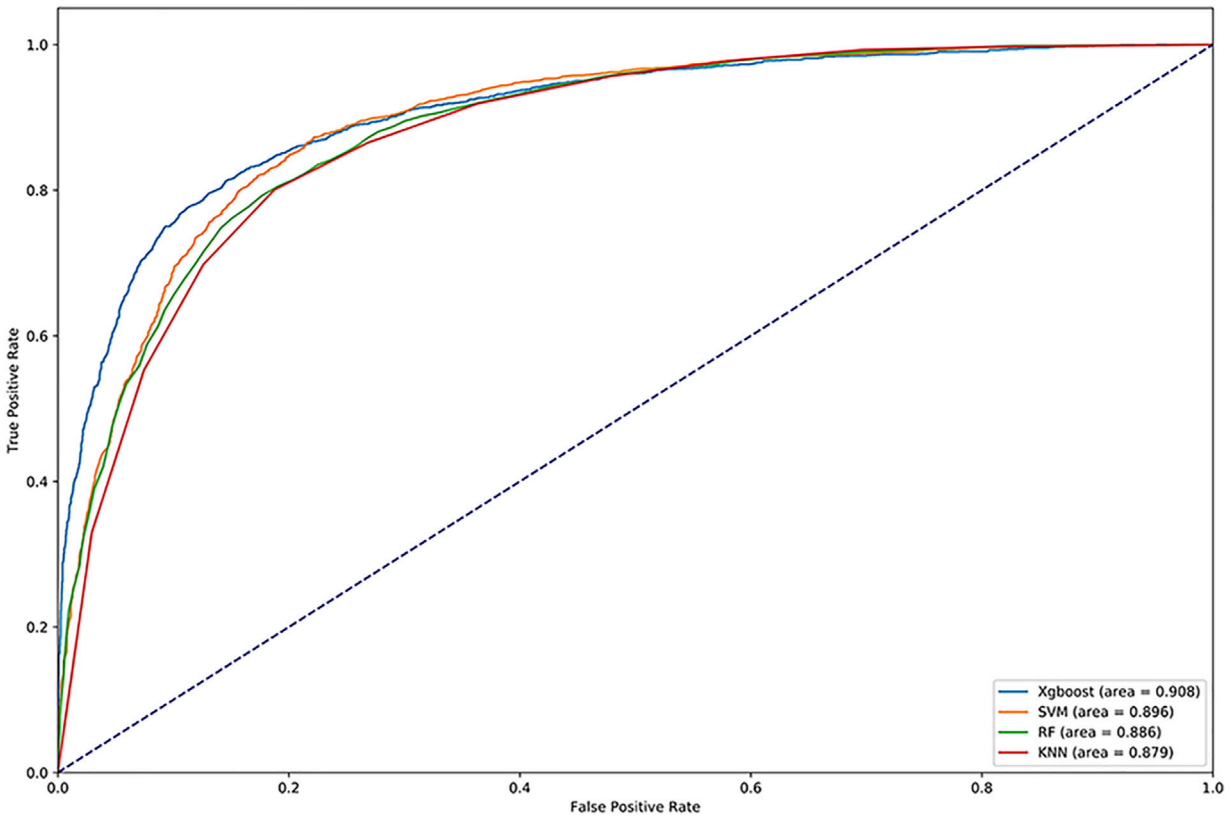


FIGURE 4 | ROC curve of vesicle transporters identified by different methods.

TABLE 5 | Performance comparison between our model and GRU.

	Acc	Sens	Spec	Precision	MCC	AUC
GRU	0.809	0.708	0.829	0.515	0.459	0.850
VTP-Identifier	0.836	0.757	0.852	0.517	0.531	0.873

Next, we combined these three methods in pairs for comparison. We found that the combination of CSP-SegPseP-SegACP and AATP was the best method, through these two methods, we extracted 1,120 dimension feature vectors. The result after using the combination methods on the training set is shown in **Figure 3B**.

3.4 Feature Selection

After dealing with the imbalance of the data set, our model has made significant progress. In this section, we reduced the dimension of the feature vector by feature selection.

In the process of machine learning, the high dimensionality of the input feature vector will have a huge impact on the model, which will make the model too complex and reduce the generalization. Therefore, when the dimension of the feature vector is high, dimensionality reduction can improve the learning ability of the machine learning model and reduce the time required to train the model.

In this work, we adopted Max-Relevance-Max-Distance algorithm (MRMD). By using AATP and CSP-SegPseP-SegACP to extract features, and then combined the features and normalized them by Z-score standardization. The dimension of the feature vector is 1,120. In this work, we used the latest version of MRMD to improve our model. MRMD has five feature ranking methods: Hits_a, Hits_h, TrustRank, PageRank and LeaderRank. TrustRank and PageRank were originally used in web search system, MRMD modified them and applied them to feature selection. LeaderRank is derived from the basic PageRank algorithm. It adds a background node to make two-way connection with all nodes. Hits is similar to PageRank and is also applied to web search, the difference is that the number of web pages processed by hits is small, and it is related to queries. We used all five methods, and then we chose Hits_h by comparing the results of cross-validation. The results are shown in **Table 3**. Finally, through MRMD, we changed the 1120-dimensional feature vector to 791 dimensions, and the accuracy was also improved.

3.5 Performance on Different Methods

Through the above processing, we obtained a good performance model. In this model, the parameters of XGBoost are: learning_rate = 0.1, n_estimators = 1,000, max_depth = 8, min_child_weight = 1, gamma = 0, subsample = 0.8, colsample_bytree = 0.8, objective = "binary:logistic", nthread = 20, scale_pos_weight = 0.6. Next, we compared the effect of our cross-validation set on different methods. In this section, we applied the data

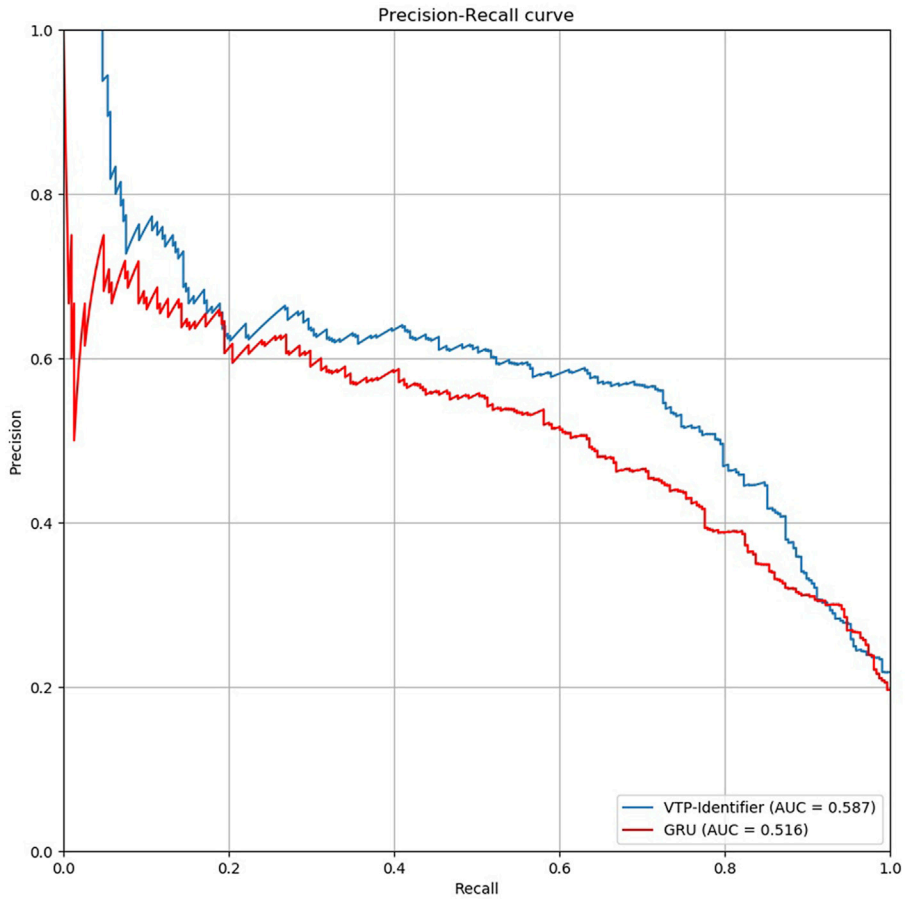


FIGURE 5 | Comparison of PR curves between our model and GRU.

processed by the same feature extraction method, imbalance processing method and feature selection method to different machine learning models.

We used Random Forest, KNN and SVM for comparison. We optimized the parameters of each classifier and set n_estimators = 100 in random forest, k = 10 in KNN, gamma = 0.5 and cost = 8 in SVM. The results are shown in **Table 4**. We drew the ROC and calculated the AUC, which are shown in **Figure 4**. Obviously, XGBoost is the best choice. Compared with other methods, XGBoost was also very efficient in the process of training the model.

Then, we used independent test sets to test the model performance. Nguyen Quoc Khanh Le used Gru neural network for deep learning in his research, we used the model provided by Nguyen Quoc Khanh Le's research result to classify vesicular transport proteins, and then compared with our model. The results are shown in **Table 5**. The PR curves of the two models are shown in **Figure 5**. Obviously, the performance of our model is better.

4 DISCUSSION

In this paper, we provide a method to identify vesicular transport proteins based on feature extraction from PSSM. In our dataset, the number of vesicular transport proteins and non-vesicular transport proteins are 2,533 and 9,086, and the number of training sets are 2,214 and 7,573. We used ENN to address the imbalance of the training data set, reduced the number of non-vesicular transport proteins from 7,573 to 4,707. We used AATP and CSP-SegPseP-SegACP to extract features from PSSM and then obtained 1,120 dimensional feature vector. Next we used MRMD to reduce the dimension of the feature vector and the dimension is reduced to 791. Finally, we sent the processed data to XGBoost and got a model to accurately identify vesicular transport proteins. The experimental comparison shows that our model is better than the previous research result. The accuracy of our model on the test set is 83.6%, which exceeds the previous research results obtained by Nguyen Quoc Khanh Le through deep learning.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/1024488648/vesicle_transporter.

AUTHOR CONTRIBUTIONS

TJZ and JYZ conceived and designed the project. YG and ZXZ conducted the experiments and analyzed the data. YG and YXZ wrote the manuscript. BZD and BG revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Fundamental Research Funds for the Central Universities (2572021BH01) and the National Natural Science Foundation of China (62172087, 62172129).

REFERENCES

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

Chen, T., and Guestrin, C. (2016). "Xgboost: A Scalable Tree Boosting System," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*.

Chou, K.-C. (2009). Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. *Cp* 6 (4), 262–274. doi:10.2174/157016409789973707

Cláudio, N., Pereira, F. J., and Barral, D. C. (2001). *Membrane Traffic and Disease*. eLS. doi:10.1002/9780470015902.a0020892

Cui, X., Yu, Z., Yu, B., Wang, M., Tian, B., and Ma, Q. (2019). UbiSitePred: A Novel Method for Improving the Accuracy of Ubiquitination Sites Prediction by Using LASSO to Select the Optimal Chou's Pseudo Components. *Chemometrics Intell. Lab. Syst.* 184, 28–43. doi:10.1016/j.chemolab.2018.11.012

Dao, F. Y. (2021). DeepYY1: a Deep Learning Approach to Identify YY1-Mediated Chromatin Loops. *Brief Bioinform* 22 (4). doi:10.1093/bib/bbaa356

Ding, S., Li, Y., Shi, Z., and Yan, S. (2014). A Protein Structural Classes Prediction Method Based on Predicted Secondary Structure and PSI-BLAST Profile. *Biochimie* 97, 60–65. doi:10.1016/j.biochi.2013.09.013

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. Applic* 32 (14), 10303–10319. doi:10.1007/s00521-019-04569-z

Fdez-Glez, J., Ruano-Ordás, D., Méndez, J. R., Fdez-Riverola, F., Laza, R., and Pavón, R. (2018). Determining the Influence of Class Imbalance for the Triage of Biomedical Documents. *Cbio* 13 (6), 592–605. doi:10.2174/1574893612666170718151238

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

Hager, H. A., Roberts, R. J., Cross, E. E., Proux-Gillardeaux, V., and Bader, D. M. (2010). Identification of a Novel Bves Function: Regulation of Vesicular Transport. *Embo J.* 29 (3), 532–545. doi:10.1038/emboj.2009.379

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

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

Jones, D. T. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices 1 1Edited by G. Von Heijne. *J. Mol. Biol.* 292 (2), 195–202. doi:10.1006/jmbi.1999.3091

Kawashima, S., and Kanehisa, M. (2000). AAindex: Amino Acid index Database. *Nucleic Acids Res.* 28 (1), 374. doi:10.1093/nar/28.1.374

Kim, H., and Park, H. (2004). Prediction of Protein Relative Solvent Accessibility with Support Vector Machines and Long-Range Interaction 3D Local Descriptor. *Proteins* 54 (3), 557–562. doi:10.1002/prot.10602

Le, N. Q. K., and Nguyen, V.-N. (2019). SNARE-CNN: a 2D Convolutional Neural Network Architecture to Identify SNARE Proteins from High-Throughput Sequencing Data. *PeerJ Comp. Sci.* 5, e177. doi:10.7717/peerj-cs.177

Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N., Chua, M. C. H., and Yeh, H.-Y. (2019). Computational Identification of Vesicular Transport Proteins from Sequences Using Deep Gated Recurrent Units Architecture. *Comput. Struct. Biotechnol. J.* 17, 1245–1254. doi:10.1016/j.csbj.2019.09.005

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Machine Learn. Res.* 18 (1), 559–563.

Li, H.-L., Pang, Y.-H., and Liu, B. (2021). BioSeq-BLM: a Platform for Analyzing DNA, RNA and Protein Sequences Based on Biological Language Models. *Nucleic Acids Res.* doi:10.1093/nar/gkab829

Liang, Y., Liu, S., and Zhang, S. (2015). Prediction of Protein Structural Classes for Low-Similarity Sequences Based on Consensus Sequence and Segmented PSSM. *Comput. Math. Methods Med.*, 2015. doi:10.1155/2015/370756

Liang, Y., and Zhang, S. (2017). Predict Protein Structural Class by Incorporating Two Different Modes of Evolutionary Information into Chou's General Pseudo Amino Acid Composition. *J. Mol. Graphics Model.* 78, 110–117. doi:10.1016/j.jmgm.2017.10.003

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., Li, C.-C., and Yan, K. (2020). DeepSVM-fold: Protein Fold Recognition by Combining Support Vector Machines and Pairwise Sequence Similarity Scores Generated by Deep Learning Networks. *Brief. Bioinform.* 21 (5), 1733–1741. doi:10.1093/bib/bbz098

Long, J., Yang, H., Yang, Z., Jia, Q., Liu, L., Kong, L., et al. (2021). Integrated Biomarker Profiling of the Metabolome Associated with Impaired Fasting Glucose and Type 2 Diabetes Mellitus in Large-Scale Chinese Patients. *Clin. Transl Med.* 11 (6), e432. doi:10.1002/ctm2.432

Orci, L., Malhotra, V., Amherdt, M., Serafini, T., and Rothman, J. E. (1989). Dissection of a Single Round of Vesicular Transport: Sequential Intermediates for Intercisternal Movement in the Golgi Stack. *Cell* 56 (3), 357–368. doi:10.1016/0092-8674(89)90239-0

Patthy, L. (1987). Detecting Homology of Distantly Related Proteins with Consensus Sequences. *J. Mol. Biol.* 198 (4), 567–577. doi:10.1016/0022-2836(87)90200-2

Qu, Y.-H., Yu, H., Gong, X.-J., Xu, J.-H., and Lee, H.-S. (2017). On the Prediction of DNA-Binding Proteins Only from Primary Sequences: A Deep Learning Approach. *PloS one* 12 (12), e0188129. doi:10.1371/journal.pone.0188129

Shao, J., and Liu, B. (2021). ProtFold-DFG: Protein Fold Recognition by Combining Directed Fusion Graph and PageRank Algorithm. *Brief Bioinform* 22 (3), 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 (3), bbaa144. doi:10.1093/bib/bbaa144

Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2020). Empirical Comparison and Analysis of Web-Based Cell-Penetrating Peptide Prediction Tools. *Brief. Bioinform.* 21 (2), 408–420. doi:10.1093/bib/bby124

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A Deep forest Model to Predict Anti-cancer Drug Response. *Methods* 166, 91–102. doi:10.1016/j.ymeth.2019.02.009

Suzuki, T., Araki, Y., Yamamoto, T., and Nakaya, T. (2006). Trafficking of Alzheimer's Disease-Related Membrane Proteins and its Participation in Disease Pathogenesis. *J. Biochem.* 139 (6), 949–955. doi:10.1093/jb/mvj121

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

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, X., Yang, Y., Liu, J., and Wang, G. (2021). The Stacking Strategy-Based Hybrid Framework for Identifying Non-Coding RNAs. *Brief Bioinform* 22 (5). doi:10.1093/bib/bbab023

Wang, Y. (2019). CrystalM: a Multi-View Fusion Approach for Protein Crystallization Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*. doi:10.1109/tcbb.2019.2912173

Wang, Y., Ding, Y., Guo, F., Wei, L., and Tang, J. (2017). Improved Detection of DNA-Binding Proteins via Compression Technology on PSSM Information. *PloS one* 12 (9), e0185587. doi:10.1371/journal.pone.0185587

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., 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., 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., 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

Yang, C. (2021). Granular Multiple Kernel Learning for Identifying RNA-Binding Protein Residues via Integrating Sequence and Structure Information. *Neural Comput. Appl.*, 1–13. 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

Yang, J.-Y., and Chen, X. (2011). Improving Taxonomy-Based Protein Fold Recognition by Using Global and Local Features. *Proteins* 79 (7), 2053–2064. doi:10.1002/prot.23025

Zeng, X., 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

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, 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, S., Ye, F., and Yuan, X. (2012). Using Principal Component Analysis and Support Vector Machine to Predict Protein Structural Class for Low-Similarity Sequences via PSSM. *J. Biomol. Struct. Dyn.* 29 (6), 1138–1146. doi:10.1080/07391102.2011.672627

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. (2019). Meta-path Methods for Prioritizing Candidate Disease miRNAs. *Ieee/acm Trans. Comput. Biol. Bioinform* 16 (1), 283–291. doi:10.1109/TCBB.2017.2776280

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

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A Novel Features Ranking Metric with Application to Scalable Visual and Bioinformatics Data Classification. *Neurocomputing* 173, 346–354. doi:10.1016/j.neucom.2014.12.123

Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Sun, Z.-J., Dao, F.-Y., Yu, X.-L., et al. (2021). Identification of Cyclin Protein Using Gradient Boost Decision Tree Algorithm. *Comput. Struct. Biotechnol. J.* 19, 4123–4131. doi:10.1016/j.csbj.2021.07.013

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.

Copyright © 2022 Gong, Dong, Zhang, Zhai, Gao, Zhang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 January 2022
doi: 10.3389/fgene.2021.803606

Screening and Identification of Survival-Associated Splicing Factors in Lung Squamous Cell Carcinoma

Min Chen[1†], *Rui Zhu*[1†], *Fangzhou Zhang*[2,3*] *and Liucun Zhu*[1*]

[1]School of Life Sciences, Shanghai University, Shanghai, China, [2]School of Materials Science and Engineering, Institute of Materials, Shanghai University, Shanghai, China, [3]Shaoxing Institute of Technology, Shanghai University, Shanghai, China

Lung squamous cell carcinoma (LUSC) is a disease with high morbidity and mortality. Many studies have shown that aberrant alternative splicing (AS) can lead to tumorigenesis, and splicing factors (SFs) serve as an important function during AS. In this research, we propose an analysis method based on synergy to screen key factors that regulate the initiation and progression of LUSC. We first screened alternative splicing events (ASEs) associated with survival in LUSC patients by bivariate Cox regression analysis. Then an association network consisting of OS-ASEs, SFs, and their targeting relationship was constructed to identify key SFs. Finally, 10 key SFs were selected in terms of degree centrality. The validation on TCGA and cross-platform GEO datasets showed that some SFs were significantly differentially expressed in cancer and paracancer tissues, and some of them were associated with prognosis, indicating that our method is valid and accurate. It is expected that our method would be applied to a wide range of research fields and provide new insights in the future.

Keywords: lung squamous cell carcinoma, alternative splicing, splicing factor, bivariate cox regression, bipartite graph

OPEN ACCESS

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

Reviewed by:
Yingying Dong,
Soochow University, China
Xuebing Li,
Tianjin Medical University General Hospital, China

***Correspondence:**
Fangzhou Zhang
zhangfzh@shu.edu.cn
Liucun Zhu
zhuliucun@shu.edu.cn

†These authors have contributed equally to this work and share first authorship

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

Received: *28 October 2021*
Accepted: *27 December 2021*
Published: *20 January 2022*

Citation:
Chen M, Zhu R, Zhang F and Zhu L (2022) Screening and Identification of Survival-Associated Splicing Factors in Lung Squamous Cell Carcinoma. Front. Genet. 12:803606. doi: 10.3389/fgene.2021.803606

INTRODUCTION

Lung cancer is one of the most common malignant tumors, and about 85% of cases are non-small cell lung cancer (NSCLC) (Wang et al., 2019). According to pathological classification, NSCLC can be divided into lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) (Cheng et al., 2019). Compared with LUAD, patients with LUSC have a poorer treatment outcome and prognosis (Li et al., 2018). In recent years, targeted therapies for specific genes have greatly improved the living conditions of patients with advanced LUAD. However, LUSC patients respond poorly to targeted therapies due to the lack of driver mutations, and the specific molecular mechanisms of LUSC pathogenesis and progression have not been systematically assessed. As a result, further exploration of the molecular mechanisms underlying the development of LUSC is essential for the development of more effective therapeutic regimens.

Alternative splicing (AS) is an important post-transcriptional regulatory mechanism. A single gene can generate more than one mRNA transcript through AS, and each mRNA transcript encodes a protein with a different structure and function (Baralle and Giudice, 2017). More than 95% of human genes experience AS under normal physiological conditions. On the one hand, the AS process regulates the tissue-specific and stage-specific expressions of specific genes during human development (Xu et al., 2002; Pan et al., 2008) and is essential for normal biological processes, such as hematopoiesis (Wong et al., 2018), brain development (Matsuda et al., 2019), and muscle

function (Nakka et al., 2018). On the other hand, abnormal AS triggers a series of tumor-related processes, including cell proliferation (Xie et al., 2019), apoptosis (Tyson-Capper and Gautrey, 2018), epithelial-mesenchymal transition (EMT) (Pradella et al., 2017), and tumor invasion and metastasis (Chen et al., 2017; Wang et al., 2017) in response to hypoxia (Han et al., 2017), thereby promoting malignant cell transformation and providing a survival advantage (Climente-González et al., 2017; Moncada et al., 2020). The AS process is regulated by splicing factors (SFs), and abnormal expression of SFs is the main contributor to overall changes in alternative splicing events (ASEs) in malignancies (David and Manley, 2010; Dvinge et al., 2016; Su et al., 2018). Therefore, exploring abnormal ASEs and SFs in malignant tumors may provide new insights into the mechanisms of tumorigenesis and progression.

Recent studies have paid more attention to assessing the clinical significance of ASEs and SFs in cancers and their potential pathogenic pathways and regulatory networks. The abnormal ASEs and SFs, which make network dysregulated, have been shown to modulate malignant transformation of cells and epithelial-mesenchymal transition (Sveen et al., 2016). Several excellent studies have also discussed the role of SFs in DNA damage (Shkreta and Chabot, 2015) or in carcinogenesis and anticancer therapies (Miura et al., 2012; Shkreta et al., 2013). However, SFs have the potential to become molecular markers and therapeutic targets for malignancies (Anczukow and Krainer, 2016; Yuan et al., 2017; Park et al., 2019). Although there is an increasing systematic analysis of AS signatures and the effect of SFs in colorectal cancer, glioblastoma, breast cancer, and ovarian cancer (Dorman et al., 2014; Suo et al., 2015; Zong et al., 2018), the analytical methods for identifying tumor-associated SFs remain deficient. Only univariate difference and survival analysis were performed in these studies (Zhu et al., 2018; Hu et al., 2019; Zhao et al., 2020). However, biological processes are complex and are mostly regulated by multiple factors rather than a single factor. It is indicated that, as a whole, some factors would have a high correlation with the tumor process, but this would show a low correlation when they are separated. Hence, we propose an analysis method based on synergy to screen key factors that regulate the initiation and progression of LUSC. We first screened the ASEs associated with overall survival (OS-ASEs) from combinations consisting of two ASEs using bivariate Cox regression and AUROC. Then an association network consisting of OS-ASEs, SFs, and their targeting relationship was constructed to identify key SFs. This method can screen a relatively complete set of OS-ASEs to a certain extent, thereby improving the completeness for subsequent screening of key SFs and providing new ideas for LUSC mechanism research.

MATERIALS AND METHODS

Data Collection and Preprocessing

Clinical information and expression levels of LUSC patients (generated by RNA-seq) were collected from The Cancer Genome Atlas (TCGA) database. Additionally, ASEs data were retrieved from the TCGASpliceSeq database (Ryan et al., 2016). In TCGASpliceSeq, the Percent Spliced In (PSI) values are computed for each possible splice event in each gene. PSI is the ratio of reads indicating the presence of a transcript element versus the total reads covering the event. The cross-platform validation set, including GSE157010, GSE3268, and GSE6044 (**Supplementary Table S1**), was downloaded from the NCBI-GEO database (Barrett et al., 2013). SFs are protein factors involved in the splicing process of pre-RNA. A total of 404 SFs were collected in this study (Wu et al., 2020), as shown in **Supplementary Table S2**.

The TCGA database included 550 LUSC samples, 501 of which were tumor samples. After removing 8 samples with no clinical information, 493 tumor samples were retained for subsequent analysis (**Supplementary Table S3**). The TCGASpliceSeq database contained a total of 46,020 ASEs for LUSC, of which 9424 ASEs were retained for subsequent analysis by removing ASE containing "null" and then excluding ASEs with variances less than 0.001 in all samples (**Supplementary Source Code S1**) (**Supplementary Source Code S2**). The distinguishable visualization UpSet plot, generated by *UpSetR* (version 1.4.0) (Wang et al., 2021), was used to quantitatively analyze the intersections among the seven types of ASEs in LUSC. The expressions of 404 SFs were extracted after being normalized by log2 (FPKM+1) (Bullard et al., 2010). SFs with expression values of 0 in half of the samples were excluded, and 398 SFs were finally retained for subsequent analysis (**Figure 1**). The GSE157010 dataset constitutes 235 LUSC tumor samples, each containing clinical information. The GSE3268 dataset represents 5 tumor samples from LUSC patients and paired normal samples. The GSE6044 dataset includes 5 normal samples and 15 tumor samples. Ten of these 15 tumor patients have not received platinum-based therapy, and the other five have. Probe IDs for each GEO dataset were converted to Ensembl ID. When multiple probes correspond to an Ensembl ID, only the probe with the highest mean is retained. The batch correction was performed to eliminate the batch effect of three datasets using normalizeBetweenArrays function of *limma* (version 3.46.0).

Methods for Screening Alternative Splicing Events Associated With Overall Survival

In order to investigate the prognostic value of ASEs in LUSC patients, all bivariate ASEs combinations were first constructed. Then Cox proportional risk hypothesis tests and bivariate Cox proportional risk regressions were performed using the survival package in R (Bradburn et al., 2003). The significance of the independent variables in the regressions was tested using likelihood ratio tests (Hazra and Gogtay, 2017). Additionally, the area under the receiver operating characteristic curve (AUROC) was used to show the sensitivity and specificity of the bivariate combination model in predicting OS (Linden, 2006). Values greater than 0.8 were considered excellent combinations. The two indicators mentioned above, the p-value of the likelihood ratio test and the AUROC, were used to screen OS-ASEs.

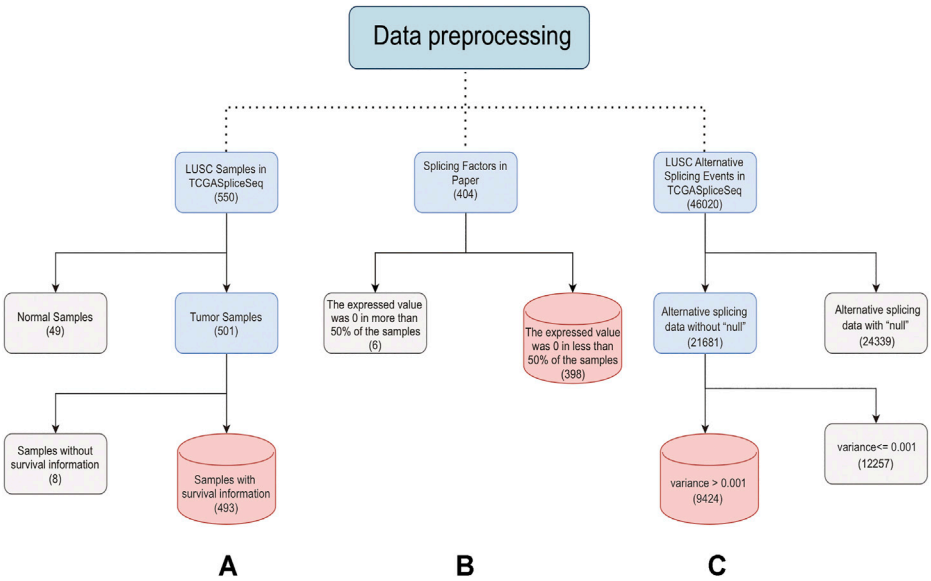


FIGURE 1 | Steps of data preprocessing. The solid line represents the preprocessing process. The light blue box represents the data to be processed. The gray box represents the rejected data. The red box represents the last retained data. The numbers in brackets represent the data amount. **(A)** is the preprocessing process of samples, **(B)** is the preprocessing process of SFs, and **(C)** is the preprocessing process of ASEs.

Methods for Association Network Construction and Analysis

Spearman correlation analysis was performed to explore the correlation between the PSI values of OS-ASEs and the expression levels of SF genes (Bishara and Hittner, 2012). The correlation networks visualization was visualized by *EVeen* (Chen T. et al., 2021) with SFs and the OS-ASEs as vertices and the Spearman significant correlation between them as edges. It is assumed that the value of a vertex in a network depends first on its position in the network. More central vertex indicates a greater impact on the structure and function of the network (Kitsak et al., 2010). The importance of a vertex in the network is usually expressed by degree centrality, which is the number of connected edges of the vertex in the network (Freeman, 1978).

Validation Methods for Alternative Splicing Events and Splicing Factors Functions

In order to identify potential mechanisms of OS-ASEs in LUSC, the survival-related genes were analyzed by Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis, which were both done by *DAVID* (Huang et al., 2009). The results of KEGG analysis were presented by bubble plots generated by *ggplot2* (version 3.3.5). The results of GO analysis were visualized by a web tool *Revigo*, which shows the cluster representatives in a two-dimensional space derived by applying multidimensional scaling to a matrix of the GO terms' semantic similarities (Supek et al., 2011).

In order to validate the function of SFs, violin plots visualized by *ggplot2* (version 3.3.5) were used for verifying the difference in the expression of SFs in tumor and normal tissues. The paired-samples *t* test was used to test the significance of the difference.

The Kaplan-Meier (KM), generated by *survival* (version 3.2-11) and *survminer* (version 0.4.9), was applied to validate the prognostic effect of SFs (**Supplementary Source Code S3**) (Dinse and Lagakos, 1982). The log-rank test was used to test the significance of differences in survival between high- and low-risk patients (Mantel, 1966). The p-value < 0.05 was considered statistically significant in this study.

RESULTS

Clinical Characteristics of the Lung Squamous Cell Carcinoma Cohort

The current study included a total of 493 LUSC patients from the TCGA database, and the characteristics and clinical information of these patients are listed in **Table 1**. There were 365 men and 128 women among these patients. With a median age of 68 (ranging from 39 to 85 years old), the mean survival time of patients was 1,044 days (ranging from 4 to 4,765 days). It is worth noting that the survival time of patients is censored data. The patient mortality rate of 43% confirms that LUSC is a tumor with a high mortality rate. The LUSC tumor staging data show that most patients are in stages I or II. Stage I tumors are usually small, without lymph nodes and distant metastases, and can be completely removed by surgery. In contrast, higher stages mean that the tumor is more progressive.

Overview of Alternative Splicing Events in the Lung Squamous Cell Carcinoma Cohort

The TCGASpliceSeq database recorded seven types of ASEs, including exon skipping (ES), mutually exclusive (ME) exons, intron retention (RI), alternative promoter (AP), alternative

TABLE 1 | Clinical characteristics of 493 LUSC patients in the TCGA database.

Characteristics	Groups	No. of patients	%
Sex	Male	365	74
	Female	128	26
Age at diagnosis	Median	68	
	Range	39–85	
	<61	107	22
	≥61	381	77
	Unknown	5	1
Vital status	Alive	268	54
	Dead	211	43
	Unknown	14	3
Stage	I	241	48.88
	II	158	32.05
	III	83	16.84
	IV	7	1.42
	Unknown	4	0.81
T category	T1	114	23
	T2	286	58
	T3	70	14
	T4	23	5
N category	N0	316	64
	N1	127	26
	N2	40	8
	N3	5	1
	NX	5	1
M category	M0	405	82
	M1	7	1
	MX	77	16
	Unknown	4	1

terminator (AT), alternative donor (AD) site, and alternative acceptor (AA) site (**Figure 2A**).

In this cohort, a total of 9424 ASEs were in 4246 genes, with 3876 ATs in 1817 genes, 2048 ESs in 1488 genes, 1761 APs in 791 genes, 656 RIs in 504 genes, 531 AAs in 468 genes, 515 ADs in 421 genes, and 37 MEs in 37 genes. Multiple ASEs can occur in a single gene (**Figure 2B**).

Screening and Analysis of Alternative Splicing Events Related to Survival

Due to the complexity of biological processes, synergistic interactions between genes are more prevalent. To accurately screen OS-ASEs, we employed the *p*-value of the likelihood ratio test in bivariate Cox proportional risk regression and AUROC as screening criteria ($P < 0.05, AUROC > 0.8$). Consequently, 1118 combinations of OS-ASEs were screened, including 953 non-redundant ASEs.

A total of 953 OS-ASEs were detected in 489 genes. More specifically, there were 689 ATs in 348 genes, 241 APs in 121 genes, 12 ESs in 12 genes, 10 RIs in 10 genes, and 1 AA in 1 gene. Two splice types, AD and ME, were not included (**Figure 3A**). Next, in order to understand the function of the genes corresponding to OS-ASEs, KEGG analysis and GO analysis were performed. KEGG analysis demonstrated that these genes were enriched in histone-lysine N-methyltransferase activity, protein tyrosine phosphatase activity, and DNA repair and apoptosis pathways. These pathways are closely associated with cancer progression (Östman et al., 2006; Wong, 2011; Jeggo et al., 2016; Husmann and Gozani, 2019). A recent study has shown that histone-lysine N-methyltransferase is a key driver for the induction of LUSC (**Figure 3B**) (Yuan et al., 2021). GO analysis revealed that these genes were enriched in both the nucleus and cytoplasm and play a role in protein binding, nucleic acid binding, and histone lysine N-methyltransferase activity. These genes are involved in important biological processes such as DNA repair, peptidyltryosine dephosphorylation, and apoptosis (**Figure 3C**).

TABLE 2 | Top 10 SFs for degree centrality.

Rank	Ensembl	Symbol
1	ENSG00000130332	LSM7
2	ENSG00000169976	SF3B5
3	ENSG00000163634	THOC7
4	ENSG00000051596	THOC3
5	ENSG00000108561	C1QBP
6	ENSG00000137168	PPIL1
7	ENSG00000079134	THOC1
8	ENSG00000139343	SNRPF
9	ENSG00000108883	EFTUD2
10	ENSG00000123154	WDR83

Construction and Analysis of the Association Network Between Splicing Factors and Alternative Splicing Events

Systems biology is the study of the composition and interrelationships of the biological systems and is widely used in the study of gene networks (Hood, 2003). For our association network, identifying key vertices is an important way to find key SFs (Zhao and Liu, 2019). The association network was formed with 489 ASEs and 398 SFs as vertices and 9414 pairs of significant correlations as edges (**Figure 4**). The degree distribution is shown in **Supplementary Figure S1**. The average degree of the top 10 vertices in this network is 69, and the average degree of the remaining vertices is 22, indicating that the top 10 SFs are associated with more ASEs and is important in this network. Therefore, we consider these 10 SFs as key SFs. (**Table 2**)

Validation of Splicing Factors

To verify the validity of the above approach, we analyzed the expression patterns of the 10 SFs in the TCGA-LUSC dataset. It is noticed that a significant difference exists in the expression of the 10 SFs between cancerous and paracancerous tissues (**Figure 5**). Moreover, patients were divided into two groups according to the expression of SFs, and the difference of survival time between them was analyzed with KM curves. It is found that 5 of these 10 SFs are significantly associated with the prognosis of LUSC patients. (**Supplementary Figure S2**).

In order to further assess the applicability of our approach, three cross-platform datasets from the GEO dataset were recruited. The GSE157010 dataset matches 9 SFs, 6 of which exhibit prognostic function (**Supplementary Figure S3**). In the

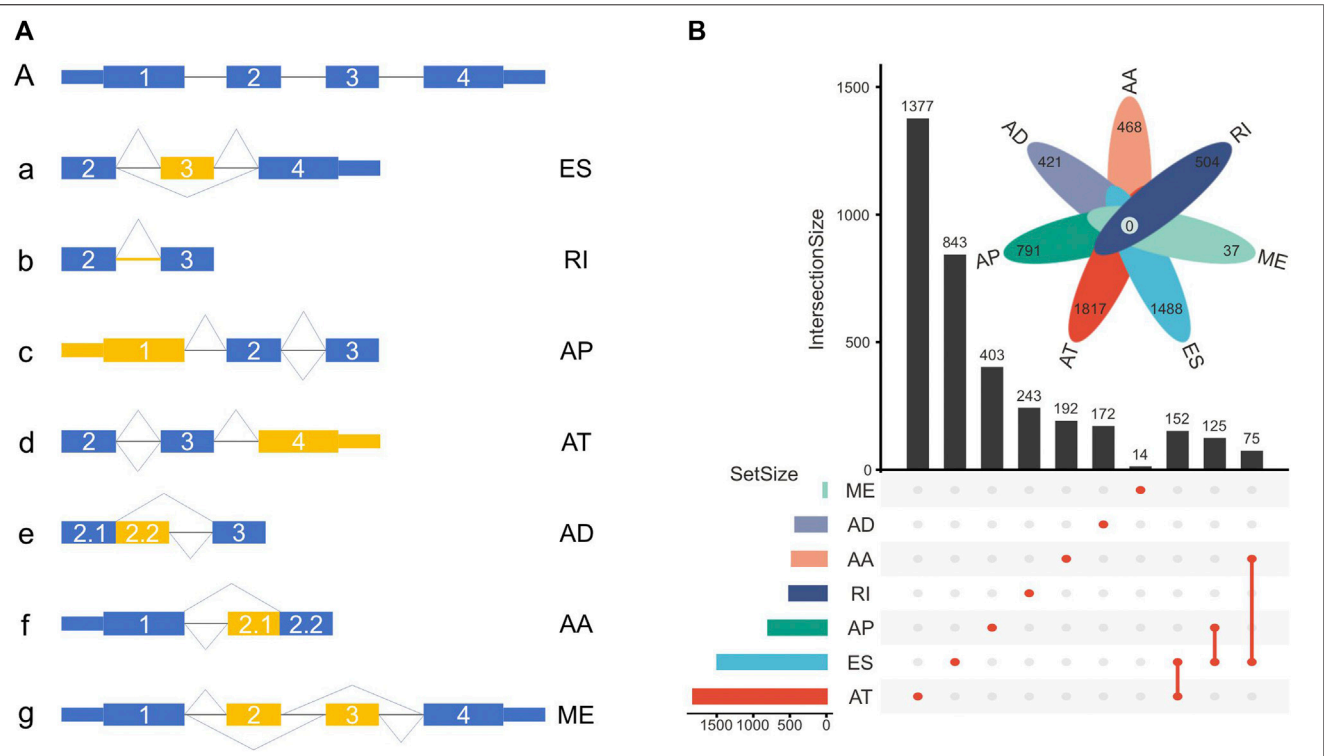


FIGURE 2 | (A) Schematic representation of ASEs, including exon skipping (ES), intron retention (RI), alternative promoter (AP), alternative terminator (AT), alternative donor (AD) site, alternative acceptor (AA) site, and mutually exclusive (ME) exon. **(B)** The number of genes with ASEs in LUSC, with 3876 ATs in 1817 genes, 2048 ESs in 1488 genes, 1761 APs in 791 genes, 656 RIs in 504 genes, 531 AAs in 468 genes, 515 ADs in 421 genes, and 37 MEs in 37 genes.

GSE3268 and GSE6044 datasets, the matched SFs are differentially expressed in normal and tumor samples (**Supplementary Figure S4**). In the GSE6044 dataset, the expression levels of SFs patients who received platinum-based therapy are slightly decreased compared with that of patients who did not, which is closer to the expression level in normal tissues (**Supplementary Figure S5**).

DISCUSSION

In this research, bivariate Cox regression and the systems biology approach were employed to detect OS-ASEs and SFs associated with LUSC. The results showed that all 10 candidates (SFs) were expressed at significantly higher levels in tumor samples than in paracancerous tissues in both the TCGA-LUSC and GEO datasets. Moreover, 7 of these SFs were associated with the overall survival time in tumor patients in one or more datasets. These results are consistent with the currently known characteristics of tumor-associated genes (Givechian et al., 2018; Qi et al., 2018). It is found that 3 of the 10 SFs are reported to be connected with lung cancer, namely LSM7, C1QBP, and THOC1. Specifically, LSM7 is a prognosis-related key gene and mediates autophagy in LUSC, with significant expression differences between tumor and normal tissues (Gatica et al., 2019; Li et al., 2020); C1QBP is involved in various cellular processes, including mRNA splicing, ribosome biosynthesis, protein synthesis in mitochondria, apoptosis, transcriptional regulation, and viral infection, and its expression correlated with the prognosis of patients with lung, breast, and colon tumors (Saha et al., 2019); THOC1 is down-regulated in lung cancer cell lines SPC-A1 and NCI-H1975, and its overexpression inhibits cell proliferation, induces G2/M cell cycle arrest, and promotes cell apoptosis (Wan et al., 2014). THOC1 also inhibits the proliferation of tumor cells in hepatocellular carcinoma and prostate cancer (Liu et al., 2015; Cai et al., 2020). The above evidence suggests that our method is reliable and accurate.

In addition, we identified 7 new SFs, 6 of which, including SF3B5, THOC7, THOC3, SNRPF, EFTUD2, and WDR83, were reported to be associated with other tumors. It has been suggested that SF3B5 is a key prognostic factor in ovarian cancer (Ouyang et al., 2021). Studies have shown a relationship between the downregulation of THOC7 and the activation of tumorigenic pathways in cervical cancer (Lando et al., 2013; Lando et al., 2015). THOC3 is involved in the THO subcomplex and is necessary for coupled mRNA transcriptional extension and nuclear export, and its expression is significantly elevated in glioma cells (Chen Z. et al., 2021). SNRPF is aberrantly expressed in human glioma. *In vitro* experiments have revealed that ubiquitin carboxy-terminal hydrolase isozyme L5 could inhibit human glioma cell migration and invasion by downregulating SNRPF (Ge et al., 2017). EFTUD2 is markedly

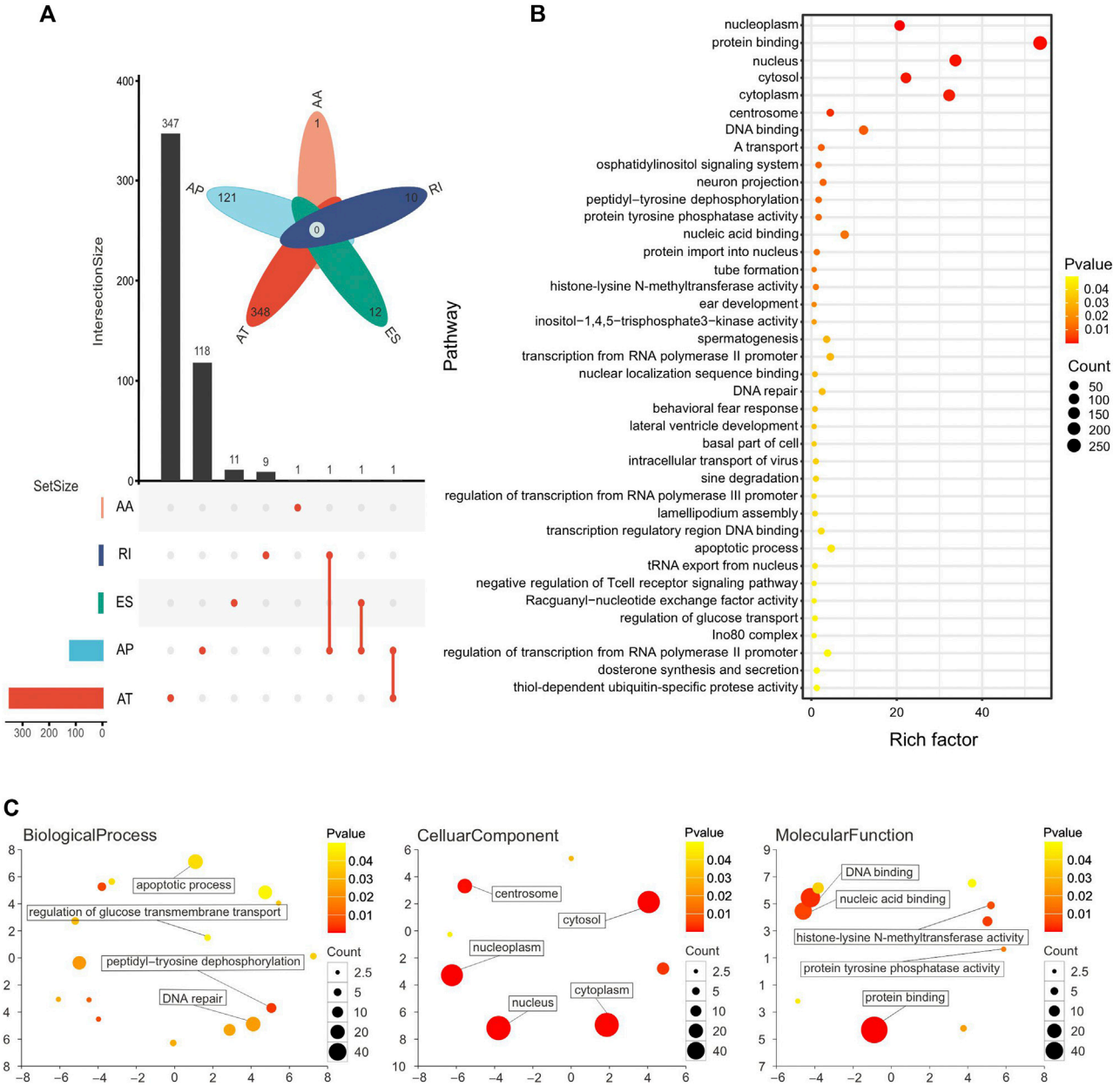


FIGURE 3 | (A) Number of genes with OS-ASEs in LUSC, with 689 ATs in 348 genes, 241 APs in 121 genes, 12 ESs in 12 genes, 10 RIs in 10 genes, and 1 AA in 1 gene. **(B)** Pathway enrichment analysis of genes with OS-ASEs. Larger dots represent more genes enriched in the pathway and vice versa. A smaller *p*-value is represented when the color of the dot is closer to blue, and a larger *p*-value is represented when the color of the dot is closer to red. **(C)** Functional enrichment analysis of genes with OS-ASEs. The scatterplot shows the cluster representatives in a two-dimensional space derived by applying multidimensional scaling to a matrix of the GO terms' semantic similarities. The dot represents all GO items, and its size is related to the number of genes enriched in that GO term. The color of dots is related to the *p*-value. A smaller *p*-value is represented when the color of the dot is closer to blue, and a larger *p*-value is represented when the color of the dot is closer to red.

overexpressed in hepatocellular carcinoma tissues. High expression of EFTUD2 in hepatocellular carcinoma patients is associated with clinical features and is pivotal in hepatocellular carcinoma cell proliferation and cell cycle course (Lv et al., 2021). As the NAT of WDR83, the protein-coding gene, deoxyhypusine synthase, concordantly regulates the expressions of WDR83 mRNA and protein. Conversely, WDR83 also regulates deoxyhypusine synthase by antisense pairing concordantly. As a pair of protein-coding cis-sense/antisense transcripts, WDR83 and DHPS are upregulated simultaneously and correlate positively in lung cancer. They drive the pathophysiology of lung cancer by promoting cell proliferation (Su et al., 2012). Furthermore, the remaining SF PPIL1, which has not been directly reported in the literature to be associated with cancer,

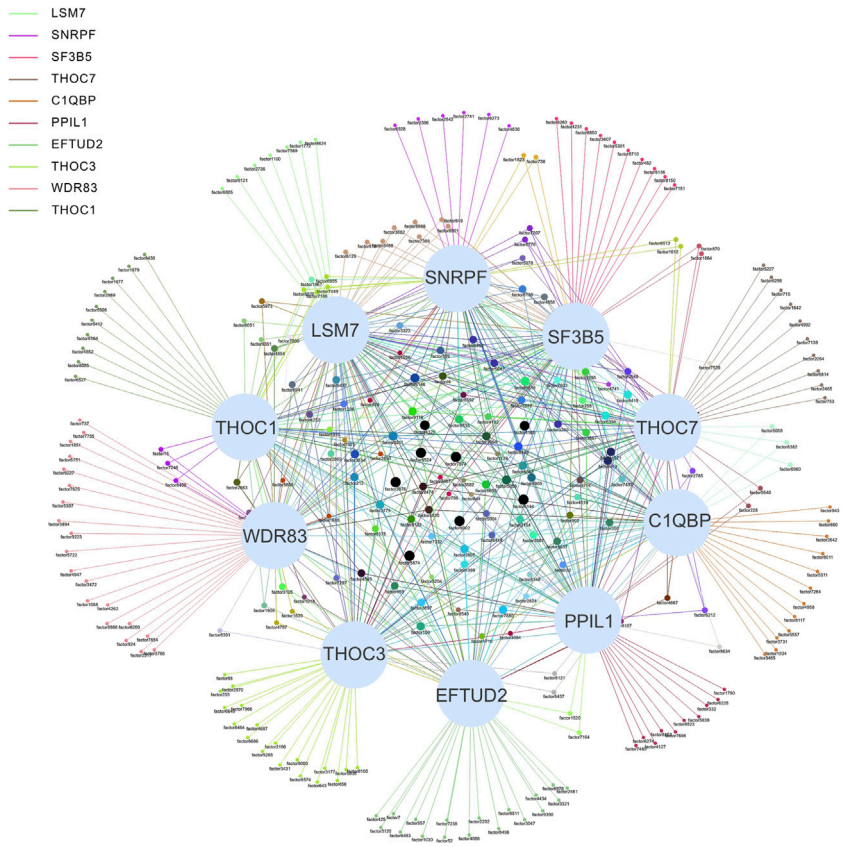


FIGURE 4 | Network diagram of the top 10 SFs and OS-ASEs. The 10 large light blue dots represent 10 SFs, the small dots represent ASEs, and the edges represent significant correlations between the two dots. Edges of different colors represent associations with different SFs. When an ASE is associated with multiple SFs, the color of the edge is a superposition of the corresponding multiple colors. The correspondence between the factor and ASE is shown in **Supplementary Table S3**.

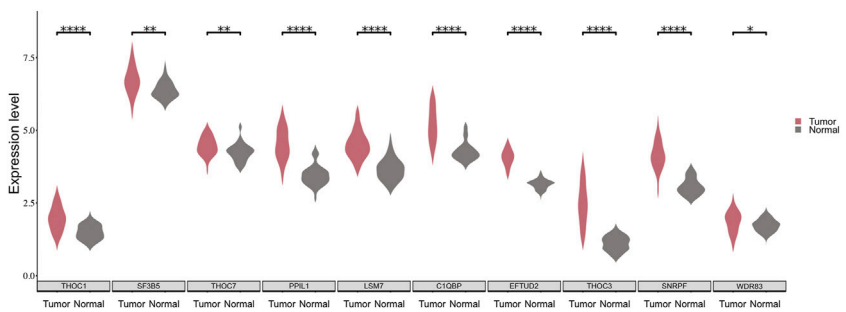


FIGURE 5 | The expression distribution of 10 SFs between cancerous and paraneoplastic tissues in the TCGA dataset. The horizontal axis represents the expression of genes. The vertical axis shows the 10 key SFs.

is a member of the peptidyl-prolyl isomerase procyclin family and is frequently overexpressed in colon cancer cells (Chai et al., 2021). In summary, it is reasonable to speculate that the 7 SFs may play a role in the development of tumors, and the relationship between these SFs and lung cancer warrants further exploration in the future.

Our analysis method can be used not only to screen for key SFs in LUSC but also to apply to a wider range of studies. From the perspective of the study object, although our method is only applied to LUSC data in this study, it is also applicable to other tumor data. From the perspective of research objectives, our method is not limited to screening SFs, but also can be used to screen regulatory factors, such as transcription factors, miRNAs or lncRNAs. For example, we can screen combinations of genes that can accurately classify tumor samples by downscaling or regression and then find key vertices by constructing a regulatory network of miRNAs that can anchor key miRNAs associated with tumors.

In conclusion, our analytical approach with a wide range of applications helps to obtain proper results and can provide new directions and perspectives for the exploration of related studies. In our study, although the specific functions and mechanisms of the 10 key SFs need to be further investigated, the available data and literature imply that they play a critical role in LUSC. Seven of these new SFs are also expected to be a new focus for future studies on SFs in LUSC. Furthermore, our proposed method will provide ideas and references for more studies. However, some limitations remain in our study. Due to the complexity of calculating multivariate combinations, we only calculated bivariate combinations, but multifactor combinations were not further explored. In subsequent studies, we will further improve our methods and extend to more scientific questions to provide novel focuses for future research.

CONCLUSION

Abnormal AS is widely considered a novel indicator of carcinogenic processes, and SFs play a vital role in this process. Consequently, our aim is to screen key SFs that regulate carcinogenesis and progression. All combinations consisting of two ASEs were first constructed and screened using bivariate Cox regression and AUROC. Next, an association network of OS-ASEs and SFs was constructed by the Spearman correlation. Based on topological properties, we screened the top 10 SFs in terms of degree centrality. Finally, literature and data validation were performed on these 10 SFs. The data validation showed that 10 SFs were all significantly differentially expressed in both cancerous and paracancerous tissues of LUSC patients. Moreover, 5 of these SFs showed prognostic effects. It has been reported that 8 of these SFs are closely associated with tumors. In addition, cross-platform validations of GEO were carried out, and similar results were obtained. These findings can serve as a reference for subsequent experimental studies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: TCGA database (https://portal.gdc.cancer.gov/projects/TCGA-LUSC), TCGASpliceSeq database (https://bioinformatics.mdanderson.org/TCGASpliceSeq/index.jsp).

AUTHOR CONTRIBUTIONS

MC and RZ contributed to the data acquisition, figure drawing and drafting of the manuscript. LZ and FZ contributed to the data analysis and interpretation, conception and design of the study. All authors read and approved the final manuscript.

FUNDING

This work was supported by grants from National Natural Science Foundation of China (Grant No. 32070395).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.803606/full#supplementary-material

REFERENCES

Anczuków, O., and Krainer, A. R. (2016). Splicing-factor Alterations in Cancers. *RNA* 22, 1285–1301. doi:10.1261/rna.057919.116

Baralle, F. E., and Giudice, J. (2017). Alternative Splicing as a Regulator of Development and Tissue Identity. *Nat. Rev. Mol. Cel Biol.* 18, 437–451. doi:10.1038/nrm.2017.27

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for Functional Genomics Data Sets-Update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Bishara, A. J., and Hittner, J. B. (2012). Testing the Significance of a Correlation with Nonnormal Data: Comparison of Pearson, Spearman, Transformation, and Resampling Approaches. *Psychol. Methods* 17, 399–417. doi:10.1037/a0028087

Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. (2003). Survival Analysis Part II: Multivariate Data Analysis - an Introduction to Concepts and Methods. *Br. J. Cancer* 89, 431–436. doi:10.1038/sj.bjc.6601119

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *BMC Bioinformatics* 11, 94. doi:10.1186/1471-2105-11-94

Cai, S., Bai, Y., Wang, H., Zhao, Z., Ding, X., Zhang, H., et al. (2020). Knockdown of THOC1 Reduces the Proliferation of Hepatocellular Carcinoma and Increases the Sensitivity to Cisplatin. *J. Exp. Clin. Cancer Res.* 39, 135. doi:10.1186/s13046-020-01634-7

Chai, G., Webb, A., Li, C., Antaki, D., Lee, S., Breuss, M. W., et al. (2021). Mutations in Spliceosomal Genes PPIL1 and PRP17 Cause Neurodegenerative Pontocerebellar Hypoplasia with Microcephaly. *Neuron* 109, 241–256. doi:10.1016/j.neuron.2020.10.035

Chen, L., Yao, Y., Sun, L., Zhou, J., Miao, M., Luo, S., et al. (2017). Snail Driving Alternative Splicing of CD44 by ESRP1 Enhances Invasion and Migration in Epithelial Ovarian Cancer. *Cell. Physiol. Biochem.* 43, 2489–2504. doi:10.1159/000484458

Chen, T., Zhang, H., Liu, Y., Liu, Y.-X., and Huang, L. (2021a). EVenn: Easy to Create Repeatable and Editable Venn Diagrams and Venn Networks Online. *J. Genet. Genomics* 48, 863–866. doi:10.1016/j.jgg.2021.07.007

Chen, Z., Wu, H., Yang, H., Fan, Y., Zhao, S., and Zhang, M. (2021b). Identification and Validation of RNA-binding Protein-related Gene Signature Revealed Potential Associations with Immunosuppression and Drug Sensitivity in Glioma. *Cancer Med.* 10, 7418–7439. doi:10.1002/cam4.4248

Cheng, Z., Yu, C., Cui, S., Wang, H., Jin, H., Wang, C., et al. (2019). circTP63 Functions as a ceRNA to Promote Lung Squamous Cell Carcinoma Progression by Upregulating FOXM1. *Nat. Commun.* 10, 3200. doi:10.1038/s41467-019-11162-4

Climente-González, H., Porta-Pardo, E., Godzik, A., and Eyras, E. (2017). The Functional Impact of Alternative Splicing in Cancer. *Cel Rep.* 20, 2215–2226. doi:10.1016/j.celrep.2017.08.012

David, C. J., and Manley, J. L. (2010). Alternative Pre-mRNA Splicing Regulation in Cancer: Pathways and Programs Unhinged. *Genes Dev.* 24, 2343–2364. doi:10.1101/gad.1973010

Dinse, G. E., and Lagakos, S. W. (1982). Nonparametric Estimation of Lifetime and Disease Onset Distributions from Incomplete Observations. *Biometrics* 38, 921. doi:10.2307/2529872

Dorman, S. N., Viner, C., and Rogan, P. K. (2014). Splicing Mutation Analysis Reveals Previously Unrecognized Pathways in Lymph Node-Invasive Breast Cancer. *Sci. Rep.* 4. doi:10.1038/srep07063

Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R. K. (2016). RNA Splicing Factors as Oncoproteins and Tumour Suppressors. *Nat. Rev. Cancer* 16, 413–430. doi:10.1038/nrc.2016.51

Freeman, L. C. (1978). Centrality in Social Networks Conceptual Clarification. *Social Networks* 1, 215–239. doi:10.1016/0378-8733(78)90021-7

Gatica, D., Hu, G., Liu, X., Zhang, N., Williamson, P. R., and Klionsky, D. J. (2019). The Pat1-Lsm Complex Stabilizes ATG mRNA during Nitrogen Starvation-Induced Autophagy. *Mol. Cel* 73, 314–324. doi:10.1016/j.molcel.2018.11.002

Ge, J., Hu, W., Zhou, H., Yu, J., Sun, C., and Chen, W. (2017). Ubiquitin Carboxyl-Terminal Hydrolase Isozyme L5 Inhibits Human Glioma Cell Migration and Invasion via Downregulating SNRPF. *Oncotarget* 8, 113635–113649. doi:10.18632/oncotarget.23071

Givechian, K. B., Wnuk, K., Garner, C., Benz, S., Garban, H., Rabizadeh, S., et al. (2018). Identification of an Immune Gene Expression Signature Associated with Favorable Clinical Features in Treg-Enriched Patient Tumor Samples. *Npj Genomic Med.* 3, 1–12. doi:10.1038/s41525-018-0054-7

Han, J., Li, J., Ho, J. C., Chia, G. S., Kato, H., Jha, S., et al. (2017). Hypoxia Is a Key Driver of Alternative Splicing in Human Breast Cancer Cells. *Sci. Rep.* 7, 4108. doi:10.1038/s41598-017-04333-0

Hazra, A., and Gogtay, N. (2017). Biostatistics Series Module 9: Survival Analysis. *Indian J. Dermatol.* 62, 251–257. doi:10.4103/ijd.IJD_201_17

Hood, L. (2003). Systems Biology: Integrating Technology, Biology, and Computation. *Mech. Ageing Dev.* 124, 9–16. doi:10.1016/S0047-6374(02)00164-1

Hu, Y.-X., Zheng, M.-J., Zhang, W.-C., Li, X., Gou, R., Nie, X., et al. (2019). Systematic Profiling of Alternative Splicing Signature Reveals Prognostic Predictor for Cervical Cancer. *J. Transl. Med.* 17, 379. doi:10.1186/s12967-019-02140-x

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Husmann, D., and Gozani, O. (2019). Histone Lysine Methyltransferases in Biology and Disease. *Nat. Struct. Mol. Biol.* 26, 880–889. doi:10.1038/s41594-019-0298-7

Jeggo, P. A., Pearl, L. H., and Carr, A. M. (2016). DNA Repair, Genome Stability and Cancer: A Historical Perspective. *Nat. Rev. Cancer* 16, 35–42. doi:10.1038/nrc.2015.4

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., et al. (2010). Identification of Influential Spreaders in Complex Networks. *Nat. Phys* 6, 888–893. doi:10.1038/nphys1746

Lando, M., Fjeldbo, C. S., Wilting, S. M., Snoek, B. C., Aarnes, E.-K., Forsberg, M. F., et al. (2015). Interplay between Promoter Methylation and Chromosomal Loss in Gene Silencing at 3p11-P14 in Cervical Cancer. *Epigenetics* 10, 970–980. doi:10.1080/15592294.2015.1085140

Lando, M., Wilting, S. M., Snipstad, K., Clancy, T., Bierkens, M., Aarnes, E.-K., et al. (2013). Identification of Eight Candidate Target Genes of the Recurrent 3p12-P14 Loss in Cervical Cancer by Integrative Genomic Profiling. *J. Pathol.* 230, 59–69. doi:10.1002/path.4168

Li, W., Li, X., Gao, L.-N., and You, C.-G. (2020). Integrated Analysis of the Functions and Prognostic Values of RNA Binding Proteins in Lung Squamous Cell Carcinoma. *Front. Genet.* 11. doi:10.3389/fgene.2020.00185

Li, Y., Gu, J., Xu, F., Zhu, Q., Ge, D., and Lu, C. (2018). Transcriptomic and Functional Network Features of Lung Squamous Cell Carcinoma through Integrative Analysis of GEO and TCGA Data. *Sci. Rep.* 8, 1–12. doi:10.1038/s41598-018-34160-w

Linden, A. (2006). Measuring Diagnostic and Predictive Accuracy in Disease Management: An Introduction to Receiver Operating Characteristic (ROC) Analysis. *J. Eval. Clin. Pract.* 12, 132–139. doi:10.1111/j.1365-2753.2005.00598.x

Liu, C., Yue, B., Yuan, C., Zhao, S., Fang, C., Yu, Y., et al. (2015). Elevated Expression of Thoc1 Is Associated with Aggressive Phenotype and Poor Prognosis in Colorectal Cancer. *Biochem. Biophysical Res. Commun.* 468, 53–58. doi:10.1016/j.bbrc.2015.10.166

Lv, C., Li, X. J., Hao, L. X., Zhang, S., Song, Z., Ji, X. D., et al. (2021). Over-activation of EFTUD2 Correlates with Tumor Propagation and Poor Survival Outcomes in Hepatocellular Carcinoma. *Clin. Transl. Oncol.* doi:10.1007/s12094-021-02673-y

Mantel, N. (1966). Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration. *Cancer Chemother. Rep.* 50, 163–170. Available at: : https://pubmed.ncbi.nlm.nih.gov/5910392/ (Accessed September 27, 2021).

Matsuda, T., Namura, A., and Oinuma, I. (2019). Dynamic Spatiotemporal Patterns of Alternative Splicing of an F-Actin Scaffold Protein, Afadin, during Murine Development. *Gene* 689, 56–68. doi:10.1016/j.gene.2018.12.020

Miura, K., Fujibuchi, W., and Unno, M. (2012). Splice Isoforms as Therapeutic Targets for Colorectal Cancer. *Carcinogenesis* 33, 2311–2319. doi:10.1093/carcin/bgs347

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., et al. (2020). Integrating Microarray-Based Spatial Transcriptomics and Single-Cell RNA-Seq Reveals Tissue Architecture in Pancreatic Ductal Adenocarcinomas. *Nat. Biotechnol.* 38, 333–342. doi:10.1038/s41587-019-0392-8

Nakka, K., Ghigna, C., Gabellini, D., and Dilworth, F. J. (2018). Diversification of the Muscle Proteome through Alternative Splicing. *Skeletal Muscle* 8, 8. doi:10.1186/s13395-018-0152-3

Östman, A., Hellberg, C., and Böhmer, F. D. (2006). Protein-tyrosine Phosphatases and Cancer. *Nat. Rev. Cancer* 6, 307–320. doi:10.1038/nrc1837

Ouyang, Y., Xia, K., Yang, X., Zhang, S., Wang, L., Ren, S., et al. (2021). Alternative Splicing Acts as an Independent Prognosticator in Ovarian Carcinoma. *Sci. Rep.* 11, 10413. doi:10.1038/s41598-021-89778-0

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing. *Nat. Genet.* 40, 1413–1415. doi:10.1038/ng.259

Park, S., Brugiolo, M., Akerman, M., Das, S., Urbanski, L., Geier, A., et al. (2019). Differential Functions of Splicing Factors in Mammary Transformation and Breast Cancer Metastasis. *Cel Rep.* 29, 2672–2688. doi:10.1016/j.celrep.2019.10.110

Pradella, D., Naro, C., Sette, C., and Ghigna, C. (2017). EMT and Stemness: Flexible Processes Tuned by Alternative Splicing in Development and Cancer Progression. *Mol. Cancer* 16. doi:10.1186/s12943-016-0579-2

Qi, C., Chen, Y., Zhou, Y., Huang, X., Li, G., Zeng, J., et al. (2018). Delineating the Underlying Molecular Mechanisms and Key Genes Involved in Metastasis of Colorectal Cancer via Bioinformatics Analysis. *Oncol. Rep.* 39, 2297–2305. doi:10.3892/or.2018.6303

Ryan, M., Wong, W. C., Brown, R., Akbani, R., Su, X., Broom, B., et al. (2016). TCGASpliceSeq a Compendium of Alternative mRNA Splicing in Cancer. *Nucleic Acids Res.* 44, D1018–D1022. doi:10.1093/nar/gkv1288

Saha, S., Kim, K., Islam, S. M., Cho, S.-G., and Gil, M. (2019). Systematic Multiomics Analysis of Alterations in C1QBP mRNA Expression and Relevance for Clinical Outcomes in Cancers. *Jcm* 8, 513. doi:10.3390/jcm8040513

Shkreta, L., Bell, B., Revil, T., Venables, J. P., Prinos, P., Elela, S. A., et al. (2013). Cancer-associated Perturbations in Alternative Pre-messenger RNA Splicing. *Cancer Treat. Res.* 158, 41–94. doi:10.1007/978-3-642-31659-3_3

Shkreta, L., and Chabot, B. (2015). The RNA Splicing Response to DNA Damage. *Biomolecules* 5, 2935–2977. doi:10.3390/biom5042935

Su, H., Hu, J., Huang, L., Yang, Y., Thenoz, M., Kuchmiy, A., et al. (2018). SHQ1 Regulation of RNA Splicing Is Required for T-Lymphoblastic Leukemia Cell Survival. *Nat. Commun.* 9, 4281. doi:10.1038/s41467-018-06523-4

Su, W.-Y., Li, J.-T., Cui, Y., Hong, J., Du, W., Wang, Y.-C., et al. (2012). Bidirectional Regulation between WDR83 and its Natural Antisense Transcript DHPS in Gastric Cancer. *Cell Res* 22, 1374–1389. doi:10.1038/cr.2012.57

Suo, C., Hrydziuszko, O., Lee, D., Pramana, S., Saputra, D., Joshi, H., et al. (2015). Integration of Somatic Mutation, Expression and Functional Data Reveals Potential Driver Genes Predictive of Breast Cancer Survival. *Bioinformatics* 31, 2607–2613. doi:10.1093/bioinformatics/btv164

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). Revigo Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One* 6, e21800. doi:10.1371/journal.pone.0021800

Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A., and Skotheim, R. I. (2016). Aberrant RNA Splicing in Cancer; Expression Changes and Driver Mutations of Splicing Factor Genes. *Oncogene* 35, 2413–2427. doi:10.1038/onc.2015.318

Tyson-Capper, A., and Gautrey, H. (2018). Regulation of Mcl-1 Alternative Splicing by hnRNP F, H1 and K in Breast Cancer Cells. *RNA Biol.* 15, 1448–1457. doi:10.1080/15476286.2018.1551692

Wan, J., Zou, S., Hu, M., Zhu, R., Xu, J., Jiao, Y., et al. (2014). Thoc1 Inhibits Cell Growth via Induction of Cell Cycle Arrest and Apoptosis in Lung Cancer Cells. *Mol. Med. Rep.* 9, 2321–2327. doi:10.3892/mmr.2014.2088

Wang, C., Tan, S., Liu, W.-R., Lei, Q., Qiao, W., Wu, Y., et al. (2019). RNA-seq Profiling of Circular RNA in Human Lung Adenocarcinoma and Squamous Cell Carcinoma. *Mol. Cancer* 18, 134. doi:10.1186/s12943-019-1061-8

Wang, F., Fu, X., Chen, P., Wu, P., Fan, X., Li, N., et al. (2017). SPSB1-mediated HnRNP A1 Ubiquitylation Regulates Alternative Splicing and Cell Migration in EGF Signaling. *Cel Res* 27, 540–558. doi:10.1038/cr.2017.7

Wang, Y., Yang, F., Shang, J., He, H., and Yang, Q. (2021). Integrative Analysis Reveals the Prognostic Value and Functions of Splicing Factors Implicated in Hepatocellular Carcinoma. *Sci. Rep.* 11, 1–14. doi:10.1038/s41598-021-94701-8

Wong, A. C. H., Rasko, J. E. J., and Wong, J. J.-L. (2018). We Skip to Work: Alternative Splicing in normal and Malignant Myelopoiesis. *Leukemia* 32, 1081–1093. doi:10.1038/s41375-018-0021-4

Wong, R. S. (2011). Apoptosis in Cancer: From Pathogenesis to Treatment. *J. Exp. Clin. Cancer Res.* 30, 87. doi:10.1186/1756-9966-30-87

Wu, Z., Chen, H., Liang, Y., Luo, W., Deng, F., and Zeng, F. (2020). Alternative Splicing Implicated in Immunity and Prognosis of colon Adenocarcinoma. *Int. Immunopharmacology* 89, 107075. doi:10.1016/j.intimp.2020.107075

Xie, R., Chen, X., Chen, Z., Huang, M., Dong, W., Gu, P., et al. (2019). Polypyrimidine Tract Binding Protein 1 Promotes Lymphatic Metastasis and Proliferation of Bladder Cancer via Alternative Splicing of MEIS2 and PKM. *Cancer Lett.* 449, 31–44. doi:10.1016/j.canlet.2019.01.041

Xu, Q., Modrek, B., and Lee, C. (2002). Genome-wide Detection of Tissue-specific Alternative Splicing in the Human Transcriptome. *Nucleic Acids Res.* 30, 3754–3766. doi:10.1093/nar/gkf492

Yuan, G., Flores, N. M., Hausmann, S., Lofgren, S. M., Kharchenko, V., Angulo-Ibanez, M., et al. (2021). Elevated NSD3 Histone Methylation Activity Drives Squamous Cell Lung Cancer. *Nature* 590, 504–508. doi:10.1038/s41586-020-03170-y

Yuan, J.-h., Liu, X.-n., Wang, T.-t., Pan, W., Tao, Q.-f., Zhou, W.-p., et al. (2017). The MBNL3 Splicing Factor Promotes Hepatocellular Carcinoma by Increasing PXN Expression through the Alternative Splicing of lncRNA-PXN-AS1. *Nat. Cel Biol.* 19, 820–832. doi:10.1038/ncb3538

Zhao, J., Chang, L., Gu, X., Liu, J., Sun, B., and Wei, X. (2020). Systematic Profiling of Alternative Splicing Signature Reveals Prognostic Predictor for Prostate Cancer. *Cancer Sci.* 111, 3020–3031. doi:10.1111/cas.14525

Zhao, X., and Liu, Z.-P. (2019). Analysis of Topological Parameters of Complex Disease Genes Reveals the Importance of Location in a Biomolecular Network. *Genes* 10, 143. doi:10.3390/genes10020143

Zhu, J., Chen, Z., and Yong, L. (2018). Systematic Profiling of Alternative Splicing Signature Reveals Prognostic Predictor for Ovarian Cancer. *Gynecol. Oncol.* 148, 368–374. doi:10.1016/j.ygyno.2017.11.028

Zong, Z., Li, H., Yi, C., Ying, H., Zhu, Z., and Wang, H. (2018). Genome-wide Profiling of Prognostic Alternative Splicing Signature in Colorectal Cancer. *Front. Oncol.* 8, 537. doi:10.3389/fonc.2018.00537

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.

Copyright © 2022 Chen, Zhu, Zhang and Zhu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 January 2022
doi: 10.3389/fgene.2021.827161

Immunoglobulin Classification Based on FC* and GC* Features

Hao Wan[1], *Jina Zhang*[2], *Yijie Ding*[3], *Hetian Wang*[4]* *and Geng Tian*[2]*

[1]Institute of Advanced Cross-field Science, College of Life Science, Qingdao University, Qingdao, China, [2]Geneis (Beijing) Co., Ltd., Beijing, China, [3]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, [4]Beidahuang Industry Group General Hospital, Harbin, China

Immunoglobulins have a pivotal role in disease regulation. Therefore, it is vital to accurately identify immunoglobulins to develop new drugs and research related diseases. Compared with utilizing high-dimension features to identify immunoglobulins, this research aimed to examine a method to classify immunoglobulins and non-immunoglobulins using two features, FC* and GC*. Classification of 228 samples (109 immunoglobulin samples and 119 non-immunoglobulin samples) revealed that the overall accuracy was 80.7% in 10-fold cross-validation using the J48 classifier implemented in Weka software. The FC* feature identified in this study was found in the immunoglobulin subtype domain, which demonstrated that this extracted feature could represent functional and structural properties of immunoglobulins for forecasting.

Keywords: immunoglobulin classification, machine learning, key feature extraction, MRMD, autoprop

OPEN ACCESS

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

Reviewed by:
Ru Xiaoqing,
University of Tsukuba, Japan
Pengfei Zhu,
Tianjin University, China

***Correspondence:**
Hetian Wang
hetw2021@163.com
Geng Tian
tiang@geneis.cn

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

Received: *01 December 2021*
Accepted: *22 December 2021*
Published: *24 January 2022*

Citation:
Wan H, Zhang J, Ding Y, Wang H and Tian G (2022) Immunoglobulin Classification Based on FC and GC* Features. Front. Genet. 12:827161. doi: 10.3389/fgene.2021.827161*

1 INTRODUCTION

Immunoglobulins, or antibodies, are a group of proteins secreted by B lymphocytes that recognize invading antigens and bind to antigens with high affinity and specificity to neutralize toxic substances. In general, antibodies are composed of two identical polypeptide chains, each with a light chain and a heavy chain (Narciso et al., 2011). They can be divided functionally into variable (V) domains, which bind to antigens, and constant (C) domains, which activate, complement, or bind to Fc receptors (Schroeder and Cavacini, 2010). To predict the structure of immunoglobulins, (Lepore et al., 2017) developed the PIGSPro Server, an updated version of the popular PIGS Server.

Immunoglobulins have a pivotal role in disease regulation. Therefore, human and nonhuman polyclonal immunoglobulins have been used in therapeutics for many years. Five monoclonal immunoglobulins ranked in the top 10 blockbuster biotherapeutics drugs (Norman et al., 2020). Patients with primary immune deficiencies greatly benefit from the intravenous or subcutaneous administration of human immunoglobulin preparations (Perez et al., 2017). The advanced development of medicine is urged by its finite supply, which requires more identification of valuable therapeutic immunoglobulins. However, biochemical experiments are time-consuming with enzymes to fragment immunoglobulin molecules (Schroeder and Cavacini, 2010) or X-ray crystallography to obtain accurate structures (Narciso et al., 2011).

Machine learning can identify desired proteins from a large number of sequences within a short time to guide the experimental discovery process (Guo et al., 2020; Liu et al., 2020; Song G. et al., 2021; Cheng et al., 2021; Deng et al., 2021; Dong et al., 2021; Guo et al., 2021; Tang et al., 2021; Yu et al., 2021; Zhao et al., 2021). Over the past decades, researchers have developed many machine learning–based techniques for protein sequence analysis (Zhai et al., 2020; Zeng et al., 2020; Chen et al., 2021; Li et al., 2021). The bioinformatics approach of identifying immunoglobulins is to convert protein sequences into numerical vectors to reveal the internal structures of proteins. The

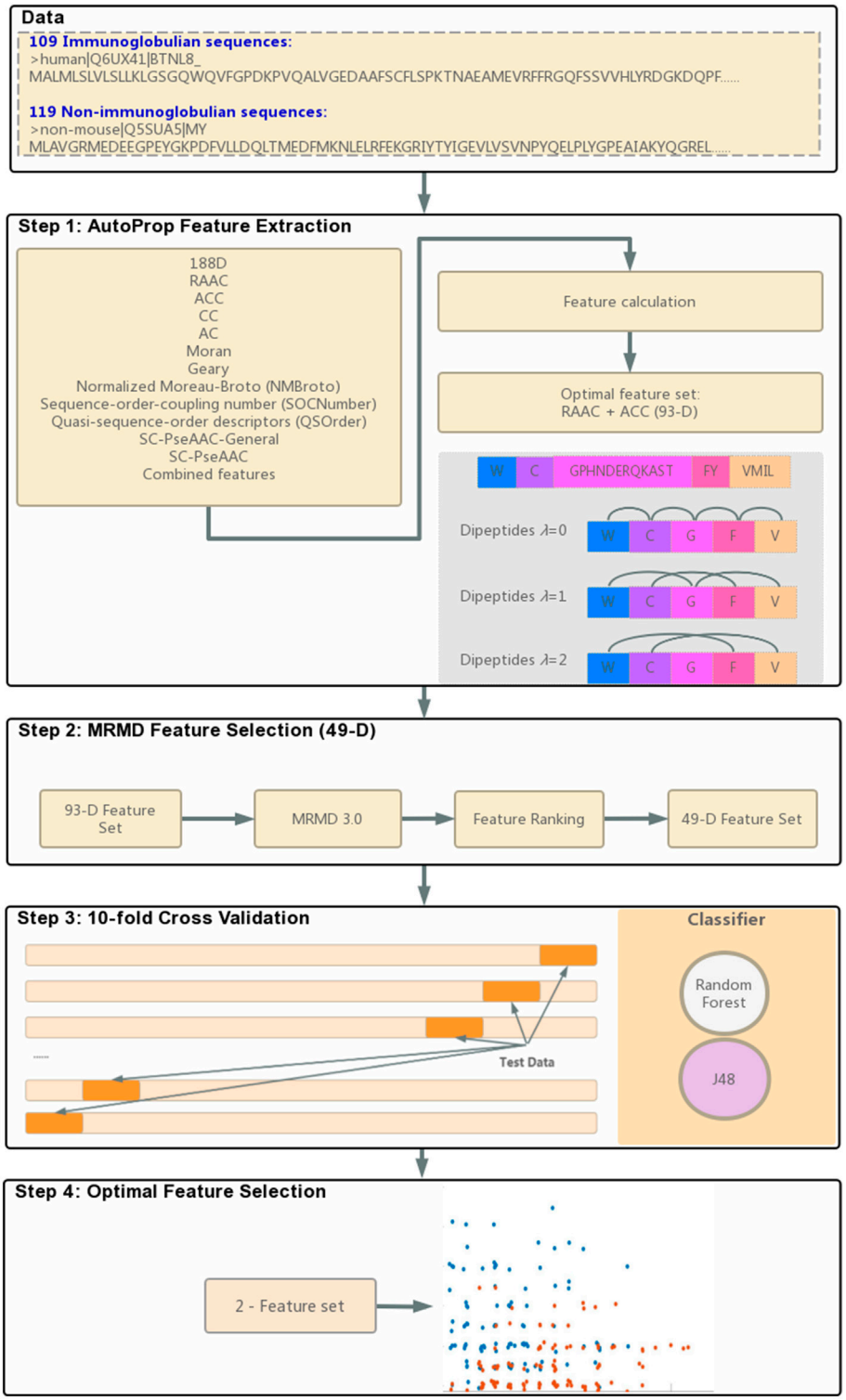


FIGURE 1 | Flowchart of identifying immunoglobulins.

critical aspects of protein identification can be listed as follows: feature extraction, feature selection, and machine learning. Feature extraction methods include *n-gram* feature type: amino acid composition (AAC), Dipeptides (Dip), Tripeptides, where frequencies of *n*-length peptides are used as feature vectors (Ding et al., 2011; Gautam et al., 2013; Diener et al., 2016; Rahman et al., 2018; Liu et al., 2019; Lv et al., 2019; Fu et al., 2020; Wang H. et al., 2021; Wang J. et al., 2021; Zhai et al., 2020; Shao and Liu, 2021; Yang et al., 2021; Zhang et al., 2021). In addition, pseudo–amino acid composition (PseAAC) is also a widely adopted feature extraction method, including physicochemical properties between residues (Hansen et al., 2008; Sanders et al., 2011; Gautam et al., 2013; Chen et al., 2016; Diener et al., 2016; Khan et al., 2020; Awais et al., 2021; Naseer et al., 2021).

Many feature types and complex classification methods may generate redundant information (Song B. et al., 2021). Therefore, some studies began to eliminate redundant parts to improve the predictive performance of classification models. This process is also called feature selection. MRMD (Zou et al., 2016; Ao et al., 2020; Li et al., 2020a; Li et al., 2020b; Meng et al., 2020) and ANOVA (Anderson, 2001; Lv et al., 2019) are standard feature selection methods. For optimal feature identification, (Feng et al., 2021) uses the PCA and MCE methods to make the features orthogonal and obtain the core feature set with the minimum 10-dimensional attributes for PPR gene identification and realized 97.9% accuracy. (Li et al., 2020b) used a 19-dimensional feature model to classify anticancer peptide sequences. (Ao et al., 2020) used a 10-dimensional feature model to classify antioxidant proteins and realized 90.44% accuracy. (Meng et al., 2020) used a 6-dimensional feature model to classify cell wall lytic enzymes.

However, very few tools have been developed for immunoglobulin identification. (Tang et al., 2016) used the pseudo amino acid composition (PseAAC) feature extraction approach to realize over 96% prediction accuracy in their pioneering work on immunoglobulin identification. (Gong et al., 2021) used the CC–PSSM and monoTriKGap feature extraction, MRMD feature selection, and single dimension reduction methods to realize 92.1% immunoglobulin identification accuracy by two-dimensional features. However, the link between optimal features and functional structures of immunoglobulins remains to be investigated.

To obtain a diverse feature set, this study integrated 188-D physicochemical properties, auto-cross covariance (ACC) information, and dipeptide compositions of reduced amino acids. Dimensions were reduced using the max-relevance-max-distance (MRMD) method and the single dimension reduction method. The RF and J48 classifiers implemented in Weka software were used to identify immunoglobulins. Finally, two features can correctly predict immunoglobulins, FC* and GC*. The entire modeling process is illustrated in **Figure 1**. The FC* feature identified in this study was found in immunoglobulin subtype domain IPR003599, which demonstrated that this extracted feature could represent functional and structural properties of immunoglobulins for forecasting.

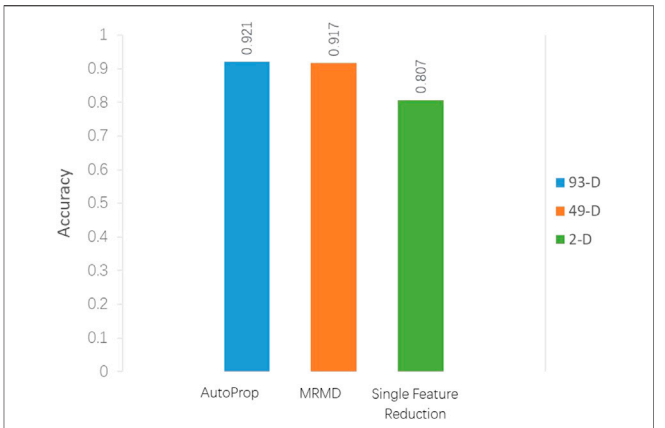


FIGURE 2 | Classification accuracy comparison between models with different feature selection methods.

2 MATERIALS AND METHODS

2.1 Datasets

Data for this study were collected by (Tang et al., 2016), which contain 228 samples (109 immunoglobulin samples and 119 non-immunoglobulin samples) extracted from the Universal Protein Resource (UniProt).

2.2 RAAC

Polypeptide chains fold to tertiary structures based on the physicochemical properties of residues (Tang et al., 2016). Analyzing the occurrence frequency of residue compositions cannot visualize three-dimensional protein structures. The reduced amino acid cluster (RAAC) method, replacing protein sequences with less than 20 amino acid alphabets based on a specific reducing scheme, can reduce sequence complexity. With removing non-essential information, functionally conserved regions will be displayed more clearly. Recent work presented 3D protein structures of ectonucleotide pyrophosphatase with a 1D view using the RAAC method (Solis, 2015; Zheng et al., 2019).

There are many choices of reduced schemes, and different decisions could produce distinctive protein classification results. For example, the RAACBook web server provided 74 types of reduced amino acid alphabets derived from over 1,000 published articles in PubMed (Zheng et al., 2019). Bins within the scheme are related to the chemical properties of amino acids. Dayhoff classes (AGPST, DENQ, HKR, ILMV, FWY, and C) are most used. Also, S and T are frequently together, and so are K and R, D, and E (Susko and Roger, 2007).

We used the AutoProp (Feng et al., 2020) to screen out the optimal reduced scheme of the immunoglobulin and non-immunoglobulin sequences. GPHNDERQKAST, FY, VMIL, C, and W (**Figure 1** Step 1) were used. Under this reduced scheme, the 20 amino acid alphabets were represented by five alphabets: G, F, V, C, and W. For instance, any amino acid that is a G, P, H, N, D, E, R, Q, K, A, S, or T is then treated as character G. For any amino acid F and Y, it is then treated as character F, and so forth.

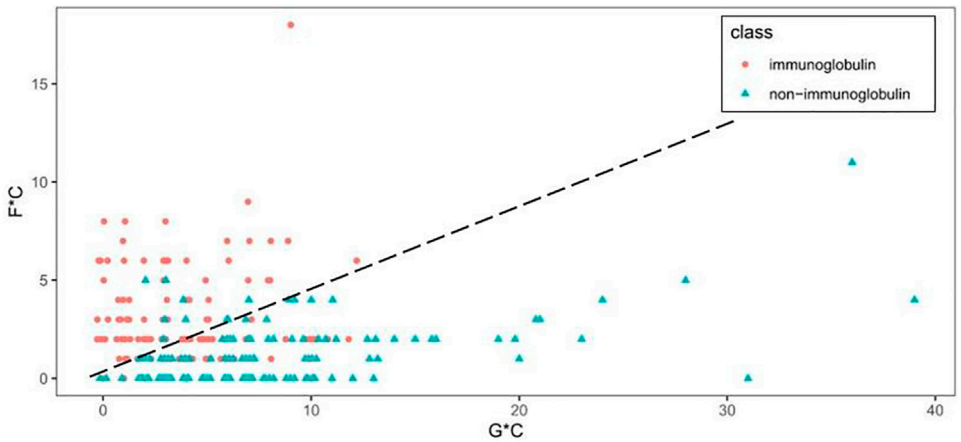


FIGURE 3 | Scatter plot of GC* and FC* features.

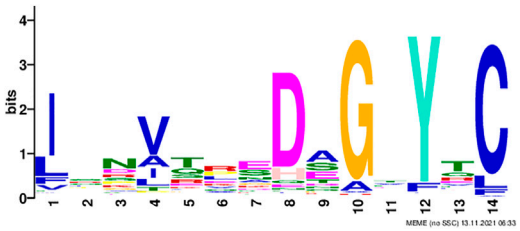


FIGURE 4 | Motif discovered among immunoglobulin sequences using the MEME tool; the height of the letter indicates its relative frequency at the given position within the motif.

2.3 Feature Extraction

A sequence can be represented by sequential form and discrete form. Homolog sequences can be compared with the BLAST or FASTA program benchmark datasets for traditional sequence comparison methods. However, the similarity-based way is unsuitable for distantly related sequences (Wei et al., 2014; Chen et al., 2016; Jin et al., 2019; Manavalan et al., 2019; Hong et al., 2020; Tang et al., 2020; Wang et al., 2020; Ding et al., 2021a; Ding et al., 2021b; Huang et al., 2021; Shao et al., 2021). By converting amino acid codes to a series of discrete numerical vectors, the discrete form can overcome this drawback and be used by machine learning for protein classification. Sometimes, proteins can be classified according to fewer features, while BLAST cannot.

Different numerical values of protein codes mean different feature descriptors. Feature descriptors provided by AutoProp include 188D, ACC, PseAAC, and another nine methods (**Figure 1** Step 1). Also, AutoProp provides combined features between those methods. The built-in classifiers will then calculate the accuracy percentage of each feature and decide the optimal feature.

For our data, the optimal feature is the combined features of RAAC and ACC. RAAC features also represent dipeptides of reduced amino acid, like CV, C*V (λ-gap = 1), and C**V (λ-gap = 2). The following formula was used to calculate the values of those features:

$$f_u = \frac{n_u^\lambda}{\sum n_u^\lambda},$$

where λ = 0,1,2, and n_u^λ denotes the number of λ-gap dipeptides of type u in a protein sequence.

ACC means the autocross covariance (ACC) transformation and contains auto covariance (AC) and cross-covariance (CC) and is introduced to transform protein sequences into fixed-length vectors (Feng et al., 2020). With its ability to identify sequence homologies, ACC has been successfully used for protein family classification and protein interaction prediction (Dong et al., 2009).

2.4 MRMD

The main disadvantage of the sequence word frequency vector is that they are usually huge. Therefore, dimension reduction, also called feature selection, is chosen for protein classification. The MRMD method, which is the max-relevance-max-distance–based dimensionality reduction method, is more considered for relationships among features and stability of feature selection. Cross-validation and the ROC curve are usually used to evaluate classification accuracy. The MRMD method can reduce feature dimensions with few accuracy drops (Zou et al., 2016; He et al., 2020; Tao et al., 2020).

2.5 Performance Measurement

We used three metrics to evaluate model performance. Indicators include sensitivity (SE), specificity (SP), and Accuracy (Jiang et al., 2013; Wang X. et al., 2021). Calculation methods are described as follows:

$$\mathrm{SE} = \frac{TP}{TP + FN},$$

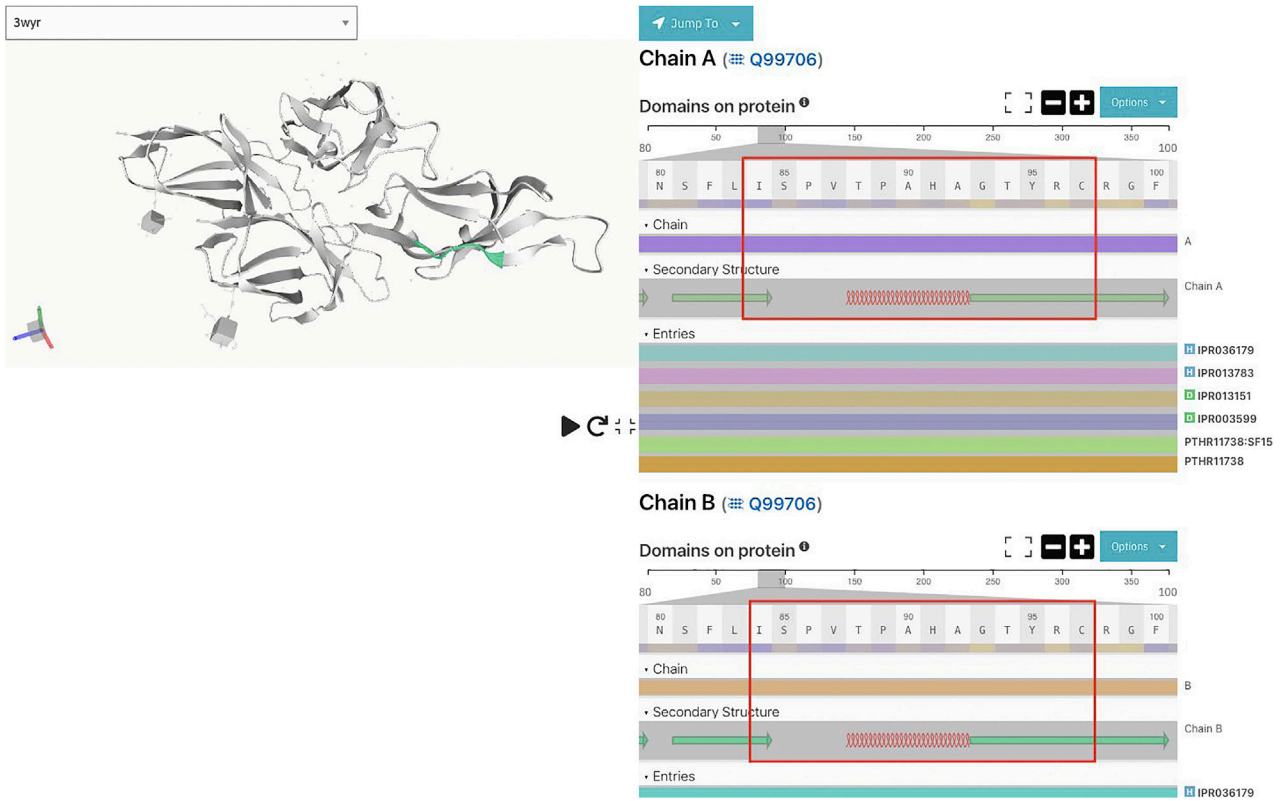


FIGURE 5 | Shared motif and its secondary structure (from PDB entry 3wyr) using InterproScan.

$$\mathrm{SP} = \frac{TN}{TN + FP},$$

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + FN + TN + NP},$$

where TN, TP, FN, and FP refer to the numbers of correctly predicted non-immunoglobulin proteins, correctly predicted immunoglobulin proteins, incorrectly predicted non-immunoglobulin proteins, and incorrectly predicted immunoglobulin proteins, respectively. Sensitivity (SE) is also known as recall, and it measures the percentage that positive samples can be expected correctly over all the samples. SP indicators measure the probability of negative samples classified as non-immunoglobulins, and Accuracy is used to evaluate the overall performance of a prediction model.

3 RESULTS AND DISCUSSION

3.1 Classification Results Under Different Features

Props returned 93D best features, and the frequency of dipeptides (λ-gap = 0, 1, 2) is saved in features 1–75, followed by 18 ACC features. The classification accuracy was 92.1% in the RF classifier and 10-fold cross-validation using Weka software. The MRMD method further reduced the dimension to 49D, and accuracy was 91.7% using the same classifier. It can be seen that MRMD reduces nearly half of the feature dimension, but the accuracy is only dropped by 0.4% (**Figure 2**). After continuous attempts to reduce features, the optimal two features (GC* and FC*) are finally obtained; the classification accuracy was 80.3% using the J48 classifier in Weka.

3.2 2D Features Scatter Distribution

Figure 3 shows the scatter plot of GC* and FC* features. What stands out in **Figure 3** is that immunoglobulin and non-immunoglobulin samples can be distinguished. Immunoglobulins are scattered on the upper left with higher FC* values, and non-immunoglobulins are found in the lower right with higher GC* values. For 118 out of 119 non-immunoglobulin samples, the FC* value is equal to or less than 5. Among these, the FC* value of 49 samples is zero. The GC* value for immunoglobulin samples is less than or equal to 12.

3.3 Interpretation of Feature FC*

We noticed 49 out of 119 non-immunoglobulin samples had an FC* value of zero, whereas only four immunoglobulin samples had an FC* value of zero. Using motif search website MEME Suite 5.4.1 (Bailey and Elkan, 1994; Bailey et al., 2009) and running 109 immunoglobulin sequences, results showed that 107 out of 109 immunoglobulin samples had a motif, "ISNVTREDAGTYTC" (**Figure 4**). Based on the reduced scheme, Y was treated as F.

Immunoglobulin sequences were subjected to InterProScan (Zdobnov and Apweiler 2001) to understand the motif structure

better to map protein domains. Results showed that the finding motif belonged to immunoglobulin subtype domain IPR003599.

Also, secondary structure predictions of the motif using JPred (Drozdetskiy et al., 2015) predict that the shared motif comprises alpha helices and beta sheets separated by disordered regions (**Figure 5**).

4 CONCLUSION

The present research aimed to examine a method to classify immunoglobulins and non-immunoglobulins using two features, GC* and FC*. Classification of 228 samples (109 immunoglobulin samples and 119 non-immunoglobulin samples) revealed that the overall accuracy was 80.7% in the J48 classifier and 10-fold cross-validation using Weka software. The FC* feature identified in this study was found in immunoglobulin subtype domain IPR003599, which demonstrated that this extracted feature could represent functional and structural properties of immunoglobulins for forecasting.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: DOI: 10.1039/c5mb00883b.

AUTHOR CONTRIBUTIONS

Conceptualization, HeW and GT; data collection or analysis, HaW, JZ and YD; validation, HaW; writing—original draft preparation, HaW; writing—review and editing, HeW and GT. All authors have read and agreed to the published version of the manuscript.

FUNDING

The work was supported by the Special Science Foundation of Quzhou (2021D004).

REFERENCES

Anderson, M. J. (2001). A New Method for Non-parametric Multivariate Analysis of Variance. *Austral Ecol.* 26 (1), 32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x

Ao, C., Zhou, W., Gao, L., Dong, B., and Yu, L. (2020). Prediction of Antioxidant Proteins Using Hybrid Feature Representation Method and Random forest. *Genomics* 112 (6), 4666–4674. doi:10.1016/j.ygeno.2020.08.016

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

Bailey, T. L., and Elkan, C. (1994). Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: Tools for Motif Discovery and Searching. *Nucleic Acids Res.* 37, W202–W208. doi:10.1093/nar/gkp335

Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. *Biomed. Res. Int.* 2016, 1654623. doi:10.1155/2016/1654623

Chen, Y., Ma, T., Yang, X., Wang, J., Song, B., and Zeng, X., (2021). MUFFIN: Multi-Scale Feature Fusion for Drug–Drug Interaction Prediction. *Bioinformatics* 37 (17), 2651–2658. doi:10.1093/bioinformatics/btab169

Cheng, Y., Gong, Y., Liu, Y., Song, B., and Zou, Q. (2021). Molecular Design in Drug Discovery: a Comprehensive Review of Deep Generative Models. *Brief Bioinform* 22 (6). doi:10.1093/bib/bbab344

Deng, L., Huang, Y., Liu, X., and Liu, H., (2021). Graph2MDA: a Multi-Modal Variational Graph Embedding Model for Predicting Microbe-Drug Associations. *Bioinform.* btab792. doi:10.1093/bioinformatics/btab792

Diener, C., Garza Ramos Martínez, G., Moreno Blas, D., Castillo González, D. A., Corzo, G., Castro-Obregon, S., et al. (2016). Effective Design of Multifunctional Peptides by Combining Compatible Functions. *Plos Comput. Biol.* 12 (4), e1004786. doi:10.1371/journal.pcbi.1004786

Ding, H., Liu, L., Guo, F.-B., Huang, J., and Lin, H. (2011). Identify Golgi Protein Types with Modified Mahalanobis Discriminant Algorithm and Pseudo Amino Acid Composition. *Ppl* 18 (1), 58–63. doi:10.2174/092986611794328708

Ding, Y., Yang, C., Tang, J., and Guo, F. (2021a). *Identification of Protein-Nucleotide Binding Residues via Graph Regularized K-Local Hyperplane Distance Nearest Neighbor Model*. Berlin, Germany: Applied Intelligence.

Ding, Y., Tang, J., and Guo, F. (2021b). Protein Crystallization Identification via Fuzzy Model on Linear Neighborhood Representation. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18 (5), 1986–1995. doi:10.1109/tcbb.2019.2954826

Dong, J., Zhao, M., Liu, Y., Su, Y., and Zeng, X., (2021). Deep Learning in Retrosynthesis Planning: Datasets, Models and Tools. *Brief. Bioinform.* bbab391. doi:10.1093/bib/bbab391

Dong, Q., Zhou, S., and Guan, J. (2009). A New Taxonomy-Based Protein Fold Recognition Approach Based on Autocross-Covariance Transformation. *Bioinformatics* 25 (20), 2655–2662. doi:10.1093/bioinformatics/btp500

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. (2015). JPred4: a Protein Secondary Structure Prediction Server. *Nucleic Acids Res.* 43 (W1), W389–W394. doi:10.1093/nar/gkv332

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

Feng, C., Zou, Q., and Wang, D. (2021). Using a Low Correlation High Orthogonality Feature Set and Machine Learning Methods to Identify Plant Pentatricopeptide Repeat Coding Gene/protein. *Neurocomputing* 424, 246–254. doi:10.1016/j.neucom.2020.02.079

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

Gautam, A., Chaudhary, K., Kumar, R., Sharma, A., Kapoor, P., Tyagi, A., et al. (2013). In Silico approaches for Designing Highly Effective Cell Penetrating Peptides. *J. Transl Med.* 11, 74. doi:10.1186/1479-5876-11-74

Gong, Y., Peng, D., Liao, B., and Zou, Q., (2021). Accurate Prediction and Key Feature Recognition of Immunoglobulin. *Appl. Sciences-Basel* 11 (15), 6894. doi:10.3390/app11156894

Guo, Y., Yan, K., Lv, H., and Liu, B. (2021). PreTP-EL: Prediction of Therapeutic Peptides Based on Ensemble Learning. *Brief. Bioinform.* 22 (6), bbab358. doi:10.1093/bib/bbab358

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

Hansen, M., Kilk, K., and Langel, U. (2008). Predicting Cell-Penetrating Peptides. *Adv. Drug Deliv. Rev.* 60 (4-5), 572–579. doi:10.1016/j.addr.2007.09.003

He, S., Guo, F., Zou, Q., and Ding, H., (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, S., He, X., Wang, G., and Bao, E., (2021). AlignGraph2: Similar Genome-Assisted Reassembly Pipeline for PacBio Long Reads. *Brief Bioinform* 22 (5), bbab022. doi:10.1093/bib/bbab022

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, Q., Meng, Z., Pham, T. D., Chen, Q., Wei, L., and Su, R. (2019). DUNet: A Deformable Network for Retinal Vessel Segmentation. *Knowledge-Based Syst.* 178, 149–162. doi:10.1016/j.knosys.2019.04.025

Khan, Y. D., Alzahrani, E., Alghamdi, W., and Ullah, M. Z., (2020). Sequence-based Identification of Allergen Proteins Developed by Integration of PseAAC and Statistical Moments via 5-Step Rule. *Curr. Bioinformatics* 15 (9), 1046–1055. doi:10.2174/1574893615999200424085947

Lepore, R., Olimpieri, P. P., Messih, M. A., and Tramontano, A. (2017). PIGSPro: Prediction of immunoGlobulin Structures V2. *Nucleic Acids Res.* 45 (W1), W17–W23. doi:10.1093/nar/gkx334

Li, H.-L., Pang, Y.-H., and Liu, B. (2021). BioSeq-BLM: a Platform for Analyzing DNA, RNA and Protein Sequences Based on Biological Language Models. *Nucleic Acids Res.* 49, e129. doi:10.1093/nar/gkab829

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

Li, Q., Dong, B., Wang, D., and Wang, S. (2020a). Identification of Secreted Proteins from Malaria Protozoa with Few Features. *Ieee Access* 8, 89793–89801. doi:10.1109/access.2020.2994206

Li, Q., Zhou, W., Wang, D., Wang, S., and Li, Q. (2020b). Prediction of Anticancer Peptides Using a Low-Dimensional Feature Model. *Front. Bioeng. Biotechnol.* 8, 892. doi:10.3389/fbioe.2020.00892

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

Lv, Z., Jin, S., Ding, H., and Zou, Q. (2019). A Random Forest Sub-golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features. *Front. Bioeng. Biotechnol.* 7, 215. doi:10.3389/fbioe.2019.00215

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (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

Meng, C., Wu, J., Guo, F., Dong, B., and Xu, L. (2020). CWLy-Pred: A Novel Cell wall Lytic Enzyme Identifier Based on an Improved MRMD Feature Selection Method. *Genomics* 112 (6), 4715–4721. doi:10.1016/j.ygeno.2020.08.015

Narciso, J. E. T., Uy, I. D. C., Cabang, A. B., Chavez, J. F. C., Pablo, J. L. B., Padilla-Concepcion, G. P., et al. (2011). Analysis of the Antibody Structure Based on High-Resolution Crystallographic Studies. *New Biotechnol.* 28 (5), 435–447. doi:10.1016/j.nbt.2011.03.012

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

Norman, R. A., Ambrosetti, F., Bonvin, A. M. J. J., Colwell, L. J., Kelm, S., Kumar, S., et al. (2020). Computational Approaches to Therapeutic Antibody Design: Established Methods and Emerging Trends. *Brief. Bioinform.* 21 (5), 1549–1567. doi:10.1093/bib/bbz095

Perez, E. E., Orange, J. S., Bonilla, F., Chinen, J., Chinn, I. K., Dorsey, M., et al. (2017). Update on the Use of Immunoglobulin in Human Disease: A Review of Evidence. *J. Allergy Clin. Immunol.* 139 (3), S1–S46. doi:10.1016/j.jaci.2016.09.023

Rahman, M. S., Rahman, M. K., Kaykobad, M., and Rahman, M. S. (2018). isGPT: An Optimized Model to Identify Sub-golgi Protein Types Using SVM and Random Forest Based Feature Selection. *Artif. Intelligence Med.* 84, 90–100. doi:10.1016/j.artmed.2017.11.003

Sanders, W. S., Johnston, C. I., Bridges, S. M., Burgess, S. C., and Willeford, K. O. (2011). Prediction of Cell Penetrating Peptides by Support Vector Machines. *Plos Comput. Biol.* 7 (7), e1002101. doi:10.1371/journal.pcbi.1002101

Schroeder, H. W., and Cavacini, L. (2010). Structure and Function of Immunoglobulins. *J. Allergy Clin. Immunol.* 125 (2), S41–S52. doi:10.1016/j.jaci.2009.09.046

Shao, J., and Liu, B. (2021). ProtFold-DFG: Protein Fold Recognition by Combining Directed Fusion Graph and PageRank Algorithm. *Brief Bioinform* 22 (3), 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 (3), bbaa144. doi:10.1093/bib/bbaa144

Solis, A. D. (2015). Amino Acid Alphabet Reduction Preserves Fold Information Contained in Contact Interactions in Proteins. *Proteins* 83 (12), 2198–2216. doi:10.1002/prot.24936

Song, G., Wang, G., Luo, X., Cheng, Y., Song, Q., Wan, J., et al. (2021). An All-To-All Approach to the Identification of Sequence-specific Readers for Epigenetic DNA Modifications on Cytosine. *Nat. Commun.* 12 (1), 795. doi:10.1038/s41467-021-20950-w

Song, B., Li, F., Liu, Y., and Zeng, X. (2021). Deep Learning Methods for Biomedical Named Entity Recognition: a Survey and Qualitative Comparison. *Brief. Bioinform.* 22 (6), bbab282. doi:10.1093/bib/bbab282

Susko, E., and Roger, A. J. (2007). On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Mol. Biol. Evol.* 24 (9), 2139–2150. doi:10.1093/molbev/msm144

Tang, H., Chen, W., and Lin, H. (2016). Identification of Immunoglobulins Using Chou's Pseudo Amino Acid Composition with Feature Selection Technique. *Mol. Biosyst.* 12 (4), 1269–1275. doi:10.1039/c5mb00883b

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

Tang, Y.-J., Pang, Y.-H., and Liu, B., 2021, DeepIDP-2L: Protein Intrinsically Disordered Region Prediction by Combining Convolutional Attention Network and Hierarchical Attention Network. *Bioinformatics* 2021, btab810. doi:10.1093/bioinformatics/btab810

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., Tang, J., 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. Bioinformatics* 22 (5), bbaa409. doi:10.1093/bib/bbaa409

Wang, J., Liu, X., Shen, S., and Deng, L. (2021). DeepDDS: Deep Graph Neural Network with Attention Mechanism to Predict Synergistic Drug Combinations. *Brief. Bioinform.* bbab390. doi:10.1093/bib/bbab390

Wang, X., Liu, J., Yang, Y., and Wang, G. (2021). The Stacking Strategy-Based Hybrid Framework for Identifying Non-coding RNAs. *Brief Bioinform* 22 (5). doi:10.1093/bib/bbab023

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

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 (17), 11387–11399. doi:10.1007/s00521-020-05573-4

Yu, L., Su, Y., Liu, Y., and Zeng, X. (2021). Review of Unsupervised Pretraining Strategies for Molecules Representation. *Brief. Funct. Genomics* 20 (5), 323–332. doi:10.1093/bfgp/elab036

Zeng, X., Song, X., Ma, T., Pan, X., Zhou, Y., Hou, Y., et al. (2020). Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning. *J. Proteome Res.* 19 (11), 4624–4636. doi:10.1021/acs.jproteome.0c00316

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, J., Zhang, Z., Pu, L., Tang, J., and Guo, F. (2021). AIEpred: An Ensemble Predictive Model of Classifier Chain to Identify Anti-inflammatory Peptides. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18 (5), 1831–1840. doi:10.1109/tcbb.2020.2968419

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-the J. Biol. Databases Curation.* baz131. doi:10.1093/database/baz131

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A Novel Features Ranking Metric with Application to Scalable Visual and Bioinformatics Data Classification. *Neurocomputing* 173, 346–354. doi:10.1016/j.neucom.2014.12.123

Conflict of Interest: Authors JZ and GT were employed by the company Geneis (Beijing) Co., Ltd.

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

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.

Copyright © 2022 Wan, Zhang, Ding, Wang and Tian. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

METHODS
published: 24 January 2022
doi: 10.3389/fgene.2021.827522

Predicting Ischemic Stroke Outcome Using Deep Learning Approaches

Gang Fang, Zhennan Huang and Zhongrui Wang*

Institute of Computing Science and Technology, Guangzhou University, Guangzhou, China

Predicting functional outcomes after an Ischemic Stroke (IS) is highly valuable for patients and desirable for physicians. This facilitates physicians to set reasonable goals for patients and cooperate with patients and relatives effectively, and furthermore to reach common after-stroke care decisions for recovery and make exercise plans to facilitate rehabilitation. The objective of this research is to apply three current Deep Learning (DL) approaches for 6-month IS outcome predictions, using the openly accessible International Stroke Trial (IST) dataset. Furthermore, another objective of this research is to compare these DL approaches with machine learning (ML) for performing in clinical prediction. After comparing various ML methods (Deep Forest, Random Forest, Support Vector Machine, etc.) with current DL frameworks (CNN, LSTM, Resnet), the results show that DL doesn't outperform ML significantly. DL methods and reporting used for analyzing structured medical data should be developed and improved.

Keywords: machine learning, ischemic stroke, deep learning, IST, IS outcome

OPEN ACCESS

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

Reviewed by:
Yansen Su,
Anhui University, China
Cheng Zhang,
Peking University, China

****Correspondence:***
Gang Fang
gangf@gzhu.edu.cn

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

Received: *02 December 2021*
Accepted: *15 December 2021*
Published: *24 January 2022*

Citation:
Fang G, Huang Z and Wang Z (2022) Predicting Ischemic Stroke Outcome Using Deep Learning Approaches. Front. Genet. 12:827522. doi: 10.3389/fgene.2021.827522

INTRODUCTION

Stroke is one of the leading causes of death and permanent disability in the last 20 years globally (Global Burden of Disease Collaborative Network, 2018; World Health Organization, 2018). In China, the number of patients diagnosed with stroke each year is approximately 2 million, and the mortality rate is 11.48% (Chen et al., 2017). Stroke is mainly subtyped into ischemic (85%) and hemorrhagic types (15%) (Caplan, 2016). IS occurs when a cerebral artery is blocked (Park, 2017). Long-term physical disabilities after IS can create enormous mental and financial burdens for families and society. Proper exercise and early rehabilitation definitely improve recovery of patients and reduce disabilities (Veerbeek et al., 2011). Predicting a patient's functional outcomes precisely after a stroke will help physicians in managing an appropriate long-term plan for early rehabilitation. In addition, it guides clinicians in setting realistic goals, provides accurate information to patients and their caregivers, and facilitates the creation of an early discharge plan (Veerbeek et al., 2011). Now, endovascular treatment (EVT) is widely used for IS. Accurate prediction of functional outcomes and reperfusion may potentially improve stroke care, as it can guide selecting the most beneficial treatment option for the individual patient: to perform or to refuse EVT. Recently, clinical variables and radiological image biomarkers are utilized in studies on outcome prediction strategies in ischemic stroke patients after EVT (Venema et al., 2017; Van Os, 2018). More works have been devoted to predicting functional outcomes after stroke (Stinear, 2010; Meyer et al., 2015; Lin et al., 2020). Several medical communities have created and developed scores and methods that can predict the patient's functional outcomes after a stroke effectively by only using data readily collected at admission (Ntaios et al., 2012; Hilbert et al., 2019). The score statistically analyzes the data and identifies the most relevant predictors from a set of covariates selected by domain experts. The method uses deep learning to predict the functional outcome of patients with acute IS after EVT.

Recently, machine learning methods have been ubiquitously used to solve complex problems in many scientific fields, especially in medical science. Medical diagnosis and prognosis prediction are fulfilled in this way (Lin et al., 2018; Van Os, 2018; Debs et al., 2020; Fang et al., 2020).

Recently, DL frameworks have attained great success in various applications, particularly in image processing and natural language processing (NLP) (Hinton et al., 2012; Krizhenvsky et al., 2012), leading to the hot wave of DL (Goodfellow et al., 2016). Though DL frameworks are powerful, they have apparent deficiencies. For example, large scale training data is always required for training, restricting the direct application of DL to tasks with smaller scale data. It is well known that DL is a supervised learning. But nowadays the data of many real tasks are still not sufficiently and correctly labeled due to the high cost of labeling. Because of this, DL frameworks always perform inferiorly in tasks with poor quality data. DL frameworks, especially modern deep neural networks, always possess too many hyper-parameters, and careful tuning of them directly can mainly influence the learning performance of DL. Recently, it is used to diagnose and predict prognosis in the clinical medical field (Ge et al., 2019; Hilbert et al., 2019; Debs et al., 2020). But it is seldomly used to analyze structured clinical medical data. In this paper, currently used DL frameworks are tested to predict stroke outcomes. Furthermore, several ML methods, especially Deep Forest (DF) (Zhou and Feng, 2019), are used to analyze IST dataset and are compared with several DL frameworks. The DF is proposed based on gcForest (multi-Grained Cascade Forest), which is a novel ensemble method of decision tree. This method generates a deep forest ensemble, with a cascade structure which enables gcForest to do representation learning (Zhou and Feng, 2019). Its representational learning ability can be further enhanced by multi-grained scanning when the inputs are with high dimensionality, potentially enabling gcForest to be contextual or structural aware (Zhou and Feng, 2019). In their experiments, the training time cost of DF is smaller than that of DL; even so, DF attains highly superior performance to DL. Herein, the DF and other ML methods are compared with DL to analyze structured clinical medical data. The results show that there are no evidences of superior performance of DL over ML.

MATERIALS AND METHODS

Data

The data used in this paper is The International Stroke Trial (IST) dataset. The IST, including the pilot phase between 1991 and 1993, was conducted between 1991 and 1996 and is a large, prospective, randomized controlled trial, with 100% complete baseline data and over 99% complete follow-up data. The objective of the trial is to know whether early administration of aspirin, heparin, both, or neither influenced the clinical course and outcome of acute IS (Sandercock et al., 2011). The dataset analyzed in this study is downloaded from the IST website. Patients in this trial are identified only by an anonymous code. They were treated more than 20 years ago, and many have died. Hospitals are also identified by an anonymous code. There are no identifying data such as name, address, or social security numbers appearing. Patient age has been rounded to the nearest whole number. Thus, usage of the dataset definitely can't present material risk to confidentiality of patients.

The following baseline data: time from onset to randomization, gender, age, aspirin administration within 3 days prior to randomization, systolic blood pressure at randomization, presence or absence of atrial fibrillation (AF), level of consciousness, and neurological deficit, are all included in the dataset. Neurological deficits are classified as one of the Oxfordshire Community Stroke Project (OCSP) categories: posterior circulation syndrome (POCS), partial anterior circulation syndrome (PACS), total anterior circulation syndrome (TACS), and lacunar syndrome (LACS). A total of 19,435 patients from 467 hospitals in 36 countries are randomized within 48 h of symptoms onset, of whom 13,020 take a CT scan before randomization, 5,569 are first scanned after randomization, and 846 were not scanned at all. Entries with missing data are deleted, with 18,128 entries left. We exclude patients who are not finally diagnosed as IS. The variable of 6-month outcome is taken as a target. It is represented as 1-dead, 2-dependent, 3-not recovered, 4-recovered, and 8 or 9-missing status. The entries of 6-month outcome with missing status are also deleted. Six-month outcome of 2-dependent and 3-not recovered are merged as one category (not recovered) due to their similarity, and then the target includes three categories (0-*dead*, 1-*not recovered*, 2-*recovered*). At last, 16,403 patients are left. The data of these 16,403 patients finally diagnosed as IS are used to predict the outcome of IS using ML and DL.

METHODS

This paper investigates the ability of some supervised ML methods to predict IS outcomes. Classic ML methods such as support vector machine (SVM) (Cristianini and Shawe-Taylor, 2000), random forest (RF) (Liaw and Wiener, 2002), and deep forest (DF) (Zhou and Feng, 2019) are explored for comparison to DL frameworks such as convolutional neural network (CNN) (LeCun et al., 1998), long- and short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997), and residual neural network (Resnet) (He et al., 2016). Developing logistic regression models is the usual approach to analyze the stroke outcomes; however, an alternative of ML methods has been proposed, particularly for large-scale and multi-institutional data. The prominent advantage of ML is that it can easily incorporate newly available data and improve prediction performance (Hamed et al., 2014). Nowadays, DL frameworks are prevalent and succeed in the field of image processing and natural language processing (NLP). In this paper, classical ML methods are compared to popular DL frameworks to exhibit their respective performances.

The workflow of the study consists of three sections. Firstly, features collected at the beginning of and on 14 days of randomization in the refined IST dataset (including 16,403 patients) are used. Features, such as date and comments, are

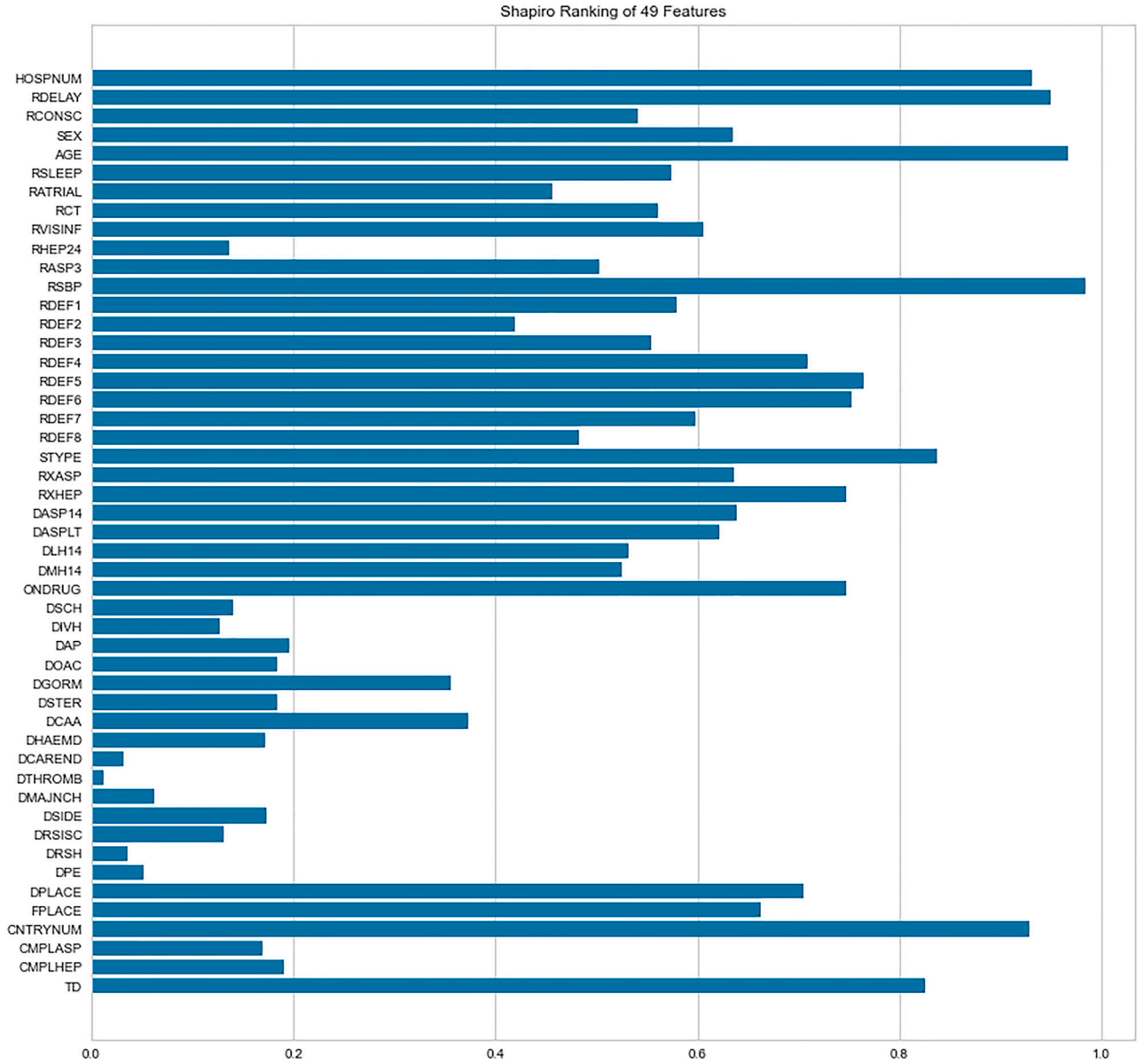


FIGURE 1 | Importance of features ranked by Shapiro-Wilk algorithm.

removed manually (features are definitely not related to IS outcome). Six-month outcome is kept as the target feature in the dataset. Features that overlap with 6-month outcome are deleted manually. Then, 50 features are kept and used. These features are utilized to predict long-term prognosis (6-month outcome) of acute IS. Based on previous research (Fang et al., 2020), feature selection carried out using recursive feature elimination with cross-validation (RFECV) don't eliminate explicitly less important features in the whole IST dataset. Thus, all initially chosen features are used to predict. Secondly, a simple CNN framework which consists of three convolutional layers and two fully connected layers is built, and the first convolutional layer is one dimensional convolution. The used LSTM framework is a two-layer LSTM with just one direction and added into a one-dimensional convolutional layer before it. The last layer of the LSTM framework is a fully connected layer. The Resnet lacking bottleneck blocks which consists of eight residual blocks is also added into a one-dimensional convolutional layer as the first layer. This manipulation allows these DL frameworks to accept and process structured clinical medical data, such as IST. ML methods (SVM, RF, Multinomial-Naïve-Bayes, AdaBoost, and DF) are carried out immediately to compare with these DL frameworks. The SVM classifier use linear kernel (with the parameter max_iter = 10,000), and the other ML methods are carried out with default parameters. To implement these methods for this study, we use the libraries of scikit-learn 1.0.1 (Pedregosa

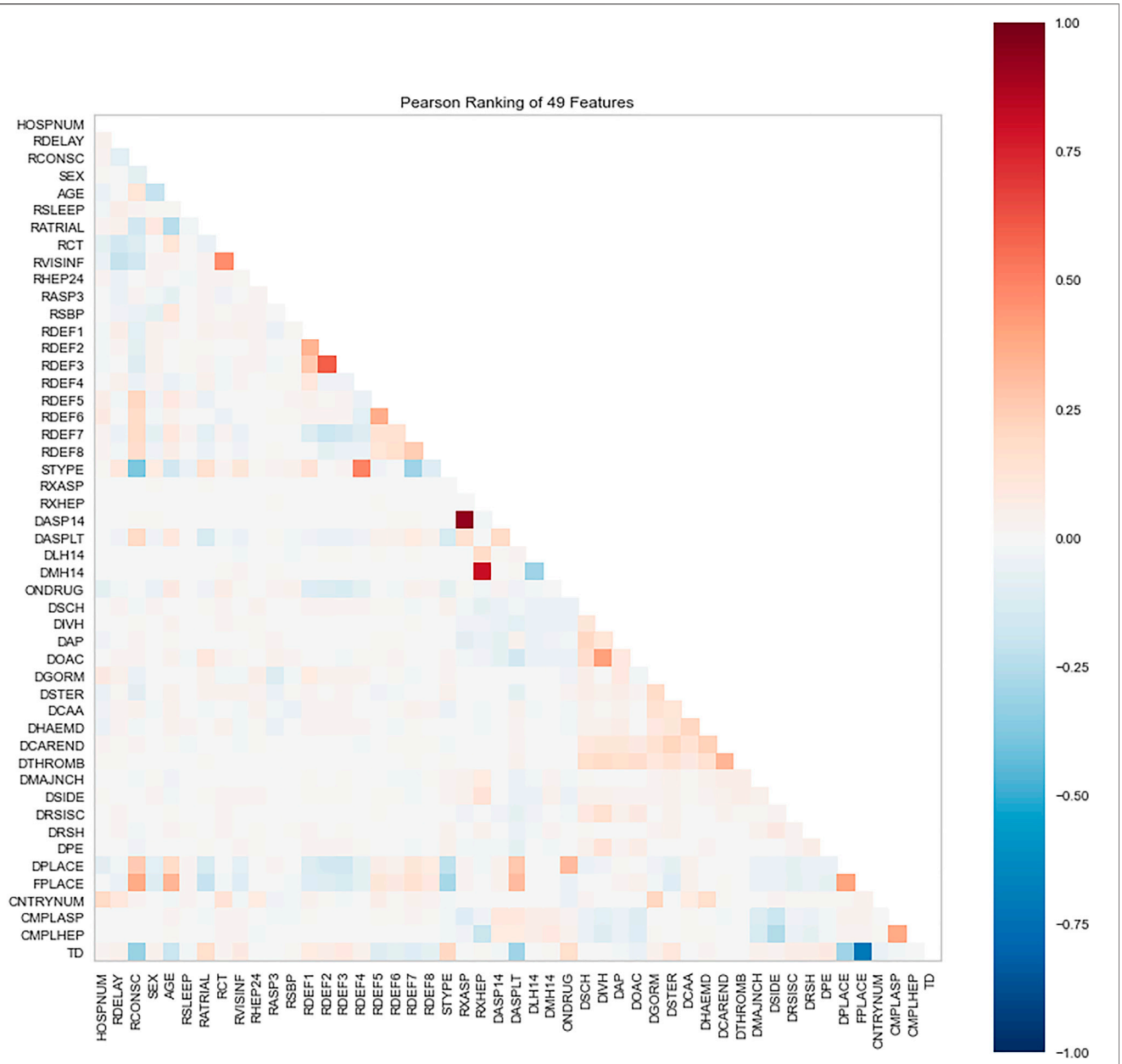


FIGURE 2 | Pearson correlations between features in the dataset.

et al., 2011) and PyTorch neural networks API (PyTorch., 2021). Thirdly, all these methods are implemented for comparison in predicting accuracy and other metrics.

RESULTS AND ANALYSIS

Because we only consider IS, 50 features are initially selected in the data of all kept 16,403 patients. The feature of 6-month outcome (OCCODE) is kept as target (including 3 categories: 0-*dead*, 1-*not recovered*, 2-*recovered*). The other 49 features include CNTRYNUM, HOSPNUM, SEX, AGE, DPLACE, FPLACE, RDELAY, RCONSC, RATRIAL, RSLEEP, RASP3, RSBP, RXASP, RXHEP, DASP14, DASPLT, RCT, RVISINF, DLH14, DMH14, neurological deficit symptom (RDEF1, RDEF2, RDEF3 … … , etc.), STYPE, ONDRUG, DCAA, DOAC, TD, etc. Readers can be referred to **Supplementary Materials** for the detailed explanation of these features. Shapiro-Wilk algorithm is used to rank the importance of these features, and Pearson correlations between features are analyzed too. Shapiro-Wilk algorithm is a normal distribution assessing algorithm that regard the instances with respect to the feature, which is improved by Royston to process large data (Shapiro and Wilk, 1965; Royston, 1982). Except OCCODE, the other 49 features are ranked by the algorithm (**Figures 1**, **2**).

	precision	recall	f1-score	support
0	0.99	1.00	0.99	849
1	0.78	0.99	0.88	2540
2	0.59	0.03	0.06	712
accuracy			0.83	4101
macro avg	0.79	0.67	0.64	4101
weighted avg	0.79	0.83	0.76	4101

FIGURE 3 | Performance of CNN after 5 epochs of training.

	precision	recall	f1-score	support
0	0.98	0.99	0.99	849
1	0.78	0.99	0.87	2540
2	0.54	0.02	0.04	712
accuracy			0.82	4101
macro avg	0.77	0.67	0.63	4101
weighted avg	0.78	0.82	0.75	4101

FIGURE 4 | Performance of LSTM after 5 epochs of training.

	precision	recall	f1-score	support
0	0.99	0.99	0.99	849
1	0.78	1.00	0.87	2540
2	0.25	0.00	0.01	712
accuracy			0.82	4101
macro avg	0.67	0.66	0.62	4101
weighted avg	0.73	0.82	0.75	4101

FIGURE 5 | Performance of Resnet after 5 epochs of training.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	849
1	0.80	0.95	0.87	2540
2	0.49	0.18	0.27	712
accuracy			0.82	4101
macro avg	0.77	0.71	0.71	4101
weighted avg	0.79	0.82	0.79	4101

FIGURE 6 | Performance of DF.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	849
1	0.80	0.96	0.87	2540
2	0.49	0.15	0.23	712
accuracy			0.83	4101
macro avg	0.76	0.70	0.70	4101
weighted avg	0.79	0.83	0.79	4101

FIGURE 7 | Performance of RF.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	849
1	0.79	0.98	0.87	2540
2	0.46	0.05	0.08	712
accuracy			0.82	4101
macro avg	0.74	0.68	0.65	4101
weighted avg	0.77	0.82	0.76	4101

FIGURE 8 | Performance of SVM.

The Shapiro-Wilk results show that DTHROMB (Thrombolysis) and DCAREND (Carotid surgery) are the two least important features. The reason for this is that these therapies were seldom carried out in the 1990s. The Pearson analysis shows that high correlations between features are not common in the dataset. The highest correlated features are DASP14 (Aspirin given for 14 days or till death or discharge) and RXASP (Trial aspirin allocated) which are related to aspirin usage. The second highest correlated features are RXHEP (Trial heparin allocated) and DMH14 (Medium dose heparin given for 14 days or till death/discharge) which are related to heparin usage. After this all these 49 features (except OCCODE) are adopted to predict the outcome of IS using ML and DL.

Firstly, all selected 49 features of the IST dataset which consists of 16,403 patients are processed by DL frameworks. The dataset is divided into training set including 12,302 patients and test set including 4,101 patients randomly. When processed by CNN, 5 epochs of training are carried out and attain an accuracy of 0.826 in test set. Other metrics including precision, recall, and f1-score are also considered (**Figure 3**). When processed by LSTM, 5 epochs of training are also carried out and attain an accuracy of 0.821 in test set. Other metrics are shown in **Figure 4**. When processed by Resnet, 5 epochs of training are carried out and attain an accuracy of 0.821 in test set. Other metrics are shown in **Figure 5**. In this study, all DL frameworks are trained with fewer epochs because more epochs of training lead to overfitting.

After this, test sets including all 4,101 patients are processed by ML approaches. First by DF, it attains an accuracy of 0.824 in test set. Other metrics including precision, recall, and f1-score are also considered (**Figure 6**). The performances of SVM and RF are showed in **Figures 7**, **8**. For performances of other ML methods, readers can be referred to **Supplementary Materials** for more details.

The results show that DL frameworks don't outperform ML methods in any aspects when predicting IS outcomes in IST dataset. On the contrary, ML methods, especially DF, outperform DL in predicting IS outcomes of *recovered*. It attains a higher precision,

	precision	recall	f1-score	support
0	1.00	0.96	0.98	849
1	0.82	0.91	0.86	2540
2	0.49	0.31	0.38	712
accuracy			0.82	4101
macro avg	0.77	0.73	0.74	4101
weighted avg	0.80	0.82	0.80	4101

FIGURE 9 | Performance of CNN with 44 features after 20 epochs of training.

	precision	recall	f1-score	support
0	0.99	1.00	0.99	849
1	0.80	0.96	0.87	2540
2	0.49	0.12	0.20	712
accuracy			0.82	4101
macro avg	0.76	0.69	0.69	4101
weighted avg	0.78	0.82	0.78	4101

FIGURE 10 | Performance of LSTM with 44 features after 20 epochs of training.

recall, and f1-score in predicting the outcomes of *recovered* (represented as 2, **Figure 6**). All methods, especially DL, don't work well in predicting the outcomes of *recovered*. The reason of this lies in the heterogeneity of data in this category. In other words, there are more variables that can exert influence on the recovery of IS patients.

Based on previous Shapiro-Wilk analysis, the less important features whose Shapiro-Wilk ranking value is less than 0.1 are eliminated. These features include DTHROMB (Thrombolysis), DCAREND (Carotid surgery), DRSH (Recurrent stroke within 14 days, Haemorrhagic stroke), DPE (Other events within 14 days, Pulmonary embolism), and DMAJNCH (Major non-cerebral haemorrhage). Then 44 features are left for predicting the outcomes of IS. The predicting performances of DL frameworks are compared to ML methods with these features. When processed by CNN, after 20 epochs of training it attains an accuracy of 0.817 in test set. Other metrics including precision, recall, and f1-score are also considered (**Figure 9**). After 20 epochs of training LSTM attains an accuracy of 0.823 in test set. Other metrics are shown in **Figure 10**. After 20 epochs of training Resnet attains an accuracy of 0.827 in test set. The accuracy doesn't decrease because the eliminated 5 features are less important and not related to the 6th outcome. Other metrics are shown in **Figure 11**.

Subsequently, test sets including all 4,101 patients with 44 features are processed by ML approaches. DF attains an accuracy of 0.828 in the test set. Other metrics including precision, recall, and f1-score are considered (**Figure 12**). The performances of SVM and RF are shown in **Figures 13**, **14**. For performances of other ML methods, readers can be referred to **Supplementary Materials** for more.

	precision	recall	f1-score	support
0	1.00	0.99	1.00	849
1	0.78	0.99	0.88	2540
2	0.62	0.04	0.07	712
accuracy			0.83	4101
macro avg	0.80	0.67	0.65	4101
weighted avg	0.80	0.83	0.76	4101

FIGURE 11 | Performance of Resnet with 44 features after 20 epochs of training.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	849
1	0.81	0.95	0.87	2540
2	0.52	0.19	0.28	712
accuracy			0.83	4101
macro avg	0.77	0.71	0.72	4101
weighted avg	0.80	0.83	0.80	4101

FIGURE 12 | Performance of DF with 44 features.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	849
1	0.80	0.96	0.87	2540
2	0.51	0.15	0.23	712
accuracy			0.83	4101
macro avg	0.77	0.70	0.70	4101
weighted avg	0.79	0.83	0.79	4101

FIGURE 13 | Performance of RF with 44 features.

	precision	recall	f1-score	support
0	1.00	0.96	0.98	849
1	0.77	1.00	0.87	2540
2	0.43	0.01	0.02	712
accuracy			0.82	4101
macro avg	0.73	0.65	0.62	4101
weighted avg	0.76	0.82	0.75	4101

FIGURE 14 | Performance of SVM with 44 features.

The results show that there is no decrease in predicting performance using both ML and DL after eliminating the five least important features. But compared to previous results, after 20 epochs of training Resnet attains a higher precision than before in

	precision	recall	f1-score	support
0	1.00	1.00	1.00	849
1	0.85	0.80	0.83	2540
2	0.42	0.50	0.45	712
accuracy			0.79	4101
macro avg	0.75	0.77	0.76	4101
weighted avg	0.81	0.79	0.80	4101

FIGURE 15 | Performance of Resnet with 44 features after 100 epochs of training.

predicting the outcomes of *recovered* with 44 features (represented as 2, **Figure 11**). Considering this observation, more epochs of training are carried out to attempt to explain this. After 100 epochs of training, the predicting accuracy of Resnet is 0.791 in test set. But it gets higher recall and f1-score than before in predicting the outcomes of *recovered* (**Figure 15**). After 500 epochs of training, the predicting accuracy of Resnet is 0.794 in test set and the other two DL frameworks overfit (accuracy of CNN and LSTM is 0.740 and 0.769 respectively). But there are no increases in recall and f1-score when predicting the outcomes of *recovered* (represented as 2, **Supplementary Figure S1**). After 100 epochs training of Resnet, the overall predicting accuracy decreased. But macro and weighted average f1-score increased (**Figure 15**) and are better than before (**Figures 3–5**, **Figures 9–11**). Macro and weighted average f1-score are an important index for performance of multi-classification tasks. It is suggested that Resnet will work better when trained appropriately, but it doesn't outperform ML methods especially DF significantly in this case (**Figures 12**, **15**). When trained 500 epochs, it starts overfitting (**Supplementary Figure S1**). For more information readers can be referred to **Supplementary Materials**.

DISCUSSION

In this study, classic ML algorithms and current DL frameworks are adopted to predict the outcomes of IS in IST dataset. Both methods attain considerable accuracy. The performances of ML and DL are also compared. The results show that adapted DL frameworks don't outperform ML in predicting capability, although Resnet raised the weighted average f1-score after trained by 100 epochs (**Figure 15**). The main reason of this lies in that the used DL frameworks are developed and employed for processing image and serial data. They are seldom used in censored and structured medical clinical data. In this study, three DL frameworks, CNN, LSTM, and Resnet, are adapted to process this sort of data and predicting the outcomes of IS. The structure of the adapted CNN is similar to LeNet-5 (LeCun et al., 1998) with an added one-dimensional convolutional layer as the first convolutional layer. The used LSTM and Resnet are also added to a one-dimensional convolutional layer as the first layer. In this way, these DL frameworks can admit and process tabulated data, such as structured medical data. CNN attains the accuracy of 0.83 when trained with less epochs, but it gets less f1-score (**Figure 3**). This suggested that it doesn't work well in multiclassification task, so does LSTM (**Figure 4**). After eliminating the 5 least important features and after trained with more epochs (100 epochs), Resnet gets a higher weighted average f1-score (**Figure 15**). The first reason is that the left 44 features are more important to the outcomes of IS. The second reason is that Resnet is a fairly complex DL framework. It adopts residual shortcut connection to overcome degradation problems. When trained appropriately Resnet can capture some intrinsic qualities of the tabulated data and work better in a multiclassification task. In this study, the used Resnet is similar to Resnet18 which possesses fewer layers. Next, deeper Resnet framework and more powerful computing workstations will be adopted to study this issue.

To investigate the predicting capability of DL in the IST dataset, the performances of classic ML algorithms are compared to them. The results show that DL doesn't surpass ML. Resnet raises f1-score after 100 epochs training with the selected 44 features. After eliminating the 5 least important features, the DF and RF raise the f1-score a little and both attain the accuracy of 0.83 (**Figures 12**, **13**). And moderate f1-scores are attained in previous training and test (**Figures 6**, **7**). This means the left 44 features are more important and the 2 ML classifiers are robust to be used in this sort of data. The newly proposed DF is used to be compared with DL frameworks. In our experiments, DF doesn't achieve highly competitive performance to deep neural networks, although the training time cost of DF is smaller than that of deep learning. The reason of this lies in that the used features were collected in the early 1990s. Some important features may be neglected, and this reduces the predicting ability of ML and DL. Next, deeper DL frameworks will be adopted to investigate the performance of them. Furthermore, some new features and variables will be collected to enhance the performance of the machine learning and deep learning approaches.

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

GF wrote the manuscript and designed the study. ZH and ZW wrote some codes.

FUNDING

The study is supported by National Natural Science Foundation of China with grant number 61972107.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.827522/full#supplementary-material

REFERENCES

Caplan, L. R. (2016). *Caplan's Stroke*: Massachusetts. Cambridge University Press.

Chen, Z., Jiang, B., Ru, X., Sun, H., Sun, D., Liu, X., et al. (2017). Mortality of Stroke and its Subtypes in China: Results from a Nationwide Population-Based Survey. *Neuroepidemiology* 48, 95–102. doi:10.1159/000477494

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. London: Cambridge University Press.

Debs, N., Rasti, P., Victor, L., Cho, T.-H., Frindel, C., and Rousseau, D. (2020). Simulated Perfusion MRI Data to Boost Training of Convolutional Neural Networks for Lesion Fate Prediction in Acute Stroke. *Comput. Biol. Med.* 116, 103579. doi:10.1016/j.compbiomed.2019.103579

Fang, G., Liu, W., and Wang, L. (2020). A Machine Learning Approach to Select Features Important to Stroke Prognosis. *Comput. Biol. Chem.* 88, 107316. doi:10.1016/j.compbiolchem.2020.107316

Ge, Y., Wang, Q., Wang, L., Wu, H., Peng, C., Wang, J., et al. (2019). Predicting post-stroke Pneumonia Using Deep Neural Network Approaches. *Int. J. Med. Inform.* 132, 103986. doi:10.1016/j.ijmedinf.2019.103986

Global Burden of Disease Collaborative Network (2018). *Global Burden of Disease Study 2017*. USA: IHME, 1–7.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Hamed, A., Richard, D., Bernard, Y., and Peter, M. (2014). Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-arterial Therapy. *PLOS ONE* 9 (2), e88225. 10.1371/journal.pone.0088225.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition[C]," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 27–30, 2016, 770–778. doi:10.1109/CVPR.2016.90

Hilbert, A., Ramos, L. A., van Os, H. J. A., Olabarriaga, S. D., Tolhuisen, M. L., Wermer, M. J. H., et al. (2019). Data-efficient Deep Learning of Radiological Image Data for Outcome Prediction after Endovascular Treatment of Patients with Acute Ischemic Stroke. *Comput. Biol. Med.* 115, 103516. doi:10.1016/j.compbiomed.2019.103516

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal. Process. Mag.* 29 (6), 82–97. doi:10.1109/msp.2012.2205597

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Krizhenvsky, A., Sutskever, I., and Hinton, G. (2012). "ImageNet Classification with Deep Convolutional Neural Networks," in NIPS, Lake Tahoe, NV, December 3–6, 2012, 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based Learning Applied to Document Recognition. *Proc. IEEE* 86 (11), 2278–2324. doi:10.1109/5.726791

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R. News* 2, 18–22. https://www.researchgate.net/publication/228451484"

Lin, W.-Y., Chen, C.-H., Tseng, Y.-J., Tsai, Y.-T., Chang, C.-Y., Wang, H.-Y., et al. (2018). Predicting post-stroke Activities of Daily Living through a Machine Learning-Based Approach on Initiating Rehabilitation. *Int. J. Med. Inform.* 111, 159–164. doi:10.1016/j.ijmedinf.2018.01.002

Lin, C.-H., Hsu, K.-C., Johnson, K. R., Fann, Y. C., Tsai, C.-H., Sun, Y., et al. (2020). Evaluation of Machine Learning Methods to Stroke Outcome Prediction Using a Nationwide Disease Registry. *Comp. Methods Programs Biomed.* 190, 105381. doi:10.1016/j.cmpb.2020.105381

Meyer, M. J., Pereira, S., McClure, A., Teasell, R., Thind, A., Koval, J., et al. (2015). A Systematic Review of Studies Reporting Multivariable Models to Predict Functional Outcomes after post-stroke Inpatient Rehabilitation. *Disabil. Rehabil.* 37, 1316–1323. doi:10.3109/09638288.2014.963706

Ntaios, G., Faouzi, M., Ferrari, J., Lang, W., Michel, P., and Vemmos, K. (2012). An Integer-Based Score to Predict Functional Outcome in Acute Ischemic Stroke: The ASTRAL Score. *Neurology* 79, 2293–2294. doi:10.1212/wnl.0b013e31827a3c0a

Park, J. (2017). *Acute Ischemic Stroke*. Unitedstates: Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. https://arxiv.org/pdf/1201.0490v4.pdf"

PyTorch (2021). Available at: https://pytorch.org (Accessed Feb 24, 2021)

Royston, J. P. (1982). An Extension of Shapiro and Wilk's W Tests for Normality to Large Samples. *Appl. Statist.* 31 (1), 115–124. doi:10.2307/2347973

Sandercock, P. A., Niewada, M., and Członkowska, A. (2011). The International Stroke Trial Database. *Trials* 12 (12), 101. doi:10.1186/1745-6215-12-101

Shapiro, S. S., and Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52, 591–611. doi:10.1093/biomet/52.3-4.591

Stinear, C. (2010). Prediction of Recovery of Motor Function after Stroke. *Lancet Neurol.* 9, 1228–1232. doi:10.1016/s1474-4422(10)70247-7

Van Os, H. J. A., Ramos, L. A., Hilbert, A., van Leeuwen, M., van Walderveen, M. A. A., Kruyt, N. D., et al. (2018). Predicting Outcome of Endovascular Treatment for Acute Ischemic Stroke: Potential Value of Machine Learning Algorithms. *Front. Neurol.* 9 (9), 1–8. doi:10.3389/fneur.2018.00784

Veerbeek, J. M., Kwakkel, G., van Wegen, E. E. H., Ket, J. C. F., and Heymans, M. W. (2011). Early Prediction of Outcome of Activities of Daily Living after Stroke. *Stroke* 42 (5), 1482–1488. doi:10.1161/strokeaha.110.604090

Venema, E., Mulder, M. J. H. L., Roozenbeek, B., Broderick, J. P., Yeatts, S. D., Khatri, P., et al. (2017). Selection of Patients for Intra-arterial Treatment for Acute Ischaemic Stroke: Development and Validation of a Clinical Decision Tool in Two Randomised Trials. *BMJ* 357, j1710. doi:10.1136/bmj.j1710

World Health Organization (2018). *WHO - The Top 10 Causes of Death*. Geneva: Maggio, 24.

Zhou, Z. H., and Feng, J. (2019). Deep Forest: Towards an Alternative to Deep Neural Networks. *Natl. Sci. Rev.* 6 (1), 74–86. https://arxiv.org/pdf/1702.08835v2.pdf"

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.

Copyright © 2022 Fang, Huang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 January 2022
doi: 10.3389/fgene.2022.825318

MIMRDA: A Method Incorporating the miRNA and mRNA Expression Profiles for Predicting miRNA-Disease Associations to Identify Key miRNAs (microRNAs)

Xianbin Li[1†], *Hannan Ai*[1,2,3†*], *Bizhou Li*[1], *Chaohui Zhang*[1], *Fanmei Meng*[1] *and Yuncan Ai*[1*]

[1]State Key Laboratory for Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China, [2]Department of Electrical and Computer Engineering, The Grainger College of Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [3]National Center for Quality Supervision and Inspection of Automatic Equipment, National Center for Testing and Evaluation of Robots (Guangzhou), CRAT, SINOMACH-IT, Guangzhou, China

OPEN ACCESS

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

Reviewed by:
Xiangxiang Zeng,
Hunan University, China
Xiangrong Liu,
Xiamen University, China

***Correspondence:**
Yuncan Ai
Lssayc@mail.sysu.edu.cn
Hannan Ai
Allenai94@icloud.com

[†]These authors have contributed equally to this work

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

Received: 30 November 2021
Accepted: 10 January 2022
Published: 27 January 2022

Citation:
Li X, Ai H, Li B, Zhang C, Meng F and Ai Y (2022) MIMRDA: A Method Incorporating the miRNA and mRNA Expression Profiles for Predicting miRNA-Disease Associations to Identify Key miRNAs (microRNAs). Front. Genet. 13:825318. doi: 10.3389/fgene.2022.825318

Identifying cancer-related miRNAs (or microRNAs) that precisely target mRNAs is important for diagnosis and treatment of cancer. Creating novel methods to identify candidate miRNAs becomes an imminent Frontier of researches in the field. One major obstacle lies in the integration of the state-of-the-art databases. Here, we introduce a novel method, MIMRDA, which incorporates the **mi**RNA and **m**RNA expression profiles for predicting mi**R**NA-**d**isease **a**ssociations to identify key miRNAs. As a proof-of-principle study, we use the MIMRDA method to analyze TCGA datasets of 20 types (BLCA, BRCA, CESE, CHOL, COAD, ESCA, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PAAD, PRAD, READ, SKCM, STAD, THCA and UCEC) of cancer, which identified hundreds of top-ranked miRNAs. Some (as Category 1) of them are endorsed by public databases including TCGA, miRTarBase, miR2Disease, HMDD, MISIM, ncDR and mTD; others (as Category 2) are supported by literature evidences. miR-21 (representing Category 1) and miR-1258 (representing Category 2) display the excellent characteristics of biomarkers in multi-dimensional assessments focusing on the function similarity analysis, overall survival analysis, and anti-cancer drugs' sensitivity or resistance analysis. We compare the performance of the MIMRDA method over the Limma and SPIA packages, and estimate the accuracy of the MIMRDA method in classifying top-ranked miRNAs via the Random Forest simulation test. Our results indicate the superiority and effectiveness of the MIMRDA method, and recommend some top-ranked key miRNAs be potential biomarkers that warrant experimental validations.

Keywords: MIMRDA prediction method, microRNA (or miRNA), miRNA-disease association, survival analysis, drug resistance, drug sensitivity

INTRODUCTION

Cancer-related microRNAs (miRNAs) targeting mRNAs affect cell differentiation, proliferation, migration and apoptosis, leading to initiation or prevention of cancer (Evan and Vousden 2001; Bartel 2004; Esquela-Kerscher and Slack 2006). Identifying cancer-related miRNAs to be biomarkers roots in the promising diagnosis and treatment of cancer (Rupaimoole and Slack, 2017; Chen et al., 2019; Zhao, Chen, and Yin 2019). Methods and databases have been developed over decades, including but not limited to miRGen (Megraw et al., 2007), miR2Disease (Jiang, et al., 2009), MiRCancer (Xie et al., 2013), HMDD (Li L. et al., 2014), HMDD 3.0 (Huang, et al., 2019), miRWalk (Dweep and Gretz 2015), dbDEMC (Yang et al., 2017), ncDR (Dai et al., 2017), mTD (Chen et al., 2017a), MISIM (Li et al., 2019), miRbase (Kozomara et al., 2019), DBMDA (Zheng et al., 2020) and miRTarBase (Huang et al., 2020). Creating novel methods to identify candidate miRNAs has become an imminent Frontier of researches in the field.

There are two approaches: the complex network-based methods and the machine learning-based methods (Chen, et al., 2019). The former approach relies on the complex network that integrated miRNA similarity network, disease similarity network and known miRNA-disease relationship network to predict miRNA-disease connections (Jiang et al., 2010). This family includes WBSMDA (Chen et al., 2016a), RWRMDA (Chen et al., 2012), HGIMDA (Chen et al., 2016b) and PBMDA (You et al., 2017). These methods constructed local networks of the miRNA and disease similarity to infer global networks; but the prediction with limited information is of poor quality. The hypergeometric distribution or binomial distribution was fundamentally assumed in most methods, similar to that of the Limma package (Ritchie et al., 2015) and the SPIA package (Tarca et al., 2009). The latter approach applies machine learning (supervised or semi-supervised) techniques to predict miRNA-disease connections. Some examples are the SVM classifier (Xu et al., 2011), HDMP (Xuan et al., 2013), RLSMDA (Chen and Yan 2014), RBMMDA (Chen et al., 2015), MCMDA (Li C. et al., 2017) and RKNNMDA (Chen et al., 2017b). These methods performed better in some cases. Yet, the need for fine-tuning parameters inevitably hinders applications in complex biological systems.

Three works pioneered a new direction through incorporating the miRNA and mRNA expression profiles. One was to construct a relationship network between miRNAs and their target mRNAs (disease-genes) by utilizing the limited miRNA and mRNA expression profiles (Xu et al., 2014). Another was to construct a subnetwork between the disease similarity and the miRNA similarity derived from multiple data-sources (Liu et al., 2017). The third was to construct an mRNA-miRNA-lncRNA network prognostic for triple-negative breast cancer (Huang et al., 2021). However, problems remain challenging due to insufficient relationships between miRNAs and mRNAs (disease-genes) in databases.

The major gap in the field is how to integrate sophisticated databases to identify key miRNAs associated with diseases. This article introduces a novel method, MIMRDA, which incorporates the **mi**RNA and **m**RNA expression profiles for predicting mi**R**NA-**d**isease **a**ssociations to identify key miRNAs. As a proof-of-principle study, we use the MIMRDA method to analyze TCGA datasets of 20 types of cancer (comprising 10,449 samples), followed by functional cross-verification through utilizing multiple sophisticated databases including miR2Disease (Jiang, et al., 2009), HMDD 3.0 (Huang, et al., 2019), ncDR (Dai et al., 2017), mTD (Chen et al., 2017a), MISIM 2.0 (Li et al., 2019) and miRTarBase (Huang et al., 2020). We evaluate the superiority of the MIMRDA method to the Limma and SPIA packages (Tarca et al., 2009; Ritchie et al., 2015). We estimate the accuracy of the MIMRDA method in classifying top-ranked miRNAs via the Random Forest simulation test. We discuss some top-ranked key miRNAs with experimental evidences drawn from literature, suggesting their potential to be biomarkers for clinical applications.

MATERIALS AND METHODS

Design and Implementation of the MIMRDA Method

The miRNA-disease association prediction method (MIMRDA) incorporated the expression profiles of both miRNAs and mRNAs to identify key miRNAs. The demo R code was freely available at https://github.com/eshinesimida/MIMRDA. The datasets from TCGA (https://portal.gdc.cancer.gov/) and miRTarBase (Huang et al., 2020) were used as starting-points, followed by multiple steps for predicting and verifying the key miRNAs that were significantly related to at least one type of cancer (**Figure 1**, top-box). Key miRNAs were predicted at the significance level of global probability P_G, for which the Differentially Expressed miRNAs (DE_miRNAs) and their target mRNAs (DE_mRNAs) were essentially measured (**Figure 1**, bottom-box). The sequential procedures were outlined below.

Firstly, we counted the total number of DE_mRNAs (***N***) that were identified from a TCGA dataset by using the Limma package (Ritchie et al., 2015) at the significance level of BH-adj. Pval <0.01. **Secondly**, we estimated the probability P_{miRNA} based on DE_miRNAs in the TCGA dataset by using the Limma package (Ritchie et al., 2015) at the significance level of BH-adj. Pval <0.01. **Thirdly**, we extracted the miRNAs and their target mRNAs, whose associations had been experimentally pre-validated in the miRTarBase database (Chou et al., 2018; Huang et al., 2020), while counting the total number of mRNAs (***M***) and the total number of DE_mRNAs (***k***), as well as the number of DE_mRNAs (***m***) that were precisely targeted by the *i*th miRNA (*i* being the current step in the iteration) from the miRTarBase database. **Fourthly**, we estimated the probability P_{NDE_miRNA} with an over-representation analysis (ORA) via the SPIA package (Tarca et al., 2009), assuming that the number of DE_miRNAs (that precisely targeted DE_mRNAs) followed a hypergeometric distribution with three parameters (***N***, ***M*** and ***k***). These parameters included the total number of DE_mRNAs (***N***) observed in a given TCGA dataset, the total number of mRNAs (***M***) plus the number of DE_mRNAs (***k***) observed in the miRTarBase database, and the number of mRNAs (***m***) that

A. Selection of the DE_miRNAs/DE_mRNAs pairs	**From the starting-point databases (`miRTarBase` and `TCGA`)**
	• `miRTarBase` database : The pre-validated miRNAs/mRNAs pairs
	==> List of miRNAs/mRNAs pairs
	• `TCGA` database : The downloaded datasets of selected types of cancer
(applied the `Limma` package)	==> List of differentially expressed, DE_miRNAs/DE_mRNAs, pairs
	OUTPUT-1: List of the intersected pairs of DE_miRNAs/DE_mRNAs
B. Prediction of key miRNAs	**By calculation of a global probability value, P_G**
(applied the `SPIA` package) (applied the `Limma` package)	$P_{NDE_miRNA}(x > k) = 1 - \sum_{x=0}^{k-1} \frac{\binom{m}{x}\binom{M-m}{N-x}}{\binom{M}{N}}$
	$c = P_{NDE_miRAN} * P_{miRNA}$
	$P_G = c - c\ln(c)$
	OUTPUT-2: The predicted key miRNAs
C. Verification of the predicted key miRNAs	**Against the pre-validated databases of the known types of associations**
	• The miRNA-disease association databases (`miR2Disease` and `HMDD`)
	• The miRNA-drug association databases (`ncDR` and `mTD`)
	• The miRNA function similarity database (`MISIM`)
	• The Kaplan-Meier (KM) survival analysis based on `TCGA` database
	OUTPUT-3: The verified key miRNAs

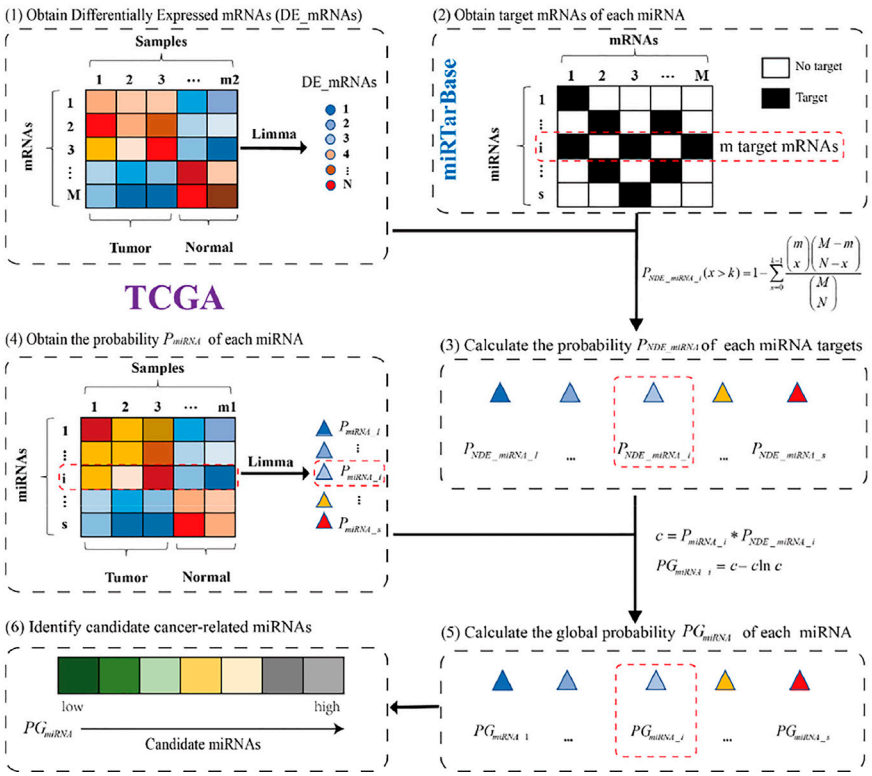


FIGURE 1 | Workflow of the MIMRDA method. Multiple steps for predicting and verifying key miRNAs (top-box). Sequential procedures for calculating a global probability P_G value (bottom-box). The probability P_{miRNA} is estimated by using the Limma package for DE_miRNAs from a TCGA dataset. The probability P_{NDE_miRNA} is estimated with the formula, which incorporates the expression profiles of miRNAs and their target mRNAs from both TCGA dataset and miRTarBase database. The global probability P_G is adjusted by the Fisher's product of P_{NDE_miRNA} and P_{miRNA}. Symbols: Total number of DE_mRNAs (***N***) present in a given TCGA dataset; Total number of mRNAs (***M***) and the number of DE_mRNAs (***k***), as well as the number of mRNAs (***m***) that are precisely targeted by the *i*th miRNA (*i* being the current step in the iteration) present in the miRTarBase database. See the main text for details.

TABLE 1 | Datasets of 20 types of cancer downloaded from TCGA.

Cancer	Total	Tumor	Normal	Details
BLCA	453	416	37	Bladder Urothelial Carcinoma
BRCA	1,282	1,120	162	Breast invasive carcinoma
CESE	317	309	8	Cervical squamous cell carcinoma
CHOL	71	51	20	Cholangiocarcinoma
COAD	570	477	93	Colon adenocarcinoma
ESCA	251	186	65	Esophageal carcinoma
HNSC	612	530	82	Head and Neck squamous cell carcinoma
KICH	190	113	71	Kidney Chromophobe
KIRC	985	543	442	Kidney renal clear cell carcinoma
KIRP	380	292	88	Kidney renal papillary cell carcinoma
LIHC	469	380	89	Liver hepatocellular carcinoma
LUAD	877	603	274	Lung adenocarcinoma
LUSC	765	511	254	Lung squamous cell carcinoma
PAAD	221	185	36	Pancreatic adenocarcinoma
PRAD	623	505	118	Prostate adenocarcinoma
READ	192	173	19	Rectum adenocarcinoma
SKCM	477	474	3	Skin Cutaneous Melanoma
STAD	544	443	101	Stomach adenocarcinoma
THCA	615	515	100	Thyroid carcinoma
UCEC	605	553	52	Uterine Corpus Endometrial Carcinoma

were precisely targeted by the *i*th miRNA (*i* being the current step in the iteration) observed in the miRTarBase database. Statistically, the probability P_{NDE_miRNA} value represented the probability of observing the DE_miRNAs for a given number of times or higher, just by chance. **Finally**, we generated the global probability (P_G) by adjusting the Fisher's product of P_{NDE_miRNA} and P_{miRNA}. The global probability P_G value was used not only to rank DE_miRNAs, but also to choose a desired level of type I error. Small P_G values could occur by chance when multiple testing were simultaneously performed. The FDR-adjusted P_G value was used for controlling the false discovery rate (FDR).

Case Studies: Evaluating the MIMRDA Method

As a proof-of-principle study, we employed the MIMRDA method to analyze TCGA datasets of 20 types of cancer, comprising 10,499 samples (**Table 1**). The miRNAs and mRNAs expression profiles along with clinical information were downloaded at the TCGA data portal (https://portal.gdc.cancer.gov/) (as of April 30, 2020). The Limma package (Ritchie et al., 2015) was deployed to extract differentially expressed mRNAs (DE_mRNAs) and miRNAs (DE_miRNAs), respectively, from each dataset. The Benjamini–Hochberg adjusted *p*-value (BH-adj.*p*-value) < 0.01 was used to select significantly, differentially expressed entities (DE_mRNAs and DE_miRNAs).

Cross-Verification of key miRNAs Against the miRNA-Disease Association Databases (miR2Disease and HMDD)

The miR2Disease database (http://www.miR2Disease.org) was manually curated, containing miRNAs related to human diseases (Jiang, et al., 2009). Each entry contained information about the miRNA-disease association, including miRNA ID, disease name, brief description of the relationship, miRNA expression pattern, miRNA expression detection method, target genes that were experimentally pre-verified in literature. This database currently comprised 3,273 entries, involving 349 miRNAs related to 163 human diseases (as of April 30, 2021). The HMDD 3.0 database (Huang, et al., 2019) currently contained 5,430 types of relationship between 495 miRNAs and 383 diseases (as of April 30, 2021), which was employed to infer the miRNA-disease associations. The miRNA-disease pairs were downloaded (as of April 30, 2021) at http://www.cuilab.cn/hmdd for analysis.

Cross-Verification of key miRNAs Against the Function Similarity Database (MISIM)

The MISIM 2.0 database (http://www.lirmed.com/misim/) (Li et al., 2019) integrated the co-expression similarity, GO function similarity and disease similarity. It was applied to manifest the functional similarity of miRNAs as a tool for the miRNA function analysis (Wang et al., 2010). We deployed the known miRNA-disease interactions to evaluate the functional similarity of miRNAs because miRNAs with similar functions should tentatively associate with similar diseases (Chen D. et al., 2018; Che et al., 2019; Zheng et al., 2020).

Cross-Verification of key miRNAs via the Kaplan-Meier (KM) Survival Analysis Based on TCGA Database

The Kaplan-Meier (KM) method (Saluja et al., 2019) was used to evaluate the prognostic survival rate of key miRNAs. The median values of miRNAs expression were calculated. miRNAs with expression values higher than the median value were considered to be highly expressed, and vice versa. The TCGA

database (with clinical information of patients) was employed to screen the significantly, differentially expressed miRNAs (DE_miRNAs) and determine whether such miRNAs were related to the overall survival (OS). The hazard ratio (HR) and p-value were estimated to evaluate the direct relationship between miRNA and prognostic survival. A p-value < 0.05 was considered statistically significant.

Cross-Verification of key miRNAs Against the miRNA-Drug Association Databases (ncDR and mTD)

An miRNA targeting mRNAs caused sensitivity or resistance to anti-cancer drugs. We applied top-20 ranked miRNAs to search against two databases, ncDR (Dai et al., 2017) and mTD (Chen et al., 2017b), looking for candidate matches, thus predicted possible resistance or sensitivity to anti-cancer drugs. These two databases currently contained 5,661 and 3,669 miRNAs-drugs interactions for all diseases (as of October 2021), respectively, which provided information about the dysfunctions of non-coding RNAs (ncRNAs), leading to resistance or sensitivity to anti-cancer drugs.

Comparison on the Performance of the MIMRDA Method Over Existing Methods

No similar methods was available for side-by-side comparisons. We compared the number distribution of top-ranked miRNAs identified by the MIMRDA method (P_G), the Limma package (P_{miRNA}) and the SPIA package (P_{NDE_miRNA}), respectively, at the significance level of adj. Pval <0.01 since the MIMRDA method rooted in the usage of the Limma package (Ritchie et al., 2015) and the SPIA package (Tarca et al., 2009) (see **Figure 1**). For simplicity, we focused on comparing the number distribution of top-100 ranked miRNAs obtained by these three methods from each dataset of each type of cancer. The more the identified disease-related miRNAs were flagged, the better the method performed.

Evaluating the Performance of MIMRDA via the Random Forest Simulation Test

To evaluate the accuracy of the MIMRDA method in classifying top-ranked miRNAs, we employed a machine learning method, i.e., the five-fold cross-validation Random Forest (RF), for simulation test (Speiser et al., 2019). Samples of each dataset from each type of cancer were divided (at a ratio of 4:1) into the training and testing sets, respectively. The five-fold cross-validation RF simulation generated a predicted value. We obtained an AUC value by comparing the predicted value with an actual value, and thus compared the MIMRDA method top-ranked (top_5, top_10, top_15, top_20) miRNAs with the randomly selected (random_5, random_10, random_15, random_20) miRNAs, both after the RF simulations. These processes were repeated 1,000 times in order to get a set of AUC values. We then used the AUC-based statistics analysis to evaluate the accuracy of the MIMRDA method in classifying the top-ranked miRNAs. The larger the AUC value was, the better the accuracy of the method classified. The difference was considered statistically significant at p-value < 0.001.

RESULTS

Identification of miRNAs and Their Target mRNAs

The miRNAs and their target mRNAs were extracted from the miRTarBase database (Huang et al., 2020) with the experimentally pre-validated miRNA-target associations. The number distribution of miRNAs and mRNAs, respectively, indicates that the majority of miRNAs have 200–300 target mRNAs (**Figure 2A**), while the majority of target mRNAs have 20–50 miRNAs (**Figure 2B**); Top-10 ranked miRNAs have more than 1,000 target mRNAs (**Figure 2C**), while top-10 ranked target mRNAs have more than 250 miRNAs (**Figure 2D**). These data suggest that such diverse samples are appropriate for subsequent analysis.

Identification of the Differentially Expressed miRNAs and mRNAs

We screened the differentially expressed miRNAs (DE_miRNAs) and target mRNAs (DE_mRNAs) from each dataset by using the Limma package (Ritchie et al., 2015) at the significance level of BH-adj. Pval <0.01. The percentage distribution of top-ranked (top-10, 20, 30, 40, 50) miRNAs indicates that most miRNAs are significantly essential in biology (**Figure 3**). Note that the percentage of top-ranked miRNAs is a proportion of the top-ranked miRNAs out of the total cancer-related miRNAs that were identified from the given datasets of a cancer type. For instance, surveyed against the HMDD database, we obtained the top-10 ranked miRNAs from the BLCA datasets, of which only nine miRNAs were identified to be truly associated with BLCA, thus yielding a percentage of 90%. The percentage distribution of such top-50 ranked miRNAs suggests an accuracy greater than 70% in BLCA, BRCA, LIHC, LUAD, LUSC, PRAD and STAD datasets, and an accuracy less than 40% in CHOL, KICH, KIRP, PAAD, SKCM and THCA datasets. Similar surveys with the top-10 ranked miRNAs suggest an accuracy greater than 60% in the majority of datasets. These data indicate the effectiveness of the MIMRDA method in identifying key miRNAs that were significantly, differentially expressed in the datasets from 20 types of known cancer, suggesting that they are closely related to the 20 types of known cancer (see **Table 1**).

The Impacts of key miRNAs on Multiple Types of Cancer

We extracted top-20 ranked miRNAs from each dataset and searched them against the miRNA-disease association databases (miR2Disease and HMDD) whose biological functions had been

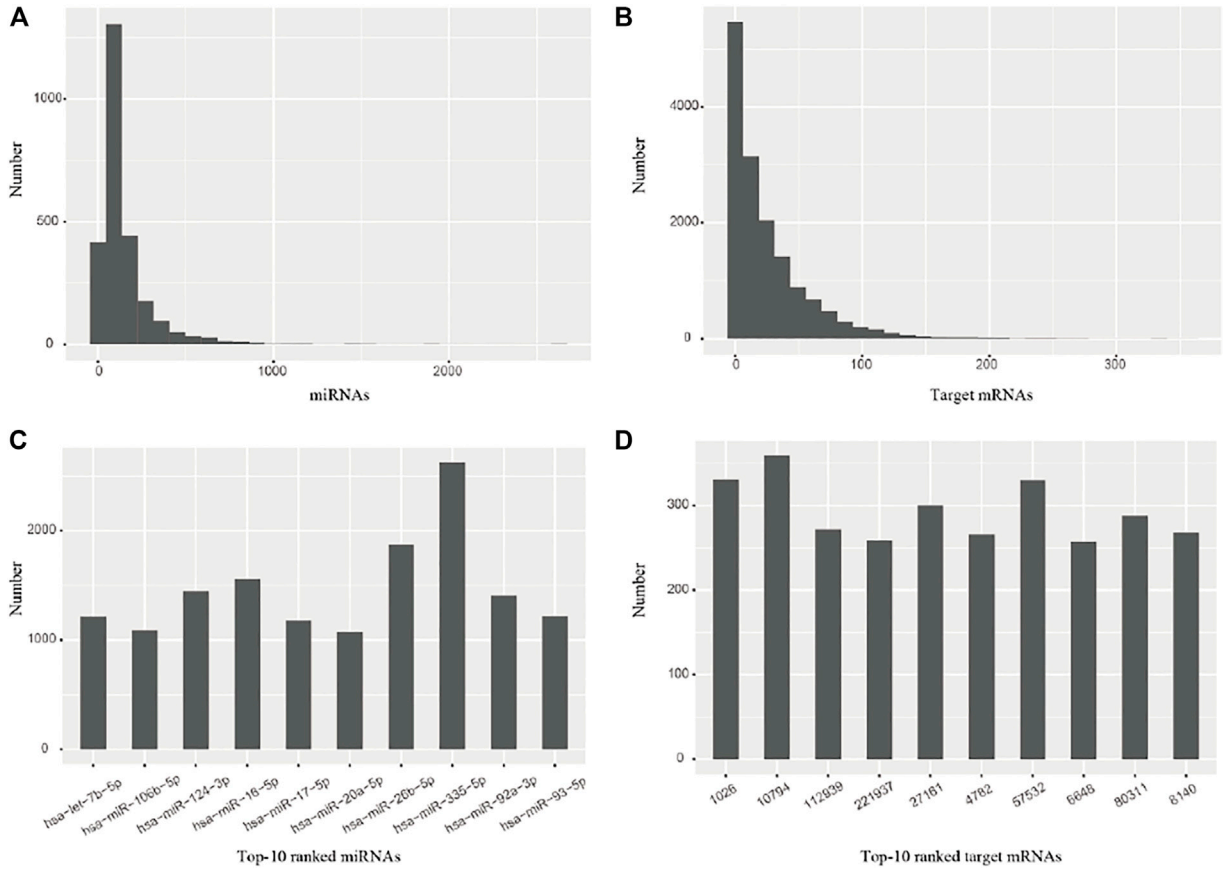


FIGURE 2 | The number distribution of miRNAs and their target mRNAs. **(A)** miRNAs **(B)** Target mRNAs **(C)** Top-10 ranked miRNAs. **(D)** Top-10 ranked target mRNAs.

pre-verified clinically or experimentally. The results (**Figure 4**) indicated that more than 50% of the top-20 ranked miRNAs were related to 14 types of cancer (BLCA, BRCA, CESC, COAD, ESCA, HNSC, LIHC, LUAD, LUSC, PRAD, STAD, THCA and UCEC), despite that certain top-20 ranked miRNAs were not related to any cancer type at all. We identified perfect matches (defined as Category 1), including 1) 18 miRNAs were from BRCA, LIHC, LUAD, LUSC and STAD; 2) 17, 16, 15, 14, 13, 12, 11, 11, 10 miRNAs separately were from BLCA, PRAD, UCEC, CESC, COAD, THCA, ESCA, HNSC, READ; and 3) less than 10 miRNAs were from CHOL, KICH, KIRC, KIRP, PAAD and SKCM. Strikingly, the MIMRDA method suggested that certain top-20 ranked miRNAs (e.g., **miR-1258** and **miR-4686**) were related to cancer, but they were beyond (i.e., they were not matched with) the current version of miR2Disease and HMDD databases. We defined these candidate miRNAs as Category 2, which warrant validations in future experiments.

Among the 198 miRNAs out of the top-20 ranked miRNAs screened from the 20 types of cancer, 85 miRNAs were related to multiple types of cancer whereas the rest 113 miRNAs were related to one cancer type (**Figure 4**). Those key miRNAs related to multiple types of cancer will be discussed (in Discussion) later with accumulated experimental evidences drawn from literature. Here, we highlight certain cases that were related to single type of cancer. 1) Four (miR-148b, miR-185, miR-671 and miR-18a) were related to BLCA, and ranked 5th, 7th, 9th and 14th, respectively. 2) Five (miR-145, miR-125b01, miR-99a, miR-6507 and miR-100) were related to BRCA, and ranked 8th, 15th, 16th, 17th and 20th, respectively. 3) One (miR-215) was related to CESC, and ranked 6th. 4) Two (miR-218-1 and miR-218-2) ranked 15th and 16th were related to CHOL. 5) Eight (miR-74a, miR-6803, miR-6887, miR-6749, miR-542, miR-125a, miR-6756 and miR-197) were related to COAD, and ranked 6th, 11th, 12th, 15th, 16th, 17th, 18th and 19th, respectively. 6) Three (miR-30c-2, miR-30c-1 and miR-877) were related to ESCA, and ranked 18th, 19th and 20th, respectively. 7) Four (miR-5089, miR-4510, miR-503 and miR-195) were related to HNSC, and ranked 3rd, 8th, 14th and 15th, respectively. 8) Seven (miR-135b, miR-874, miR-130a, miR-124-2, miR-124-3, miR-3065 and miR-22) were related to KICH, and ranked 4th, 6th, 8th, 12th, 17th, 18th and 19th, respectively. 9) Five (miR-2355, miR-584, miR-362, miR-629 and miR-20) were related to KIRC, and ranked 11th, 13th, 14th, 17th and 20th, respectively. 10) Seven (miR-216b, miR-4508, miR-891a, miR-489, miR-124-1, miR-377 and miR-6863) were related to KIRP, and ranked 6th, 7th, 8th, 10th, 13th, 16th and 19th, respectively. 11) Two (miR-4686 and let-7c)

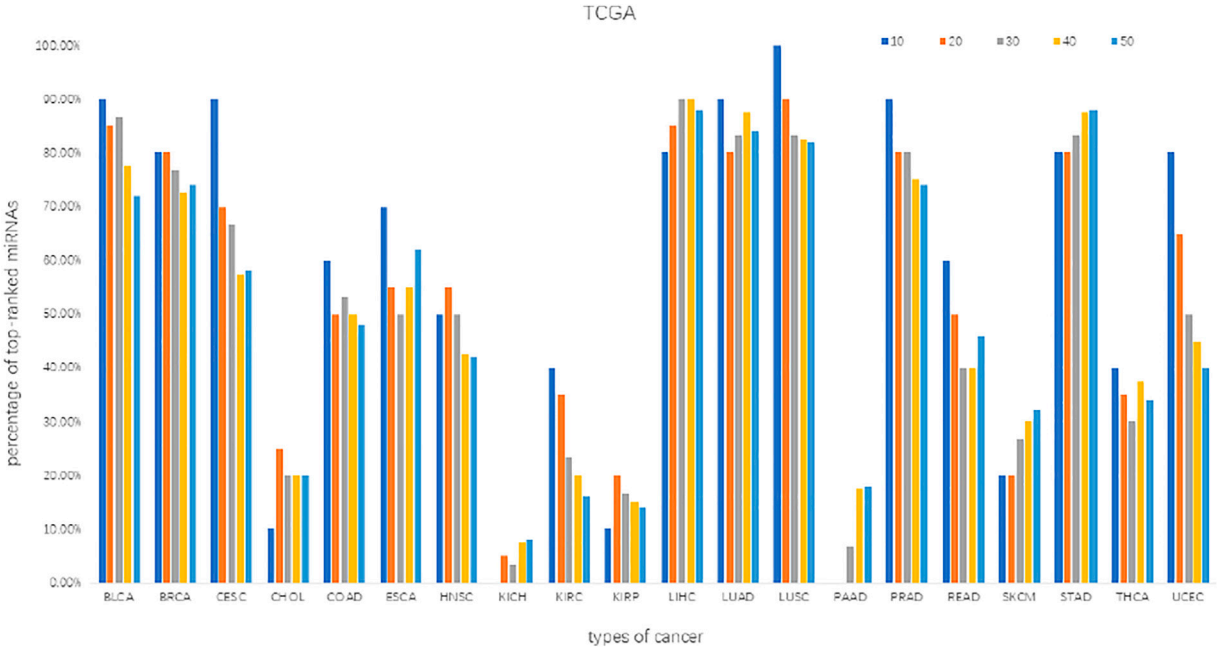


FIGURE 3 | The percentage distribution of top-ranked miRNAs screened from the datasets of 20 types of cancer.

were related to LIHC, and ranked 3rd and 9th, respectively. (**xii**) Six (miR-7-1, let-7a-2, let-7a-1, let-7a-3, miR-4529 and miR-310a) were related to LUAD, and ranked 8th, 12th, 14th, 15th, 18th and 19th, respectively. (**xiii**) Three (miR-205, miR-30d and miR-944) were related to LUSC, and ranked 2nd, 3rd and 11th, respectively. (**xiv**) Nine (miR-6788, miR-5196, miR-574, let-7d, miR-346, miR-6726, miR-6849, miR-1224 and miR-766) were related to READ, and ranked 2nd, 6th, 7th, 9th, 13th, 15th, 16th, 18th and 19th. (**xv**) One (miR-98) was related to STAD, and ranked 13th. (**xvi**) Remarkably, no miRNAs was related to UCEC at all. Taken together, these data suggest that the MIMRDA method is effective in identifying key miRNAs from specific type of cancer.

Verification of key miRNAs via the Biological Function Similarity Analysis

We applied MISIM 2.0 database to annotate the top-20 ranked miRNAs from each dataset of the 20 types of cancer (**Figure 5**). The findings revealed that the majority of top-20 ranked miRNAs were annotated, including 19 in CHOL and STAD; 18 in CESC, ESCA, KIRC, LUSC and PRAD; 17 in BLCA, KICH, THCA and UCEC; 16 in KIRP, LIHC and LUAD; 15 in BRCA and HNSC; and 14 in COAD and READ. However, none of the top-20 ranked miRNAs was annotated in PAAD and SKCM. Meanwhile, the function similarity network of the top-20 ranked miRNAs indicated that the majority of miRNAs were highly related to one another in biological functions, as the red line represents that the correlation coefficient is greater than 0.5 (**Figure 5**). For instance, the top-10 ranked miRNAs are corresponding to the enriched biological functions (FDR <0.05), which are mainly involved in cell cycle, proliferation, inflammation, death and apoptosis (**Figure 5**). And these functions have been experimentally pre-verified to be closely associated with various types of cancer (Evan and Vousden 2001; Taniguchi and Karin 2018). These results suggest that such key miRNAs possess highly coupled linkages, which drive the essential biological functions at the system-level, thereby enhancing their potential of clinical applications.

Verification of key miRNAs via the Kaplan-Meier (KM) Survival Analysis

The top-3 ranked miRNAs demonstrated drastic variations on the survival of patients (**Figure 6**), which impacted the prognostic survival of patients in BLCA, BRCA, CESC, ESCA, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PAAD, READ, STAD, THCA and UCEC. Two categories have strongly functioned in a positive or negative manner, respectively. 1) with strong POSITIVE impacts: **miR-21** (HR = 0.43, log_rank p = 0.0063 in KIRP; HR = 0.62, log_rank p = 0.0048 in BLCA); **miR-92a** (HR = 0.58, log_rank p = 2e-04 in BLCA); **miR-148b** (HR = 0.63, log_rank p = 0.0043 in BLCA); **miR-182** (HR = 0.51, log_rank p = 0.0021 in UCEC); **miR-206** (HR = 0.47, log_rank p = 1.4e-06 in KICH); **miR-490** (HR = 0.34, log_rank p = 3.6e-10 in LIHC); **miR-934** (HR = 0.37, log_rank p = 2.1e-11 in KIRC); **miR-1258** (HR = 0.44, log_rank p = 2.6e-06 in LIHC); **miR-4686** (HR = 0.35, log_rank p = 7.8e-10 in LIHC); and **miR-4709** (HR = 0.24, log_rank p = 0.0026 in THCA). 2) with strong NEGATIVE impacts: **miR-21** (HR =

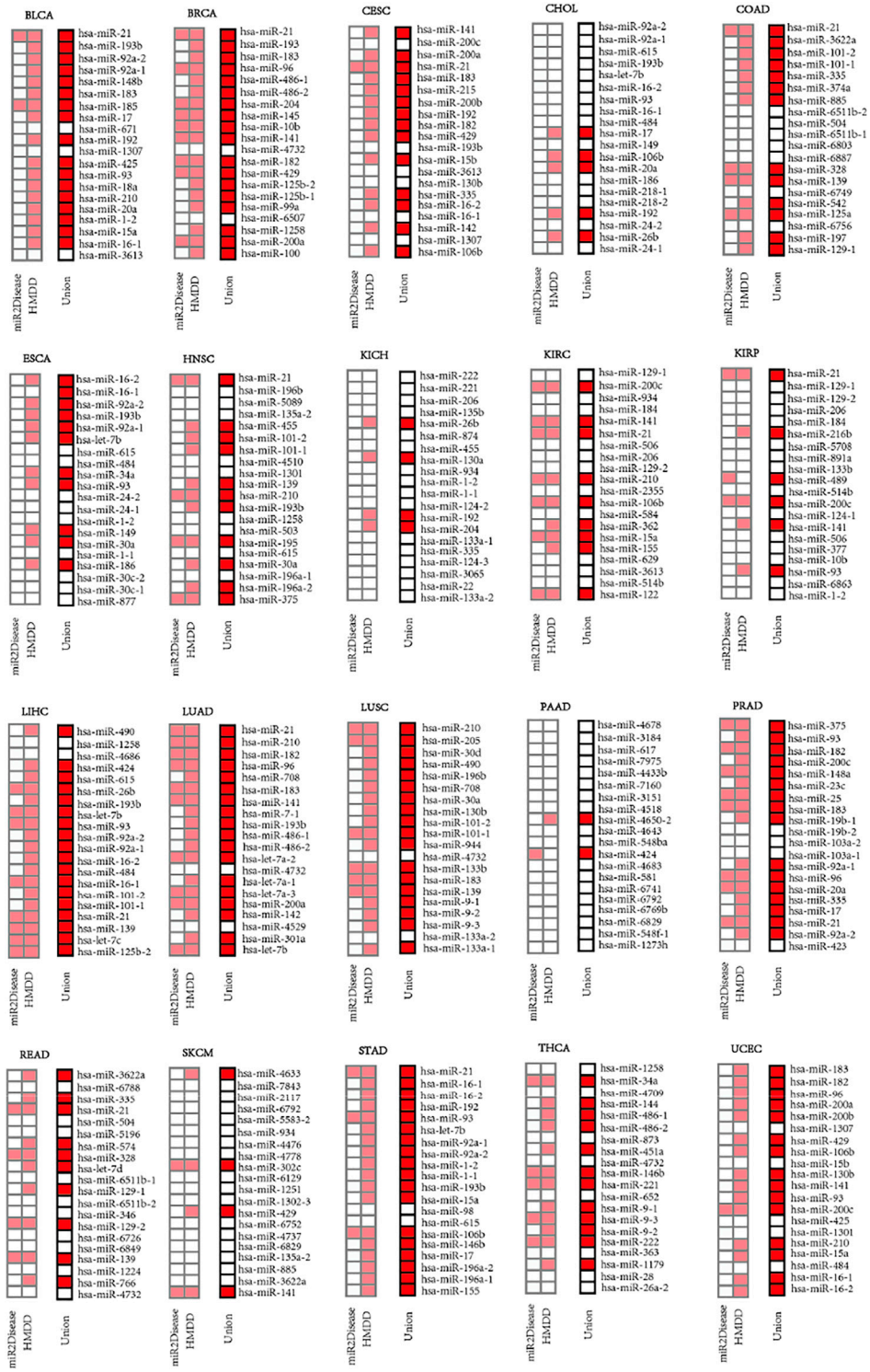


FIGURE 4 | Top-20 ranked miRNAs on the lists of priorities (squares in light red or red color) for 20 types of cancer when searched against the miR2Disease and HMDD databases, respectively.

1.63, log_rank p = 0.004 in BRCA; HR = 1.59, log_rank p = 0.0028 in LUAD); **miR-92a** (HR = 2.65, log_rank p = 0071 in ESCA); **miR-139** (HR = 1.80, log_rank p = 0.0021 in BRCA); **miR-200c** (HR = 1.66, log_rank p = 0.0066 in KIRC); **miR-221** (HR = 2.32, log_rank p = 3e-08 in KICH); **miR-222** (HR = 2.09, log_rank p = 1.4e-06 in KICH); **miR-617** (HR = 2.27, log_rank



BLCA			BRCA			CESC		
Term	Count	FDR	Term	Count	FDR	Term	Count	FDR
Immune Response	12	1.58E-08	Inflammation	8	1.87E-04	Regulation of Stem Cell	11	3.33E-08
Angiogenesis	10	1.07E-07	Regulation of Stem Cell	8	1.98E-05	Cell Cycle	8	1.13E-04
Apoptosis	10	4.36E-06	Cell Cycle	7	2.55E-04	Cell Proliferation	8	8.95E-05
Cell Cycle	10	7.16E-07	Cell Death	7	1.92E-04	EMT	8	1.13E-04
Cell Death	9	4.47E-06	Aging	6	5.54E-04	Inflammation	6	6.76E-04
Cell Proliferation	8	5.66E-05	Cell Proliferation	6	1.82E-03	Apoptosis	6	3.97E-03
HSP	8	7.30E-06	Cell Differentiation	6	3.19E-04	Cell Death	6	3.19E-04
Inflammation	8	4.78E-04	EMT	6	2.15E-03	Immune Response	6	8.08E-03
Onco-miRNAs	8	5.92E-07	Osteoblast Differentiation	6	4.77E-06	Nephrotoxicity	6	1.57E-06
Aging	7	1.34E-04	Apoptosis	5	3.18E-02	TSM	6	1.67E-03

HSP: Hormone-mediated Signaling Pathway, EMT: Epithelial-to-Mesenchymal Transition, TSM: Tumor Suppressor miRNAs



CHOL			COAD			ESCA		
Term	Count	FDR	Term	Count	FDR	Term	Count	FDR
Apoptosis	12	6.40E-08	Inflammation	6	3.47E-02	Cell Death	13	4.87E-10
Angiogenesis	9	2.03E-06	Apoptosis	5	0.1328	Cell Proliferation	12	9.97E-09
Cell Cycle	9	1.08E-05	EMT	4	0.2363	Apoptosis	11	1.34E-06
Cell Death	9	7.32E-06	Immune Response	4	0.3035	Cell Cycle	11	1.33E-07
Cell Proliferation	9	8.92E-06	Cell Death	3	0.4514	Angiogenesis	10	2.08E-07
Inflammation	9	9.32E-05	HSP	3	0.3062	HSP	10	8.15E-08
HSP	8	9.96E-06	TSM	3	0.3459	Bone Regeneration	9	1.22E-08
Immune Response	8	1.83E-04	T-Cell Differentiation	2	0.2531	Immune Response	8	3.13E-04
Onco-miRNAs	8	8.19E-07	DNA Damage Repair	2	0.31926	Inflammation	8	1.14E-03
Regulation of Stem Cell	8	7.27E-05	Lipid Metabolism	2	0.5549	Aging	8	2.63E-05

HSP: Hormone-mediated Signaling Pathway, EMT: Epithelial-to-Mesenchymal Transition, TSM: Tumor Suppressor miRNAs

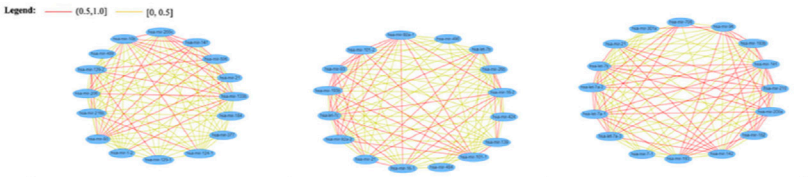


HNSC			KICH			KIRC		
Term	Count	FDR	Term	Count	FDR	Term	Count	FDR
Apoptosis	8	4.43E-04	Inflammation	14	5.02E-10	EMT	9	8.00E-06
Immune Response	8	2.07E-04	Cell Proliferation	10	1.05E-06	Cell Cycle	8	6.06E-05
Inflammation	7	3.80E-03	HSP	10	6.58E-08	Cell Death	7	3.63E-04
Cell Cycle	6	5.59E-03	Regulation of Stem Cell	9	8.48E-06	Inflammation	7	2.32E-03
Onco-miRNAs	5	1.67E-03	Muscle Development	9	5.37E-10	Regulation of Stem Cell	7	3.83E-04
Cell Proliferation	4	1.15E-01	Cell Cycle	8	1.09E-04	Aging	6	1.03E-03
EMT	4	1.28E-01	Apoptosis	7	3.13E-03	Angiogenesis	6	1.15E-03
HSP	4	4.71E-02	Brain Development	7	7.35E-06	Cell Proliferation	6	2.69E-03
Lipid Metabolism	4	2.37E-02	Cell Differentiation	7	9.81E-05	Apoptosis	5	3.62E-02
Aging	3	2.21E-01	Osteogenesis	7	1.22E-04	Cell Differentiation	5	4.16E-03

HSP: Hormone-mediated Signaling Pathway, EMT: Epithelial-to-Mesenchymal Transition, TSM: Tumor Suppressor miRNAs

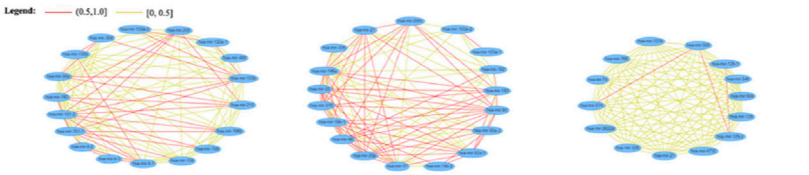
FIGURE 5 | Biological function similarity analysis of the top-20 ranked miRNAs.

p = 0.0018 in PADD); **miR-3184** (HR = 2.27, log_rank p = 0.0018 in PADD); **miR-3622a** (HR = 1.82, log_rank p = 0.13 in READ); **miR-4678** (HR = 2.27, log_rank p = 0.0018 in PADD); and **miR-6788** (HR = 1.70, log_rank p = 0.18 in READ). Remarkably, these key miRNAs have been pre-verified by clinical information of patients in the TCGA database and



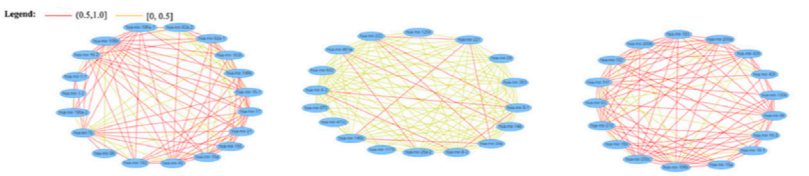
KIRP			LIHC			LUAD		
Term	Count	FDR	Term	Count	FDR	Term	Count	FDR
Cell Death	7	9.52E-04	Apoptosis	10	1.22E-06	Cell Cycle	9	4.29E-06
Cell Proliferation	7	1.02E-03	Cell Proliferation	8	1.80E-05	Cell Death	8	3.21E-05
EMT	7	1.20E-03	Immune Response	8	4.27E-05	Cell Differentiation	8	3.83E-06
Inflammation	7	4.11E-03	Aging	7	4.81E-05	EMT	8	4.60E-05
Angiogenesis	6	2.38E-03	Angiogenesis	7	5.56E-05	Inflammation	8	2.49E-04
Cell Cycle	6	5.59E-03	Cell Death	7	1.41E-04	Aging	7	7.83E-05
Regulation of Stem Cell	6	4.55E-03	HSP	7	3.34E-05	Cell Proliferation	6	2.25E-03
HSP	5	8.21E-03	TSM	7	5.56E-05	Regulation of Stem Cell	6	2.12E-03
Muscle Development	4	2.80E-03	Cell Cycle	6	1.97E-03	TSM	6	8.20E-04
Brain Development	4	1.25E-02	T-Cell Differentiation	6	1.83E-06	Apoptosis	5	4.59E-02

HSP: Hormone-mediated Signaling Pathway, EMT: Epithelial-to-Mesenchymal Transition, TSM: Tumor Suppressor miRNAs



LUSC			PRAD			READ		
Term	Count	FDR	Term	Count	FDR	Term	Count	FDR
Inflammation	12	2.08E-07	Immune Response	13	2.33E-09	Inflammation	5	0.2377
Apoptosis	10	2.05E-05	Cell Cycle	12	7.90E-09	EMT	4	0.3574
HSP	9	1.49E-06	Cell Death	10	8.98E-07	Cell Death	4	0.3181
Immune Response	8	3.87E-04	Apoptosis	10	1.42E-05	Cell Proliferation	3	0.6061
Regulation of Stem Cell	8	1.75E-04	Inflammation	9	1.69E-04	Glucose Metabolism	3	0.1872
Angiogenesis	6	2.99E-03	Angiogenesis	8	3.37E-05	T-Cell Differentiation	3	0.0765
Aging	5	1.72E-02	Onco-miRNAs	8	8.87E-07	Aging	2	0.8622
Cell Motility	5	1.97E-04	Regulation of Stem Cell	8	1.18E-04	Angiogenesis	2	0.8985
Cell Differentiation	5	1.11E-02	Adipocyte Differentiation	7	2.10E-05	Cranulopoiesis	2	0.214
Cell Death	5	3.64E-02	Aging	7	2.55E-04	Innate Immunity	2	0.681

HSP: Hormone-mediated Signaling Pathway, EMT: Epithelial-to-Mesenchymal Transition, TSM: Tumor Suppressor miRNAs



STAD			THCA			UCEC		
Term	Count	FDR	Term	Count	FDR	Term	Count	FDR
Cell Cycle	11	4.09E-08	Aging	10	1.05E-07	Cell Cycle	10	8.54E-07
Cell Proliferation	11	3.58E-08	Apoptosis	9	6.50E-05	Cell Proliferation	10	6.68E-07
Immune Response	11	9.55E-08	Inflammation	8	5.40E-04	Regulation of Stem Cell	10	6.16E-07
Apoptosis	10	3.90E-06	Regulation of Stem Cell	8	8.88E-05	Cell Differentiation	8	8.17E-06
Cell Death	10	3.79E-07	Cell Proliferation	7	5.81E-04	Apoptosis	7	3.18E-03
HSP	10	3.21E-08	Cell Cycle	6	4.09E-03	EMT	7	8.46E-04
Aging	9	1.02E-06	HSP	6	8.78E-04	Immune Response	7	1.51E-03
Angiogenesis	9	1.22E-06	Brain Development	5	1.05E-03	TSM	7	2.31E-04
Inflammation	9	5.37E-05	Cell Differentiation	5	5.57E-03	Cell Death	6	4.16E-03
Onco-miRNAs	7	6.81E-06	Hematopoiesis	5	5.99E-03	Inflammation	6	2.18E-02

HSP: Hormone-mediated Signaling Pathway, EMT: Epithelial-to-Mesenchymal Transition, TSM: Tumor Suppressor miRNAs

FIGURE 5 | (Continued).

the miRNA-disease association databases (miR2Disease and HMDD); some of them are in line with the accumulated evidences drawn from literature as discussed (in Discussion) later, which enhance their potential of clinical applications. To our knowledge, most of them are uncovered for the first time, thus deserving to be exploited through future experiments.

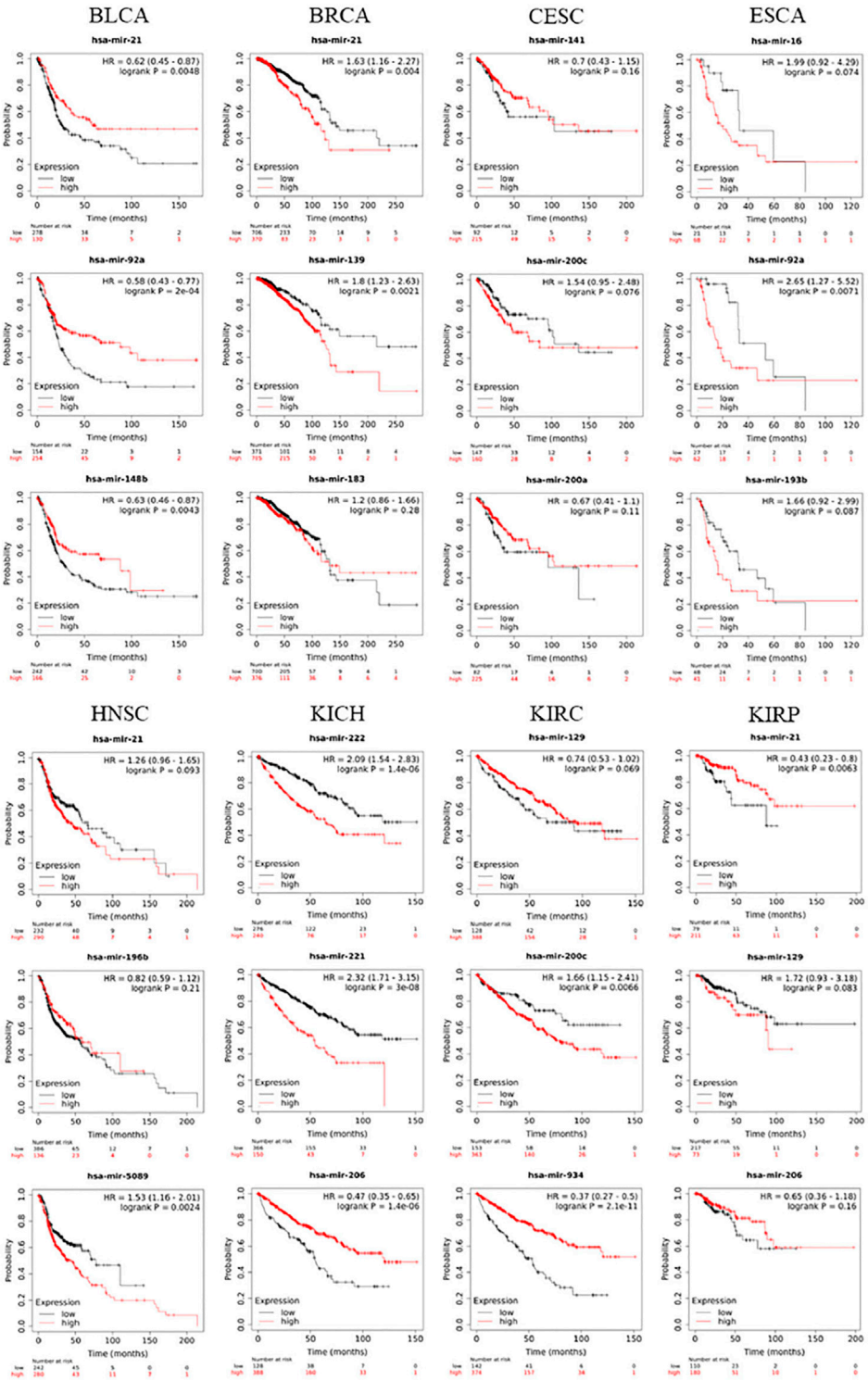


FIGURE 6 | The Kaplan-Meier survival analysis of top-3 ranked miRNAs extracted from the datasets of 20 types of cancer.

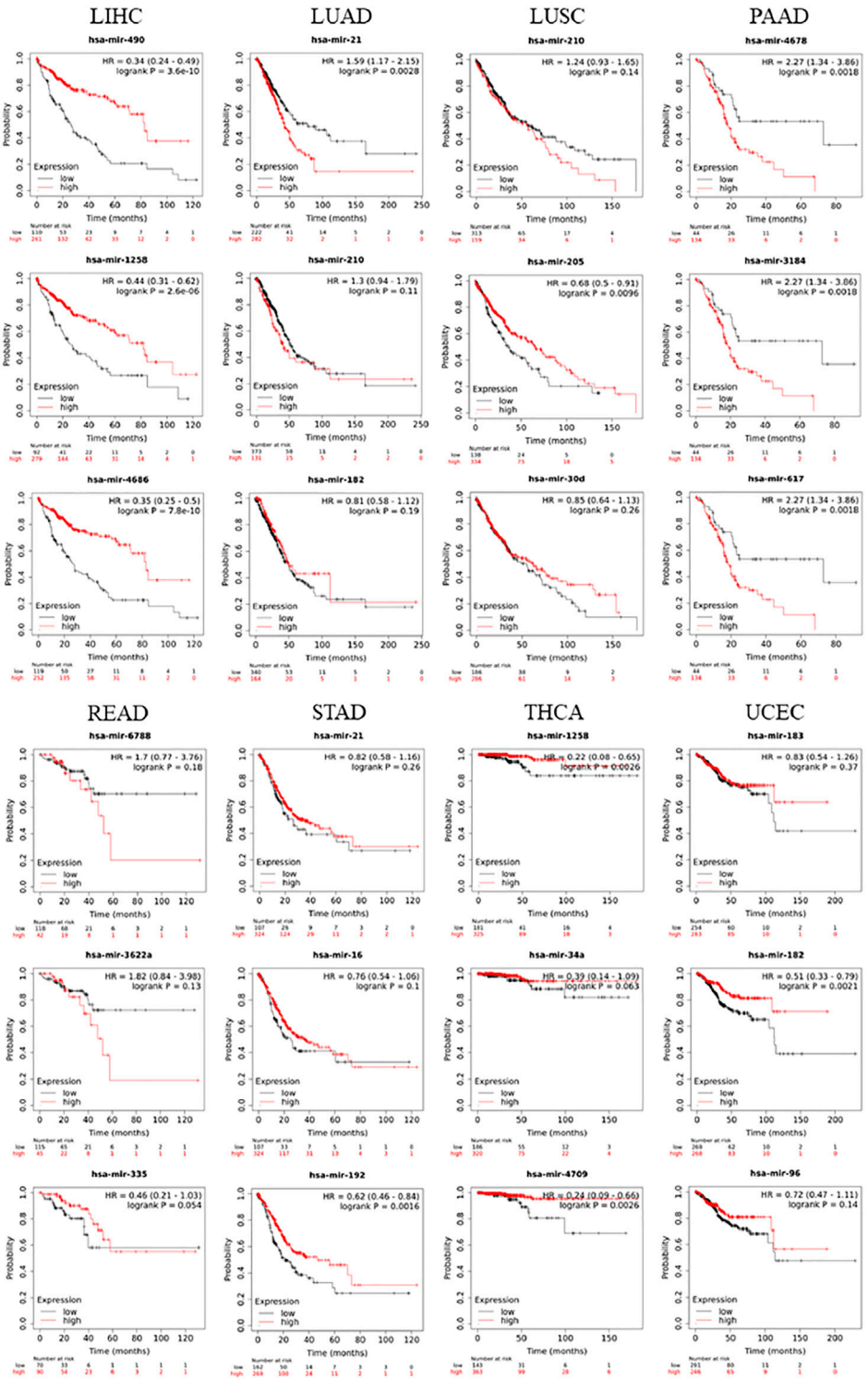


FIGURE 6 | (Continued).

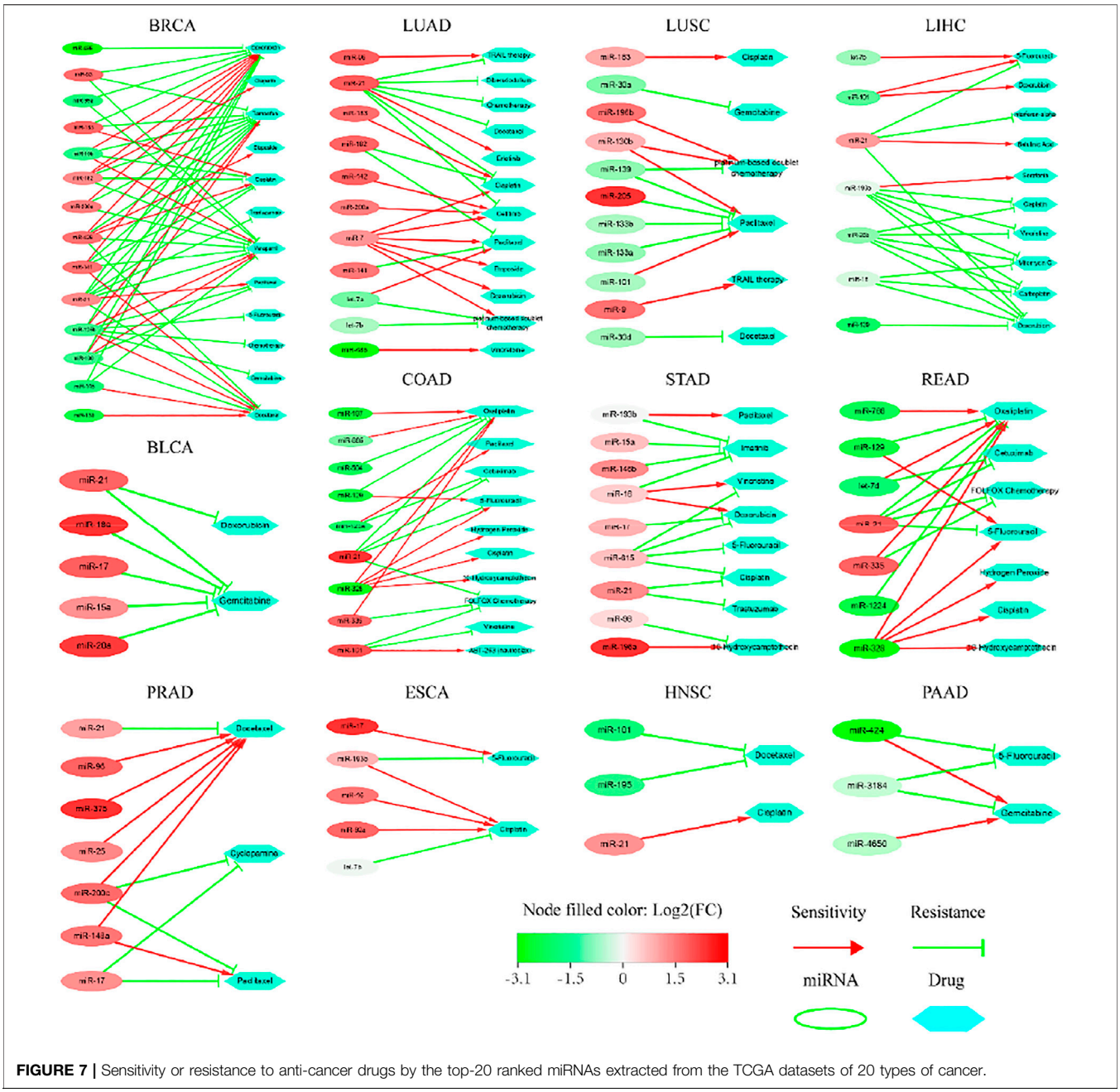


FIGURE 7 | Sensitivity or resistance to anti-cancer drugs by the top-20 ranked miRNAs extracted from the TCGA datasets of 20 types of cancer.

Verification of key miRNAs via the Analysis of Sensitivity or Resistance to Anti-Cancer Drugs

We submitted the top-20 ranked miRNAs to ncDR and mTD, respectively, searching for candidate matches. The results are outlined (**Figure 7**) below. 1) 14, 9, 11, 7, 11, 7, 7, 9 miRNAs impacted drug resistance or sensitivity in BRCA, COAD, LUAD, LIHC, LUSC, PRAD, READ and STAD, respectively; 2) 5, 5, 3 and 3 miRNAs impacted drug sensitivity or resistance in BLCA, ESCA, HNSC and PAAD, respectively; but 3) none of the miRNAs impacted drug resistance or sensitivity in CESC, CHOL, KICH, KIRC, SKCM, THCA and UCEC. We remind that a possible reason for these fewer matches probably lies in that there are relatively fewer records on these cases in the current version of two databases.

Our data suggest that the abnormal expression of key miRNAs impacted the sensitivity or resistance to anti-cancer drugs; some miRNAs promoted drug sensitivity whereas others increased drug resistance (**Figure 7**). We highlighted certain cases as follows. 1) One miRNA impacted a number of drugs, which produced different sensitivity or resistance; and vice versa. It was reported that the overexpression of miR-182 in breast cancer caused resistance to Olaparib, Verapamil, Tamoxifen and Cisplatin, but increased sensitivity to Doxorubicin (Kovalchuk et al., 2008). Here, we

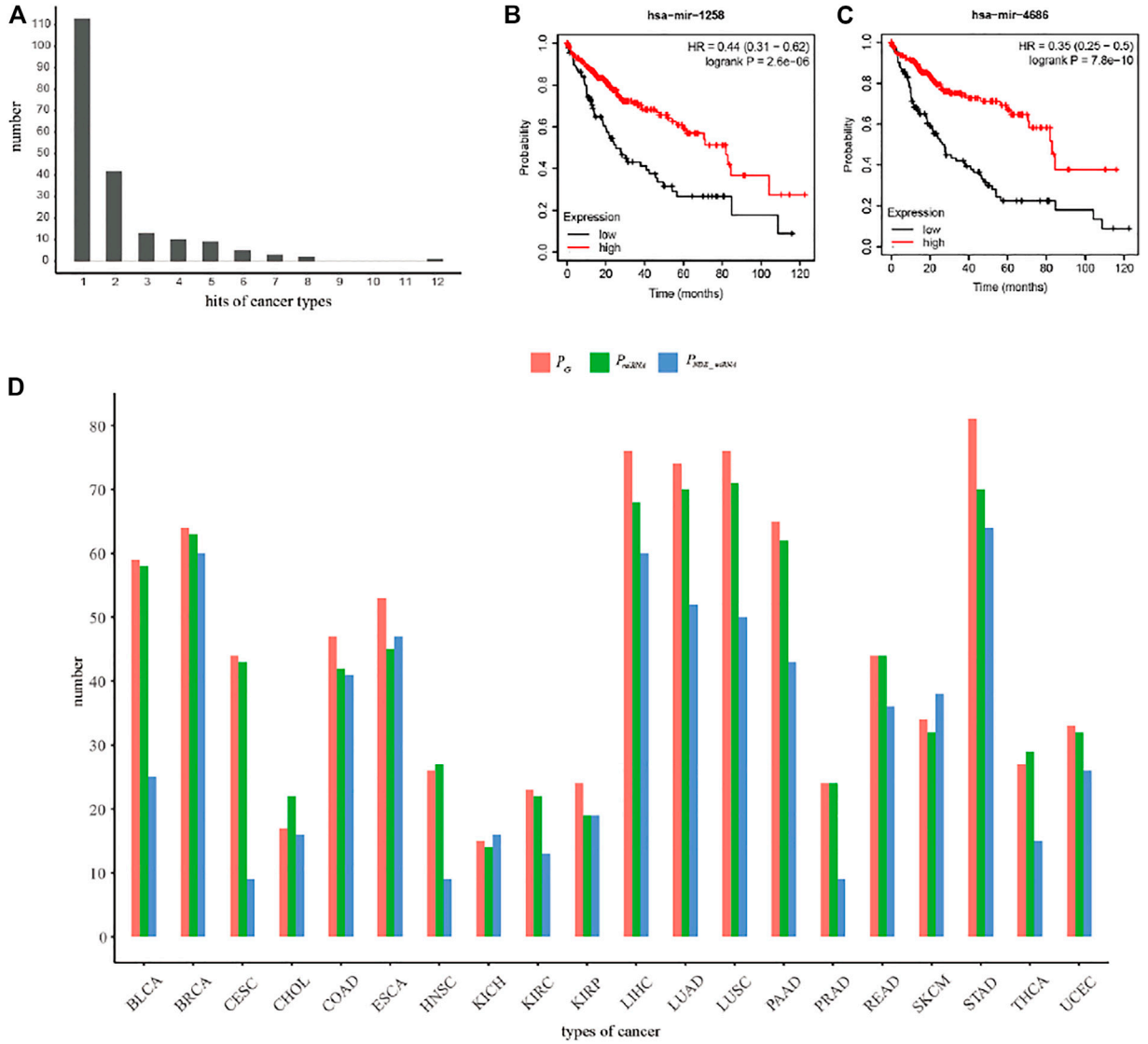


FIGURE 8 | The performance comparison of the MIMRDA method over other methods. **(A)** The number distribution of top-20 ranked cancer-related miRNAs. **(B)** The survival analysis of miR-1258 in LIHC. **(C)** The survival analysis of miR-4686 in LIHC. **(D)** The performance comparison among the MIMRDA method (P_G), the Limma package (P_{miRNA}) and the SPIA package (P_{NDE_miRNA}) based on the top-100 ranked miRNAs identified from the TCGA datasets of 20 types of cancer.

found more cases. All overexpressed miRNAs in bladder cancer promoted resistance to Gemcitabine. Low expression of miR-129 in colon cancer induced resistance to Oxaliplatin, but increased sensitivity to 5-Fluorouracil. Overexpression of miR-193b in esophageal cancer promoted resistance to 5-Fluorouracil, but increased sensitivity to Cisplatin. Overexpression of miR-200c in prostate cancer promoted sensitivity to Docetaxel, but increased resistance to Cyclopamine and Paclitaxel. Overexpression of miR-7 in lung adenocarcinoma weakened resistance to 6 drugs. Overexpression of miR-130, but low expression of miR-101, promoted sensitivity; while low expression of miR-139, miR-133a, miR-133b, but overexpression of miR-205, increased resistance to Paciltaxel. Overexpression of most miRNAs in gastric cancer was associated with drug sensitivity or resistance. Low expression of most miRNAs in liver cancer was associated with sensitivity or resistance. Low expression of miR-101 and miR-195 increased resistance to Docetaxel, but overexpression of miR-21 promoted sensitivity to Cisplatin in the cancer of head and neck. Low expression of miR-424 in pancreatic cancer promoted sensitivity to Gemcitabine, but increased resistance to 5-Fluorouraci. 2) Strikingly, **miR-21** appeared frequently in multiple datasets. Abnormal expression of miR-21 impacted sensitivity or resistance to multiple drugs in BRCA, BLCA, PRAD, LUAD, STAD, HNSC, LIHC and READ. The mechanisms underlying these candidates remained elusive. Collectively, these key miRNAs have complex impacts on the above anti-cancer drugs, which not only illustrate their potential roles in tumorigenesis, but also provide a new perspective for precision medicine.

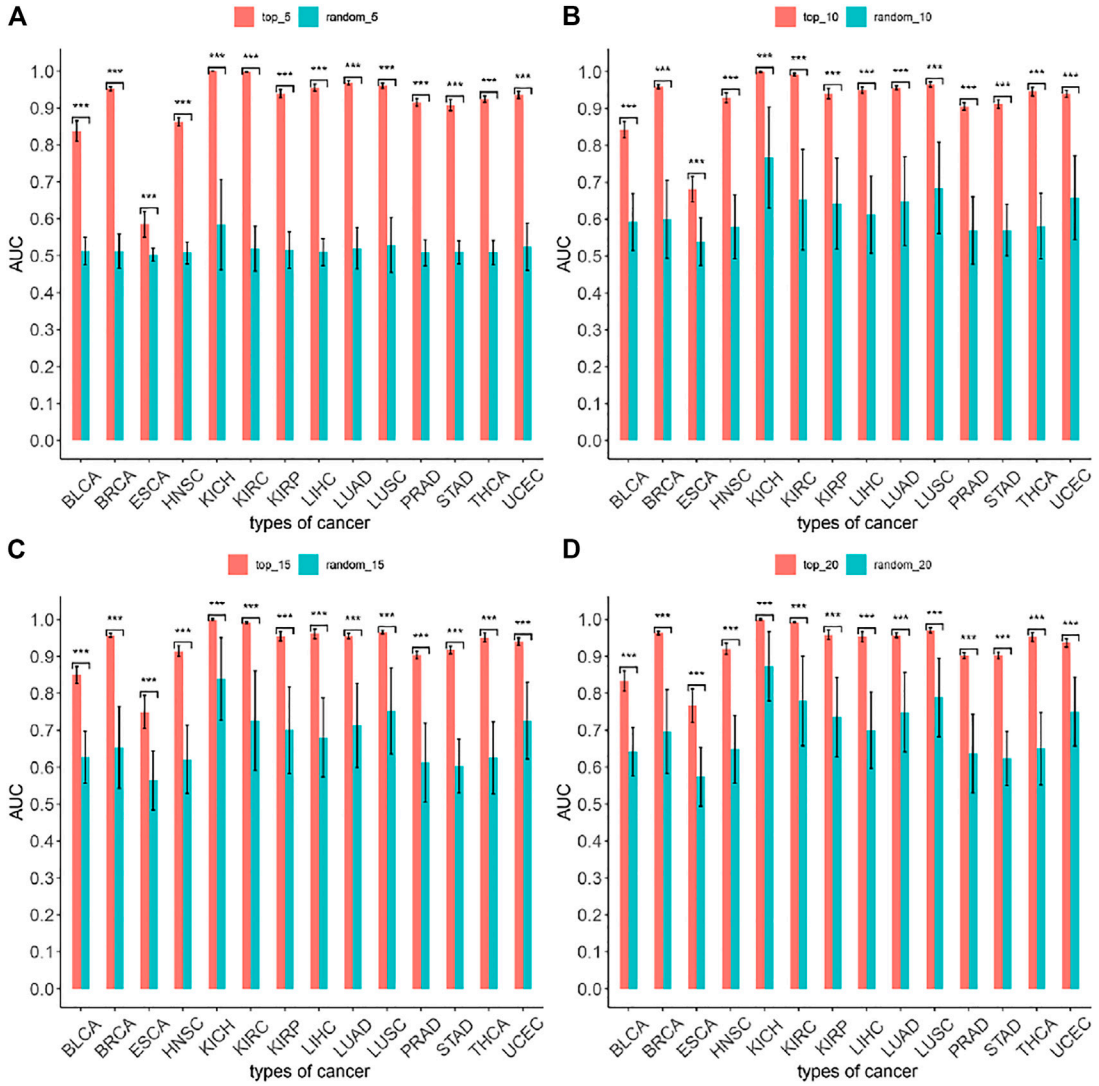


FIGURE 9 | Evaluation on the performance of the MIMRDA method via the Random Forest simulation test. The top-ranked miRNAs identified by the MIMRDA method are compared with the randomly selected miRNAs, both after the five-fold cross-validation Random Forest simulations. **(A)** Top_5 ranked miRNAs *vs* random_5 miRNAs. **(B)** Top_10 ranked miRNAs *vs* random_10 miRNAs. **(C)** Top_15 ranked miRNAs *vs* random_15 miRNAs. **(D)** Top_20 ranked miRNAs *vs* random_20 miRNAs. *p*-value < 0.001***.

Comparison on the Performance of the MIMRDA Method Over Existing Methods

To illustrate the superiority of the MIMRDA method, we compared the miRNAs that were identified by the MIMRDA method, the Limma package (Ritchie et al., 2015) and the SPIA package (Tarca et al., 2009), respectively. For simplicity, we focused on the top-100 ranked miRNAs that were extracted from each dataset of each type of cancer (**Figure 8**). Note that since the classical approaches utilized the known disease-related miRNAs to establish training sets to prioritize miRNAs (Ritchie et al., 2015), it is impossible to use those prioritization methods based on the expression values of genes (or miRNAs), or an overall performance metrics. Hence, we compared the number distribution of candidate miRNAs (i.e., the known disease-related miRNAs). A method performs better if more disease-related miRNAs are found. Obviously, the MIMRDA method identified more miRNAs related to the known types of cancer, which solidifies the superiority of the MIMRDA method to the counterpart methods. Remarkably, as representatives in the second category, who are not matched with the aforementioned two databases, miR-1258 (**Figure 8B**) and miR-4686 (**Figure 8C**) have shown perfect survival rates, which warrant future experimental validations.

Evaluation on the Performance of the MIMRDA Method via the Random Forest Simulation Test

The five-fold cross-validation Random Forest simulation test (see Materials and Methods) was applied to evaluate the accuracy of

the MIMRDA method in classifying top-ranked miRNAs. The results indicate that the MIMRDA method is significantly (p-value < 0.001) better than the random selection in terms of the overall AUC values (**Figure 9**), suggesting the effectiveness and reliable ability of the MIMRDA method in classifying the top-ranked miRNAs.

DISCUSSION

The proposed MIMRDA method identified hundreds of top-ranked miRNAs from TCGA datasets of 20 types of cancer, and recommended them warrant further validations. We employed miR2Disease (Jiang, et al., 2009) and HMDD 3.0 (Huang, et al., 2019) to infer the miRNA-disease associations based on the pre-verified evidences. We deployed MISIM 2.0 (Li et al., 2019) to infer the function similarity of key miRNAs based on the pre-verified function similarities. We applied ncDR (Dai et al., 2017) and mTD (Chen et al., 2017a) to infer the sensitivity or resistance to anti-cancer drugs based on the pre-verified miRNAs-drug associations. Such that our findings were cross-verified to one another. We conclude that most of the top-ranked key miRNAs are the cancer-related miRNAs deposited in miRTarBase (Huang et al., 2020) and TCGA (https://portal.gdc.cancer.gov/) databases, while some are supported by literature evidences. We highlight some key miRNAs that are well supported by the accumulated experimental evidences recaptured from literature, thus highlighting their potential to be biomarkers, which should be valuable to the community.

Firstly, the majority of top-ranked miRNAs (as Category 1, e.g., **miR-21**) are endorsed by the pre-verified relationship of miRNAs-cancer in the state-of-the-art databases (**Figures 4–7**), suggesting that they are truly cancer-related miRNAs and have high potentials to be biomarkers. Here are some examples highlighted with the experimental evidences drawn from literature. 1) **miR-16** inhibited the proliferation and migration of gastric cancer cells by targeting SALL4 (Jiang and Wang 2018). 2) **miR-21** was up-regulated in gastric cancer, and its dysfunction had a critical role in gastric cancer growth and dissemination by regulating PTEN and PDCD4, plus by modulating the pathways involved in mediating cell growth, migration, invasion and apoptosis (Li Y. et al., 2014). **miR-21** and **miR-155** promoted the development of non-small cells by down-regulating SOCS1, SOCS6 and PTEN (Xue et al., 2016). **miR-21** significantly reduced or increased epithelial-mesenchymal transition (Dai et al., 2019). Overexpression of **miR-21** in non-small cell lung cancer up-regulated the expression of cyclin D1 and cyclin E1, respectively (Dai et al., 2019). 3) **miR-34a** was overexpressed and used as a potential target for thyroid cancer (Shabani et al., 2018). 4) **miR-182** targeted CTTN in non-small cell carcinoma to inhibit the formation of aggressive pseudopodia in lung cancer, inhibiting the metastasis of lung cancer (Li et al., 2018). 5) **miR-192-5p** was down-regulated in gastric cancer, as a potential diagnostic target (Tavakolian et al., 2020). 6) **miR-210** promoted the development of lung cancer by targeting LOXL since down-regulation of LOXL4 significantly inhibited the proliferation, migration and invasion of lung cancer cells in A549 and H1650 cell lines (Xie et al., 2019). 7) **miR-335** exhibited a tumor suppressor effect by inhibiting Twsit1 in colorectal cancer (Wang et al., 2017), whereas miR-3065-3p promoted stemness and metastasis by targeting CRLF1 in colorectal cancer (Li et al., 2021). 8) **miR-490-5p** was related to tumor size, tumor metastasis stage and survival rate of HCC patients because miR-490-5p inhibited HCC cell metastasis by regulating E2F2 and ECT2 (Fang et al., 2018). Therefore, such experimental evidences in literature are in line with our findings of some top-ranked key miRNAs.

Secondly, some top-ranked key miRNAs (as Category 2, e.g., **miR-1258**) are not matched with the above databases, but they were well supported by the experimental evidences drawn from literature. For instance, among the top-20 ranked miRNAs, two (**miR-1258** and **miR-4686**) were not matched with miR2Disease and HMDD, respectively, despite that the rest 18 related to LIHC did match. However, we found that miR-1258 and miR-4686 were down-regulated in tumor samples when comparing 375 samples of liver cancer with 50 normal samples (data not shown). We performed the KM survival analysis of miR-1258 and miR-4686 (**Figures 7B,C**), respectively, based on the miRNA expression profiles in 375 samples of primary liver cancer alongside clinical information from TCGA database, and found the significant (p-value < 0.001) survival. Our data suggest that miR-1258 and miR-4686 are likely the potential prognosis factors in LIHC. In fact, miR-1258 was reported significantly down-regulated in liver cancer samples that closely related to the poor survival of patients (Hu et al., 2016), which is consistent with our data. Moreover, loss of miR-1258 led to the initiation and development of liver cancer by targeting CKS1B (Hu et al., 2016); while overexpression of miR-1258 inhibited the growth, proliferation and tumorigenicity of liver cancer cells by increasing G0/G1 cell cycle arrest and promoting cell apoptosis (Hu et al., 2016); and miR-1258 exerted anti-cancer function by targeting TMPRSS4 in thyroid cancer (Wang and Cai 2020). Taken together, our findings coincide with the experimental evidences drawn from literature, and suggest that **miR-1258** has the potential to be developed as an independent prognosis factor in liver cancer.

Thirdly, some top-ranked key miRNAs are related to multiple types of cancer, whereas others are related to a single type of cancer (**Figure 4**). For instance, miR-16-1, **miR-21**, **miR-93**, miR-141, **miR-183** and **miR-193b** present in 7, 12, 8, 7, 7 and 8 types of cancer, respectively, thus impacting the carcinogenesis of multiple types of cancer. Here are examples highlighted. 1) **miR-21** is related to 12 types of cancer (BLCA, BRCA, CESC, CODA, HNSC, KIRC, KIRP, LIHC, LUAD, PRAD, READ and STAD). In fact, miR-21 was experimentally verified to be highly correlated with cancer initiation and metastasis (Liu H. et al., 2018; Wang et al., 2019). 2) **miR-93** is related to 8 types of cancer (BLCA, CHOL, ESCA, KIRP, LIHC, PRAD, STAD and UCEC). In fact, **miR-93** was reported to be closely associated with lung cancer (Li J.-Q. et al., 2017), prostate cancer (Liu J.-J. et al., 2018) and liver cancer (Xu et al., 2018). 3) **miR-183** is related to 7 types of cancer (BLCA, BRCA, CESC, LUAD, LUSC, PRAD and UCEC). In fact, the abnormal expression of **miR-183** initiated multiple types of cancer (Chen X. et al., 2018; Trinh et al., 2019; Li

et al., 2020). 4) **miR-193b** is related to 8 types of cancer (BLCA, CESC, CHOL, ESCA, HNSC, LIHC, LUAD and STAD). In fact, miR-193b was reported to be closely associated with breast cancer (Hulin et al., 2017), liver cancer (Yin et al., 2018) and gastric cancer (Song et al., 2018). Besides, some top-ranked key miRNAs were recaptured in details earlier (see Results) to be uniquely related to a single type of cancer. Taken together, we conclude that some top-ranked key miRNAs are either poly- or mono-valence against multiple types or single type of cancer, respectively.

Finally, the majority of top-ranked key miRNAs are positively or negatively involved in the overall prognostic survival, in the context of specific type of cancer (**Figure 6**). The mechanisms underlying such survival rates remained elusive, but are partly supported by the accumulated experimental evidences drawn from literature. Here are examples highlighted. 1) Abnormal expression of **miR-16** inhibited cell apoptosis by regulating the expression of RECK and SOX6, promoted cell growth and ultimately led to the occurrence of esophageal cancer (Zhu et al., 2014). 2) **miR-21** regulated cell proliferation and sensitivity to Adriamycin in bladder cancer cells (Tao et al., 2011). Overexpression of **miR-21** was highly correlated with poor prognosis of breast cancer (Yan et al., 2008). Overexpression of **miR-21** in T24 cells promoted cell proliferation and resistance to Adriamycin, and resulted in the up-regulation of BLC2, which prevented the apoptosis of T24 cells induced by Adriamycin, favoring the carcinogenic effect of **miR-21** in bladder cell carcinoma (Tao et al., 2011). **miR-21** and PTEN expression had negative correlation *in vivo* in T24 cells (Tao et al., 2011). Low expression of **miR-21** was correlated with poor prognosis of bladder cancer (Zhang et al., 2015). Overexpression of **miR-21** was highly related to the initiation and development of cancer of head and neck (Arantes et al., 2017). **miR-21** promoted the proliferation and metastasis of breast cancer cells by targeting LZTFL1 (Wang et al., 2019). 3) **miR-92a** might be a target for the clinical diagnosis of bladder cancer. Low expression of **miR-92a** was correlated with the poor prognosis of bladder cancer (Motawi et al., 2016). **miR-92a** inhibited the expression of tumor suppressor CDH1. Overexpression of miR-92a restored the metastatic activity of miR-92a, suggesting that miR-92a promoted the migration of esophageal cancer cells by partly inhibiting CDH1. Patients with up-regulated miR-92a were prone to lymph-node metastasis and had a poor prognosis (Chen et al., 2011). 4) **miR-139-3p** exerted a tumor suppressor effect in breast cancer by targeting RAB1A, and might serve as a potential biomarker for prognosis of breast cancer (Zhang et al., 2019). 5) Overexpression of **miR-141** led to the occurrence of cervical cancer (Gómez-Gómez et al., 2013). 6) The serum **miR-148b** markers might have a clinical value in the diagnosis of bladder cancer (Jiang et al., 2015). 7) **miR-183** was dysregulated in breast cancer, related to the expression of estrogen receptor and HER2/neu receptor (Lowery et al., 2010). 8) **miR-193b**/KRAS was expressed in a stage-dependent manner; KRAS was regarded as a direct target of miR-193b; and the upregulation of miR-193b increased the percentage of apoptosis. miR-193b was a biomarker for the treatment of esophageal cancer (Kang et al., 2019). 9) **miR-196a** and **miR-196b** produced cell-specific responses to target genes and downstream pathways, which significantly impacted the cell proliferation, migration and invasion (Álvarez-Teijeiro et al., 2017). Abnormal expression of **miR-196b** presented in the initiation of head and neck cancer. **miR-196b** was a biomarker for early diagnosis of head and neck cancer. 10) **miR-200a** was down-regulated in cervical cancer (Bozgeyik et al., 2020). **miR-200c** inhibited the metastasis and growth of cervical cancer cells via targeting MAP4K4 (Mei et al., 2018). **miR-200c** controlled cell cycle progression and cell growth by down-regulating the G1-S regulator CDK2, and had anti-cancer impacts in ccRCC (Wang et al., 2015). 11) **miR-206** was one of the most critical tumor suppressor miRNAs in ccRCC, which induced cell cycle arrest and inhibited the proliferation of ccRCC cells via targeting CDK4, CDK9 and CCND1 (Xiao et al., 2016). 12) **miR-221** and **miR-222** discriminated the renal cell carcinoma subtypes and tumor cell (Di Meo et al., 2018). 13) **miR-934** was a diagnostic and prognostic biomarker of clear renal cell carcinoma (Liang et al., 2017). Taken together, we conclude that the candidacy of certain key miRNAs identified in this study are supported by experimental evidences recaptured from literature, which provide informative cues for future validations to develop them to be biomarkers ultimately used for the diagnosis and treatment of multiple types of cancer.

We would like to mention possible limitations of our method. We incorporated the mRNA and miRNA expression profiles from the TCGA datasets to identify key miRNAs (microRNAs), rather than utilized other kinds of ncRNAs datasets, such as lncRNAs (Ou-Yang et al., 2019; Lan et al., 2020; Wu et al., 2021) and circRNAs (Liu et al., 2021). Utilizing lncRNAs and circRNAs will be another possible direction of identifying the cancer-related ncRNAs by integrating complex network-based and machine learning-based methods in the future work.

CONCLUSION

We introduced the MIMRDA method, which incorporated the expression profiles of miRNAs and target mRNAs for predicting the miRNA-disease association to identified key miRNAs (microRNAs). As a proof-of-principle study, we deployed the MIMRDA method to analyze 10,499 samples from TCGA datasets of 20 types of cancer, and identified hundreds of key miRNAs. Most of them were significantly related to at least one type of cancer under study, which were supported by the pre-verified miRNA-disease/drug association databases. We indicated the superiority of the MIMRDA method to the Limma and SPIA packages, and the accuracy of the method in classifying top-ranked miRNAs. Our results recommended some top-ranked key miRNAs be experimentally validated as biomarkers 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 authors.

AUTHOR CONTRIBUTIONS

XL and YA designed the project. XL drafted codes and conducted primary computations. HA revised algorithms, clarified codes and inspected demos. XL, HA, BL, FM and YA analyzed data. HA, FM and YA interpreted results. HA, YA, XL, BL, CZ and FM wrote manuscript. All authors approved the final manuscript.

FUNDING

This work was supported by a grant-in-aid from the National Science and Technology Major Programs of China (No. 2014ZX0801105B-002), Supercomputing Program of National Natural Science Foundation of China (No. 201603-534).

ACKNOWLEDGMENTS

We sincerely thank numerous colleagues at Sun Yat-sen University and Guangzhou Medical University for their kind consultations on clinical cases of multiple types of cancer under study. We sincerely acknowledge valuable discussions on the manuscript with Drs. Wenbin Liu and Peng Xu at Institute of Computational Science and Technology, Guangzhou University.

REFERENCES

Álvarez-Teijeiro, S., Menéndez, S. T., Villaronga, M. Á., Rodrigo, J. P., Manterola, L., de Villalaín, L., et al. (2017). Dysregulation of Mir-196b in Head and Neck Cancers Leads to Pleiotropic Effects in the Tumor Cells and Surrounding Stromal Fibroblasts. *Sci. Rep.* 7, 17785. doi:10.1038/s41598-017-18138-8

Arantes, L. M. R. B., Laus, A. C., Melendez, M. E., De Carvalho, A. C., Sorroche, B. P., De Marchi, P. R. M., et al. (2017). MiR-21 as Prognostic Biomarker in Head and Neck Squamous Cell Carcinoma Patients Undergoing an Organ Preservation Protocol. *Oncotarget* 8, 9911–9921. doi:10.18632/oncotarget.14253

Bartel, D. P. (2004). MicroRNAs. *Cell* 116, 281–297. doi:10.1016/s0092-8674(04)00045-5

Bozgeyik, E., Tepe, N. B., and Bozdag, Z. (2020). Identification of microRNA Expression Signature for the Diagnosis and Prognosis of Cervical Squamous Cell Carcinoma. *Pathol. - Res. Pract.* 216, 153159. doi:10.1016/j.prp.2020.153159

Che, K., Guo, M., Wang, C., Liu, X., and Chen, X. (2019). Predicting miRNA-Disease Association by Latent Feature Extraction with Positive Samples. *Genes* 10, 80. doi:10.3390/genes10020080

Chen, D., Li, S. G., Chen, J. Y., and Xiao, M. (2018a). MiR-183 Maintains Canonical Wnt Signaling Activity and Regulates Growth and Apoptosis in Bladder Cancer via Targeting AXIN2. *Eur. Rev. Med. Pharmacol. Sci.* 22, 4828–4836. doi:10.26355/eurrev_201808_15618

Chen, X., Clarence Yan, C., Zhang, X., Li, Z., Deng, L., Zhang, Y., et al. (2015). RBMMMDA: Predicting Multiple Types of Disease-microRNA Associations. *Sci. Rep.* 5, 13877. doi:10.1038/srep13877

Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). RWRMDA: Predicting Novel Human microRNA-Disease Associations. *Mol. Biosyst.* 8, 2792–2798. doi:10.1039/c2mb25180a

Chen, X., Wang, C.-C., Yin, J., and You, Z.-H. (2018b). Novel Human miRNA-Disease Association Inference Based on Random Forest. *Mol. Ther. - Nucleic Acids* 13, 568–579. doi:10.1016/j.omtn.2018.10.005

Chen, X., Wu, Q.-F., and Yan, G.-Y. (2017a). RKNNMDA: Ranking-Based KNN for miRNA-Disease Association Prediction. *RNA Biol.* 14, 952–962. doi:10.1080/15476286.2017.1312226

Chen, X., Xie, D., Zhao, Q., and You, Z.-H. (2019). MicroRNAs and Complex Diseases: from Experimental Results to Computational Models. *Brief Bioinform* 20, 515–539. doi:10.1093/bib/bbx130

Chen, X., Xie, W.-B., Xiao, P.-P., Zhao, X.-M., and Yan, H. (2017b). mTD: A Database of microRNAs Affecting Therapeutic Effects of Drugs. *J. Genet. Genomics* 44, 269–271. doi:10.1016/j.jgg.2017.04.003

Chen, X., Yan, C. C., Zhang, X., You, Z.-H., Deng, L., Liu, Y., et al. (2016a). WBSMDA: within and between Score for miRNA-Disease Association Prediction. *Sci. Rep.* 6, 21106. doi:10.1038/srep21106

Chen, X., Yan, C. C., Zhang, X., You, Z.-H., Huang, Y.-A., and Yan, G.-Y. (2016b). HGIMDA: Heterogeneous Graph Inference for miRNA-Disease Association Prediction. *Oncotarget* 7, 65257–65269. doi:10.18632/oncotarget.11251

Chen, X., and Yan, G.-Y. (2014). Semi-supervised Learning for Potential Human microRNA-Disease Associations Inference. *Sci. Rep.* 4, 5501. doi:10.1038/srep05501

Chen, Z.-l., Zhao, X.-h., Wang, J.-w., Li, B.-z., Wang, Z., Sun, J., et al. (2011). microRNA-92a Promotes Lymph Node Metastasis of Human Esophageal Squamous Cell Carcinoma via E-Cadherin. *J. Biol. Chem.* 286, 10725–10734. doi:10.1074/jbc.M110.165654

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., et al. (2018). miRTarBase Update 2018: a Resource for Experimentally Validated microRNA-Target Interactions. *Nucleic Acids Res.* 46, D296–d302. doi:10.1093/nar/gkx1067

Cretoiu, D., Chen, F., Zheng, Y., Zhang, D., Qian, B., Ji, H., et al. (2019). miR-21 Regulates Growth and EMT in Lung Cancer Cells via PTEN Akt GSK3 Beta Signaling. *Front. Biosci.* 24, 1426–1439. doi:10.2741/4788

Dai, E., Yang, F., Wang, J., Zhou, X., Song, Q., An, W., et al. (2017). ncDR: a Comprehensive Resource of Non-coding RNAs Involved in Drug Resistance. *Bioinformatics* 33, 4010–4011. doi:10.1093/bioinformatics/btx523

Di Meo, A., Saleeb, R., Wala, S. J., Khella, H. W., Ding, Q., Zhai, H., et al. (2018). A miRNA-Based Classification of Renal Cell Carcinoma Subtypes by PCR and *In Situ* Hybridization. *Oncotarget* 9, 2092–2104. doi:10.18632/oncotarget.23162

Dweep, H., and Gretz, N. (2015). miRWalk2.0: a Comprehensive Atlas of microRNA-Target Interactions. *Nat. Methods* 12, 697. doi:10.1038/nmeth.3485

Esquela-Kerscher, A., and Slack, F. J. (2006). Oncomirs - microRNAs with a Role in Cancer. *Nat. Rev. Cancer* 6 (4), 259–269. doi:10.1038/nrc1840

Evan, G. I., and Vousden, K. H. (2001). Proliferation, Cell Cycle and Apoptosis in Cancer. *Nature* 411, 342–348. doi:10.1038/35077213

Fang, Z. Q., Li, M. C., Zhang, Y. Q., and Liu, X. G. (2018). MiR-490-5p Inhibits the Metastasis of Hepatocellular Carcinoma by Down-Regulating E2F2 and ECT2. *J. Cel. Biochem.* 119, 8317–8324. doi:10.1002/jcb.26876

Gómez-Gómez, Y., Organista-Nava, J., and Gariglio, P. (2013). Deregulation of the miRNAs Expression in Cervical Cancer: Human Papillomavirus Implications. *Biomed. Res. Int.* 2013, 407052. doi:10.1155/2013/407052

Hu, M., Wang, M., Lu, H., Wang, X., Fang, X., Wang, J., et al. (2016). Loss of miR-1258 Contributes to Carcinogenesis and Progression of Liver Cancer through Targeting CDC28 Protein Kinase Regulatory Subunit 1B. *Oncotarget* 7, 43419–43431. doi:10.18632/oncotarget.9728

Huang, H.-Y., Lin, Y.-C. -D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2020). miRTarBase 2020: Updates to the Experimentally Validated microRNA-Target Interaction Database. *Nucleic Acids Res.* 48, D148–D154. doi:10.1093/nar/gkz896

Huang, Y., Wang, X., Zheng, Y., Chen, W., Zheng, Y., Li, G., et al. (2021). Construction of an mRNA-miRNA-lncRNA Network Prognostic for Triple-Negative Breast Cancer. *Aging* 13, 1153–1175. doi:10.18632/aging.202254

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3.0: a Database for Experimentally Supported Human microRNA-Disease Associations. *Nucleic Acids Res.* 47, D1013–D1017. doi:10.1093/nar/gky1010

Hulin, J.-A., Tommasi, S., Elliot, D., Hu, D. G., Lewis, B. C., and Mangoni, A. A. (2017). MiR-193b Regulates Breast Cancer Cell Migration and Vasculogenic Mimicry by Targeting Dimethylarginine Dimethylaminohydrolase 1. *Sci. Rep.* 7, 13996. doi:10.1038/s41598-017-14454-1

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of Disease microRNAs through a Human Phenome-microRNAome Network. *BMC Syst. Biol.* 4 (Suppl. 1), S2. doi:10.1186/1752-0509-4-s1-s2

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a Manually Curated Database for microRNA Deregulation in Human Disease. *Nucleic Acids Res.* 37, D98–D104. doi:10.1093/nar/gkn714

Jiang, X., Du, L., Wang, L., Li, J., Liu, Y., Zheng, G., et al. (2015). Serum microRNA Expression Signatures Identified from Genome-wide microRNA Profiling Serve as Novel Noninvasive Biomarkers for Diagnosis and Recurrence of Bladder Cancer. *Int. J. Cancer* 136, 854–862. doi:10.1002/ijc.29041

Jiang, X., and Wang, Z. (2018). miR-16 Targets SALL4 to Repress the Proliferation and Migration of Gastric Cancer. *Oncol. Lett.* 16, 3005–3012. doi:10.3892/ol.2018.8997

Kang, M., Li, Y., Zhu, S., Zhang, S., Guo, S., and Li, P. (2019). MicroRNA-193b Acts as a Tumor Suppressor Gene in Human Esophageal Squamous Cell Carcinoma via Target Regulation of KRAS. *Oncol. Lett.* 17, 3965–3973. doi:10.3892/ol.2019.10039

Kovalchuk, O., Filkowski, J., Meservy, J., Ilnytskyy, Y., Tryndyak, V. P., Chekhun, V. F., et al. (2008). Involvement of microRNA-451 in Resistance of the MCF-7 Breast Cancer Cells to Chemotherapeutic Drug Doxorubicin. *Mol. Cancer Ther.* 7, 2152–2159. doi:10.1158/1535-7163.Mct-08-0021

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA Sequences to Function. *Nucleic Acids Res.* 47, D155–D162. doi:10.1093/nar/gky1141

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2020). LDICDL: LncRNA-Disease Association Identification Based on Collaborative Deep Learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* [Epub ahead of print] 1. doi:10.1109/tcbb.2020.3034910

Li, C., Lyu, J., and Meng, Q. H. (2017a). MiR-93 Promotes Tumorigenesis and Metastasis of Non-small Cell Lung Cancer Cells by Activating the PI3K/Akt Pathway via Inhibition of LKB1/PTEN/CDKN1A. *J. Cancer* 8, 870–879. doi:10.7150/jca.17958

Li, J.-Q., Rong, Z.-H., Chen, X., Yan, G.-Y., and You, Z.-H. (2017b). MCMDA: Matrix Completion for MiRNA-Disease Association Prediction. *Oncotarget* 8, 21187–21199. doi:10.18632/oncotarget.15061

Li, J., Zhang, S., Wan, Y., Zhao, Y., Shi, J., Zhou, Y., et al. (2019). MISIM v2.0: a Web Server for Inferring microRNA Functional Similarity Based on microRNA-Disease Associations. *Nucleic Acids Res.* 47, W536–w541. doi:10.1093/nar/gkz328

Li, L., Zhou, L., Li, Y., Lin, S., and Tomuleasa, C. (2014a). MicroRNA-21 Stimulates Gastric Cancer Growth and Invasion by Inhibiting the Tumor Suppressor Effects of Programmed Cell Death Protein 4 and Phosphatase and Tensin Homolog. *J. Buon* 19, 228–236.

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014b). HMDD v2.0: a Database for Experimentally Supported Human microRNA and Disease Associations. *Nucl. Acids Res.* 42, D1070–D1074. doi:10.1093/nar/gkt1023

Li, Y., Xun, J., Wang, B., Ma, Y., Zhang, L., Yang, L., et al. (2021). miR-3065-3p Promotes Stemness and Metastasis by Targeting CRLF1 in Colorectal Cancer. *J. Transl Med.* 19, 429. doi:10.1186/s12967-021-03102-y

Li, Y., Zeng, Q. a., Qiu, J., Pang, T., Ye, F., Huang, L., et al. (2020). MiR-183-5p Promotes Proliferation, Metastasis and Angiogenesis in Breast Cancer Cells through Negatively Regulating Four and a Half LIM Protein 1. *J. Breast Cancer* 23, 355–372. doi:10.4048/jbc.2020.23.e47

Li, Y., Zhang, H., Gong, H., Yuan, Y., Li, Y., Wang, C., et al. (2018). miR-182 Suppresses Invadopodia Formation and Metastasis in Non-small Cell Lung Cancer by Targeting Cortactin Gene. *J. Exp. Clin. Cancer Res.* 37, 141. doi:10.1186/s13046-018-0824-1

Liang, B., Zhao, J., and Wang, X. (2017). A Three-microRNA Signature as a Diagnostic and Prognostic Marker in clear Cell Renal Cancer: An In Silico Analysis. *PLoS One* 12, e0180660. doi:10.1371/journal.pone.0180660

Liu, H., Cheng, L., Cao, D., and Zhang, H. (2018a). Suppression of miR-21 Expression Inhibits Cell Proliferation and Migration of Liver Cancer Cells by Targeting Phosphatase and Tensin Homolog (PTEN). *Med. Sci. Monit.* 24, 3571–3577. doi:10.12659/msm.907038

Liu, J.-J., Zhang, X., and Wu, X.-H. (2018b). miR-93 Promotes the Growth and Invasion of Prostate Cancer by Upregulating its Target Genes TGFBR2, ITGB8, and LATS2. *Mol. Ther. - Oncolytics* 11, 14–19. doi:10.1016/j.omto.2018.08.001

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2017). Inferring microRNA-Disease Associations by Random Walk on a Heterogeneous Network with Multiple Data Sources. *Ieee/acm Trans. Comput. Biol. Bioinf.* 14, 905–915. doi:10.1109/tcbb.2016.2550432

Liu, Z., Ji, C., Ni, J.-C., Wang, Y.-T., Qiao, L., and Zheng, C.-H. (2021). Convolution Neural Networks Using Deep Matrix Factorization for Predicting circRNA-Disease Association. *IEEE/ACM Trans. Comput. Biol. Bioinform.* [Epub ahead of print], 1 doi:10.1109/tcbb.2021.3138339

Lowery, A. J., Miller, N., Dwyer, R. M., and Kerin, M. J. (2010). Dysregulated miR-183 Inhibits Migration in Breast Cancer Cells. *BMC Cancer* 10, 502. doi:10.1186/1471-2407-10-502

Megraw, M., Sethupathy, P., Corda, B., and Hatzigeorgiou, A. G. (2007). miRGen: a Database for the Study of Animal microRNA Genomic Organization and Function. *Nucleic Acids Res.* 35, D149–D155. doi:10.1093/nar/gkl904

Mei, J., Wang, D. H., Wang, L. L., Chen, Q., Pan, L. L., and Xia, L. (2018). MicroRNA-200c Suppressed Cervical Cancer Cell Metastasis and Growth via Targeting MAP4K4. *Eur. Rev. Med. Pharmacol. Sci.* 22, 623–631. doi:10.26355/eurrev_201802_14286

Motawi, T. K., Rizk, S. M., Ibrahim, T. M., and Ibrahim, I. A.-R. (2016). Circulating microRNAs, miR-92a, miR-100 and miR-143, as Non-invasive Biomarkers for Bladder Cancer Diagnosis. *Cell Biochem Funct* 34, 142–148. doi:10.1002/cbf.3171

Ou-Yang, L., Huang, J., Zhang, X.-F., Li, Y.-R., Sun, Y., He, S., et al. (2019). lncRNA-Disease Association Prediction Using Two-Side Sparse Self-Representation. *Front. Genet.* 10, 476. doi:10.3389/fgene.2019.00476

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

Rupaimoole, R., and Slack, F. J. (2017). MicroRNA Therapeutics: towards a new era for the Management of Cancer and Other Diseases. *Nat. Rev. Drug Discov.* 16, 203–222. doi:10.1038/nrd.2016.246

Saluja, R., Cheng, S., Santos, K. A., and Chan, K. K. W. (2019). Estimating hazard Ratios from Published Kaplan-Meier Survival Curves: A Methods Validation Study. *Res. Syn Meth* 10, 465–475. doi:10.1002/jrsm.1362

Shabani, N., Razaviyan, J., Paryan, M., Tavangar, S. M., Azizi, F., Mohammadi-Yeganeh, S., et al. (2018). Evaluation of miRNAs Expression in Medullary Thyroid Carcinoma Tissue Samples: miR-34a and miR-144 as Promising Overexpressed Markers in MTC. *Hum. Pathol.* 79, 212–221. doi:10.1016/j.humpath.2018.05.019

Song, B., Du, J., Song, D.-f., Ren, J.-c., and Feng, Y. (2018). Dysregulation of NCAPG, KNL1, miR-148a-3p, miR-193b-3p, and miR-1179 May Contribute to the Progression of Gastric Cancer. *Biol. Res.* 51, 44. doi:10.1186/s40659-018-0192-5

Speiser, J. L., Miller, M. E., Tooze, J., and Ip, E. (2019). A Comparison of Random forest Variable Selection Methods for Classification Prediction Modeling. *Expert Syst. Appl.* 134, 93–101. doi:10.1016/j.eswa.2019.05.028

Taniguchi, K., and Karin, M. (2018). NF-κB, Inflammation, Immunity and Cancer: Coming of Age. *Nat. Rev. Immunol.* 18, 309–324. doi:10.1038/nri.2017.142

Tarca, A. L., Draghici, s., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., et al. (2009). A Novel Signaling Pathway Impact Analysis. *Bioinformatics* 25, 75–82. doi:10.1093/bioinformatics/btn577

Tavakolian, S., Goudarzi, H., and Faghihloo, E. (2020). Evaluating the Expression Level of miR-9-5p and miR-192-5p in Gastrointestinal Cancer: Introducing Novel Screening Biomarkers for Patients. *BMC Res. Notes* 13, 226. doi:10.1186/s13104-020-05071-9

Trinh, T. L., Kandell, W. M., Donatelli, S. S., Tu, N., Tejera, M. M., Gilvary, D. L., et al. (2019). Immune Evasion by TGFβ-Induced miR-183 Repression of MICA/B Expression in Human Lung Tumor Cells. *Oncoimmunology* 8, e1557372. doi:10.1080/2162402x.2018.1557372

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the Human microRNA Functional Similarity and Functional Network Based on microRNA-Associated Diseases. *Bioinformatics* 26, 1644–1650. doi:10.1093/bioinformatics/btq241

Wang, H., Tan, Z., Hu, H., Liu, H., Wu, T., Zheng, C., et al. (2019). microRNA-21 Promotes Breast Cancer Proliferation and Metastasis by Targeting LZTFL1. *BMC Cancer* 19, 738. doi:10.1186/s12885-019-5951-3

Wang, J., Wang, X., Liu, F., and Fu, Y. (2017). microRNA-335 Inhibits Colorectal Cancer HCT116 Cells Growth and Epithelial-Mesenchymal Transition (EMT)

Process by Targeting Twist1. *Pharmazie* 72, 475–481. doi:10.1691/ph.2017.7489

Wang, L.-J., and Cai, H.-Q. (2020). miR-1258: a Novel microRNA that Controls TMPRSS4 Expression Is Associated with Malignant Progression of Papillary Thyroid Carcinoma. *Endokrynol Pol.* 71, 146–152. doi:10.5603/EP.a2020.0009

Wang, X., Chen, X., Han, W., Ruan, A., Chen, L., Wang, R., et al. (2015). miR-200c Targets CDK2 and Suppresses Tumorigenesis in Renal Cell Carcinoma. *Mol. Cancer Res.* 13, 1567–1577. doi:10.1158/1541-7786.Mcr-15-0128

Wu, Q.-W., Xia, J.-F., Ni, J.-C., and Zheng, C.-H. (2021). Gaerf: Predicting lncRNA-Disease Associations by Graph Auto-Encoder and Random forest. *Brief Bioinform* 22 (5), bbaa391. doi:10.1093/bib/bbaa391

Xiao, H., Xiao, W., Cao, J., Li, H., Guan, W., Guo, X., et al. (2016). miR-206 Functions as a Novel Cell Cycle Regulator and Tumor Suppressor in clear-cell Renal Cell Carcinoma. *Cancer Lett.* 374, 107–116. doi:10.1016/j.canlet.2016.01.032

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). miRCancer: a microRNA-Cancer Association Database Constructed by Text Mining on Literature. *Bioinformatics* 29, 638–644. doi:10.1093/bioinformatics/btt014

Xie, S., Liu, G., Huang, J., Hu, H. B., and Jiang, W. (2019). miR-210 Promotes Lung Adenocarcinoma Proliferation, Migration, and Invasion by Targeting Lysyl Oxidase-like 4. *J. Cel Physiol* 234, 14050–14057. doi:10.1002/jcp.28093

Xu, C., Ping, Y., Li, X., Zhao, H., Wang, L., Fan, H., et al. (2014). Prioritizing Candidate Disease miRNAs by Integrating Phenotype Associations of Multiple Diseases with Matched miRNA and mRNA Expression Profiles. *Mol. Biosyst.* 10, 2800–2809. doi:10.1039/c4mb00353e

Xu, F., Zha, G., Wu, Y., Cai, W., and Ao, J. (2018). Overexpressing lncRNA SNHG16 Inhibited HCC Proliferation and Chemoresistance by Functionally Sponging Hsa-miR-93. *Onco Targets Ther.* 11, 8855–8863. doi:10.2147/ott.S182005

Xu, J., Li, C.-X., Lv, J.-Y., Li, Y.-S., Xiao, Y., Shao, T.-T., et al. (2011). Prioritizing Candidate Disease miRNAs by Topological Features in the miRNA Target-Dysregulated Network: Case Study of Prostate Cancer. *Mol. Cancer Ther.* 10, 1857–1866. doi:10.1158/1535-7163.Mct-11-0055

Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., et al. (2013). Prediction of microRNAs Associated with Human Diseases Based on Weighted K Most Similar Neighbors. *PLoS One* 8, e70204. doi:10.1371/journal.pone.0070204

Xue, X., Liu, Y., Wang, Y., Meng, M., Wang, K., Zang, X., et al. (2016). MiR-21 and MiR-155 Promote Non-small Cell Lung Cancer Progression by Downregulating SOCS1, SOCS6, and PTEN. *Oncotarget* 7, 84508–84519. doi:10.18632/oncotarget.13022

Yan, L.-X., Huang, X.-F., Shao, Q., Huang, M.-Y., Deng, L., Wu, Q.-L., et al. (2008). MicroRNA miR-21 Overexpression in Human Breast Cancer Is Associated with Advanced Clinical Stage, Lymph Node Metastasis and Patient Poor Prognosis. *RNA* 14, 2348–2360. doi:10.1261/rna.1034808

Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., et al. (2017). dbDEMC 2.0: Updated Database of Differentially Expressed miRNAs in Human Cancers. *Nucleic Acids Res.* 45, D812–D818. doi:10.1093/nar/gkw1079

Yin, W., Nie, Y., Chen, L., Wang, Q., Liu, S., He, X., et al. (2018). Deregulation of microRNA-193b Affects the Proliferation of Liver Cancer via Myeloid Cell Leukemia-1. *Oncol. Lett.* 15, 2781–2788. doi:10.3892/ol.2017.7690

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., et al. (2017). PBMDA: A Novel and Effective Path-Based Computational Model for miRNA-Disease Association Prediction. *Plos Comput. Biol.* 13, e1005455. doi:10.1371/journal.pcbi.1005455

Zhang, H.-H., Qi, F., Cao, Y.-H., Zu, X.-B., and Chen, M.-F. (2015). Expression and Clinical Significance of microRNA-21, Maspin and Vascular Endothelial Growth Factor-C in Bladder Cancer. *Oncol. Lett.* 10, 2610–2616. doi:10.3892/ol.2015.3540

Zhang, W., Lu, Q., Wu, D., Li, P., Xu, B., Qing, W., et al. (2011). microRNA-21 Modulates Cell Proliferation and Sensitivity to Doxorubicin in Bladder Cancer Cells. *Oncol. Rep.* 25, 1721–1729. doi:10.3892/or.2011.1245

Zhang, W., Xu, J., Wang, K., Tang, X., and He, J. (2019). miR-139-3p Suppresses the Invasion and Migration Properties of Breast Cancer Cells by Targeting RAB1A. *Oncol. Rep.* 42, 1699–1708. doi:10.3892/or.2019.7297

Zhao, Y., Chen, X., and Yin, J. (2019). Adaptive Boosting-Based Computational Model for Predicting Potential miRNA-Disease Associations. *Bioinformatics* 35, 4730–4738. doi:10.1093/bioinformatics/btz297

Zheng, K., You, Z.-H., Wang, L., Zhou, Y., Li, L.-P., and Li, Z.-W. (2020). DBMDA: A Unified Embedding for Sequence-Based miRNA Similarity Measure with Applications to Predict and Validate miRNA-disease Associations. *Mol. Ther. - Nucleic Acids* 19, 602–611. doi:10.1016/j.omtn.2019.12.010

Zhu, Y., Xia, Y., Niu, H., and Chen, Y. (2014). MiR-16 Induced the Suppression of Cell Apoptosis while Promote Proliferation in Esophageal Squamous Cell Carcinoma. *Cell Physiol Biochem* 33, 1340–1348. doi:10.1159/000358701

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.

Copyright © 2022 Li, Ai, Li, Zhang, Meng and Ai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 January 2022
doi: 10.3389/fgene.2021.818841

A SNARE Protein Identification Method Based on iLearnPlus to Efficiently Solve the Data Imbalance Problem

Dong Ma, Zhihua Chen, Zhanpeng He and Xueqin Huang*

Institute of Computing Science and Technology, Guangzhou University, Guangdong, China

Machine learning has been widely used to solve complex problems in engineering applications and scientific fields, and many machine learning-based methods have achieved good results in different fields. SNAREs are key elements of membrane fusion and required for the fusion process of stable intermediates. They are also associated with the formation of some psychiatric disorders. This study processes the original sequence data with the synthetic minority oversampling technique (SMOTE) to solve the problem of data imbalance and produces the most suitable machine learning model with the iLearnPlus platform for the identification of SNARE proteins. Ultimately, a sensitivity of 66.67%, specificity of 93.63%, accuracy of 91.33%, and MCC of 0.528 were obtained in the cross-validation dataset, and a sensitivity of 66.67%, specificity of 93.63%, accuracy of 91.33%, and MCC of 0.528 were obtained in the independent dataset (the adaptive skip dipeptide composition descriptor was used for feature extraction, and LightGBM with proper parameters was used as the classifier). These results demonstrate that this combination can perform well in the classification of SNARE proteins and is superior to other methods.

Keywords: SNARE protein identification, ASDC features, SMOTE, data imbalance, machine learning

OPEN ACCESS

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

Reviewed by:
Xiangxiang Zeng,
Hunan University, China
Xun Wang,
China University of Petroleum, Huadong, China

****Correspondence:***
Zhihua Chen
czhgd@gzhu.edu.cn

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

Received: *20 November 2021*
Accepted: *14 December 2021*
Published: *28 January 2022*

Citation:
Ma D, Chen Z, He Z and Huang X (2022) A SNARE Protein Identification Method Based on iLearnPlus to Efficiently Solve the Data Imbalance Problem. Front. Genet. 12:818841. doi: 10.3389/fgene.2021.818841

INTRODUCTION

Soluble N-ethylmaleimide-sensitive fusion protein attachment protein receptor (SNARE) proteins are a small superfamily of proteins. They have an uncomplicated domain structure, and a feature of them is the SNARE motif—an evolutionarily conserved heptanucleotide repeat consisting of 60–70 amino acids. (Jahn and Scheller, 2006) They can be divided into Q-SNAREs and R-SNAREs pursuant to the structural characteristics of SNAREs. Functionally, SNAREs are most likely associated with various aspects of membrane transport specificity, and they are a key element in membrane fusion and are necessary for stable fusion intermediates. (Schoch et al., 2001) SNARE proteins are involved in membrane vesicle transport, such as synaptic transmission between nerve cells (synaptic vesicle transport) and plant disease resistance (disease resistance signaling). In addition, SNAREs are also implicated in the formation of some mental disorders. (Wang et al., 2018)

It is relatively complex to explore the function of a particular protein in the field of biology, the general prediction method is based on Protein-Protein-Interaction (PPI) (Hu et al., 2011; Zhai et al., 2020; Sundar and Narmadha, 2021) and protein structure information (Kinjo and Nakamura, 2012; Sharma and Srivastava, 2021). In the subsequent process, the specific function of detection through the complex biological experiment needs to be clear, which greatly increases the difficulty and the

resources required of the properties that determine protein function, thus reducing the efficiency due to unavoidable time consumption.

In recent years, with the development of machine learning, many methods have achieved good results in various fields, such as Nature Language Processing (NLP) and computer vision (Jin et al., 2021). In addition, the classification task is one of the most basic applications in machine learning, and relevant research is has matured. (Ke et al., 2017) Nguyen Quoc Khanh Le, et al. (Le et al., 2019) employed PSSM profiles and 2D CNN to identify SNARE proteins. Su, Xin, et al. (Su et al., 2019) applied the multiscale convolutional network to the identification of antimicrobial peptides, so it is appropriate to apply machine learning to protein classification tasks.

In this paper, multiple feature extraction algorithms are used to extract different features, obtain the best performance descriptor through performance comparison, and then perform data enhancement processing on the extracted features of this descriptor to address the problem of sample imbalance in the data to a certain extent. Finally, the processed feature data and raw data of the independent test set were used to train the classifier to obtain the eventual model.

MATERIALS AND METHODS

The task of protein sequence classification models based on machine learning generally includes five main steps: protein sequence data collection, feature extraction and processing, classifier construction and optimization, model performance evaluation, and result visualization. (Liu et al., 2019; Guo et al., 2020; Tao et al., 2020; Chen Z et al., 2021; Li et al., 2021) The details of the first three steps determine whether the classification performance is satisfactory, while the last two steps are only a further explanation of the experimental results and determined by objective evaluation indicators, so the sequence classification task is mainly carried out using the first three steps. **Figure 1** illustrates the research flow of this paper.

Datasets

The research object of this paper are SNARE proteins, which are generally downloaded from the UniProt database. As the research object is a specific type of protein, less sequence data can be obtained for a specific protein compared to other non-specific types of common proteins, which leads to the final dataset being easily unbalanced, i.e., the number of nonspecific proteins in the dataset is greater than the number of specific proteins. The dataset used in this study was from other similar tasks. (Le and Nguyen, 2019) The number of SNARE proteins in this dataset was only one-tenth that of non-SNARE proteins, including 697 SNARE proteins as positive samples and 7,378 vesicle transport proteins as negative samples. During the experiment, 90% of them were extracted for the training of the model, and the rest were used as independent validation sets to evaluate the generalization ability of the model.

Feature Extraction and Processing

Biological sequence data are generally stored in a FASTA file format, and each sequence data is represented by the letter of the nucleotide or amino acid constituting the molecule. As the number of molecules composing the biological sequence is not fixed, the length of the sequence is inconsistent. However, traditional machine learning models can only deal with fixed-dimension data in digital format, so it is necessary to encode source sequence data into restricted-length digital data to meet the input requirements of the model, which is the feature extraction of sequential data. Descriptors are used in the first step of biological sequence analysis. They extract various biological sequence features from multiple perspectives, such as amino acid composition, biochemical characteristics, and residue composition, with different emphases and features. Consequently, these algorithms may have different performances for various sequence analysis tasks. Typically, the most appropriate algorithm for a given task needs to be obtained by testing various feature extraction algorithms on the dataset and comparing the performance of each algorithm.

Treatment of Data Imbalance

As mentioned above, the number of positive samples in the dataset used in this paper is only one 10th of the number of negative samples, which will lead to unbalanced recognition of positive and negative samples in training process and affect the final classification results (Zou et al., 2016; Cheng et al., 2018; Azad et al., 2019; Priya and Sivaraj, 2021; Shao et al., 2021). The model trained with unbalanced data will be more inclined to fit the negative instances with a large number, which will lead to the degradation of the model's classification performance for the small number of positive samples. Since there are more negative samples in the dataset than positive samples, if the source files are directly used for training, the classifier will learn too many negative samples, thus reduce the recognition ability of the model for positive samples, but this is contrary to our main purpose. Therefore, it is necessary to adopt some strategies to alleviate the problem of sample imbalance. The relatively small number of specific proteins in nature and the widespread sample imbalance in the field of biological sequence classification had also led to abundant research on the processing of unbalanced data. (Chao et al., 2019; Kaur et al., 2019; Yang et al., 2020; Ao et al., 2021a; Shao and Liu, 2021) The most common are oversampling and downsampling. Oversampling is balanced by adding redundant samples to a small number of positive samples, and the strategy can improve the recognition ability of positive samples to a certain extent, but it simply repeats positive examples and overemphasizes existing positive examples, which would urge the risk of overfitting positive examples. In the downsampling method, only a portion of the negative samples is selected for lower sampling to reduce the number of negative samples. However, this method can only improve the model's classification ability of positive samples to a certain extent. Because a few of the counterexample data are discarded, their influence in the overall sample is reduced, which may result in a large deviation model, and greatly affect the overall performance.

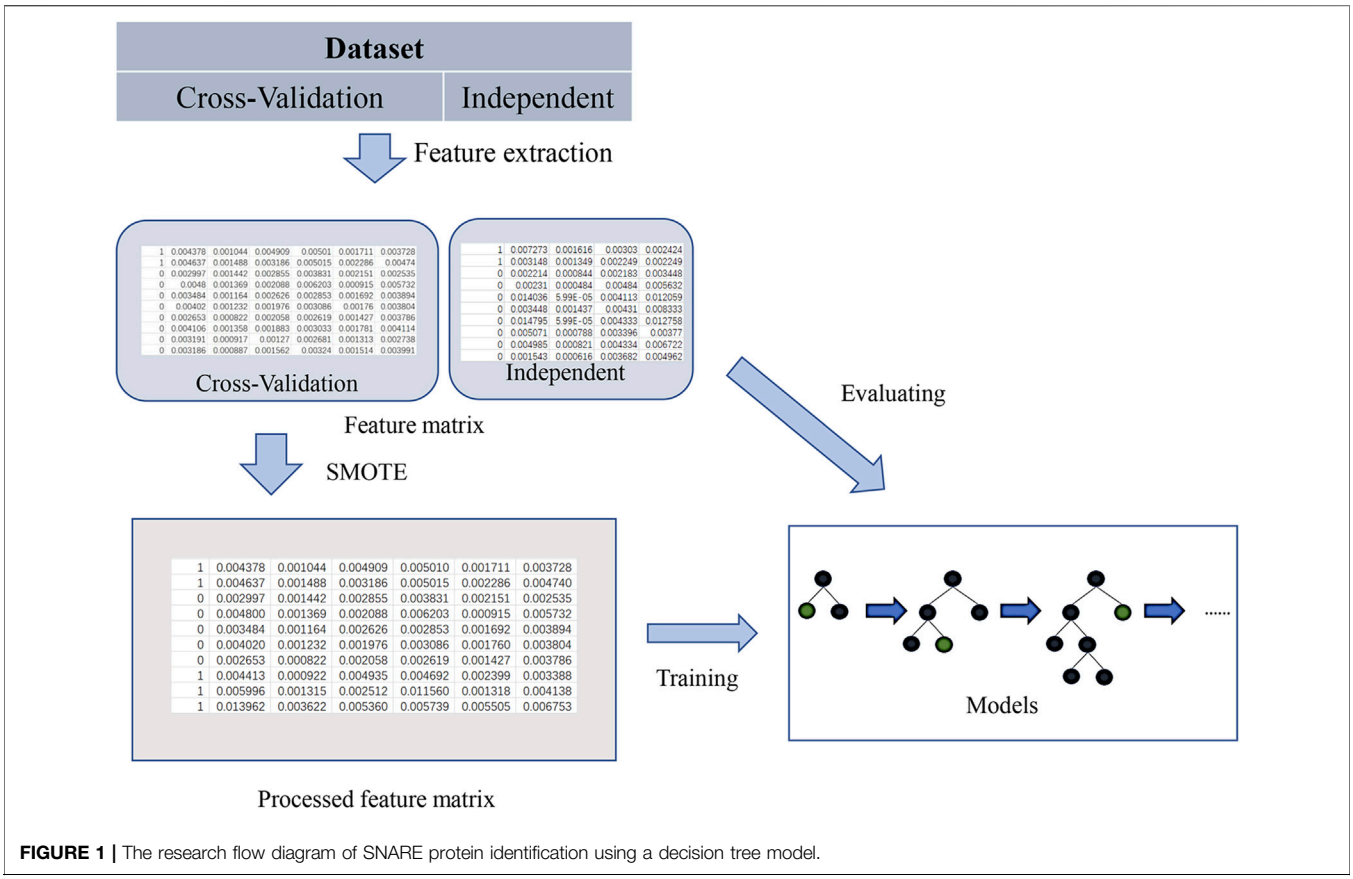


FIGURE 1 | The research flow diagram of SNARE protein identification using a decision tree model.

Considering the serious imbalance between positive and negative samples in this dataset, only one unbalanced strategy may not work well; it needs to be sampled up and down simultaneously. This article uses a combination of sampling partial negative samples and the synthetic minority oversampling technique (SMOTE) to generate new positive samples to address sample imbalance. (Chawla et al., 2002; Riaz and Li, 2019; Zhang C H et al., 2020; Zhao et al., 2020)

SMOTE is an oversampling technique that balances the quantity gap between two categories by finding the nearest neighbor of certain data in a positive example and then using the K-nearest neighbor algorithm to generate new positive samples. For each sample x in the positive sample, calculate the K positive samples x_k {k = 1, 2, K} closest to x, and determine the sampling ratio n according to the unbalanced proportion of samples. For the k nearest neighbor samples of each sample x, n samples are randomly selected, and the newly constructed sample x_{new} can be obtained through the following formula:

$$x_{new} = x + rand(0, 1) * |x - x_n| \tag{1}$$

In the experiment, part of the negative sample is treated with simple undersampling at first. SMOTE is used to generate positive samples to ensure that the number of positive and negative samples is consistent. Then, a balanced dataset of sample size can be obtained, which will be used in subsequent model training experiments.

RESULT AND DISCUSSION

Evaluation Indexes

To objectively evaluate the performance of various algorithms, some convincing indicators of these algorithms need to be compared after the experiment (Wei et al., 2017; Wei et al., 2018; Wei et al., 2019; Wang et al., 2020; Ding et al., 2021; Shang et al., 2021; Wu and Yu, 2021; Yang et al., 2021). Next, the algorithm with the best performance is selected for subsequent research according to these indices. Similarly, common metrics are used to compare the performance of each algorithm. The four values of TP, FP, TN, and FN (representing true positive, false-positive, true negative, and false negative values, respectively) can be obtained for the classifier test (Jiang et al., 2013; Cheng et al., 2016; Xiao et al., 2019; Zhang L et al., 2020; Huang et al., 2020; Li and Liu, 2020; Liu et al., 2020; Mo et al., 2020; Tang et al., 2020; Han et al., 2021; Wang et al., 2021; Xu et al., 2021). Accuracy, MCC, sensitivity, and specificity can then be calculated based on these values.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

TABLE 1 | Feature dimensions of partial feature extraction algorithms and AUROC performance under multiple classifiers.

	Feature dimension	RandomForest (Breiman, 2001)	LightGBM (Ke et al., 2017)	XGBoost (Chen and Guestrin, 2016)
ASDC (Wei et al., 2018)	400	**0.8599**	**0.8829**	**0.8839**
QSOrder (Chou, 2000)	44	0.8401	0.864	0.8604
DDE (Saravanan and Gautham, 2015)	400	0.824	0.8604	0.849
CKSAAP (Chen et al., 2007)	1,600	0.8337	0.8664	0.8588
AAC (Bhasin and Raghava, 2004)	20	0.8467	0.8514	0.8428

The meaning of the bold values is the feature extraction algorithm that performs best under a particular classification algorithm.

TABLE 2 | Model performance under different n values.

n	Cross-validation				Independent			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
628	82.97	82.954	82.486	0.6522	94.2	60.84	63.69	0.3102
1,256	95.148	89.886	92.516	0.8523	73.91	76.56	76.33	0.3152
2,510	98.486	91.832	95.158	0.9055	76.81	86.34	85.5	0.4492
3,764	99.07	94.394	96.73	0.936	65.22	93.22	90.83	0.5071
5,019	99.302	94.682	96.99	0.941	62.32	94.04	91.33	0.5081
6,640	99.292	94.414	96.852	0.9384	59.42	94.99	91.95	0.5149

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

Selection of the Descriptors

In this paper, the iLearnPlus platform (Chen Z et al., 2021) was used to compare the performance of various extraction algorithms: multiple descriptors were applied to obtain the feature vectors of the source FASTA file, followed by training and testing the obtained features using several classification algorithms and analyzing the performance of different feature extraction algorithms. To eliminate the influence of other subjective factors, the area under the receiver operating characteristic curve (AUROC) index was adopted to evaluate the performance of the algorithm.

Accuracy and MCC are widely adopted to measure model performance in classification problems. These two values can be regulated by artificially setting thresholds so that the specific performance of each algorithm cannot be truly reflected. The AUROC index takes TPR [TP/(TP + FN)] and FPR[FP/(FP + TN)] as the horizontal and vertical coordinates to obtain the area under the curve. The larger the area is, the higher the coincidence degree between the prediction label of the model and the source label is. It is necessary to take the AUROC as the evaluation standard so that the algorithm with the best overall performance can be selected.

According to the experiment, several feature extraction algorithms and classifiers with better performance can be obtained. Some experimental results are shown in **Table 1**.

The experimental results show that the performance of adaptive skip dipeptide composition (ASDC), CKSAAP, and QSOrder feature extraction algorithms outperform other algorithms. Among them, the optimal algorithm is the ASDC, and the subsequent multiple numbers also use ASDC to extract features.

ASDC is a feature extraction algorithm based on GDC (G-gap dipeptide composition) algorithm. Dipeptide composition is the fraction of any two adjacent residues as a dipeptide pair, and it measures the correlation of any two adjacent residues in the peptide sequence. GDC encapsulates the composition and local order information of any two spacer residues in the peptide sequence, it has a hyperparameter g to determine the gap between two adjacent residues. And ASDC calculates all values of g and accumulates them. For a given protein read R with L length, the feature vector for ASDC is represented by:

$$ASDC = \left(fv_1, fv_2 ..., fv_{400}\right) \quad (6)$$

where fv_i is calculated by

$$fv_i = \frac{\sum_{g=1}^{L-1} O_i^g}{\sum_{i=1}^{400} \sum_{g=1}^{L-1} O_i^g} \quad (7)$$

where g represents the g-gap (g = 1, 2, L-1) dipeptide and fv_i is the occurrence frequency of the ith (i = 1, 2, 400) adaptive skip dipeptide. It is worth mentioning that if the cumulative term with g is removed from **Eq. 7**, it becomes the formula for the GDC features.

Since there are approximately 8,000 samples in the dataset, the 400 dimension is relatively moderate. Another is that ASDC considers the frequency of any two unconnected amino acids in the whole protein and can capture all the information of dipeptide composition. It also shows that the SNARE proteins have a high correlation with their dipeptide composition. This information may bring biological assistance to the final SNARE protein recognition.

TABLE 3 | The performance of the three classifiers on the independent test set (n = 2,510).

n = 2,510	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
RandomForest	63.77	**90.92**	**88.6**	0.444
LightGBM	**76.81**	86.31	85.5	**0.4492**
XGBoost	73.91	86.99	85.87	0.4412

The meaning of the bold values is the feature extraction algorithm that performs best under a particular classification algorithm.

TABLE 4 | The performance of the three classifiers on the independent test set (n = 5,019).

n = 5,019	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
RandomForest	46.38	95.66	91.45	0.435
LightGBM	**60.87**	**95.39**	**92.44**	**0.5386**
XGBoost	**60.87**	94.58	91.7	0.5132

The meaning of the bold values is the feature extraction algorithm that performs best under a particular classification algorithm.

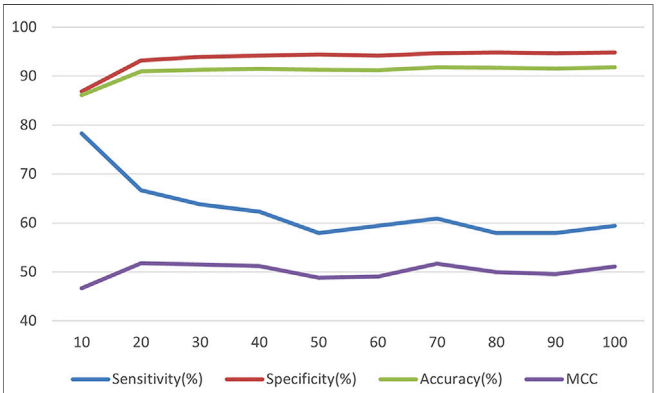


FIGURE 2 | The relationship between the number of leaves and model performance.

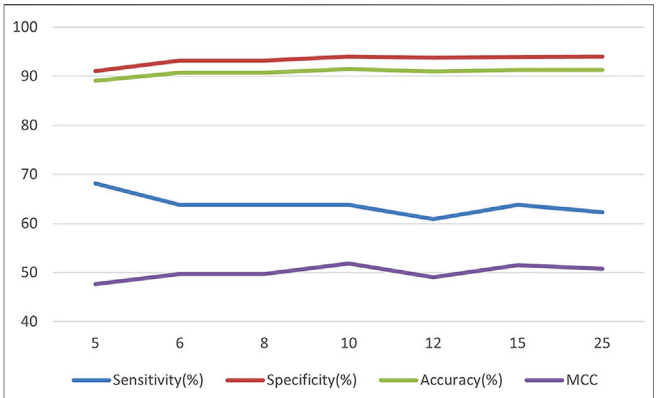


FIGURE 3 | The relationship between the number of maxdepth and model performance.

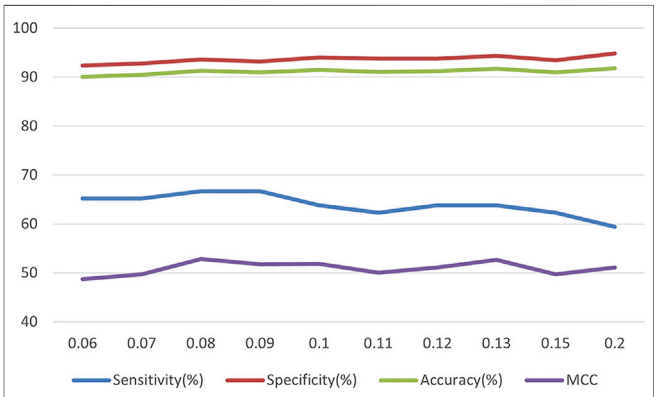


FIGURE 4 | The relationship between learning rate and model performance.

However, the results also showed that several other algorithms performed only slightly worse than ASDC, so it was considered that features stitched together after using multiple feature extraction algorithms could be used to train the model. After experimental verification, when the feature data extracted by algorithms such as ASDC and QSOrder were spliced together and then used to train the model, it was found that instead of improving the results, there was a slight decrease. In response to this result, it is believed that the data dimensionality is too large, and the resulting redundant data will not only have a positive effect on the training of the model but also degrade the model performance. Therefore, the spliced features were subsequently selected again, and relevant experiments were conducted. However, the model trained with these data still performed poorly on the independent set. After comparing the feature vectors extracted by the feature extraction algorithms used, it was concluded that the main reason was that the feature values obtained by each algorithm did not fall within the same range of values. For example, the feature matrix extracted by the QSOrder algorithm is a sparse matrix containing a large number of 0 or very close to 0 values, and there are some negative numbers in the DDE features, which when mixed together may affect the direction of the model iteration and thus the final results.

Unbalanced Processing

In the step of dealing with the data imbalance problem, n negative samples are first downsampled from the original dataset to ensure that n is greater than the number of positive samples 628. Then, the SMOTE algorithm is used to expand the number of positive samples to n to build a balanced dataset. When n = 628, the strategy is equivalent to complete downsampling, and when n = 6,640 (the total number of negative samples), the strategy is equivalent to complete oversampling, so the value of n is in the range (628, 6,640). After sampling the negative samples, all data were tested with the same independent test set to determine their generalization ability.

To analyze the effect of the number of down samples n on the classification performance, several sets of parameters were set for experiments in this paper, and the best performing n value was selected based on the results. n values were set, and the related performance is shown in **Table 2**. To partially eliminate the error caused by the randomness of the data, no put-back sampling was

TABLE 5 | Comparison with the experimental results of 2D CNN in the same setting.

Classifier	Cross-validation				Independent			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
2D CNN (Tao et al., 2020)	76.6	93.5	89.7	0.7	65.8	90.3	87.9	0.46
This Methods	98.168	90.736	94.718	0.8974	81.58	94.84	93.54	0.6839

performed in downsampling, and the set of negative samples sampled was denoted by S (n). Then, there was S(n)⊂S(m), where n < m.

Parameter Optimization

In the recognition problem, it is also very important to select the appropriate classifier. There are also multiple classifiers in the field, each with a different focus, so their performance in a particular task may be different. Therefore, to select a classifier that best fits the task, we follow the same approach as in the selection of the descriptors subsection, where different classifiers are used to train and classify the same feature data, and the best performing algorithm is selected for subsequent experiments. After using three mainstream classifiers, the model performance corresponding to the parameters of Part n is shown in **Tables 3** and **4**. It can be concluded that LightGBM with n = 5,019 is the best performer and most in line with this task. LightGBM (Light Gradient Boosting Machine) is a framework for implementing the GBDT (Gradient Boosting Decision Tree) algorithm, which is an iterative decision tree algorithm consisting of multiple decision trees. LightGBM improves on the traditional GBDT algorithm in many ways, such as using a Histogram-based decision tree algorithm and using a leaf-wise strategy instead of level-wise.

In this experiment, the number of leaf nodes, the maximum depth of the tree and the learning rate of the LightGBM algorithm were adjusted (Ao et al., 2021b). First, we compared the impact of the number of leaf nodes of the tree on the performance of the algorithm when the maximum depth of the tree was not limited. The result is shown in **Figure 2** (The MCC values in the figure have been normalized with the other three indicators for plotting purposes, and the following similar charts have been followed in the same way). Through a series of comparative experiments, the number of leaf nodes can be set to 31 while considering the efficiency of the algorithm operation.

This is followed by choosing the depth of the tree given the number of leaf nodes, as there is a maxdepth>2leaves-1 constraint, and the leaf value has been set to 31; the maximum depth of the tree cannot be less than 5 (log2 (31 + 1). The result is shown in **Figure 3**. Similarly, the optimal maxdepth can be chosen as 10.

Then, it is time to adjust the learning rate and compare the impact of changes in the learning rate on performance, and the results are shown in **Figure 4**. In the end, the optimal parameters are leaves = 31, maxdepth = 10, and learning rate = 0.08.

Comparison With the Other Method

In comparison with 2D CNN, the data of this paper needed to be modified because the data allocation differed. It used a cross-validation set of 644 positive and 2,234 counterexamples and an independent dataset of 38 positive and 349 counterexamples. Similar experiments were conducted using this setup in this paper. In this sequence classification task, the focus is on the classification performance of the SNARE protein, which in the model performance evaluation is the size of the specificity. The experimental results are shown in **Table 5**. It can be found that all the metrics performed better except for the specificity on the cross-validation set, which was slightly weaker than 2D CNN, and the method had an AUROC value of 0.9671 under the independent set, which further proves that the algorithm in this paper has a high generalization capability. The main reason for this result is that the original paper used more positive samples for training the model, with fewer positive examples remaining to evaluate the applicability of the model. However, a set partitioning ratio of 9:1 (cross validation dataset: independent dataset) was applied in this experiment, and although this may lead to some performance loss, the best results obtained in the independent dataset were still good: sensitivity of 66.67%, specificity of 93.63%, accuracy of 91.33%, and MCC of 0.528.

CONCLUSION

In this paper, we used the SMOTE algorithm with different parameters to address the sample imbalance of the dataset. The results show that this strategy can obtain a better result in terms of managing sample imbalance. In this process, ASDC as the feature extraction algorithm and LightGBM as the classification algorithm by comparing the results of various algorithms and descriptors. The combination obtained the best performance, and compared to other advanced neural networks, it achieved a significant improvement in all the typical measurement indexes. Under the same experimental setup, the method in this paper improves the accuracy by 5.64% in the independent test set and 0.2239 in the MCC metric relative to 2D-CNN. For the future research, graph neural networks (Zeng et al., 2020; Chen Y et al., 2021) and unsupervised learning (Xu et al., 2019a; Xu et al., 2019b) can be considered for performance improvement.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/khanhlee/snare-cnn.

AUTHOR CONTRIBUTIONS

DM conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and tables, performed the computation work, authored or reviewed drafts of the paper. ZC conceived and designed the experiments, authored or reviewed drafts of the paper. ZH and XH performed the experiments, analyzed the data, revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Chinese National Natural Science Foundation under Grant 61876047.

REFERENCES

Ao, C., Yu, L., and Zou, Q. (2021). Prediction of Bio-Sequence Modifications and the Associations with Diseases. *Brief. Funct. Genomics* 20 (1), 1–18. doi:10.1093/bfgp/elaa023

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.)* S1046-2023 (21), 00142. doi:10.1016/j.ymeth.2021.05.016

Azad, M. T. A., Qulsum, U., and Tsukahara, T. (2019). Comparative Activity of Adenosine Deaminase Acting on RNA (ADARs) Isoforms for Correction of Genetic Code in Gene Therapy. *Cgt* 19 (1), 31–39. doi:10.2174/1566523218666181114122116

Bhasin, M., and Raghava, G. P. S. (2004). Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. *J. Biol. Chem.* 279 (22), 23262–23266. doi:10.1074/jbc.m401932200

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

Chao, L., Wei, L., and Zou, Q. (2019). SecProMTB: A SVM-Based Classifier for Secretory Proteins of *Mycobacterium tuberculosis* with Imbalanced Data Set. *Proteomics* 19, e1900007.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *jair* 16, 321–357. doi:10.1613/jair.953

Chen, K., Kurgan, L. A., and Ruan, J. (2007). Prediction of Flexible/rigid Regions from Protein Sequences Using K-Spaced Amino Acid Pairs. *BMC Struct. Biol.* 7 (1), 25–13. doi:10.1186/1472-6807-7-25

Chen, T., and Guestrin, C. (2016). in Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, California, San Francisco, USA, August 13 - 17, 2016 Association for Computing Machinery. 785–794.

Chen, Y., Ma, T., Yang, X., Wang, J., Song, B., and Zeng, X. (2021). MUFFIN: Multi-Scale Feature Fusion for Drug-Drug Interaction Prediction. *Bioinformatics* 37, 2651–2658. doi:10.1093/bioinformatics/btab169

Chen Z, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., et al. (2021). iLearnPlus: a Comprehensive and Automated Machine-Learning Platform for Nucleic Acid and Protein Sequence Analysis, Prediction and Visualization. *Nucleic Acids Res.* 49 (10), e60. doi:10.1093/nar/gkab122

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., 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

Chou, K.-C. (2000). Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochem. biophysical Res. Commun.* 278 (2), 477–483. doi:10.1006/bbrc.2000.3815

Ding, Y., Yang, C., Tang, J., and Guo, F. (2021). Identification of Protein-Nucleotide Binding Residues via Graph Regularized K-Local Hyperplane Distance Nearest Neighbor Model. *Applied Intelligence*, 1–15. doi:10.1007/s10489-021-02737-0

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*. 1. btab011. doi:10.1093/bioinformatics/btab011

Hu, L., Huang, T., Shi, X., Lu, W.-C., Cai, Y.-D., and Chou, K.-C. (2011). Predicting Functions of Proteins in Mouse Based on Weighted Protein-Protein Interaction Network and Protein Hybrid Properties. *PloS one* 6 (1), e14556. doi:10.1371/journal.pone.0014556

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

Jahn, R., and Scheller, R. H. (2006). SNAREs - Engines for Membrane Fusion. *Nat. Rev. Mol. Cel Biol* 7 (9), 631–643. doi:10.1038/nrm2002

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

Kaur, H., Pannu, H. S., and Malhi, A. K. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning. *ACM Comput. Surv.* 52 (4), 1–36. doi:10.1145/3343440

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* 30, 3146–3154.

Kinjo, A. R., and Nakamura, H. (2012). Composite Structural Motifs of Binding Sites for Delineating Biological Functions of Proteins. *PloS one* 7 (2), e31437. doi:10.1371/journal.pone.0031437

Le, N. Q. K., and Nguyen, V.-N. (2019). SNARE-CNN: a 2D Convolutional Neural Network Architecture to Identify SNARE Proteins from High-Throughput Sequencing Data. *PeerJ Comp. Sci.* 5, e177. doi:10.7717/peerj-cs.177

Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N., Chua, M. C. H., and Yeh, H.-Y. (2019). Computational Identification of Vesicular Transport Proteins from Sequences Using Deep Gated Recurrent Units Architecture. *Comput. Struct. Biotechnol. J.* 17, 1245–1254. doi:10.1016/j.csbj.2019.09.005

Li, C.-C., and Liu, B. (2020). MotifCNN-fold: Protein Fold Recognition Based on Fold-specific Features Extracted by Motif-Based Convolutional Neural Networks. *Brief. Bioinform.* 21 (6), 2133–2141. doi:10.1093/bib/bbz133

Li, H.-L., Pang, Y.-H., and Liu, B. (2021). BioSeq-BLM: a Platform for Analyzing DNA, RNA and Protein Sequences Based on Biological Language Models. *Nucleic Acids Res.* 1. gkab829. doi:10.1093/nar/gkab829

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., Li, C.-C., and Yan, K. (2020). DeepSVM-fold: Protein Fold Recognition by Combining Support Vector Machines and Pairwise Sequence Similarity Scores Generated by Deep Learning Networks. *Brief. Bioinform.* 21 (5), 1733–1741. doi:10.1093/bib/bbz098

Mo, F., Luo, Y., Fan, D., Zeng, H., Zhao, Y., Luo, M., et al. (2020). Integrated Analysis of mRNA-Seq and miRNA-Seq to Identify C-MYC, YAP1 and miR-3960 as Major Players in the Anticancer Effects of Caffeic Acid Phenethyl Ester in Human Small Cell Lung Cancer Cell Line. *Cgt* 20 (1), 15–24. doi:10.2174/1566523220666200523165159

Priya, R. D., and Sivaraj, R. (2021). Gene Selection in Multi-Class Imbalanced Microarray Datasets Using Dynamic Length Particle Swarm Optimization. *Cbio* 16 (5), 734–748. doi:10.2174/1574893615999201002093834

Riaz, F., and Li, D. (2019). Non-coding RNA Associated Competitive Endogenous RNA Regulatory Network: Novel Therapeutic Approach in Liver Fibrosis. *Cgt* 19 (5), 305–317. doi:10.2174/1566523219666191107113046

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

Schoch, S., Dea´k, F., Königstorfer, A., Mozhayeva, M., Sara, Y., Südhof, T. C., et al. (2001). SNARE Function Analyzed in Synaptobrevin/VAMP Knockout Mice. *Science* 294 (5544), 1117–1122. doi:10.1126/science.1064335

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 (3), 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 (3), bbaa144. doi:10.1093/bib/bbaa144

Sharma, A. K., and Srivastava, R. (2021). Protein Secondary Structure Prediction Using Character Bi-gram Embedding and Bi-LSTM. *Cbio* 16 (2), 333–338. doi:10.2174/1574893615999200601122840

Su, X., Xu, J., Yin, Y., Quan, X., and Zhang, H. (2019). Antimicrobial Peptide Identification Using Multi-Scale Convolutional Network. *BMC bioinformatics* 20 (1), 730–810. doi:10.1186/s12859-019-3327-y

Sundar, G. N., and Narmadha, D. (2021). An Automated Model for Target Protein Prediction in PPI. *Cbio* 16 (4), 601–609. doi:10.2174/1574893615999200831142241

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

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, X., Yang, Y., Liu, J., and Wang, G. (2021). The Stacking Strategy-Based Hybrid Framework for Identifying Non-coding RNAs. *Brief Bioinform* 22 (5), bbab023. doi:10.1093/bib/bbab023

Wang, Y.-N., Figueiredo, D., Sun, X.-D., Dong, Z.-Q., Chen, W.-B., Cui, W.-P., et al. (2018). Controlling of Glutamate Release by Neuregulin3 via Inhibiting the Assembly of the SNARE Complex. *Proc. Natl. Acad. Sci. USA* 115 (10), 2508–2513. doi:10.1073/pnas.1716322115

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., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019). Fast Prediction of Protein Methylation Sites Using a Sequence-Based Feature Selection Technique. *Ieee/acm Trans. Comput. Biol. Bioinf.* 16 (4), 1264–1273. doi:10.1109/tcbb.2017.2670558

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.

Xiao, X., Xu, Z.-C., Qiu, W.-R., Wang, P., Ge, H.-T., and Chou, K.-C. (2019). iPSW(2L)-PseKNC: A Two-Layer Predictor for Identifying Promoters and Their Strength by Hybrid Features via Pseudo K-Tuple Nucleotide Composition. *Genomics* 111 (6), 1785–1793. doi:10.1016/j.ygeno.2018.12.001

Xu, H., Zeng, W., Zeng, X., and Yen, G. G. (2019). An Evolutionary Algorithm Based on Minkowski Distance for many-objective Optimization. *IEEE Trans. Cybern.* 49 (11), 3968–3979. doi:10.1109/tcyb.2018.2856208

Xu, H., Zeng, W., Zhang, D., and Zeng, X. (2019). MOEA/HD: A Multiobjective Evolutionary Algorithm Based on Hierarchical Decomposition. *IEEE Trans. Cybern.* 49 (2), 517–526. doi:10.1109/tcyb.2017.2779450

Xu, Z., Luo, M., Lin, W., Xue, G., Wang, P., Jin, X., et al. (2021). DLpTCR: an Ensemble Deep Learning Framework for Predicting Immunogenic Peptide Recognized by T Cell Receptor. *Brief Bioinform* 22, bbab335. doi:10.1093/bib/bbab335

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 (17), 11387–11399. doi:10.1007/s00521-020-05573-4

Yang, X.-F., Zhou, Y.-K., Zhang, L., Gao, Y., and Du, P.-F. (2020). Predicting LncRNA Subcellular Localization Using Unbalanced Pseudo-k Nucleotide Compositions. *Cbio* 15 (6), 554–562. doi:10.2174/1574893614666190902151038

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. *Cgt* 20 (2), 84–99. doi:10.2174/1566523220666200628014530

Zhang, L., Xiao, X., and Xu, Z.-C. (2020). iPromoter-5mC: A Novel Fusion Decision Predictor for the Identification of 5-Methylcytosine Sites in Genome-wide DNA Promoters. *Front. Cel Dev. Biol.* 8, 614. doi:10.3389/fcell.2020.00614

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36 (16), 4466–4472. doi:10.1093/bioinformatics/btaa428

Zou, Q., Xie, S., Lin, Z., Wu, M., and Ju, Y. (2016). Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Res.* 5, 2–8. doi:10.1016/j.bdr.2015.12.001

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.

Copyright © 2022 Ma, Chen, He and Huang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 February 2022
doi: 10.3389/fgene.2022.856075

Analysis of Breast Cancer Based on the Dysregulated Network

Yanhao Huo[1], *Xianbin Li*[1], *Peng Xu*[1,2], *Zhenshen Bao*[1,2]* *and Wenbin Liu*[1]*

[1]*Institute of Computational Science and Technology, Guangzhou University, Guangzhou, China,* [2]*School of Computer Science of Information Technology, Qiannan Normal University for Nationalities, Duyun, China*

Breast cancer is a heterogeneous disease, and its development is closely associated with the underlying molecular regulatory network. In this paper, we propose a new way to measure the regulation strength between genes based on their expression values, and construct the dysregulated networks (DNs) for the four subtypes of breast cancer. Our results show that the key dysregulated networks (KDNs) are significantly enriched in critical breast cancer-related pathways and driver genes; closely related to drug targets; and have significant differences in survival analysis. Moreover, the key dysregulated genes could serve as potential driver genes, drug targets, and prognostic markers for each breast cancer subtype. Therefore, the KDN is expected to be an effective and novel way to understand the mechanisms of breast cancer.

Keywords: breast cancer, dysregulated network, cancer-related pathways, driver genes, drug targets, survival analysis

OPEN ACCESS

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

Reviewed by:
Cheng Zhang,
Peking University, China
Yansen Su,
Anhui University, China
Yanfeng Wang,
Zhengzhou University of Light Industry, China

***Correspondence:**
Wenbin Liu
wbliu6910@gzhu.edu.cn
Zhenshen Bao
bzsbao@163.com

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

Received: *16 January 2022*
Accepted: *28 January 2022*
Published: *15 February 2022*

Citation:
Huo Y, Li X, Xu P, Bao Z and Liu W (2022) Analysis of Breast Cancer Based on the Dysregulated Network. Front. Genet. 13:856075. doi: 10.3389/fgene.2022.856075

INTRODUCTION

According to global cancer statistics in 2020, Breast cancer has become the most common cancer, with 2.3 million new cases (Sung et al., 2021). As a heterogenetic malignancy, breast cancer can be classified into four subtypes: Luminal A, Luminal B, Basal-like, and Her2-enriched (Cheang et al., 2009; Inic et al., 2014). Although significant improvements have been achieved, a better understanding of genetic changes will lead to better diagnosis and treatment of this disease (Liang et al., 2021).

The genetic variation of driver genes has been considered as one of cancer's most critical intrinsic factors (Akhavan-Safar et al., 2021). Thus, many computational tools have been developed to identify potential driver genes. For example, MaxDriver developed by Chen et al. detect driver genes based on the maximum information flow in the heterogeneous network (Chen et al., 2013). DawnRank can directly prioritize the driver genes at the individual patient level (Hou and Jian, 2014). And Shi et al. proposed a network diffusion method to identify driver genes (Shi et al., 2016). Among these tools, DriverNet is probably the most competitive tool which considers both gene mutation and abnormal expressions of downstream genes (Bashashati et al., 2012).

Differentially expressed genes (DEGs) analysis is used to identify potential biomarkers or prognostic markers for breast cancer (Yang et al., 2019). Based on DEGs and the survival analysis of hub genes in protein-protein interaction network (PPI), Wu et al. identified that ESR1 and PGR may be potential prognostic markers of ER-positive breast cancer (Wu et al., 2020). Huan et al. found that estradiol (E2) is a biomarker of breast cancer based on the analysis of DEGs in the PPI network (Huan et al., 2014). Furthermore, Eskandari et al. constructed a gene regulatory network by common DEGs to identify the key therapeutic targets for each subtype of breast cancer (Eskandari and Motalebzadeh, 2019).

TABLE 1 | The details of datasets and network.

Datasets			Number of genes	Number of samples	Number of interactions
Gene expression (TCGA-BRCA)	Normal	—	34,127	99	—
	Tumor	Luminal A		225	—
		Luminal B		123	—
		Basal-like		97	—
		HER2-enriched		57	—
Genomic aberrations	—	gene-level copy number alteration	24,776	1081	—
		somatic mutation (SNP and INDEL)	40,543	792	—
Influence network	—	—	9728	—	146171

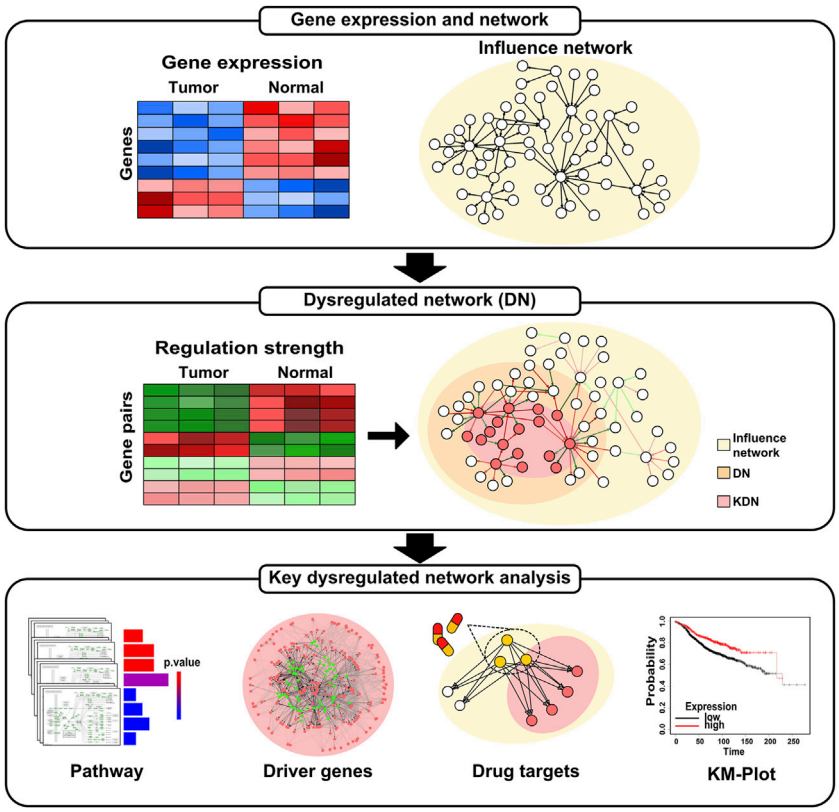


FIGURE 1 | Overview of the analysis workflow.

In summary, previous works have started from the single and independent abnormal expression of genes, but ignored the importance of changes in the interactions between genes. Actually, cancer occurs because of abnormal interactions between genes that lead to their abnormal expressions (Peng et al., 2012; Gao et al., 2013; Bao et al., 2016; Bao et al., 2020; Chai et al., 2022). In this paper, we propose a new way to measure the regulation strength between genes based on their relative expression values. Then the dysregulated network (DN) can be determined by the dysregulated interactions between normal and disease samples. Results show that not only is the key dysregulated network (KDN) enriched in many potential breast cancer related-pathways and important driver genes, but is also closely related to drug targets. Therefore, the proposed KDN provides a new tool for elucidating the underlying mechanism and potential drug repurposing for breast cancer.

MATERIALS AND METHODS

Materials

Both the gene expression dataset and genomic aberration dataset are downloaded from https://xenabrowser.net/datapages/. Gene expression dataset includes Luminal A, Luminal B, Basal-like, and HER2-enriched subtypes. Genomic aberration dataset includes gene-

TABLE 2 | Overview of the DNs.

Subtypes	Number of genes	Number of interactions	Average degree	Average betweenness
Luminal A	4971	18,771	7.28	6875
Luminal B	5847	26,853	9.29	8374
Basal-like	5573	23,307	8.36	8304
HER2-enriched	5679	24,483	8.62	8412

level copy number alteration and somatic mutation (SNP and INDEL). The somatic mutation dataset is a binary matrix containing the gene-level non-silent mutation. The influence network includes directed gene interactions from KEGG, Reactome, Panther, CellMap, and NCI Pathway Interaction Databases (Wu et al., 2010). The 29 targeted drugs are downloaded from https://www.cancer.gov/about-cancer/treatment/drugs/breast (National Cancer Institute), and their corresponding targets are obtained from https://clue.io/repurposing-app. **Table 1** presents the details of the datasets and network including the number of genes, the number of samples, and the number of interactions.

METHODS

From the perspective of gene regulatory network, it is the significantly abnormal interaction between genes that pushes cells operating from normal state to disease state. Therefore, analysis of dysregulation may help to reveal more biological insights than traditional differentially expressed genes (DEGs). Our motivation is that an upstream gene will have more influence on its downstream genes if the expression of the former is larger than that of the latter, and vice versa. Therefore, we define the regulation strength of gene i to gene j as

$$r_{ij} = \log \frac{g_i}{g_j} \tag{1}$$

where g_i and g_j is the expression value of gene i and gene j respectively. In this paper, we name the network composed of the dysregulated interactions as the dysregulated network (DN).

Then, the average absolute difference of the dysregulated strength ds_{ij} of gene i to gene j can be calculated as

$$ds_{ij} = \left| \bar{r}_{ij}^{D} - \bar{r}_{ij}^{N} \right| \tag{2}$$

where $\bar{r}_{ij}^{D}$ and $\bar{r}_{ij}^{N}$ denote the average regulation strength in the disease and normal state, respectively. Further, the dysregulation score d_i of gene i is defined as the sum of the dysregulated strength to all its downstream genes

$$d_i = \sum_{j=1}^{n_i} ds_{ij} \tag{3}$$

where n_i is the number of the direct downstream genes of gene i. A higher d_i indicates that gene i regulate more downstream genes with the higher ds_{ij}, otherwise, gene i regulate less downstream genes with the lower ds_{ij}. Finally, based on the key genes and their dysregulated interactions, the key dysregulated network (KDN) is obtained.

Figure 1 shows an overview of the analysis workflow for this study. First, we construct the DN based on gene expression data and influence network. Then, we identify the KDN of each subtype. Finally, we conduct pathway enrichment analysis, driver genes analysis, drug targets enrichment analysis, and survival analysis for the obtained KDN.

Identifying Driver Genes through DriverNet

Driver genes can be manifested through the outlying expression of genes in influence network. So, Bashashati et al. (Bashashati et al., 2012) developed a computational framework called DriverNet to identify the potential driver genes effectively. In DriverNet, a bipartite graph is constructed through genomic aberrations matrix, outlier matrix, and influence network. And, based on the bipartite graph, DriverNet could rank the genes according to the number of events (outliers). Then, a set of potential driver genes is obtained.

RESULTS AND DISCUSSION

The Dysregulated Network

For each subtype, the dysregulated interactions are determined with $\mathrm{p-value} \le 10E-4$ and $ds \ge 2$ by Limma package in R. **Table 2** presents the four dysregulated networks (DNs), including the number of genes, the number of interactions, the average degree, and the average betweenness of genes. Only about 50% of the genes and 20% of the interactions from the background influence network constitute the DN. The average degree (≈ 8) and betweenness (≈ 8000) indicate the DN is highly interconnected.

Figure 2A shows the heatmap of the dysregulated interactions in the four breast cancer subtypes. The interactions are ordered according to their observed frequency in subtypes. We color the interaction red when the disease state has a higher average regulation strength of interactions, and green otherwise. Black at the bottom represents that the interactions are not significantly dysregulated in the corresponding subtype. About 50% of the dysregulated interactions are shared by the four subtypes, and the overlapped interactions have the same dysregulation pattern. The abnormal regulation in these gene pairs may form the common mechanisms of the four subtypes. On the other hand, about 10% of the dysregulated interactions appear in only one subtype which may characterize the different phenotypes of the four subtypes at network level. Therefore, the corresponding DN may

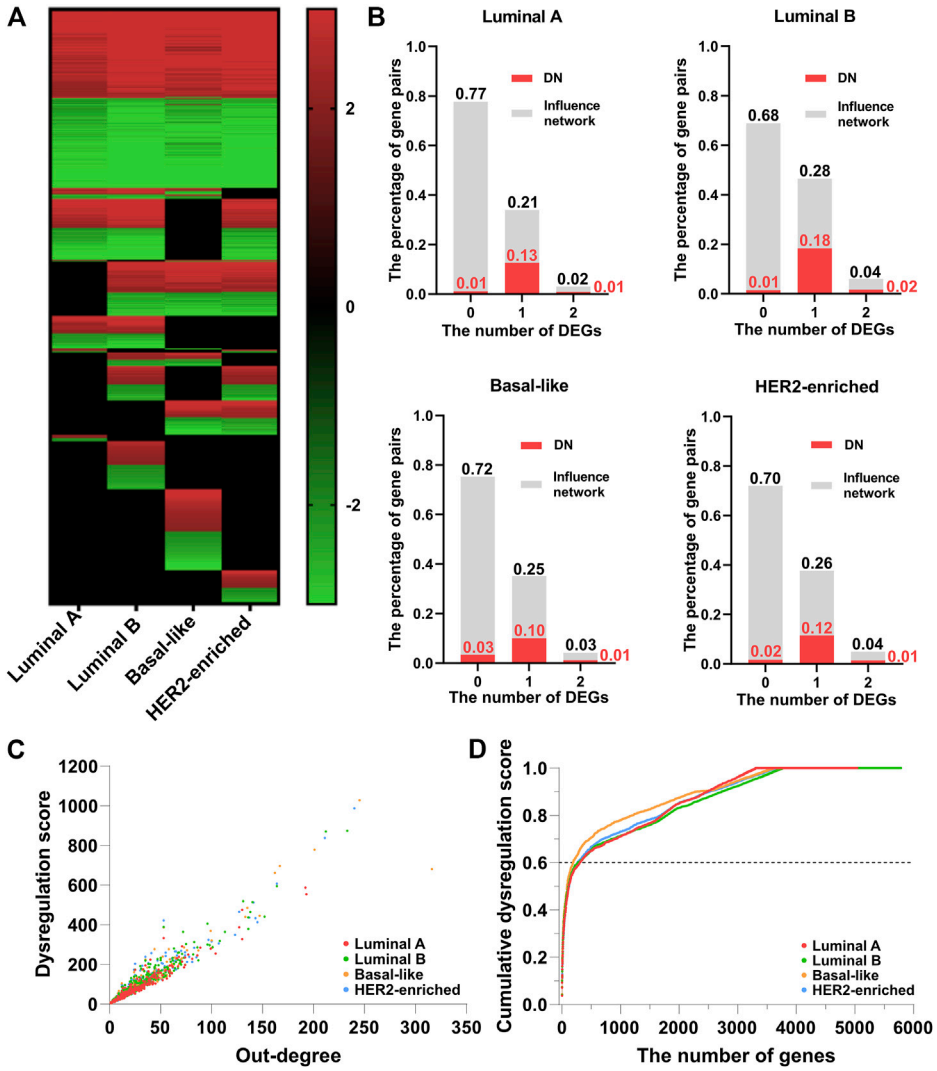


FIGURE 2 | The dysregulated network (DN) of breast cancer. **(A)** The heatmap of the dysregulated interactions in the four breast cancer subtypes. **(B)** The percentage of interactions with 0, 1, and 2 DEGs (gray color) and dysregulated interactions (red color) in the background influence network. **(C)** The scatter plot of the dysregulation score and out-degree of genes in the DN. **(D)** The relationship between the cumulative dysregulation score and the number of genes in the DN.

contribute to the development of the four subtypes of breast cancer.

As the regulation strength measures the dysregulated interactions, it is natural to ask if the dysregulated interactions are essentially caused by differentially expressed genes (DEGs). To examine the impact of DEGs ($p-value \le 10E-4$ and $FC \ge 1.5$) on the dysregulated interactions ($p-value \le 10E-4$ and $ds \ge 2$), **Figure 2B** shows the percentage of interactions with 0, 1, and 2 DEGs in the background network (gray color). And the percentage of dysregulated interactions in each group is shown with red color. Obviously, most of the dysregulated interactions, 40%–65%, come from the group with just one DEG; only a few come from the other groups. In sum, the DN only contains about half of the DEGs. Given that gene i regulates gene j, their interactions may not be significantly abnormal if their expressions change in the same way, even if one or two of them are DEGs. On the other hand, the interactions may be significantly abnormal if their expressions change inversely, even if both are not significantly differentially expressed. Therefore, dysregulated interactions can reveal the regulation abnormality of subtypes, which is hard for DEGs to detect.

Figure 2C shows the scatter plot of the dysregulation score d_i of gene i and its outdegree in the DN. First, the dysregulation score is linearly proportional to the outdegree: genes with larger outdegree tend to have larger dysregulation scores. These genes may play important roles in the DN. Secondly, only a few genes have extraordinarily large dysregulation scores. In order to determine the key genes in the DN, we sort them in descending order according to their dysregulation score. Then, the dysregulation score is normalized by the total score of all

TABLE 3 | Top 30 enrichment pathways of key genes and enrichment pathways of DEGs.

Pathway	*p*-value			
Top 30 enrichment pathways of key genes	**Luminal A**	**Luminal B**	**Basal-like**	**HER2-enriched**
Pathways in cancer	5.55E-40	4.90E-43	1.31E-35	4.20E-45
PI3K-Akt signaling pathway	1.05E-38	4.04E-34	8.13E-24	9.65E-41
Relaxin signaling pathway	1.33E-33	2.68E-31	7.43E-17	6.93E-31
Ras signaling pathway	4.66E-25	1.69E-32	6.26E-16	1.11E-27
Chemokine signaling pathway	4.28E-27	1.81E-19	2.28E-15	4.21E-21
Dopaminergic synapse	8.82E-26	3.41E-21	2.33E-14	7.04E-21
Focal adhesion	3.62E-20	1.07E-19	7.65E-14	2.66E-21
MAPK signaling pathway	9.99E-23	6.18E-26	9.58E-14	1.71E-24
Human cytomegalovirus infection	5.45E-20	1.14E-20	1.77E-13	3.04E-21
Human papillomavirus infection	5.06E-21	2.51E-17	4.80E-13	1.49E-18
Cholinergic synapse	3.94E-17	1.51E-13	1.32E-12	1.88E-16
Kaposi sarcoma-associated herpesvirus infection	2.30E-15	2.17E-19	2.12E-12	2.12E-14
Hepatitis B	5.04E-22	3.71E-20	2.79E-12	7.95E-20
cAMP signaling pathway	5.31E-18	4.87E-14	4.07E-12	3.65E-18
Circadian entrainment	7.08E-16	4.90E-12	3.27E-15	6.26E-14
Human T-cell leukemia virus 1 infection	3.46E-15	7.47E-14	5.72E-12	3.89E-16
Proteoglycans in cancer	9.06E-28	2.45E-22	9.00E-12	3.86E-17
Estrogen signaling pathway	1.45E-13	2.33E-12	7.87E-12	5.46E-16
Lipid and atherosclerosis	1.82E-15	5.79E-15	2.82E-11	1.17E-14
Thyroid hormone signaling pathway	2.68E-12	8.07E-15	5.89E-11	1.35E-12
IL-17 signaling pathway	1.59E-18	1.76E-14	6.83E-11	2.90E-14
Amphetamine addiction	9.51E-18	6.48E-12	6.34E-11	3.03E-15
Parathyroid hormone synthesis, secretion and action	5.32E-18	3.33E-11	5.24E-11	1.51E-19
AGE-RAGE signaling pathway in diabetic complications	8.22E-19	1.86E-19	2.35E-09	5.15E-18
Breast cancer	1.56E-16	1.50E-19	2.40E-09	3.57E-15
Human immunodeficiency virus 1 infection	4.08E-10	3.67E-15	7.31E-09	3.66E-13
Gastric cancer	1.55E-13	2.84E-17	2.44E-08	4.72E-14
Melanogenesis	1.48E-17	1.12E-10	2.77E-08	1.81E-12
Cocaine addiction	1.30E-12	5.58E-10	2.90E-08	7.37E-15
Rap1 signaling pathway	2.55E-20	6.69E-22	3.80E-08	1.64E-21
Growth hormone synthesis, secretion and action	5.34E-23	3.08E-18	3.81E-08	1.67E-22
Oocyte meiosis	7.24E-10	8.85E-08	1.64E-13	4.16E-10
Melanoma	1.99E-09	7.96E-14	2.66E-07	2.25E-11
Oxytocin signaling pathway	3.98E-13	3.26E-07	6.92E-10	5.36E-07
Osteoclast differentiation	1.09E-11	2.93E-17	8.64E-07	4.48E-11
Prolactin signaling pathway	3.78E-15	2.65E-15	1.95E-06	6.51E-14
Longevity regulating pathway	1.78E-20	8.69E-11	3.44E-06	1.13E-12
Morphine addiction	1.47E-08	4.55E-06	4.84E-10	1.32E-07
TNF signaling pathway	7.43E-23	9.76E-16	7.15E-06	1.93E-14
ErbB signaling pathway	2.21E-16	1.93E-14	1.62E-05	4.70E-12
Prion disease	1.29E-04	1.02E-11	4.29E-06	1.70E-14
Insulin resistance	7.64E-19	5.24E-13	1.16E-03	5.36E-09
Proteasome	—	5.97E-13	1.41E-12	6.89E-17
Parkinson disease	—	1.95E-05	8.59E-10	1.15E-09
Enrichment pathways of DEGs				
Cell cycle	2.32E-02	2.25E-14	1.44E-24	6.49E-13
Progesterone-mediated oocyte maturation	5.08E-03	1.20E-04	1.76E-07	1.47E-04
Oocyte meiosis	—	2.38E-07	3.89E-09	3.19E-06
Cellular senescence	—	3.39E-02	3.94E-05	—
Human T-cell leukemia virus 1 infection	—	5.87E-05	—	—
Homologous recombination	—	—	3.03E-03	—

genes and **Figure 2D** shows the curve of the accumulative normalized dysregulation score, which is sorted by gene order. This curve grows rapidly at first and then increases slowly as the cumulative score reaches 60%. The genes contributing to the 60% cumulative dysregulation score only include about 5% of genes in the DN (306, 274, 189, and 267 genes for Luminal A, Luminal B, Basal-like, and HER2-enriched subtypes respectively). In this paper, we refer to the network of these key genes and their dysregulated interactions as the key dysregulated network (KDN).

Key Dysregulated Genes are Enriched in Critical Breast Cancer-Related Pathways

To investigate the biological functions of the KDN, we conduct a pathway enrichment analysis on key dysregulated genes with

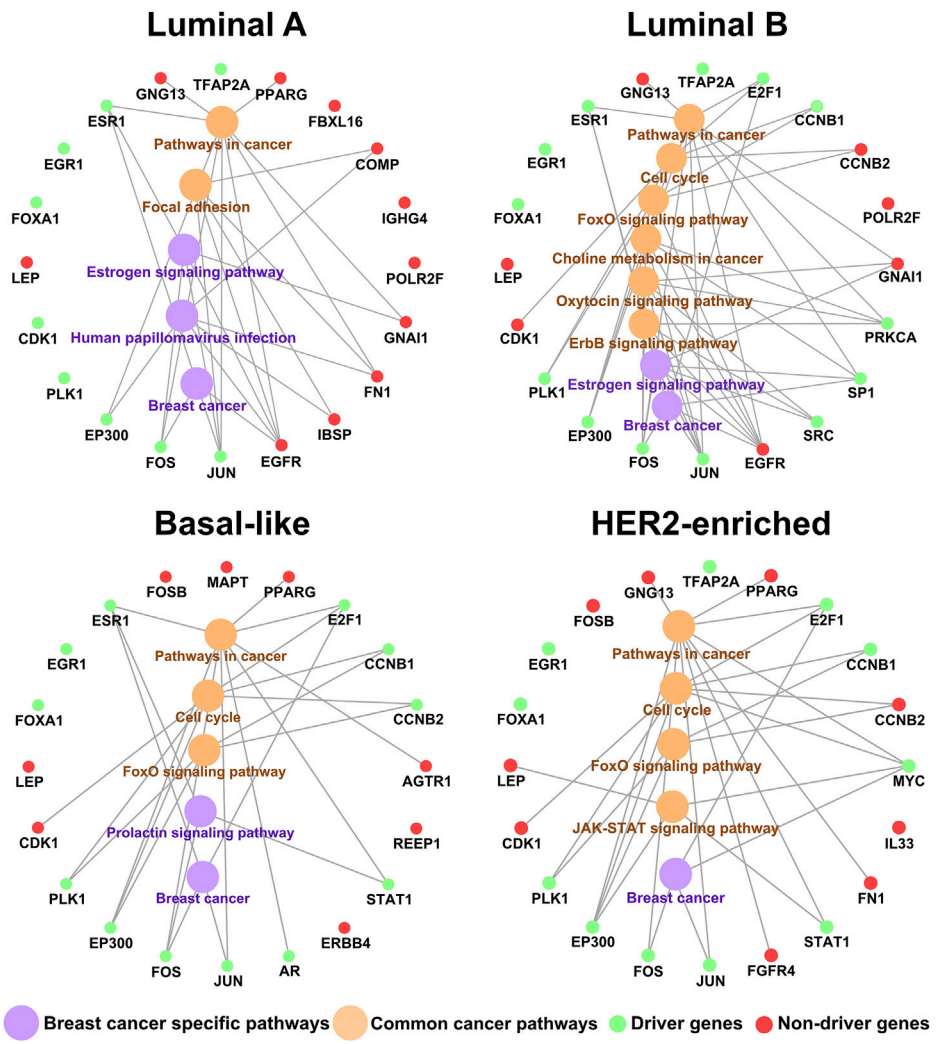


FIGURE 3 | The relation of the top 20 key genes and their enriched pathways.

p − value ≤ 0.05. We also take the same analysis of the top 300 differentially expressed genes (DEGs) obtained by Limma package. **Table 3** lists the top 30 enriched pathways by key genes and DEGs. Key genes are significantly enriched in many well-known breast cancer-related pathways including Pathway in cancer, Ras signaling pathways, MAPK signaling, Estrogen signaling, Breast cancer, Prolactin signaling pathways, etc. However, the top 300 ordinary DEGs are only enriched in very few pathways which are not the critical ones in breast cancer. This comparison suggests that the genes in the KDN are more biologically related to breast cancer than DEGs.

Furthermore, we take an enrichment analysis of the top 20 genes in the KDN. **Figure 3** shows the relation of the top 20 key genes and their enriched pathways by p − value ≤ 0.05. Green and red colors denote driver genes and non-driver genes respectively. Yellow and purple colors denote common cancer pathways and breast cancer specific pathways. Surprisingly, even the top 20 key genes are significantly enriched in some breast cancer specific pathways. Among them, the breast cancer pathway contains four subpathways and connects with many important signaling pathways, such as MAPK pathway, PI3K-Akt pathway, Notch signaling pathway, Wnt signaling pathway, P53 signaling pathway, Cell cycle pathway, etc. For the Estrogen signaling pathway, Tang et al. reported that Estrogen-triggered signaling cascades play an important role in the initiation and development of most human breast cancer (Song and Santen, 2006). In addition, Kitajima et al. also reported that Estrogen and its receptor can regulate the development and progression of breast cancer in most cases (Shin-Ichi et al., 2010). For the Prolactin signaling pathway, a 20-years prospective study has shown that Prolactin can promote proliferation and cell motility in later stage breast tumor development (Tworoger and Hankinson, 2006; Tworoger et al., 2013). For the Human papillomavirus infection pathway, HR-HPV DNA infection exists in breast cancer tissue, thus closely related to the occurrence and development of breast cancer (Wang et al., 2009). And, these pathways have crosstalk with other common cancer pathways (Chen and Wang, 2012; Sato, 2013). For

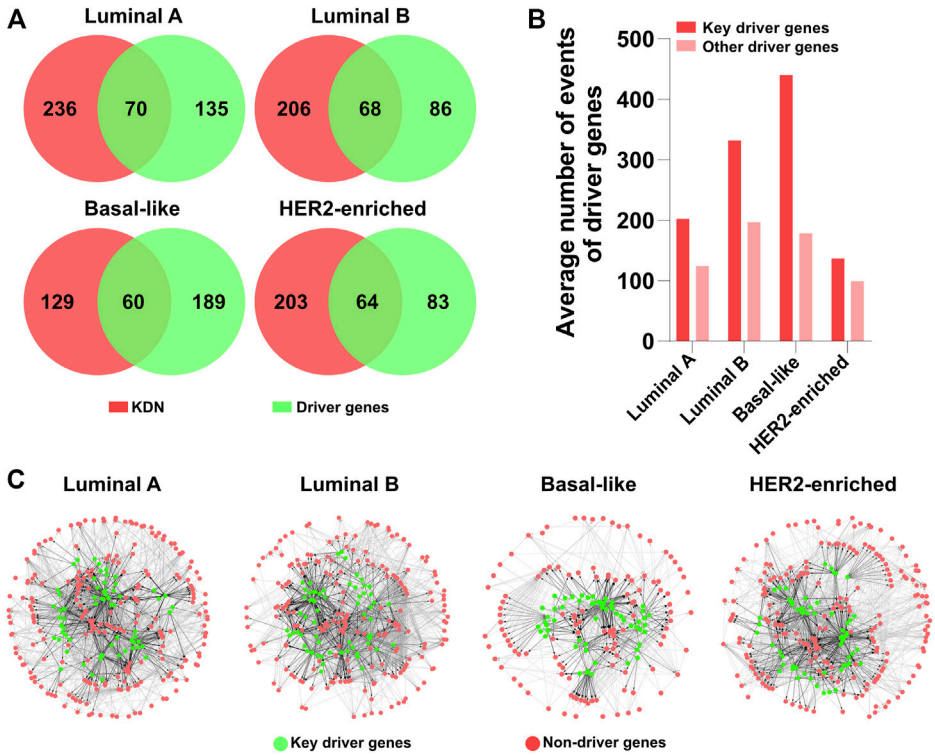


FIGURE 4 | Driver gene analysis. **(A)** Venn diagrams of the key genes and driver genes. **(B)** The average number of events of key driver genes and other driver genes. **(C)** KDN with driver genes in green color.

example, Pathways in cancer, Focal adhesion (Ocak et al., 2010; Gari et al., 2016; Gu et al., 2018; Gong et al., 2019), Cell cycle, FoxO signaling pathway (Mohd et al., 2017; Gong et al., 2020), Choline metabolism in cancer, Oxytocin signaling pathway (Cassoni et al., 1994; Pequeux, 2002; Wang et al., 2020), ErbB signaling pathway (Liu et al., 2008; Aline et al., 2015), and JAK-STAT signaling pathway (Hernández-Vargas et al., 2011; Wang et al., 2018; Na and Balko, 2019).

Based on the biological functions, the driver genes (Bashashati et al., 2012) in common top 20 key genes are highly associated with breast cancer subtypes. As shown in **Figure 3**, EGR1, EP300, FOS, JUN, FOXA1, PLK1, ESR1, and E2F1 are the driver genes for corresponding subtype. As a tumor-suppressor gene in breast cancer, overexpression of EGR1 in breast tumor cells markedly reduces transformed growth and tumorigenicity (Huang et al., 1997; Ronski et al., 2010). EP300 is recruited by the estrogen receptor alpha, a hormone inducible transcription factor, to mediate the mitogen effect of the ovarian steroid estrogen, which is a strong risk factor for breast cancer development (Wirtenberger et al., 2006). The FOS family is one of the AP-1 transcription factors, which regulated many proteins involved in breast cancer invasion (Milde-Langosch et al., 2004). Activated JUN is predominantly expressed at the invasive front in breast cancer and is associated with proliferation and angiogenesis (Vleugel et al., 2006). FOXA1 can influence the expression of a large number of genes in breast cancer associated with metabolic processes, regulation of signaling, and the cell cycle (Wolf et al., 2005). PLK1 mediates estrogen receptor (ER)-regulated gene transcription in human breast cancer cells. And PLK1-coactivated genes include classical ER target genes such as Ps2, Wisp2, and Serpina3 and are enriched in developmental and tumor-suppressive functions (Wierer et al., 2013). ESR1 encodes estrogen receptor-α, which is a major biomarker in the development of breast cancer (Yang et al., 2021). E2F1 expression is regulated by the estrogen receptor α (ERα) to mediate tamoxifen resistance in ERα-positive breast cancer cells (Montenegro and Cancer, 2014). And E2F1 can drive the metastasis of breast cancer (Hollern et al., 2019).

Driver Genes are Enriched in the Key Dysregulated Network

At the genomic level, driver genes are considered to be one of the most important factors in cancer initiation and progression. The driven mutations in the genome provoke abnormal function at protein level and impact the expression of the downstream genes. Therefore, driver genes, as an intrinsic driven regulation mechanism, should also play a critical role in the obtained DN. We apply DriverNet to identify driver genes, and it identifies 205, 154, 249, and 147 driver genes for Luminal A, Luminal B, Basal-like, and HER2-enriched subtypes respectively. We find that about 90% of the determined driver genes are observed in the dysregulation network.

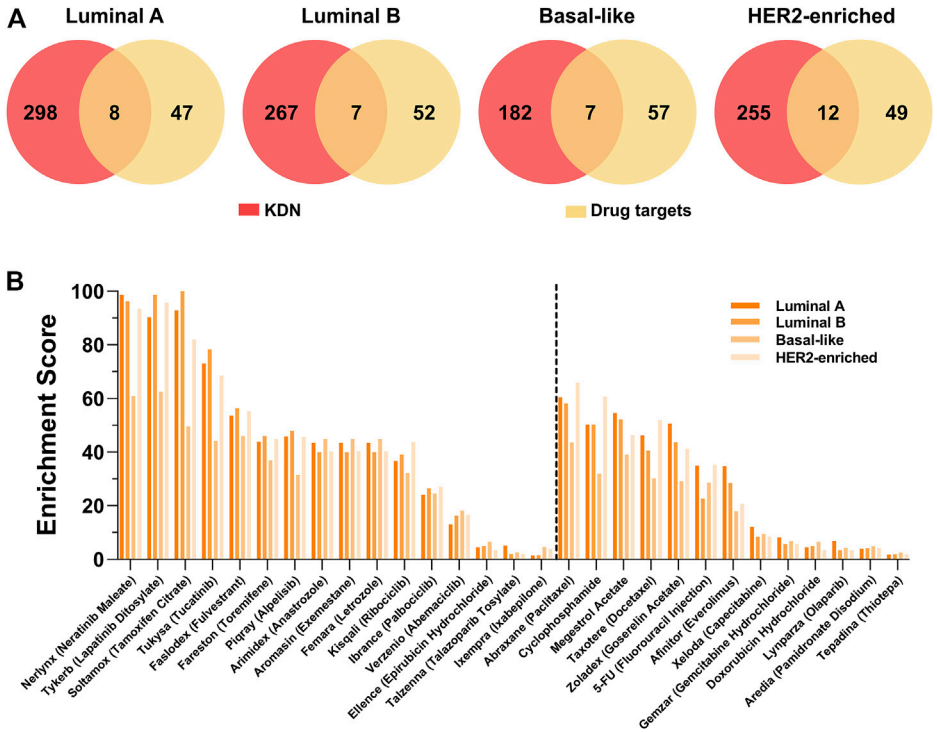


FIGURE 5 | Drug target analysis. **(A)** The Venn graph of the targets of 29 breast cancer drugs and genes in the KDN. **(B)** The enrichment scores of 29 breast cancer targeted drugs.

TABLE 4 | The drug targets in key genes.

Subtypes	Drug targets
Luminal A	ESR1, NR3C1, PRKCG, EGFR, PRKCA, ESR2, PRKCZ, PRKCB
Luminal B	ESR1, NR3C1, PRKCG, EGFR, PRKCA, ESR2, PRKCZ
Basal-like	ESR1, NR3C1, PRKCG, MAPT, PGR, BCL2, ERBB4
HER2-enriched	ESR1, NR3C1, PRKCG, EGFR, PRKCA, ESR2, PRKCZ, ERBB2, MAPT, PGR, BCL2, CYP2A6

As the key genes in DN constitute most of the dysregulation consequences, we are interested in the driver genes in the KDN. **Figure 4A** shows the Venn diagrams of the key genes and the identified driver genes in the four subtypes. In the KDN, about 20–30% of the genes are driver genes. That is, the KDN is enriched with a larger portion of driver genes. Furthermore, **Figure 4B** shows the driver genes' average number of events as defined by DriverNet (Bashashati et al., 2012), whether they are in the KDN or not. The former's average number of events is obviously higher than that of the latter, which demonstrates that key driver genes explain more abnormal expressed genes in the patient group than the latter.

In **Figure 4C**, green and red denote whether they are driver genes; black and gray denote whether the interactions originate from driver genes. The KDNs are highly connected in the central part which include some driver genes and their downstream genes, while the peripheral part is relatively sparsely connected which includes only non-driver genes. Therefore, we may hypothesize that these driver genes in the central part constitute the core tumorigenesis genes. Their mutations are the major causal factors to the corresponding subtypes. These driver genes first exert their abnormal effects on their direct downstream genes. And the downstream genes propagate the abnormal signals to other peripheral genes. Finally, the interactions between genes in the KDN contribute to the initiation and development of different breast cancer subtypes.

Breast Cancer Drug Targets are Enriched in the Key Dysregulated Network

From the perspective of the targeted therapy, the targets of drugs for breast cancer should be closely related to the KDN. **Figure 5A** shows the Venn graph of the targets of 29 breast cancer targeted drugs and the key genes. Only a few targets, such as ESR1, ESR2, EGFR, ERBB2, etc. are observed in the KDN. **Table 4** lists these targets which are targeted by 10 drugs. As most drugs' number of targets ranges

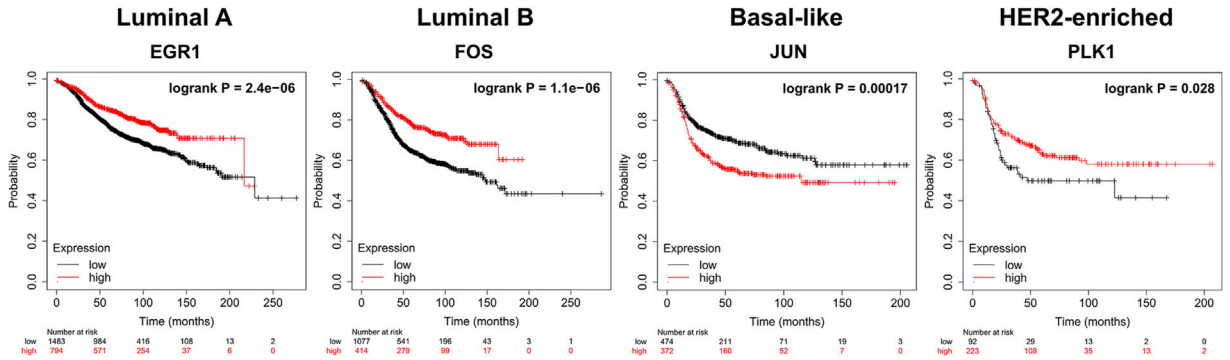


FIGURE 6 | Survival analysis (Kaplan-Meier plots) of dysregulated biomarkers. biomarkers High values are shown in red and low values are shown in black.

between 2 and 3, we use their first order neighboring genes to determine the enrichment score. The enrichment score is defined as the negative logarithm of the *p*-value of the hypergeometric test. **Figure 5B** shows the enrichment scores of the 29 drugs. Drugs solely for breast cancer are on the left, while those that can also treat other cancers are on the right. Obviously, most of the drugs are significantly enriched in the DN. In each DN, these enrichment scores are greater than 2. This demonstrates that the targets of these 29 drugs are closely related to the KDN.

To take a further look at the neighbors of targets not observed in the KDN, we find that some neighbors are the targets of other drugs in the KDN. For example, ESR1 and ESR2, targets of Soltamox (Tamoxifen Citrate) and Faslodex (Fulvestrant), are both neighbors of the other six drug targets, such as Nerlynx (Neratinib Maleate), Tykerb (Lapatinib Ditosylate), Abraxane (Paclitaxel), Cyclophosphamide, Megestrol Acetate, and Taxotere (Docetaxel). ESR1 is also a neighbor of CYP19A, which is the target of Arimidex (Anastrozole), Aromasin (Exemestane), and Femara (Letrozole). That is, most drug targets, even if not observed in the KDN, are closely associated with it. Thus, we may hypothesize that the KDN may serve as a critical level point for drugs to exert their effect and to intervene in the abnormal state of the cellular system.

The Top Dysregulated Genes may Serve as Potential Biomarkers for Survival Analysis

We apply KM-plotter to conduct the survival analysis of the top 10 dysregulated genes (http://kmplot.com/analysis/index.php?p=service&cancer=breast). For Luminal A, Luminal B, Basal-like, and HER2-enriched subtypes, 2277, 465, 846, and 315 samples are used respectively. **Figure 6** shows the results of the survival analysis with the smallest log-rank *p*-value of gene for each subtype. All *p*-values are less than 0.05. This indicates that these dysregulated genes can be used as potential prognostic markers of breast cancer subtypes.

CONCLUSION

From the perspective of biological networks, cancer is a result of the abnormal interactions between genes. In this paper, we propose a simple way to measure the regulation strength of genes based on their relative expression values. And then we construct the key dysregulated network (KDN) for the four subtypes of breast cancer. Our results show that the KDN is significantly enriched in critical breast cancer-related pathways as well as driver genes; closely associated with drug targets; and have significant differences in survival analysis. The key dysregulated genes can also serve as potential driver genes, drug targets, and prognostic markers for subtype identification. In addition, our results indicate that the key dysregulation analysis is more powerful than the traditional DEG analysis. Therefore, the KDN can be applied to other cancer studies, such as the identification of driver genes, drug repurposing, and so on.

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

YH and WL designed the project. YH and ZB drafted codes and conducted primary Computations. YH, XL, PX, ZB, and WL analyzed data. YH, ZB, and WL interpreted results. YH, XL, PX, ZB, and WL wrote manuscript. All authors approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62072128, 62002079, and 62102104).

ACKNOWLEDGMENTS

We sincerely thank numerous students and teachers at Guangzhou University for their kind consultations on clinical cases of multiple types of cancer under study.

REFERENCES

Akhavan-Safar, M., Teimourpour, B., and Kargari, M. (2021). GenHITS: A Network Science Approach to Driver Gene Detection in Human Regulatory Network Using Gene's Influence Evaluation. *J. Biomed. Inform.* 114, 103661. doi:10.1016/j.jbi.2020.103661

Aline, A. C., Pierre, H., Gérard, C., and Amar, B. J. F. I. P. (2015). Role of ErbB Receptors in Cancer Cell Migration and Invasion, *Front. Pharmacol.*, 6. 283. doi:10.3389/fphar.2015.00283

Bao, Z., Li, X., Zan, X., Shen, L., Ma, R., and Liu, W. (2016). Signalling Pathway Impact Analysis Based on the Strength of Interaction between Genes. *IET Syst. Biol.* 10, 147–152. doi:10.1049/iet-syb.2015.0089

Bao, Z., Zhu, Y., Ge, Q., Gu, W., and Bai, Y. J. I. A. (2020). *Signaling Pathway Analysis Combined with the Strength Variations of Interactions between Genes under Different Conditions*, 1.

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., et al. (2012). DriverNet: Uncovering the Impact of Somatic Driver Mutations on Transcriptional Networks in Cancer. *Genome Biol.* 13, R124. doi:10.1186/gb-2012-13-12-r124

Cassoni, P., Sapino, A., Negro, F., and Bussolati, G. (1994). Oxytocin Inhibits Proliferation of Human Breast Cancer Cell Lines. *Virchows Arch.* 425, 467–472. doi:10.1007/BF00197549

Chai, C., Wu, H. H., Abuetabh, Y., Sergi, C., and Leng, R. (2022). Regulation of the Tumor Suppressor PTEN in Triple-Negative Breast Cancer. *Cancer Lett.* 527, 41–48. doi:10.1016/j.canlet.2021.12.003

Cheang, M. C. U., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., et al. (2009). Ki67 index, HER2 Status, and Prognosis of Patients with Luminal B Breast Cancer. *J. Natl. Cancer Inst.* 101, 736–750. doi:10.1093/jnci/djp082

Chen, A. A., and Wang, J. J. C. J. O. P. (2012). Estrogen Signaling Pathway in Cancer. **28**, 570–576.

Chen, Y., Hao, J., Jiang, W., He, T., Zhang, X., Jiang, T., et al. (2013). Identifying Potential Cancer Driver Genes by Genomic Data Integration. *Sci. Rep.* 3, 3538. doi:10.1038/srep03538

Eskandari, E., and Motalebzadeh, J. J. C. G. (2019). *Transcriptomics-based Screening of Molecular Signatures Associated with Patients Overall Survival and Their Key Regulators in Subtypes of Breast Cancer*, 239.

Gao, Y.-m., Xu, P., Wang, X.-h., and Liu, W.-b. (2013). The Complex Fluctuations of Probabilistic Boolean Networks. *Biosystems* 114, 78–84. doi:10.1016/j.biosystems.2013.07.008

Gari, H. H., Degala, G. D., Ray, R., Lucia, M. S., and Lambert, J. R. (2016). PRL-3 Engages the Focal Adhesion Pathway in Triple-Negative Breast Cancer Cells to Alter Actin Structure and Substrate Adhesion Properties Critical for Cell Migration and Invasion. *Cancer Lett.* 380, 505–512. doi:10.1016/j.canlet.2016.07.017

Gong, J., Lu, X., Xu, J., Xiong, W., Zhang, H., and Yu, X. J. J. O. C. P. (2019). *Coexpression of UCA1 and ITGA2 in Pancreatic Cancer Cells Target the Expression of miR-107 through Focal Adhesion Pathway*, 234.

Gong, L., Tang, H., Luo, Z., Sun, X., Tan, X., Xie, L., et al. (2020). Tamoxifen Induces Fatty Liver Disease in Breast Cancer through the MAPK8/FoxO Pathway. *Clin. Translational Med.* 10, 137–150. doi:10.1002/ctm2.5

Gu, C., Wang, X., Long, T., Wang, X., Zhong, Y., Ma, Y., et al. (2018). FSTL1 Interacts with VIM and Promotes Colorectal Cancer Metastasis via Activating the Focal Adhesion Signalling Pathway. *Cell Death Dis* 9, 654. doi:10.1038/s41419-018-0695-6

Hernández-Vargas, H., Ouzounova, M., Calvez-Kelm, F. L., Lambert, M. P., Mckay-Chopin, S., Tavtigian, S. V., et al. (2011). Methylome Analysis Reveals Jak-STAT Pathway Deregulation in Putative Breast Cancer Stem Cells. *Epigenetics* 6, 428–439. doi:10.4161/epi.6.4.14515

Hollern, D. P., Swiatnicki, M. R., Rennhack, J. P., Misek, S. A., and Andrechek, E. R. (2019). E2F1 Drives Breast Cancer Metastasis by Regulating the Target Gene FGF13 and Altering Cell Migration. *Sci. Rep.* 9, 10718. doi:10.1038/s41598-019-47218-0

Hou, J. P., and Jian, M. J. G. M. (2014). DawnRank: Discovering Personalized Driver Genes in Cancer. Genome Medicine. 6**,** 56.doi:10.1186/s13073-014-0056-8

Huan, J., Wang, L., Xing, L., Qin, X., Feng, L., Pan, X., et al. (2014). Insights into Significant Pathways and Gene Interaction Networks Underlying Breast Cancer Cell Line MCF-7 Treated with 17β-Estradiol (E2). *Gene* 533, 346–355. doi:10.1016/j.gene.2013.08.027

Huang, R.-P., Fan, Y., De Belle, I., Niemeyer, C., Gottardis, M. M., Mercola, D., et al. (1997). Decreased Egr-1 Expression in Human, Mouse and Rat Mammary Cells and Tissues Correlates with Tumor Formation. *Int. J. Cancer* 72, 102–109. doi:10.1002/(sici)1097-0215(19970703)72:1<102:aid-ijc15>3.0.co;2-l

Inic, Z., Zegarac, M., Inic, M., Kozomara, Z., Djurisic, I., Inic, I., et al. (2014). Difference between Luminal A and Luminal B Subtypes According to Ki-67, Tumor Size, and Progesterone Receptor Negativity Providing Prognostic Information. *Clin. Med. Insights Oncol.* 8, 107–111. doi:10.4137/CMO.S18006

Liang, Y., Ye, F., Wang, Y., Li, Y., Li, Y., Song, X., et al. (2021). DGUOK-AS1 Acts as a Tumor Promoter through Regulating miR-204-5p/IL-11 axis in Breast Cancer. *Mol. Ther. - Nucleic Acids* 26, 1079–1091. doi:10.1016/j.omtn.2021.10.018

Liu, N., Zhang, J., Zhang, J., Liu, S., Liu, Y., and Zheng, D. (2008). Erbin-regulated Sensitivity of MCF-7 Breast Cancer Cells to TRAIL via ErbB2/AKT/NF-kappaB Pathway. *J. Biochem.* 143, 793–801. doi:10.1093/jb/mvn032

Milde-Langosch, K., Röder, H., Andritzky, B., Aslan, B., Hemminger, G., Brinkmann, A., et al. (2004). The Role of the AP-1 Transcription Factors C-Fos, FosB, Fra-1 and Fra-2 in the Invasion Process of Mammary Carcinomas. *Breast Cancer Res. Treat.* 86, 139–152. doi:10.1023/b:brea.0000032982.49024.71

Mohd, F., Wang, H., Uma, G., Little, P. J., Xu, J., and Zheng, W. J. I. J. O. B. S. (2017). FOXO Signaling Pathways as Therapeutic Targets in Cancer. *Int. J. Biol. Sci.* 13, 815–827.

Montenegro, M. F., and Cancer, M. C.-G. J. B. (2014). Promoting E2F1-Mediated Apoptosis in Oestrogen Receptor-α-Negative Breast Cancer Cells. *BMC Cancer* 14, 539. doi:10.1186/1471-2407-14-539

Na, L., and Balko, J. M. (2019). *Role of JAK-STAT Pathway in Cancer Signaling: Applications in Precision Medicine*. doi:10.1007/978-3-319-95228-4_26

Ocak, S., Yamashita, H., Udyavar, A. R., Miller, A. N., Gonzalez, A. L., Zou, Y., et al. (2010). DNA Copy Number Aberrations in Small-Cell Lung Cancer Reveal Activation of the Focal Adhesion Pathway. *Oncogene* 29, 6331–6342. doi:10.1038/onc.2010.362

Peng, X., Wang, X., and Liu, W. 2012). "The Influence of the basin Structure of Boolean Networks on Their Long Range Correlated Dynamics", in: IEEE International Conference on Systems Biology.

Pequeux, C. (2002). *Membrane Conformation and Transduction Pathway of Oxytocin and Vasopressin in Small Cell Lung Cancer*.

Ronski, K., Sanders, M., Burleson, J. A., Moyo, V., Benn, P., and Min, F. J. C. (2010). Early Growth Response Gene 1 (EGR1) Is Deleted in Estrogen Receptor-Negative Human Breast Carcinoma. *Cancer* 104, 925–930. doi:10.1002/cncr.21262

Sato, T. (2013). *Prolactin-Jak-Stat Signaling Pathways in Breast Cancer*. Thomas Jefferson University.

Shi, K., Gao, L., and Wang, B. (2016). Discovering Potential Cancer Driver Genes by an Integrated Network-Based Approach. *Mol. Biosyst.* 12, 2921–2931. doi:10.1039/c6mb00274a

Shin-Ichi, H., Niwa, T., and Yamaguchi, Y. J. C. S. (2010). Estrogen Signaling Pathway and its Imaging in Human Breast Cancer. 100**,** 1773–1778.

Song, R. X.-D., and Santen, R. J. (2006). Membrane Initiated Estrogen Signaling in Breast Cancer1. 75**,** 9–16.doi:10.1095/biolreprod.105.050070

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. Clin.* 71, 209–249. doi:10.3322/caac.21660

Tworoger, S., and Hankinson, S. (2006). Prolactin and Breast Cancer Risk. *Cancer Lett.* 243, 160–169. doi:10.1016/j.canlet.2006.01.032

Tworoger, S. S., Eliassen, A. H., Zhang, X., Qian, J., Sluss, P. M., Rosner, B. A., et al. (2013). A 20-year Prospective Study of Plasma Prolactin as a Risk Marker of Breast Cancer Development. *Cancer Res.* 73, 4810–4819. doi:10.1158/0008-5472.can-13-0665

Vleugel, M. M., Greijer, A. E., Bos, R., van der Wall, E., and van Diest, P. J. (2006). c-Jun Activation Is Associated with Proliferation and Angiogenesis in Invasive Breast Cancer. *Hum. Pathol.* 37, 668–674. doi:10.1016/j.humpath.2006.01.022

Wang, C. J., Gao, H. D., and Wang, L. Y. J. C. J. O. P. S. (2009). *Clinical Research of High Risk Human Papillomavirus Infection in Breast Cancer*.

Wang, S., Yao, Y., Yao, M., Fu, P., Wang, W., and Communications, B. R. (2018). Interleukin-22 Promotes Triple Negative Breast Cancer Cells Migration and Paclitaxel Resistance through JAK-STAT3/MAPKs/AKT Signaling Pathways. *Biochem. Biophysical Res. Commun.* 503, 1605–1609. doi:10.1016/j.bbrc.2018.07.088

Wang, Y., Jiang, X.-Y., and Yu, X.-Y. (2020). BRD9 Controls the Oxytocin Signaling Pathway in Gastric Cancer via CANA2D4, CALML6, GNAO1, and KCNJ5. *Transl Cancer Res. TCR* 9, 3354–3366. doi:10.21037/tcr.2020.03.67

Wierer, M., Verde, G., Pisano, P., Molina, H., Font-Mateu, J., Di croce, L., et al. (2013). PLK1 Signaling in Breast Cancer Cells Cooperates with Estrogen Receptor-dependent Gene Transcription. *Cell Rep* 3, 2021–2032. doi:10.1016/j.celrep.2013.05.024

Wirtenberger, M., Tchatchou, S., Hemminki, K., Schmutzhard, J., Sutter, C., Schmutzler, R. K., et al. (2006). Associations of Genetic Variants in the Estrogen Receptor Coactivators PPARGC1A, PPARGC1B and EP300 with Familial Breast Cancer. *Carcinogenesis* 27, 2201–2208. doi:10.1093/carcin/bgl067

Wolf, I., Bose, S., Williamson, E., Miller, C. W., and Koeffler, H. P. J. C. R. (2005). *FOXA1 Expression and Activities in Breast Cancer*, 65.

Wu, G., Feng, X., and Stein, L. (2010). A Human Functional Protein Interaction Network and its Application to Cancer Data Analysis. *Genome Biol.* 11, R53–R23. doi:10.1186/gb-2010-11-5-r53

Wu, J.-R., Zhao, Y., Zhou, X.-P., and Qin, X. (2020). Estrogen Receptor 1 and Progesterone Receptor Are Distinct Biomarkers and Prognostic Factors in Estrogen Receptor-Positive Breast Cancer: Evidence from a Bioinformatic Analysis. *Biomed. Pharmacother.* 121, 109647. doi:10.1016/j.biopha.2019.109647

Yang, K., Gao, J., and Luo, M. (2019). Identification of Key Pathways and Hub Genes in Basal-like Breast Cancer Using Bioinformatics Analysis. *Ott* 12, 1319–1331. doi:10.2147/ott.s158619

Yang, W., He, X., He, C., Peng, L., Xing, S., Li, D., et al. (2021). Impact of ESR1 Polymorphisms on Risk of Breast Cancer in the Chinese Han Population. *Clin. Breast Cancer* 21, e235–e242. doi:10.1016/j.clbc.2020.10.003

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.

Copyright © 2022 Huo, Li, Xu, Bao and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fgene.2022.889378

Clinical Prediction of Heart Failure in Hemodialysis Patients: Based on the Extreme Gradient Boosting Method

Yanfeng Wang[1†], Xisha Miao[1†], Gang Xiao[2†], Chun Huang[1], Junwei Sun[1], Ying Wang[3], Panlong Li[1] and Xu You[2]**

[1]The School of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou, China, [2]Department of Clinical Laboratory, The Third Affiliated Hospital, Southern Medical University, Guangzhou, China, [3]Department of Medical Statistics, School of Public Health, Sun Yat-sen University, Guangzhou, China

OPEN ACCESS

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

Reviewed by:
Leimin Wang,
China University of Geosciences Wuhan, China
Yansen Su,
Anhui University, China

***Correspondence:**
Panlong Li
2019021@zzuli.edu.cn
Xu You
maja_1983@126.com

[†]These authors have contributed equally to this work and share first authorship

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

Received: 04 March 2022
Accepted: 15 March 2022
Published: 26 April 2022

Citation:
Wang Y, Miao X, Xiao G, Huang C, Sun J, Wang Y, Li P and You X (2022) Clinical Prediction of Heart Failure in Hemodialysis Patients: Based on the Extreme Gradient Boosting Method. Front. Genet. 13:889378. doi: 10.3389/fgene.2022.889378

Background: Heart failure (HF) is the main cause of mortality in hemodialysis (HD) patients. However, it is still a challenge for the prediction of HF in HD patients. Therefore, we aimed to establish and validate a prediction model to predict HF events in HD patients.

Methods: A total of 355 maintenance HD patients from two hospitals were included in this retrospective study. A total of 21 variables, including traditional demographic characteristics, medical history, and blood biochemical indicators, were used. Two classification models were established based on the extreme gradient boosting (XGBoost) algorithm and traditional linear logistic regression. The performance of the two models was evaluated based on calibration curves and area under the receiver operating characteristic curves (AUCs). Feature importance and SHapley Additive exPlanation (SHAP) were used to recognize risk factors from the variables. The Kaplan–Meier curve of each risk factor was constructed and compared with the log-rank test.

Results: Compared with the traditional linear logistic regression, the XGBoost model had better performance in accuracy (78.5 vs. 74.8%), sensitivity (79.6 vs. 75.6%), specificity (78.1 vs. 74.4%), and AUC (0.814 vs. 0.722). The feature importance and SHAP value of XGBoost indicated that age, hypertension, platelet count (PLT), C-reactive protein (CRP), and white blood cell count (WBC) were risk factors of HF. These results were further confirmed by Kaplan–Meier curves.

Conclusions: The HF prediction model based on XGBoost had a satisfactory performance in predicting HF events, which could prove to be a useful tool for the early prediction of HF in HD.

Keywords: machine learning, extreme gradient boosting, heart failure prediction, hemodialysis, risk factors

INTRODUCTION

Heart failure (HF) as a clinical syndrome is one of the main causes of mortality (Ebong et al., 2014; Virani et al., 2021). More than 60 million people are affected by HF worldwide, and the number of HF patients is increasing every year (Groenewegen et al., 2020). Compared with the general population, the prevalence of HF in patients with chronic kidney disease (CKD) is much higher, especially in patients with end-stage renal disease (ESRD) (Rangaswami and McCullough, 2018; Rapa et al., 2019). Currently, hemodialysis (HD) as a renal replacement therapy (RRT) is the main treatment of ESRD (Akash et al., 2014). However, more than 40% of HD patients suffer from HF, which increase the medical care, economic burden, and mortality (House et al., 2019; Sun et al., 2019). Therefore, identification, pre-estimation, and timely interventions can improve the prognosis and prolong the survival time for the population who show a high risk of HF (Abdo, 2017).

The causes of HF in HD patients are multifactorial (Wang and Sanderson, 2011). Similar to the general population, traditional risk factors such as aging, hypertension, diabetes mellitus (DM), and atherosclerosis are associated with HF in HD patients (Liu et al., 2014; Cozzolino et al., 2018). However, excess HF prevalence and mortality in HD population are not fully accounted for by the traditional risk factors (Schefold et al., 2016). Some researchers have found that dialysis can increase the risk of HF events in ESRD patients (Rangaswami and McCullough, 2018; Samanta et al., 2019). Evidence shows that the malnutrition–inflammation syndrome in HD patients is associated with HF (Almeida et al., 2008; Chang et al., 2020). In addition, changed hemodynamic and blood pressure by HD can increase the risk of HF (Dorairajan et al., 2010). Several traditional risk prediction models were proposed to predict the risk of HF in population without HD (Ouwerkerk et al., 2014; Voors et al., 2017). However, these linear models missed specific related factors for HD people and oversimplified the complicated relationships between factors, which may lead to a decrease in performance and miss important risk factors (Mpanya et al., 2021).

Recent advances in extreme gradient boosting (XGBoost), a new integrated machine learning algorithm, have provided a robust method to identify the complex non-linear relationship between multiple variables and outcomes (Kagiyama et al., 2019). This algorithm has strong model generalization ability, fast operating speed, and high model accuracy, which has been applied in orthopedic auxiliary classification, prediction of interaction, and analysis of hypertension-related symptoms to improve accuracy in complex clinical decision-making (Chang et al., 2019; Li and Zhang, 2020; Yu et al., 2020).

Therefore, we aimed to establish a prediction model that integrated patient-specific information and non-traditional factors of HF in HD patients based on XGBoost 1) to accurately predict HF events and 2) to assess the risk of HF in patients with HD treatment.

MATERIALS AND METHODS

Study Population

This retrospective study analyzed the data of 410 ESRD patients who underwent HD treatment in the HD centers of the Third Affiliated Hospital of Southern Medical University and the Third Affiliated Hospital of Sun Yat-sen University between January 2015 and September 2019. All the patients were older than 18 years and received HD treatment for at least 3 months. Patients with the following conditions were excluded from this study: 1) history of renal transplantation; 2) malignancy, acute infection, hepatic, and pulmonary dysfunction; and 3) lack of data on demographic characteristics, laboratory examinations, or physical examination. This study was approved by the Ethics Committee of the Third Affiliated Hospital of Southern Medical University.

Data Collection

A total of 21 factors were collected at the start of HD therapy. The basic information consisted of the patient's gender, age, and history of various diseases. After fasting overnight for at least 8 h, the patient's venous blood samples were collected before dialysis therapy. The laboratory indicators included C-reactive protein (CRP), blood urea nitrogen (BUN), calcium (CA), hemoglobin (HB), phosphorus (P), cholesterol (CHOL), serum creatinine (CRE), lymphocyte (LYMPH), uric acid (UA), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), intact parathyroid hormone (IPTH), white blood cell count (WBC), platelet count (PLT), and neutrophil count (NEUT). All the abovementioned examinations were conducted at the laboratory centers in the Third Affiliated Hospital of Southern Medical University and the Third Affiliated Hospital of Sun Yat-sen University.

Outcome and Follow-Up

The occurrence of HF events (fatal or not) after HD was recorded. The HF diagnosis was made based on Framingham criteria (Groenewegen et al., 2020). Follow-up visits were performed by the HD centers. Patients who developed HF after HD were censored at the earliest date of a HF event. The HF events were recorded up to May 1, 2020.

Statistical Analysis

All statistical analyses and model establishment were performed using SPSS 26.0 for Windows and Python 3.7.6. All continuous variables were described as mean ± standard deviation (SD), and these variables were compared between groups by performing t-tests. In addition, discrete variables were expressed in numbers (n) and percentages (%), and the chi-square test was applied. All results were considered to be statistically significant within a two-sided test with $p < 0.05$.

Two classification models were established based on XGBoost and traditional linear logistic regression models. The multiple logistic regression was selected using a stepwise method. In the XGBoost model, all the characteristics of HD patients were included. The data set was randomly divided into a training set by 75% and a validation set by 25%. The performance of the models was evaluated using calibration curves and area under the receiver operating characteristic curves (AUCs). To evaluate the influence of different variables on the results, the SHAP value of important variables was calculated (Rodriguez-Perez and Bajorath, 2020). A risk curve was constructed using the

TABLE 1 | Clinical characteristics and laboratory parameters of patients in the HF group and the non-HF group.

Clinical characteristic and laboratory parameter	All patients	HF	Non-HF	*p*-value
Age (years)	56.05 (43.13–66.32)	63.1 (54.64–72.03)	52.0 (41.0–61.1)	<0.001
Gender (male) (n, %)	239 (66.71)	73 (76.04)	166 (64.59)	0.041
HTN (n, %)	281 (79.60)	88 (91.67)	193 (75.10)	<0.001
DM (n, %)	123 (34.84)	43 (44.79)	80 (31.13)	0.017
HB (g/L)	96.44 ± 21.01	97.71 ± 21.52	95.98 ± 20.96	0.423
NEUT (10^9/L)	4.39 (3.47–5.60)	4.86 (3.50–6.41)	4.26 (3.44–5.17)	0.005
WBC (10^9/L)	6.72 (5.36–8.03)	7.68 (6.07–9.00)	6.48 (5.20–7.69)	<0.001
LYMPH (10^9/L)	1.19 (0.89–1.50)	1.18 (0.90–1.54)	1.20 (0.89–1.49)	0.854
IPTH (pg/ml)	343.40 (213.43–529.19)	322.50 (192.51–465.64)	356.45 (218.37–536.82)	0.128
UA(μmol/L)	494.88 ± 146.20	513.74 ± 154.06	487.81 ± 142.83	0.861
BUN (mmol/L)	23.07 (16.39–29.09)	23.27 (18.30–31.25)	22.85 (15.96–28.69)	0.299
CRE (μmol/L)	848.60 ± 391.62	812.26 ± 372.39	862.18 ± 398.40	0.329
CA (mmol/L)	2.17 (2.04–2.32)	2.14 (1.99–2.30)	2.19 (2.07–2.33)	0.103
P (mmol/L)	1.81 (1.44–2.25)	1.81 (1.51–2.20)	1.82 (1.44–2.26)	0.854
CHOL (mmol/L)	4.15 (3.62–4.89)	4.10 (3.57–4.97)	4.17 (3.65–4.87)	0.655
TG (mmol/L)	1.51 (1.17–1.92)	1.43 (1.11–1.90)	1.52 (1.20–1.93)	0.434
HDL-C (mmol/L)	0.97 (0.87–1.11)	0.99 (0.88–1.15)	0.96 (0.86–1.10)	0.572
LDL-C (mmol/L)	2.45 (2.00–2.96)	2.46 (2.00–3.15)	2.44 (1.99–2.94)	0.668
CRP (mg/L)	5.66 (2.65–13.65)	8.40 (4.05–16.35)	5.01 (2.12–13.00)	<0.001
TIBC (mmol/L)	36.16 (33.73–39.55)	35.84 (33.48–38.09)	36.46 (33.95–40.00)	0.034
PLT (10^9/L)	201.00 (156.00–243.70)	210.00 (155.50–273.00)	198.72 (157.25–237.00)	0.177

CRP, C-reactive protein; BUN, blood urea nitrogen; CA, calcium; HB, hemoglobin; DM, diabetes mellitus; TG, triglycerides; P, phosphorus; CHOL, cholesterol; CRE, serum creatinine; LYMPH, lymphocyte; UA, uric acid; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; IPTH, intact parathyroid hormone; WBC, white blood cell count; NEUT, neutrophil count; PLT, platelet count; TIBC, total iron-binding capacity; HTN, hypertension.

Kaplan–Meier analysis, and differences between groups were compared by a log-rank test to further evaluate the performance of our model and verify the risk factors selected by the mode. To make our model be an easy-to-use tool, a nomogram was developed based on the risk factors selected by our model.

RESULTS

Demographic and Clinical Characteristics of Study Population

According to the inclusion and exclusion criteria, a total of 353 patients were finally included in this study. The basic variables of study population are shown in **Table 1**. Among these patients, 96 patients (73 males and 23 females) developed HF during the follow-up duration and 257 patients (166 males and 91 females) did not. In general, the average age of these patients was 54.92 ± 15.29, and the proportion of males was 67.71%. Compared with patients without HF, those with HF were older and had a higher ratio of hypertension and diabetes, and there was a significant difference in gender. In terms of laboratory data, for patients in the HF group, their WBC, NEUT, CRP, and TIBC levels were significantly (p <0.05) higher than those of the other group. There was no significant difference in other characteristics.

Logistic Regression Analysis

As presented in **Table 2**, the following factors were included in a multivariate stepwise logistic regression model: HTN (OR: 3.786, 95% CI: 1.640–8.739, and p = 0.002), WBC (OR: 1.115, 95% CI: 1.026–1.212, and p = 0.011), CRP (OR: 1.020, 95% CI: 1.009–1.032, and p = 0.001), and age (OR: 1.026, 95% CI: 1.008–1.044, and p = 0.004). The receiver operating characteristic (ROC) curve of the logistic regression model is shown in **Figure 1**, and the performance of this model was evaluated using the calibration curve, as shown in **Figure 2**. The area under the curve (AUC) was 0.722 for the validation group and 0.735 for the training group. In addition, accuracy, sensitivity, and specificity of the model were 74.8, 75.6, and 74.4%, respectively, evaluated by the validation group.

TABLE 2 | Multivariate logistic regression model.

Variable	B	SE	Wald	*p*-value	OR (95%CI)
HTN	1.331	0.427	9.733	0.002	3.786 (1.640–8.739)
WBC	0.109	0.043	6.532	0.011	1.115 (1.026–1.212)
CRP	0.020	0.006	11.730	0.001	1.020 (1.009–1.032)
AGE	0.025	0.009	8.113	0.004	1.026 (1.008–1.044)

HTN, hypertension; WBC, white blood cell count; CRP, C-reactive protein; OR, odds ratio; CI, confidence interval.

Extreme Gradient Boosting Model

The ROC curve and the calibration curve of the XGBoost model are presented in **Figure 1** and **Figure 2**, respectively. It was found that the accuracy, sensitivity, specificity, and AUC value were 78.5, 79.6, 78.1%, and 0.814, respectively, evaluated by the validation cohort. The feature importance ranking which represents the contribution of the corresponding feature to the model prediction based on gain is shown in **Figure 3**. It was found that HTN, age, CRP, WBC, and PLT were more important for the prediction of HC than other features.

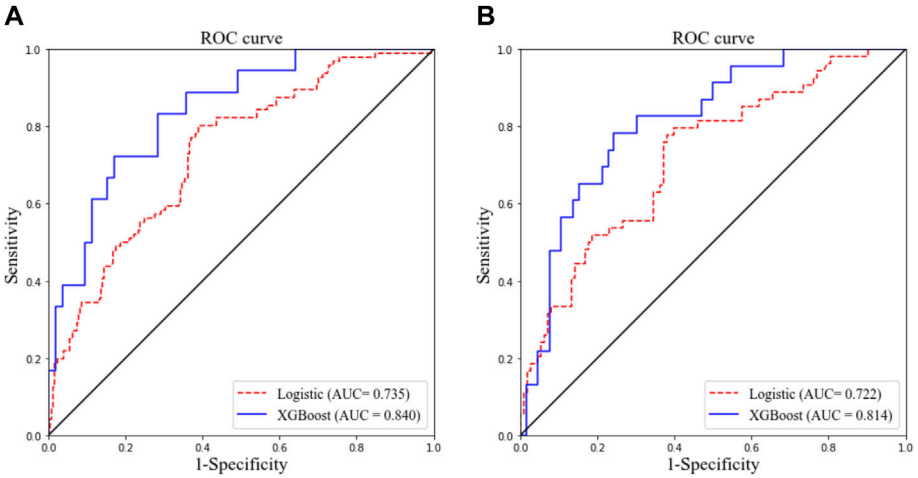


FIGURE 1 | Receiver operating characteristic curves of two models for predicting objective response in the training cohort **(A)** and validation cohort **(B)**. AUC, area under the curve.

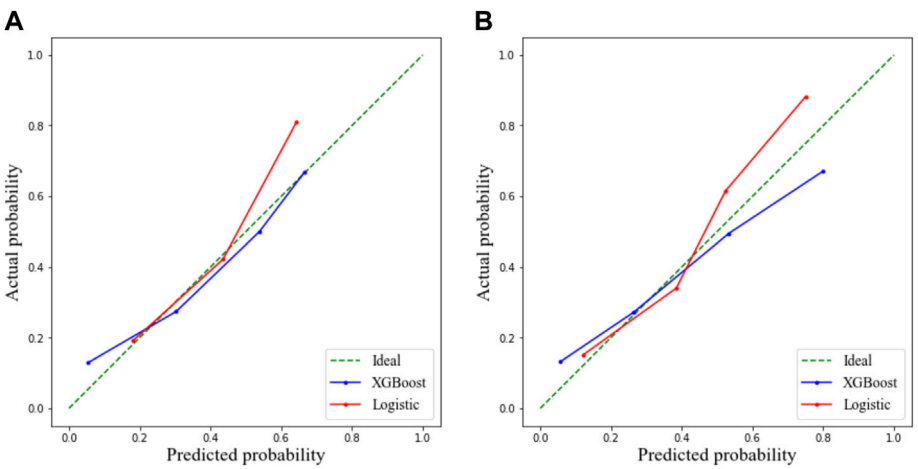


FIGURE 2 | Calibration curve demonstrating predictions from the model to the actual observed probability. **(A)** Training cohort and **(B)** validation cohort.

Interpretation of the Extreme Gradient Boosting Model

The SHAP value quantifies the contribution of data to the outcome in the XGBoost model; the size of the value reflects the influence of the corresponding eigenvalue on the outcome, and the positive and negative values reflect whether the results are promoted. Based on the established XGBoost model, the SHAP values of the five relevant factors showed how the feature affected the results, as shown in **Figure 3**.

Kaplan–Meier Analysis

The Kaplan–Meier analysis was used to confirm the effects of risk factors on the incidence of HF. The risk factors were stratified by the SHAP value when it was 0. As shown in **Figure 4**, the risk curve revealed that considering the influence of different outcome time, patients with high levels of WBC, CRP, PLT, old age, and hypertension had a higher incidence of HF, and the effect of UA was not obvious enough. The nomogram of the XGBoost model was constructed, as shown in **Figure 5**.

DISCUSSION

In this retrospective study, we established and validated a clinical prediction model to predict HF in HD patients based on the XGBoost algorithm. The results indicated that our model could effectively identify the individuals who suffer from HF using routinely clinical parameters. Our result indicated that XGBoost as a non-linear model had better performance than the logistic model and showed stronger ability in identification of risk factors.

In recent years, machine learning has been widely used in risk prediction and disease screening, which has obtained excellent performance (Kwon et al., 2020; Meng et al., 2020; Cobb et al.,

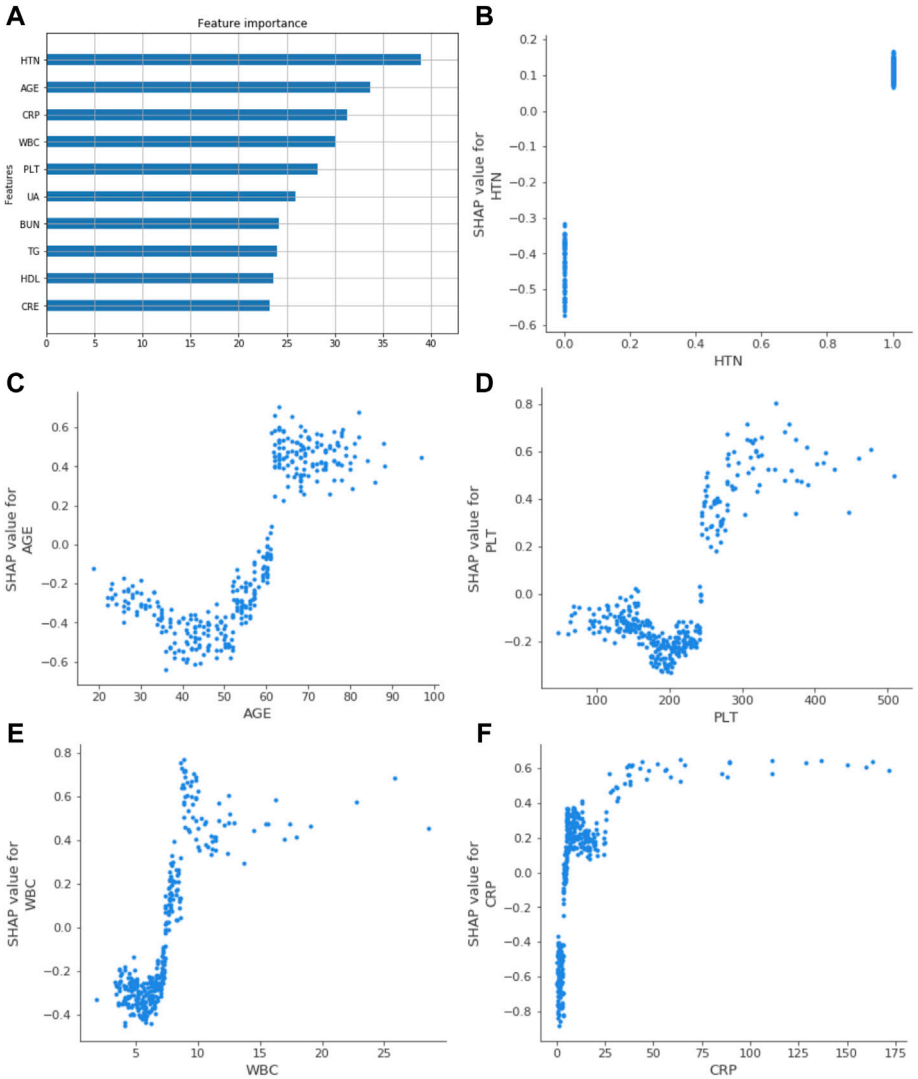


FIGURE 3 | Related factors of HF were ranked by feature gain. **(A)** showed the top ten gain features; age, HTN, WBC, PLT, and CRP had a more significant impact. **(B)**, **(C)**, **(D)**, **(E)**, and **(F)** represented the SHAP value of HTN, age, PLT, WBC, and CRP in the XGBoost model. It reflected the influence of relevant factors on the results. WBC, white blood cell count; HTN, hypertension; UA, uric acid; TG, triglycerides; BUN, blood urea nitrogen; HDL, high-density lipoprotein cholesterol; PLT, platelet count; CRE, serum creatinine; CRP, C-reactive protein.

2021; Koteluk et al., 2021; Kourou et al., 2021). Therefore, in this research, XGBoost, an integrated machine learning algorithm, was applied to identify the complex non-linear relationship between HF and clinical variables, as well as to evaluate the importance of the variables to the HF. Although traditional multivariable analysis methods have been applied for the identification of HF in HD patients (Gedfew et al., 2020; Bramania et al., 2021), to our knowledge, this was the first machine learning model for the prediction of HF. Our results showed that the performance of the XGBoost model was better than the traditional logistic regression model in prediction of HF. To make our XGBoost model interpretable, the feature importance and SHAP value were applied to evaluate the contribution of variables to HF. Except four risk factors including age, hypertension, WBC, and CRP found in the logistic regression model, a new risk factor, PLT, was found in the XGBoost model. This showed the advantages of our XGBoost mode, which could improve the accuracy of the classification and the efficiency of identification.

In addition to HF response, the performance of our model was evaluated by the Kaplan–Meier analysis. The differences between the Kaplan–Meier curves of different risk levels indicated that the risk factors selected by our XGBoost model were important prognostic factors of HF. The results of Kaplan–Meier analysis also showed that age more than 60, WBC more than 6.5×10^9, CRP more than 15 mg/L, PLT more than 250×10^9, and hypertension were independent risk factors which could increase the probability of HF. Moreover, based on the five risk factors selected by the XGBoost model and the Kaplan–Meier analysis, the nomogram (Yao et al., 2019) was designed to give an easy-to-use tool for the prediction of HF.

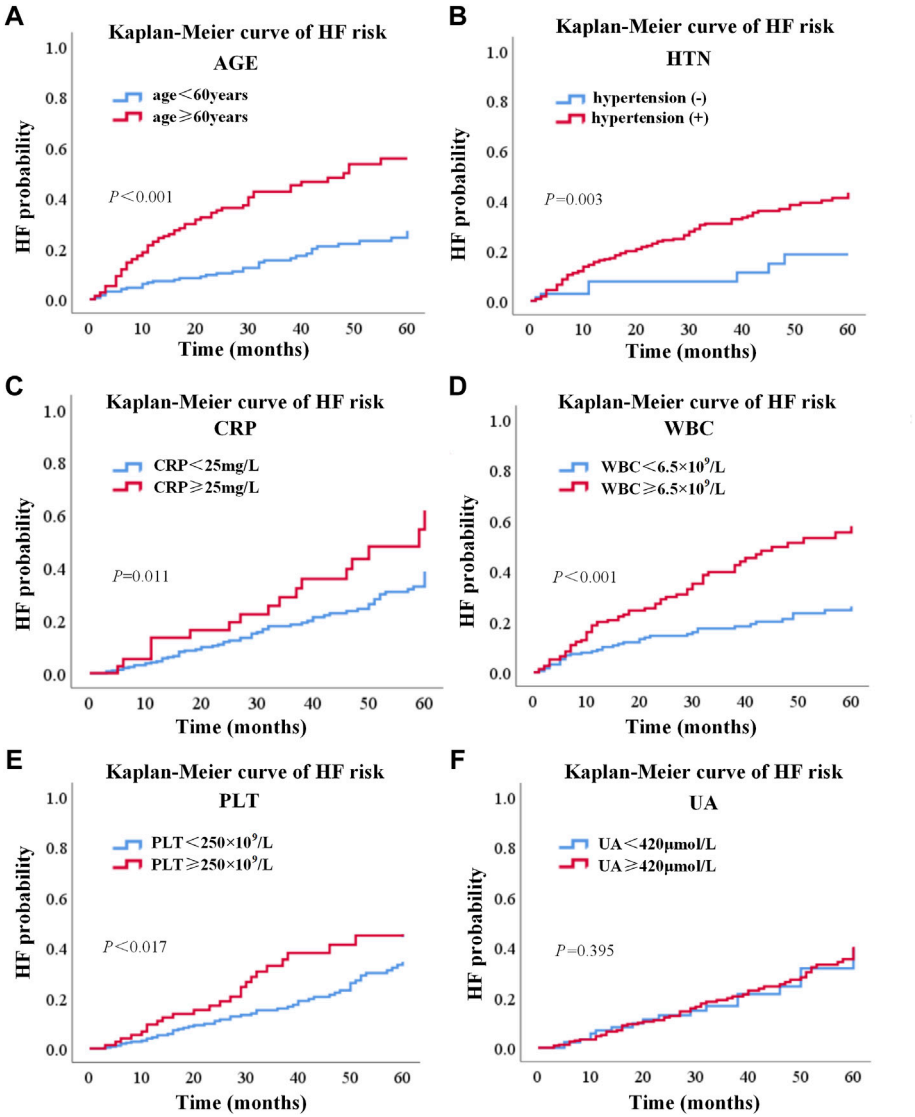


FIGURE 4 | Risk curve of six factors based on the Kaplan–Meier analysis. **(A)**, **(B)**, **(C)**, **(D)**, **(E)**, and **(F)** represented risk curves of age, hypertension, CRP, WBC, PLT, and UA, respectively. WBC, white blood cell count; UA, uric acid; CRP, C-reactive protein; PLT, platelet count.

In our study, the proportion of hypertension in HF patients was significantly greater than that in the non-HF group, which is consistent with a recent study (Bramania et al., 2021). The analysis of characteristics showed that hypertension was a risk factor for HF events, which is consistent with the results of a recent study (Cozzolino et al., 2017). Patients with hypertension already had a huge cardiovascular burden before dialysis (Van Buren, 2016). There are pieces of evidence that dialysis treatment can lead to abnormal fluctuations in blood pressure and affect patient's hemodynamics (Chou et al., 2018; Douvris et al., 2019). The vascular intima is more damaged by the abnormal fluctuations in the blood pressure, and HF is more likely to occur (Buren and Inrig, 2012).

Age is an important risk factor of HF. In old age, the shape and function of the vascular wall are changed due to oxidative stress, cell aging, and inflammation (Ghebre et al., 2016). In addition, with increasing age, body functions gradually decline, physiological compensatory function decreases, chronic diseases such as high blood pressure and diabetes occur frequently, and symptoms such as anemia, malnutrition, and chronic inflammation are prone to occur; these factors increase the risk of HF events (McHugh and Gil, 2018; Groenewegen et al., 2020). It has been reported that approximately 80% of HF patients were more than 60 years old, and the prevalence increased with age (Vigen et al., 2012). According to the results of our study, the risk of HF continued to increase from about 45 years old, which was consistent with the report from the American Heart Association (Virani et al., 2021).

WBC and CRP are widely used to reflect the inflammatory state in patients (Rienstra et al., 2012; Adamo et al., 2020).

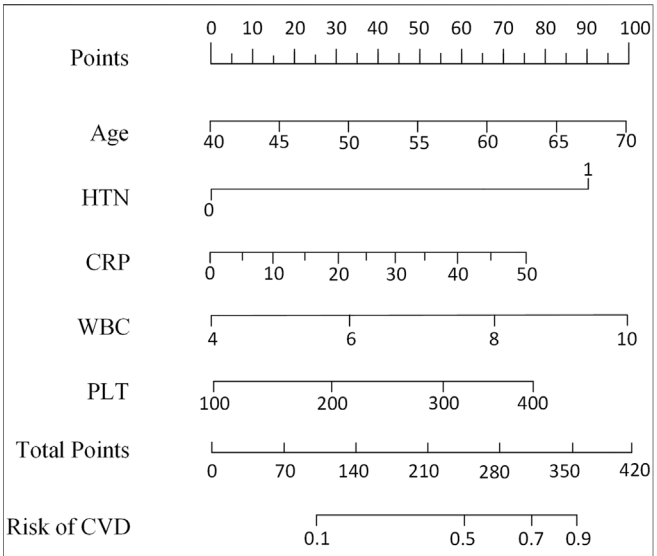


FIGURE 5 | Nomogram scaled by the coefficient of each risk factor selected by the XGBoost model. HTN, hypertension; WBC, white blood cell count; CVD, cardiovascular disease; CRP, C-reactive protein; PLT, platelet count.

Inflammation and HF are strongly interconnected, and higher levels of inflammation markers indicated an increased mortality and morbidity of HF in patients (Ekdahl et al., 2017). Some studies have confirmed that inflammation markers are implicated in the development of HF because the immune system leads to deterioration of the structure and function of the cardiovascular system by the inflammatory response (Van Linthout and Tschope, 2017; Castillo et al., 2020). A study has demonstrated that high levels of WBC are closely related to the incidence of HF (Bekwelem et al., 2011). CRP is one of the recognized risk factors for CVD events, and a meta-analysis has reported that CRP plays an important role in the development of HF (Emerging Risk Factors Collaboration et al., 2010). In this study, our result was consistent with these previous findings. In addition, our results showed that other serum biomarkers, such as an anomalous level of PLT, were related to HF. Excessive PLT can promote inflammation, negatively affect the left ventricular function, and increase the risk of atherosclerosis; these diseases would increase the probability of HF (Gary et al., 2013; Chen and Yang, 2020).

All these data required in this study can be easily obtained through routine clinical examination, not limited by many additional conditions, and can make full use of existing resources. The XGBoost model showed a better prediction effect by using higher indicators such as AUC, accuracy, sensitivity, and atherosclerosis specificity. It can be a useful tool for doctors to evaluate the HF risk of HD patients and carry out personalized intervention in advance. However, there are still some limitations to our study. We only focused on the event of HF in HD population of China; it needs to be verified by different populations and races. In addition, this study is a retrospective examination which lacks longitudinal observation. Therefore, in further research, a longitudinal data analysis can help evaluate the changes in these risk factors over time and further validate the results in our model.

In conclusion, we developed and validated a clinical prediction model based on the XGBoost algorithm for HF in end-stage renal disease patients. The model had an excellent predictive performance and provided a useful tool for the early prediction of HF in HD patients.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Third Affiliated Hospital of Southern Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

PL and XY contributed to the concept and design of the sutdy. YIW and XM organized the database. YAW, XM, and GX wrote the first draft of the manuscript. CH and JS wrote sections fo the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the Third Affiliated Hospital of Sun Yat-Sen University, Clinical Research Program (P000-277), the Science and Technology Planning Project of Guangdong Province, China (2019B020228001), the Meizhou Science and Technology Project (2019B0203003), the Open Fund of State Key Laboratory of EC Prevention and Treatment (K2020-0010 and K2020-0011), and the National Natural Science Foundation of China (U1804262, 61632002). Zhongyuan Thound Talents Program (204200510003).

REFERENCES

Abdo, A. S. (2017). Diagnosis and Management of Heart Failure in Long-Term Dialysis Patients. *Curr. Heart Fail. Rep.* 14 (5), 404–409. doi:10.1007/s11897-017-0354-8

Adamo, L., Rocha-Resende, C., Prabhu, S. D., and Mann, D. L. (2020). Reappraising the Role of Inflammation in Heart Failure. *Nat. Rev. Cardiol.* 17 (5), 269–285. doi:10.1038/s41569-019-0315-x

Akash, N. K., Giacomo, M., Enrico, R., and Claudio, R. (2014). The Role of Economies of Scale in the Cost of Dialysis across the World: a Macroeconomic Perspective. *Nephrol. Dial. Transpl.* 29 (4), 885–891. doi:10.1093/ndt/gft528

Almeida, H. R. M., Santos, E. M. C., Dourado, K., Mota, C., and Peixoto, R. (2008). Malnutrition Associated with Inflammation in the Chronic Renal Patient on Hemodialysis. *Rev. Assoc. Med. Bras (1992)* 64 (9), 837–844. doi:10.1590/1806-9282.64.09.837

Bekwelem, W., Lutsey, P. L., Loehr, L. R., Agarwal, S. K., Astor, B. C., Guild, C., et al. (2011). White Blood Cell Count, C-Reactive Protein, and Incident Heart Failure in the Atherosclerosis Risk in Communities (ARIC) Study. *Ann. Epidemiol.* 21 (10), 739–748. doi:10.1016/j.annepidem.2011.06.005

Bramania, P. K., Ruggajo, P. J., and Furia, F. F. (2021). Prevalence and Predictors of Heart Failure Among Patients on Maintenance Hemodialysis Therapy at Muhimbili National Hospital in Tanzania: a Cross-Sectional Study. *Egypt. Heart J.* 73 (1), 97. doi:10.1186/s43044-021-00223-z

Buren, P., and Inrig, J. K. (2012). Hypertension and Hemodialysis: Pathophysiology and Outcomes in Adult and Pediatric Populations. *Pediatr. Nephrol.* 27 (3), 339–350. doi:10.1007/s00467-011-1775-3

Castillo, E. C., Vazquez-Garza, E., Yee-Trejo, D., Garcia-Rivas, G., and Torre-Amione, G. (2020). What Is the Role of the Inflammation in the Pathogenesis of Heart Failure? *Curr. Cardiol. Rep.* 22 (11), 139. doi:10.1007/s11886-020-01382-2

Chang, J. F., Wu, C. C., Hsieh, C. Y., Li, Y. Y., Wang, T. M., and Liou, J. C. (2020). A Joint Evaluation of Impaired Cardiac Sympathetic Responses and Malnutrition-Inflammation Cachexia for Mortality Risks in Hemodialysis Patients. *Front. Med. (Lausanne).* 7, 99. doi:10.3389/fmed.2020.00099

Chang, W., Liu, Y., Xiao, Y., Xu, X., Zhou, S., Lu, X., et al. (2019). Probability Analysis of Hypertension-Related Symptoms Based on XGBoost and Clustering Algorithm. *Appl. Sci.* 9 (6), 1215. doi:10.3390/app9061215

Chen, T., and Yang, M. (2020). Platelet-to-lymphocyte Ratio Is Associated with Cardiovascular Disease in Continuous Ambulatory Peritoneal Dialysis Patients. *Int. Immunopharmacol* 78, 106063. doi:10.1016/j.intimp.2019.106063

Chou, J. A., Streja, E., Nguyen, D. V., Rhee, C. M., Obi, Y., Inrig, J. K., et al. (2018). Intradialytic Hypotension, Blood Pressure Changes and Mortality Risk in Incident Hemodialysis Patients. *Nephrol. Dial. Transpl.* 33 (1), 149–159. doi:10.1093/ndt/gfx037

Cobb, J. S., Seale, M. A., and Janorkar, A. V. (2021). Evaluation of Machine Learning Algorithms to Predict the Hydrodynamic Radii and Transition Temperatures of Chemo-Biologically Synthesized Copolymers. *Comput. Biol. Med.* 128, 104134. doi:10.1016/j.compbiomed.2020.104134

Cozzolino, M., Galassi, A., Pivari, F., Ciceri, P., and Conte, F. (2017). The Cardiovascular Burden in End-Stage Renal Disease. *Contrib. Nephrol.* 191, 44–57. doi:10.1159/000479250

Cozzolino, M., Mangano, M., Stucchi, A., Ciceri, P., Conte, F., and Galassi, A. (2018). Cardiovascular Disease in Dialysis Patients. *Nephrol. Dial. Transplant.* 33 (Suppl. l_3), iii28–iii34. doi:10.1093/ndt/gfy174

Dorairajan, S., Chockalingam, A., and Misra, M. (2010). Myocardial Stunning in Hemodialysis: what Is the Overall Message? *Hemodial Int.* 14 (4), 447–450. doi:10.1111/j.1542-4758.2010.00495.x

Douvris, A., Zeid, K., Hiremath, S., Bagshaw, S. M., Wald, R., Beaubien-Souligny, W., et al. (2019). Mechanisms for Hemodynamic Instability Related to Renal Replacement Therapy: a Narrative Review. *Intensive Care Med.* 45 (10), 1333–1346. doi:10.1007/s00134-019-05707-w

Ebong, I. A., Goff, D. C., Jr., Rodriguez, C. J., Chen, H., and Bertoni, A. G. (2014). Mechanisms of Heart Failure in Obesity. *Obes. Res. Clin. Pract.* 8 (6), e540–e548. doi:10.1016/j.orcp.2013.12.005

Ekdahl, K. N., Soveri, I., Hilborn, J., Fellstrom, B., and Nilsson, B. (2017). Cardiovascular Disease in Haemodialysis: Role of the Intravascular Innate Immune System. *Nat. Rev. Nephrol.* 13 (5), 285–296. doi:10.1038/nrneph.2017.17

Emerging Risk Factors CollaborationKaptoge, S., Di Angelantonio, E., Lowe, G., Pepys, M. B., Thompson, S. G., Collins, R., et al. (2010). C-reactive Protein Concentration and Risk of Coronary Heart Disease, Stroke, and Mortality: an Individual Participant Meta-Analysis. *The Lancet* 375 (9709), 132–140. doi:10.1016/S0140-6736(09)61717-7

Gary, T., Pichler, M., Belaj, K., Hafner, F., Gerger, A., Froehlich, H., et al. (2013). Platelet-to-lymphocyte Ratio: a Novel Marker for Critical Limb Ischemia in Peripheral Arterial Occlusive Disease Patients. *PLoS One* 8 (7), e67688. doi:10.1371/journal.pone.0067688

Gedfew, M., Ayenew, T., Mengst, B., Yirga, T., Zelalem, M., Worku, Y., et al. (2020). Incidence and Predictors of Congestive Heart Failure Among Hemodialysis Patients at Felege Hiote Referral Hospital, Northwest Ethiopia, 2020: Retrospective Cohort Study. *Res. Rep. Clin. Cardiol.* 11, 65–79. doi:10.2147/rrcc.s274942

Ghebre, Y. T., Yakubov, E., Wong, W. T., and Krishnamurthy, P. (2016). Vascular Aging: Implications for Cardiovascular Disease and Therapy. *Transl Med.* 6 (04). doi:10.4172/2161-1025.1000183

Groenewegen, A., Rutten, F. H., Mosterd, A., and Hoes, A. W. (2020). Epidemiology of Heart Failure. *Eur. J. Heart Fail.* 22 (8), 1342–1356. doi:10.1002/ejhf.1858

House, A. A., Wanner, C., Sarnak, M. J., Pina, I. L., McIntyre, C. W., Komenda, P., et al. (2019). Heart Failure in Chronic Kidney Disease: Conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney Int.* 95 (6), 1304–1317. doi:10.1016/j.kint.2019.02.022

Kagiyama, N., Shrestha, S., Farjo, P. D., and Sengupta, P. P. (2019). Artificial Intelligence: Practical Primer for Clinical Research in Cardiovascular Disease. *J. Am. Heart Assoc.* 8 (17), e012788. doi:10.1161/JAHA.119.012788

Koteluk, O., Wartecki, A., Mazurek, S., Kolodziejczak, I., and Mackiewicz, A. (2021). How Do Machines Learn? Artificial Intelligence as a New Era in Medicine. *J. Pers Med.* 11 (1), 32. doi:10.3390/jpm11010032

Kourou, K., Manikis, G., Poikonen-Saksela, P., Mazzocco, K., Pat-Horenczyk, R., Sousa, B., et al. (2021). A Machine Learning-Based Pipeline for Modeling Medical, Socio-Demographic, Lifestyle and Self-Reported Psychological Traits as Predictors of Mental Health Outcomes after Breast Cancer Diagnosis: An Initial Effort to Define Resilience Effects. *Comput. Biol. Med.* 131, 104266. doi:10.1016/j.compbiomed.2021.104266

Kwon, E., Cho, M., Kim, H., and Son, H. S. (2020). A Study on Host Tropism Determinants of Influenza Virus Using Machine Learning. *Curr. Bioinform* 15 (2), 121–134. doi:10.2174/1574893614666191104160927

Li, S., and Zhang, X. (2020). Research on Orthopedic Auxiliary Classification and Prediction Model Based on XGBoost Algorithm. *Neural Comput. Appl.* 32 (59), 1971–1979. doi:10.1007/s00521-019-04378-4

Liu, M., Li, X. C., Lu, L., Cao, Y., Sun, R. R., Chen, S., et al. (2014). Cardiovascular Disease and its Relationship with Chronic Kidney Disease. *Eur. Rev. Med. Pharmaco* 18 (19), 2918–2926.

McHugh, D., and Gil, J. (2018). Senescence and Aging: Causes, Consequences, and Therapeutic Avenues. *J. Cel Biol* 217 (1), 65–77. doi:10.1083/jcb.201708092

Meng, C., Zhang, J., Ye, X., Guo, F., and Zou, Q. (2020). Review and Comparative Analysis of Machine Learning-Based Phage Virion Protein Identification Methods. *Biochim. Biophys. Acta Proteins Proteom* 1868 (6), 140406. doi:10.1016/j.bbapap.2020.140406

Mpanya, D., Celik, T., Klug, E., and Ntsinjana, H. (2021). Machine Learning and Statistical Methods for Predicting Mortality in Heart Failure. *Heart Fail. Rev.* 26 (3), 545–552. doi:10.1007/s10741-020-10052-y

Ouwerkerk, W., Voors, A. A., and Zwinderman, A. H. (2014). Factors Influencing the Predictive Power of Models for Predicting Mortality And/or Heart Failure Hospitalization in Patients with Heart Failure. *JACC Heart Fail.* 2 (5), 429–436. doi:10.1016/j.jchf.2014.04.006

Rangaswami, J., and McCullough, P. A. (2018). Heart Failure in End-Stage Kidney Disease: Pathophysiology, Diagnosis, and Therapeutic Strategies. *Semin. Nephrol.* 38 (6), 600–617. doi:10.1016/j.semnephrol.2018.08.005

Rapa, S. F., Di Iorio, B. R., Campiglia, P., Heidland, A., and Marzocco, S. (2019). Inflammation and Oxidative Stress in Chronic Kidney Disease-Potential

Therapeutic Role of Minerals, Vitamins and Plant-Derived Metabolites. *Int. J. Mol. Sci.* 21 (1), 263. doi:10.3390/ijms21010263

Rienstra, M., Sun, J. X., Magnani, J. W., Sinner, M. F., Lubitz, S. A., Sullivan, L. M., et al. (2012). White Blood Cell Count and Risk of Incident Atrial Fibrillation (From the Framingham Heart Study). *Am. J. Cardiol.* 109 (4), 533–537. doi:10.1016/j.amjcard.2011.09.049

Rodriguez-Perez, R., and Bajorath, J. (2020). Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions. *J. Comput. Aided Mol. Des.* 34 (10), 1013–1026. doi:10.1007/s10822-020-00314-0

Samanta, R., Chan, C., and Chauhan, V. S. (2019). Arrhythmias and Sudden Cardiac Death in End Stage Renal Disease: Epidemiology, Risk Factors, and Management. *Can. J. Cardiol.* 35 (9), 1228–1240. doi:10.1016/j.cjca.2019.05.005

Schefold, J. C., Filippatos, G., Hasenfuss, G., Anker, S. D., and von Haehling, S. (2016). Heart Failure and Kidney Dysfunction: Epidemiology, Mechanisms and Management. *Nat. Rev. Nephrol.* 12 (10), 610–623. doi:10.1038/nrneph.2016.113

Sun, C. Y., Sung, J. M., Wang, J. D., Li, C. Y., Kuo, Y. T., Lee, C. C., et al. (2019). A Comparison of the Risk of Congestive Heart Failure-Related Hospitalizations in Patients Receiving Hemodialysis and Peritoneal Dialysis - A Retrospective Propensity Score-Matched Study. *PLoS One* 14 (10). doi:10.1371/journal.pone.0223336

Van Buren, P. N. (2016). Evaluation and Treatment of Hypertension in End-Stage Renal Disease Patients on Hemodialysis. *Curr. Cardiol. Rep.* 18 (12), 125. doi:10.1007/s11886-016-0805-y

Van Linthout, S., and Tschope, C. (2017). Inflammation - Cause or Consequence of Heart Failure or Both? *Curr. Heart Fail. Rep.* 14 (4), 251–265. doi:10.1007/s11897-017-0337-9

Vigen, R., Maddox, T. M., and Allen, L. A. (2012). Aging of the United States Population: Impact on Heart Failure. *Curr. Heart Fail. Rep.* 9 (4), 369–374. doi:10.1007/s11897-012-0114-8

Virani, S. S., Alonso, A., Aparicio, H. J., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., et al. (2021). Heart Disease and Stroke Statistics-2021 Update: A Report from the American Heart Association. *Circulation* 143 (8), e254–e743. doi:10.1161/CIR.0000000000000950

Voors, A. A., Ouwerkerk, W., Zannad, F., van Veldhuisen, D. J., Samani, N. J., Ponikowski, P., et al. (2017). Development and Validation of Multivariable Models to Predict Mortality and Hospitalization in Patients with Heart Failure. *Eur. J. Heart Fail.* 19 (5), 627–634. doi:10.1002/ejhf.785

Wang, A. Y., and Sanderson, J. E. (2011). Current Perspectives on Diagnosis of Heart Failure in Long-Term Dialysis Patients. *Am. J. Kidney Dis.* 57 (2), 308–319. doi:10.1053/j.ajkd.2010.07.019

Yao, Z., Zheng, Z., Ke, W., Wang, R., Mu, X., Sun, F., et al. (2019). Prognostic Nomogram for Bladder Cancer with Brain Metastases: a National Cancer Database Analysis. *J. Transl Med.* 17 (1), 411. doi:10.1186/s12967-019-2109-7

Yu, X., Zhou, J., Zhao, M., Yi, C., Duan, Q., Zhou, W., et al. (2020). Exploiting XG Boost for Predicting Enhancer-Promoter Interactions. *Curr. Bioinform* 15 (9), 1036–1045. doi:10.2174/1574893615666200120103948

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.

Copyright © 2022 Wang, Miao, Xiao, Huang, Sun, Wang, Li and You. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 16 May 2022
doi: 10.3389/fgene.2022.909040

Identifying Functions of Proteins in Mice With Functional Embedding Features

Hao Li[1†], ShiQi Zhang[2†], Lei Chen[3†], Xiaoyong Pan[4], ZhanDong Li[1], Tao Huang[5,6] and Yu-Dong Cai[7*]*

[1]College of Biological and Food Engineering, Jilin Engineering Normal University, Changchun, China, [2]Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark, [3]College of Information Engineering, Shanghai Maritime University, Shanghai, China, [4]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China, [5]Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [6]CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [7]School of Life Sciences, Shanghai University, Shanghai, China

OPEN ACCESS

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

Reviewed by:
Jing Yang,
ShanghaiTech University, China
Xuefeng Gu,
Shanghai University of Medicine and Health Sciences, China

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

[†]These authors contributed equally to this work

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

Received: 31 March 2022
Accepted: 28 April 2022
Published: 16 May 2022

Citation:
Li H, Zhang S, Chen L, Pan X, Li Z, Huang T and Cai Y-D (2022) Identifying Functions of Proteins in Mice With Functional Embedding Features. Front. Genet. 13:909040. doi: 10.3389/fgene.2022.909040

In current biology, exploring the biological functions of proteins is important. Given the large number of proteins in some organisms, exploring their functions one by one through traditional experiments is impossible. Therefore, developing quick and reliable methods for identifying protein functions is necessary. Considerable accumulation of protein knowledge and recent developments on computer science provide an alternative way to complete this task, that is, designing computational methods. Several efforts have been made in this field. Most previous methods have adopted the protein sequence features or directly used the linkage from a protein–protein interaction (PPI) network. In this study, we proposed some novel multi-label classifiers, which adopted new embedding features to represent proteins. These features were derived from functional domains and a PPI network via word embedding and network embedding, respectively. The minimum redundancy maximum relevance method was used to assess the features, generating a feature list. Incremental feature selection, incorporating RAndom k-labELsets to construct multi-label classifiers, used such list to construct two optimum classifiers, corresponding to two key measurements: accuracy and exact match. These two classifiers had good performance, and they were superior to classifiers that used features extracted by traditional methods.

Keywords: mouse protein, multi-label classification, embedding features, rakel, feature selection

1 INTRODUCTION

Protein is a major component associated with the maintenance of normal physical functions in cells (Milo, 2013). As the essential regulator and effector for almost all living creatures with cellular structures, proteins participate in physical biological processes in two major approaches (Aebersold and Mann, 2016). First, proteins contribute to the regulation of essential biological functions. According to recent publications, proteins are associated with various biological processes, including cell proliferation (Üretmen Kagıalı et al., 2017), enzyme-mediated metabolic processes (Davidi and Milo, 2017), DNA replication (Mughal et al., 2019), cell signaling via ligand binding (Hotamisligil

and Davis, 2016), and responses to internal or external stimulations (Chivasa and Slabas, 2012), all of which are quite complex and essential functions for living creatures. In addition, proteins can construct basic cellular structures (Aebersold and Mann, 2016), maintain the stability of cellular microenvironment, and participate in the formation of complex macrostructures of living creatures, such as hairs and nails. Considering the significance of proteins for living creatures, their biological functions and related detailed underlying mechanisms have been widely studied as an irreplaceable field in current biological studies.

Different kinds of proteins in humans are generated by 19823 predicted or confirmed protein-coding genes (Beck et al., 2011; Milo, 2013). For mouse, as a widely used experimental model, several proteins are translated from approximately 12300 specific protein-coding genes and their isoforms (Church et al., 2009). Therefore, considering the large number of proteins in humans and mice, exploring protein functions by analyzing all candidate proteins one by one through traditional experiments is impossible. For the systematic study of protein functions, computational methods and databases are introduced. Early in 2004, Ruepp et al. have already presented an effective and simplified annotation scheme for systematic classification of proteins (Ruepp et al., 2004). Using such annotation tools, proteins can be clustered into 24 functional categories. The final summary of these 24 categories is generated by balancing manual operative convenience, categorial specificity, and adaptability for further downstream analyses. Therefore, annotating proteins with these 24 categories may be an efficient and convenient way for the exploration of initial protein function.

However, in the presence of clusters and related annotated proteins, computational methods for classification may also be necessary for further systematic protein function explorations. In 2011, Hu et al. proposed two computational methods, namely, network-based and hybrid-property methods, to identify the functions of mouse proteins among the aforementioned 24 categories (Hu et al., 2011). The final method integrated these two methods in a way that the network-based method was initially applied to make prediction; if this method cannot provide predicted results, then the hybrid-property method would make further prediction. In addition, Huang et al. provided three computational methods for the prediction of mouse protein functions based on the 24 candidate categories (Huang et al., 2016). Considering the biochemical properties of proteins and specific functioning approaches for most proteins via protein–protein interactions (PPI), three methods were presented for functional annotation/prediction: 1) sequence similarity-based prediction, 2) weighted PPI-based prediction, and 3) sequence recoding-based prediction using PseAAC (Shen and Chou, 2008). The two above-mentioned studies all used mouse proteins and their functional categories reported in the Mouse Functional Genome Database (MfunGD, http://mips.gsf.de/genre/proj/mfungd/) (Ruepp et al., 2006). However, the above-mentioned methods were not absolute multi-label classifiers as they can only provide the category sequence. Moreover, determining predicted categories for a query protein remains a problem. This study continued doing some work in this field. Furthermore, Zhang et al. developed I-TASSER/COFACTOR method for neXtProt project to predict GO functions of proteins based on their structures and interactions (Zhang et al., 2018; Zhang et al., 2019). NetGO (https://issubmission.sjtu.edu.cn/ng2/) predicted protein functions by integrating massive sequence, text, domain/family and network information with Naïve GO term frequency, BLAST-KNN, LR-3mer, LR-InterPro, LR-ProFET, Net-KNN, LR-text and Seq-RNN (You et al., 2019; Yao et al., 2021).

This study also adopted mouse proteins and their function annotations reported in MfunGD. For each protein, we extracted features from two aspects. On the one hand, embedding features derived from functional domains of proteins were extracted, which can indicate the essential properties of proteins. The functional domains were retrieved from the InterPro database (Blum et al., 2021), and features were obtained by a natural language processing method, Word2vec (Mikolov et al., 2013; Church, 2017). On the other hand, other embedding features were obtained from a PPI network, which contained the linkage information to other proteins. We used the PPI network reported in STRING (Szklarczyk et al., 2015), and Node2vec (Grover and Leskovec, 2016) was applied to such network to obtain embedding features. Embedding features were collected to represent all mouse proteins. Afterward, a feature selection procedure, including the minimum redundancy maximum relevance (mRMR) method (Peng et al., 2005) and incremental feature selection (IFS) (Liu and Setiono, 1998), was designed to select essential embedding features. These features were inputted to RAndom k-labELsets (RAKEL) (Tsoumakas and Vlahavas, 2007) using a support vector machine (SVM) (Cortes and Vapnik, 1995) or random forest (RF) (Breiman, 2001) as the base classifier to construct the multi-label classifiers. The comparison results indicated that our classifiers were superior to classifiers using traditional protein features.

2 METHODS AND MATERIALS

This study aimed to predict the functions of mouse proteins. First, we used Word2vec and Node2vec to obtain embeddings of mouse proteins and identify the essential embedding features via the mRMR method. Then, we applied RAKEL, incorporating SVM or RF as the base classifier, to IFS to construct good multi-label classifiers.

2.1 Dataset

The original mouse proteins and their functions were sourced from a previous study (Hu et al., 2011), which were downloaded from MfunGD (Ruepp et al., 2006). The functions of proteins were determined by manual annotation of the literature and GO annotation (Ashburner and Lewis, 2002; Camon et al., 2003). After excluding proteins without functional domain or interaction information, 9655 proteins were obtained. These mouse proteins were further processed by CD-HIT (Fu et al., 2012) with cutoff of 0.4. 6950 mouse proteins were kept. These

TABLE 1 | Number of proteins in each functional category.

Index	Category	Number of Proteins		
		Training dataset	Test dataset	Overall
1	METABOLISM	1152	280	1432
2	ENERGY	247	64	311
3	CELL CYCLE AND DNA PROCESSING	473	124	597
4	TRANSCRIPTION	906	229	1135
5	PROTEIN SYNTHESIS	213	45	258
6	PROTEIN FATE (folding, modification, destination)	983	234	1217
7	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	3316	868	4184
8	REGULATION OF METABOLISM AND PROTEIN FUNCTION	414	102	516
9	CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES	915	227	1142
10	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM	1228	328	1556
11	CELL RESCUE, DEFENSE AND VIRULENCE	318	76	394
12	INTERACTION WITH THE ENVIRONMENT	501	138	639
13	SYSTEMIC INTERACTION WITH THE ENVIRONMENT	488	149	637
14	TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS	3	1	4
15	CELL FATE	550	171	721
16	DEVELOPMENT (Systemic)	421	127	548
17	BIOGENESIS OF CELLULAR COMPONENTS	287	68	355
18	CELL TYPE DIFFERENTIATION	146	39	185
19	TISSUE DIFFERENTIATION	144	37	181
20	ORGAN DIFFERENTIATION	237	53	290
21	SUBCELLULAR LOCALIZATION	3920	947	4867
22	CELL TYPE LOCALIZATION	80	15	95
23	TISSUE LOCALIZATION	82	26	108
24	ORGAN LOCALIZATION	168	44	212
Sum number of proteins in all categories		17,192	4392	21,584
Number of different proteins		5560	1390	6950

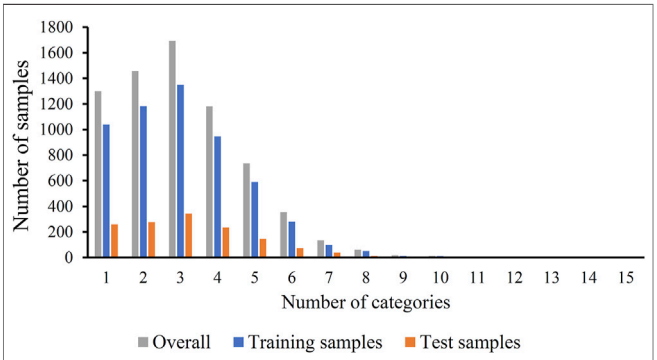


FIGURE 1 | Distribution of training, test and overall samples based on the number of categories that they belong to. Several samples belong to two or more categories.

proteins were classified into 24 functional categories, which are listed in the second column of **Table 1**. In this table, the number of proteins in each category is also provided (last column of **Table 1**). The total number of proteins in all categories was 21584, which was higher than the number of different proteins (6950), indicating that several proteins were in more than one category. Among 6950 proteins, 1299 proteins belonged to exact one functional category, whereas others belonged to two or more categories, and no proteins belonged to more than fifteen categories. The distribution of 6950 proteins based on the number of categories that they belonged to is shown in **Figure 1**. Accordingly, assigning functional labels to mouse proteins was a multi-label classification problem.

To fully evaluate the final classifiers, 6950 proteins were divided into one training dataset and one test dataset, where the training dataset contained 5560 (80%) mouse proteins and the test dataset consisted of 1390 (20%) proteins. The distribution of proteins in training and test datasets based on the number of categories that they belonged to is shown in **Figure 1**. For convenience, the training and test datasets were denoted as S_{tr} and S_{te}, respectively. The number of proteins in S_{tr} and S_{te} for each functional category is also listed in **Table 1**.

2.2 Feature Extraction

In this study, a novel feature representation scheme was presented to encode each mouse protein. This scheme extracted two types of embedding features. The first type of features was derived from functional domains of proteins, whereas the second one was obtained from a PPI network.

2.2.1 Features Derived From a Functional Domain Using Word2vec

Functional domain is a type of information, which is widely used to study various protein-related problems (Cai and Chou, 2005; Xu et al., 2008; Chen et al., 2010; Zhou et al., 2017). One-hot is the classic scheme to extract features from the functional domain. In such scheme, each protein was encoded into a binary vector. However, the model based on features obtained by this scheme was quite sensitive to some domains. Here, we adopted natural language processing to extract features. The functional domain

information of all mouse proteins was retrieved from the InterPro database (http://www.ebi.ac.uk/interpro/, October 2020) (Blum et al., 2021). A total of 16,797 domains were involved. Each mouse protein was annotated by at least one domain. Domains were regarded as words and proteins annotated by some domains as sentences. Word2vec (Mikolov et al., 2013; Church, 2017) was used to obtain embedding features of each domain. Its brief description was shown as follows.

Word2vec was widely used to generate word embeddings in natural language processing. It established the mapping of words to part-of-speech relationships and converted words into fixed-length real-valued vectors. The similarity of the words can be measured and characterized by the similarity of vector space. When using Word2vec, the word vector and sentence vector of features must be calculated. The probability of feature w_i of sentence j in category n is defined as follows:

$$P_{n,j}(w_i) = \frac{f_n(w_i)}{\sum_{n\in N} f_n(w_i)} \tag{1}$$

where $f_n(w_i)$ indicates the frequency of feature w_i in the sentence of category n. Then, the weight of feature w_i can be normalized as follows:

$$\omega_i = \frac{\exp\left(P_{n,j}(w_i) + 1\right)}{\sum \exp\left(P_{n,j}(w_i) + 1\right)} \tag{2}$$

The sentence vector of sentence j in category n is given as follows:

$$V_{n,j} = \frac{1}{f_j} \sum_{i=1}^{m} \omega_i W(w_i) \tag{3}$$

where f_j represents the frequency of features in sentence j, and $W(w_i)$ indicates the word vector of feature w_i. After calculating word vector $W(w_i)$ and sentence vector $V_{n,j}$ of feature w_i, the importance of feature w_i in the sentence can be measured by the distance between the word vector and the sentence vector of feature w_i by using the cosine distance:

$$\begin{aligned} \mathrm{dis}\left(V_{n,j}, W(w_i)\right) &= \cos\left(V_{n,j}, W(w_i)\right) \\ &= \frac{V_{n,j} W(w_i)}{\left|V_{n,j}\right| * \left|W(w_i)\right|} \end{aligned} \tag{4}$$

The feature, whose distance value was within the scale, can be selected on the basis of the ratio of feature selection to achieve the purpose of screening and distinguishing multiple categories.

This study used the Word2vec program reported in genism (https://github.com/RaRe-Technologies/gensim). This program was performed with its default parameters. As mentioned previously, each domain was called as a word. Thus, by applying the Word2vec program, a 256-D feature vector was assigned to each domain. The feature vector of a mouse protein was defined as the average vector of feature vectors of domains, which was annotated on such protein. For convenience, such features were called domain embedding features.

2.2.2 Features Derived From a Protein–Protein Interaction Network Using Node2vec

The above-mentioned embedding features of proteins were extracted from the essential properties of proteins. They cannot reflect the relationship among proteins. Recently, several network embedding algorithms, such as DeepWalk (Perozzi et al., 2014), Node2vec (Grover and Leskovec, 2016), and Mashup (Cho et al., 2016), have been proposed, which can abstract linkages in one or more networks and obtain a feature vector for each node. Features accessed in this way contained quite different information from those derived from essential properties of samples. The combination of these two types of features may fully represent each sample. To date, several models with features derived by network embedding algorithms have been set up to investigate different biological problems (Luo et al., 2017; Zhao et al., 2019; Zhou JP. et al., 2020; Pan et al., 2021a; Pan et al., 2021b; Liu et al., 2021; Zhu et al., 2021; Yang and Chen, 2022). In this study, we selected Node2vec to extract embedding features of pdsluroteins.

A network was necessary to execute Node2vec. Here, we used the PPI network reported in STRING (version 10.0) (Szklarczyk et al., 2015). The PPI information of mouse was first downloaded from STRING. Each PPI contained two proteins, encoded by Emsenbl IDs, and one confidence score. Such score was obtained by investigating several aspects of proteins, such as close neighborhood in genomes, gene fusion, occurrence across different species, gene coexpression, literature description, etc. Thus, it can widely assess the relationship among proteins. Accordingly, the PPI network used proteins as nodes, and two nodes were connected by an edge if and only if their corresponding proteins can constitute a PPI that had a confidence score larger than 0. Furthermore, we placed weight on each edge, which was defined as the confidence score of the corresponding edge. The PPI network contained 20684 nodes and 2849682 edges.

Node2vec was applied to the above-mentioned PPI network to obtain embedding features of proteins. Node2vec can be deemed as a network version of Word2vec. It produced several paths by setting each node in the network as the starting point. Each path was extended by considering the neighbors of the current end point. After generating a predefined number of paths, the nodes in each path were called as words, whereas each path was considered as a sentence. A feature vector was obtained for each node through Word2vec.

In this study, we used the Node2vec program downloaded from https://snap.stanford.edu/node2vec/. For convenience, default parameters were used. Such program was performed on the mouse PPI network. The dimension was set to 500. Finally, each mouse protein was represented by a 500-D feature vector. Features derived from PPI network via Word2vec were called network embedding features.

By combining the domain and network embedding features derived from functional domains of proteins and a PPI network, a 756-D feature vector was obtained to represent each mouse protein.

2.3 Feature Selection

The embedding features obtained by Word2vec and Node2vec were concatenated as the final representation of a protein. We obtained a 756-D vector for each protein. Evidently, some features may be important for assigning functional labels to mouse protein, whereas others were not. Therefore, using a feature selection procedure is necessary to screen out essential features. As several proteins had two or more functional labels, that is, they belonged to two or more functional categories, the original dataset, in which samples were assigned to multiple labels, was transformed into a new dataset in the following manner. If one sample had multiple labels, then this sample would be copied multiple times with different single labels. Then, each sample in the new dataset had only one label.

2.3.1 Minimum Redundancy Maximum Relevance

All features were analyzed by the mRMR method (Peng et al., 2005). Such method evaluated the importance of features by assessing their relevance to class labels and redundancies to other features. A feature list, known as the mRMR feature list, was produced by the mRMR method. This list was produced by selecting features one by one. Initially, the list was empty. A feature with maximum relevance to class labels and minimum redundancies to features already in the list was selected and appended to the list. When all features were in the list, the procedures stopped. Evidently, features with high ranks implied that they had high relevance to class labels and low redundancies to other features. Thus, some top features in such list can comprise a compact feature space for a certain classification algorithm.

The current study used the mRMR program downloaded from http://penglab.janelia.org/proj/mRMR/. It was performed with its default parameters.

2.3.2 Incremental Feature Selection

The mRMR method only generated a feature list. However, selecting the features for constructing the model remained a challenge. Here, IFS (Liu and Setiono, 1998) was used.

Given a feature list (e.g., mRMR feature list), IFS constructed all possible feature subsets. Each subset included some top features in the list. Of each feature subset, a classifier was set up and assessed by a cross-validation method (Kohavi, 1995). The feature subset with the best performance can be obtained. Features in such subset were called optimum features, whereas the classifier using these features was called the optimal classifier.

2.4 Multi-Label Classifier

As mentioned in **Section 2.1**, several proteins were in multiple functional categories. A multi-label classifier should be constructed to assign mouse proteins into functional categories. In general, two schemes were used to construct multi-label classifiers. The first one was problem transformation. It converted the original multi-label classification problem into some single-label classification problems. The second one was algorithm adaption. It extended specific single-label classifiers to deal with multi-label classification problems. This study adopted the first one to construct the multi-label classifier.

The powerful multi-label classification method, RAKEL (Tsoumakas and Vlahavas, 2007), was used to construct the multi-label classifier. Given a problem containing l labels (l=24 in this study), denoted by $L_1, L_2, \ldots, L_l$, RAKEL randomly produced m label subsets that contained k labels, where m and k were the main parameters of RAKEL. For each label subset, the power set was generated, and the members of this set were deemed as new labels. Based on the original labels of one sample, a new label in the power set was assigned to such sample. For example, suppose that the label subset contained three labels, say L_1, L_2 and L_3 and a sample had three labels, say L_1, L_3 and L_5. In this case, this sample was assigned a new label $\{L_1, L_3\}$, which was a member of the power set of the label subset. With such operation, each sample had only one label. Accordingly, a single-label classifier with a base classifier can be set up. RAKEL integrated (m) such single-label classifiers as the final multi-label classifier.

This study used "RAkEL" in Meka (http://waikato.github.io/meka/) (Read et al., 2016). Such tool obtained by the RAKEL method was used to construct multi-label classifiers. The parameters m and k were all set to 10.

2.5 Base Classifiers

In this study, RAKEL was used to construct the multi-label classifier. It needed a base classifier to construct multiple single-label classifiers, which would be integrated into the final multi-label classifier. Here, two classic base classifiers, namely, SVM (Cortes and Vapnik, 1995) and RF (Breiman, 2001), were used, which were widely applied in tackling many biological problems (Kandaswamy et al., 2011; Nguyen et al., 2015; Chen et al., 2017; Zhou JP. et al., 2020; Zhou J.-P. et al., 2020; Liang et al., 2020; Liu et al., 2021; Onesime et al., 2021; Wang et al., 2021; Zhu et al., 2021; Chen et al., 2022; Ding et al., 2022; Li et al., 2022; Wu and Chen, 2022).

2.5.1 Support Vector Machine

SVM was a supervised learning method using statistical learning theory. It can find an optimum hyperplane, which has a maximum margin between the two types of samples, in the N-dimensional space (N represented the number of features) using a Kernel technology (such as a Gaussian kernel), which can map data points to a given category for data classification prediction. The generalization error gradually decreased as the margin increased. A "one-to-one" strategy of SVM corresponded to the binary problem. When the problem extended to multiple classes, the strategy of SVM also changed to a "one-versus-the-rest" strategy.

This study used tool "SMO" integrated in Meka, which implemented a type of SVM. Moreover, this SVM was optimized by Sequential Minimization Optimization (SMO) algorithm (Platt, 1998). Default parameters were adopted. The kernel was a polynomial function and the regularization parameter C was set to 1.

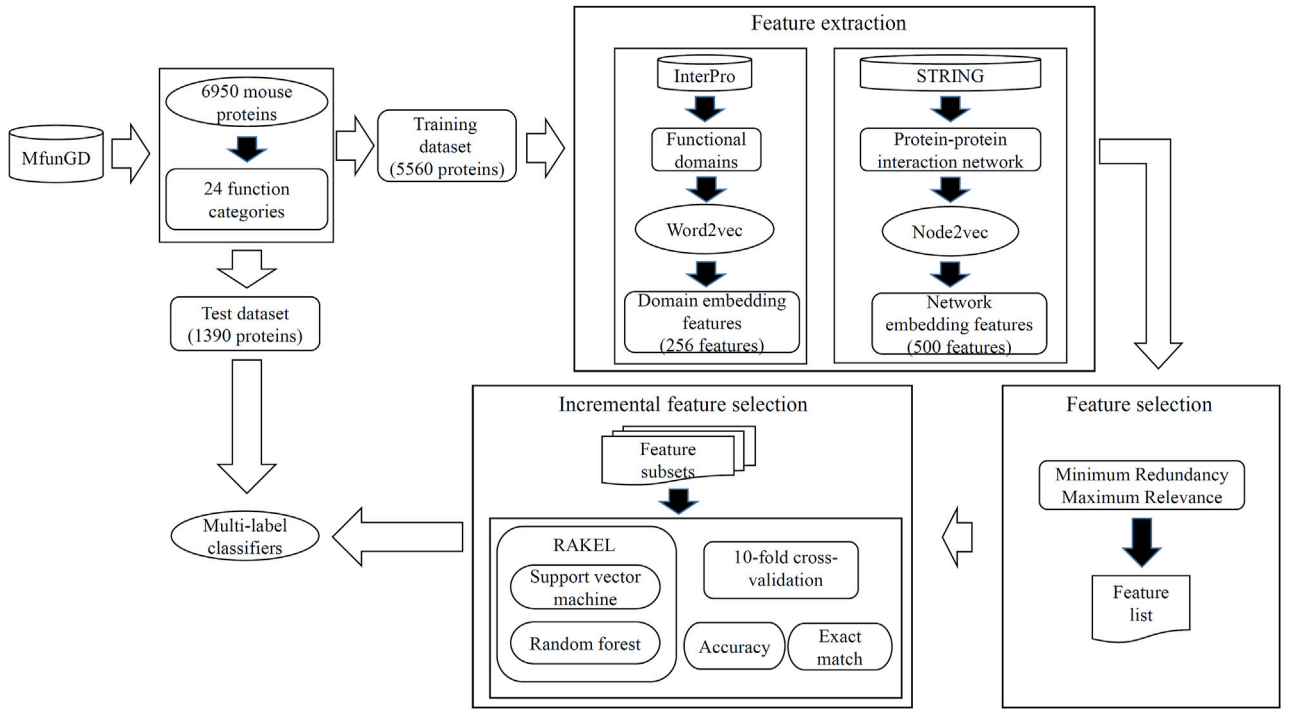


FIGURE 2 | Entire procedures to construct the multi-label classifiers for predicting functions of mouse proteins. Mouse proteins and their function annotations are retrieved from MfunGD. These proteins are randomly divided into one training dataset and one test dataset. Embedding features were derived from protein functional domains and protein–protein interaction network through Word2vec and Node2vec, respectively. A feature selection procedure is used to analyze embedding features, and essential features are fed into RAKEL to construct the multi-label classifiers. Proteins in the test dataset are fed into these classifiers to further evaluate their performance.

2.5.2 Random Forest

RF was a classic classifier used to process classification and regression problems, which was a general machine learning algorithm widely used in bioinformatics. It contained several decision tree classifiers, and subtle differences can be observed among these decision trees. RF determined its output class by aggregating votes produced by different decision trees. Compared with the decision tree, RF can avoid the overfitting problem and improve the performance.

Likewise, this study used the "RandomForest" tool integrated in Meka, which implemented RF. For convenience, default parameters were used, where the number of decision trees was set to 100.

2.6 Performance Measurement

K-fold cross-validation (Kohavi, 1995) is a widely used method to assess the performance of classifiers. In this method, samples are randomly and equally divided into K partitions. One partition is singled out as test dataset one by one, which is used to test the performance of classifier based on rest partitions. Accordingly, each sample is tested only once. The comparison of predicted labels and true labels can lead to some measurements to indicate the performance of classifiers. In this study, we selected 10-fold cross-validation to test all multi-label classifiers.

After the 10-fold cross-validation, each sample was assigned with one or more labels. Some measurements can be computed to assess the predicted results. As a multi-label classifier, accuracy and exact match were the widely used measurements (Zhou JP. et al., 2020; Zhou J.-P. et al., 2020; Pan et al., 2021b; Chen et al., 2021; Tang and Chen, 2022). They can be calculated using the following equations:

$$\begin{cases} \text{Accuracy} = \dfrac{1}{n}\sum_{i=1}^{n}\left(\dfrac{||L_i \cap L_i^{'}||}{||L_i \cup L_i^{'}||}\right) \\ \text{Exact match} = \dfrac{1}{n}\sum_{i=1}^{n}\Theta\left(L_i, L_i^{'}\right) \end{cases} \tag{5}$$

where n stands for the number of samples; L_i and $L_i^{'}$ denote the set consisting of true labels and predicted labels of the i-th sample, respectively; $\Theta(L_i, L_i^{'})$ is defined as follows:

$$\Theta\left(L_i, L_i^{'}\right) = \begin{cases} 1, & \text{If } L_i^{'} \text{ is identical to } L_i \\ 0, & \text{Otherwise.} \end{cases} \tag{6}$$

Evidently, the higher the accuracy or exact match, the higher the performance.

3 RESULTS AND DISCUSSION

In this study, some novel multi-label classifiers were proposed to identify the functions of mouse proteins. The entire procedures

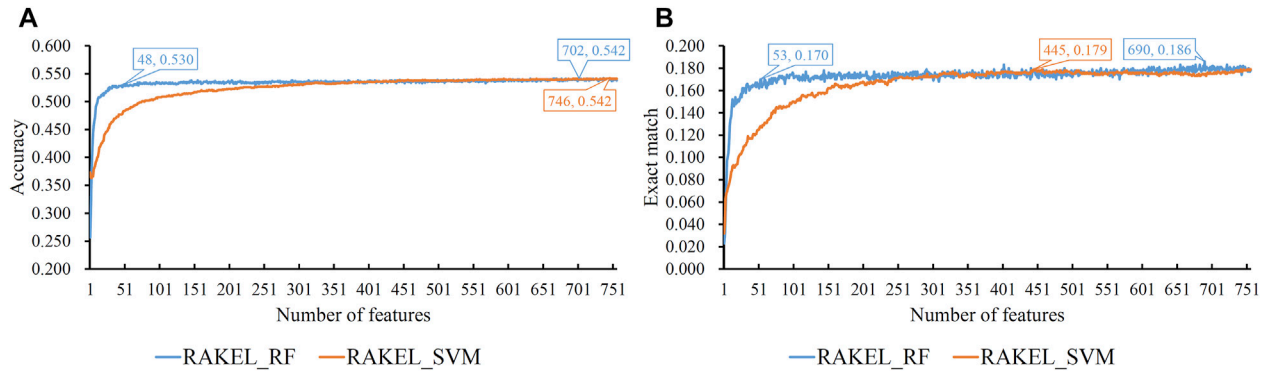


FIGURE 3 | IFS curves on embedding features using different classification methods. **(A)** Accuracy is set to the *Y*-axis. **(B)** Exact match is set to the *Y*-axis. RAKEL_RF/RAKEL_SVM indicates that RAKEL with RF/SVM as the base classifier is used to construct the multi-label classifiers.

are shown in **Figure 2**. In this section, we provided the detailed results of all procedures and made some comparisons to elaborate the unity of the classifier.

3.1 Results of the mRMR Method on Training Dataset

Each protein in S_{tr} was represented by 756 embedding features. These features were analyzed by the mRMR method, resulting in a feature list, which is called the mRMR feature list. This list is provided in **Supplementary Table S1**.

3.2 Results of IFS on Training Dataset

Based on the mRMR feature list provided in **Supplementary Table S1**, IFS was used to construct several feature subsets and set up a multi-label classifier on each feature subset. Each multi-label classifier was set up with RAKEL, and the SVM or RF was selected as the base classifier. 10-fold cross-validation was used to assess the performance of each classifier. The predicted results were assessed by calculating the accuracy and exact match, as mentioned in **Section 2.6**, which are available in **Supplementary Table S2**. Some IFS curves are plotted in **Figure 3** to show the performance of multi-label classifiers with different base classifiers and feature subsets, where the *X*-axis represented the number of features, and the *Y*-axis represented the accuracy or exact match.

As shown in **Figure 3A**, when the base classifier was RF, the highest accuracy was 0.542, which was produced by using the top 702 features in the list. Thus, we can construct an optimum multi-label classifier with these features and RF. As for another base classifier SVM, the highest accuracy was also 0.542, which was produced by using the top 746 features. An optimum multi-label classifier with SVM can be built using these features. Above two optimum classifiers provided the same accuracy. However, the exact match of the classifier with RF was 0.182 and that of the classifier with SVM was 0.179. Accordingly, the optimum multi-label classifier with RF can be deemed to be superior to the optimum multi-label classifier with SVM. When accuracy was used as the key measurement, we can construct a multi-label classifier using the top 702 features and RF. However, the efficiency of such classifier was not very high because lots of features were involved in such classifier. From **Figure 3A**, we can see that the IFS curve of RF followed a sharp increasing trend when a few features were used. It can quickly provide a quite high accuracy using much less features than SVM. By carefully checking accuracy listed in **Supplementary Table S2** and **Figure 3A**, we can find that when top 48 features were adopted, the classifier with RF can yield the accuracy of 0.530, which was only a little lower than that of the optimum classifier. Such classifier can be picked up as a tool to predict functions of query mouse proteins.

For the exact match, two IFS curves corresponding to two different base classifiers are plotted in **Figure 3B**, from which we can see that the base classifier RF generated the highest exact match of 0.186 when the top 690 features were used, whereas SVM yielded the highest exact match of 0.179 when the top 445 features were used. Evidently, the best multi-label classifier with RF was superior to the best multi-label classifier with SVM when exact match was regarded as the key measurement. Accordingly, we can construct a multi-label classifier using the top 690 features and RF. The same problem also existed for such classifier, i.e., low efficiency. It can be observed from **Figure 3B** that the IFS curve of RF was quite similar to that in **Figure 3A.** The increasing trend was much sharper at the beginning of the curve than that of IFS curve of SVM. This meant that RF can provide a high exact match using a small number of features. When top 53 features were used, the classifier with RF can produce exact match of 0.170, which was a little lower than that of the best multi-label classifier with RF. Accordingly, such classifier can be an efficient tool to identify functions of mouse proteins.

As previously mentioned, different key measurements can lead to different multi-label classifiers. For different prediction purposes, users can select the key measurement and use the corresponding classifier. The performance of above-mentioned classifiers is listed in **Tables 2**, **3**.

TABLE 2 | Accuracy of the important multi-label classifiers with different features on training and test datasets.

Method	Feature	Number of Features	Accuracy	
			Training dataset	Test dataset
RAKEL_RF	Embedding features	702	0.542	0.536
RAKEL_SVM	Embedding features	746	0.542	0.537
RAKEL_RF	Embedding features	48	0.530	0.530
RAKEL_RF	Domain features	26	0.429	0.426
RAKEL_SVM	Domain features	27	0.429	0.428
RAKEL_RF	Linkage features	233	0.462	0.460
RAKEL_SVM	Linkage features	234	0.432	0.424
RAKEL_RF	Domain and linkage features	221	0.470	0.462
RAKEL_SVM	Domain and linkage features	227	0.449	0.433

TABLE 3 | Exact match of the important multi-label classifiers with different features on training and test datasets.

Method	Feature	Number of Features	Exact match	
			Training dataset	Test dataset
RAKEL_RF	Embedding features	690	0.186	0.171
RAKEL_SVM	Embedding features	445	0.179	0.157
RAKEL_RF	Embedding features	53	0.170	0.159
RAKEL_RF	Domain features	25	0.077	0.078
RAKEL_SVM	Domain features	29	0.075	0.077
RAKEL_RF	Linkage features	158	0.130	0.123
RAKEL_SVM	Linkage features	225	0.113	0.104
RAKEL_RF	Domain and linkage features	201	0.135	0.130
RAKEL_SVM	Domain and linkage features	215	0.132	0.111

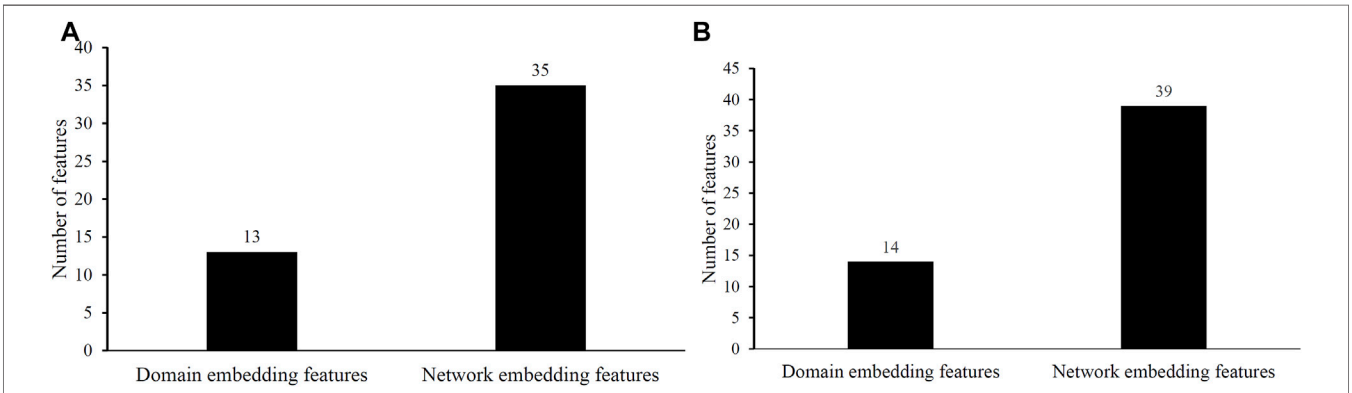


FIGURE 4 | Distribution of embedding features used in two efficient classifiers. **(A)** Distribution of embedding features used in the classifier selected by accuracy. **(B)** Distribution of embedding features used in the classifier selected by exact match.

3.3 Distribution of Embedding Features Used in Two Efficient Classifiers

Two efficient classifiers were constructed as mentioned above, which can be efficient tools for identification of protein functions. 48 and 53 embedding features were involved in these two classifiers, respectively. Their distributions on domain and network embedding features are shown in **Figure 4**. For the classifier with 48 features, 13 were domain embedding features, whereas 35 were network embedding features. As for that with 53 features, similar results can be observed (14 for domain embedding features and 39 for network embedding features). These results indicated that network embedding features gave more contributions for constructing two classifiers. However, domain embedding features were also important. Their combination was one important reason why these two classifiers yielded such good performance.

3.4 Performance of Classifiers on Test Dataset

Based on accuracy and exact match, three multi-label classifiers were built, respectively. These classifiers were further evaluated

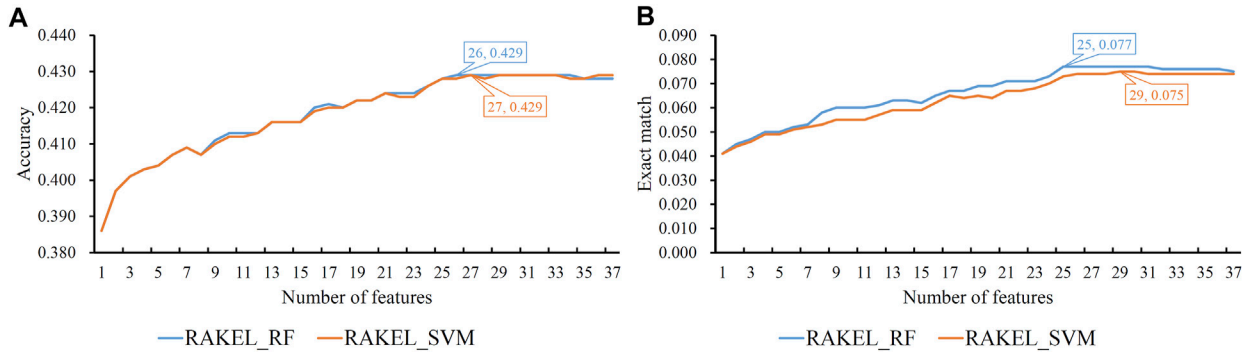


FIGURE 5 | IFS curves on domain features using different classification methods. **(A)** Accuracy is set to the *Y*-axis. **(B)** Exact match is set to the *Y*-axis. RAKEL_RF/RAKEL_SVM indicates that RAKEL with RF/SVM as the base classifier is used to construct the multi-label classifiers.

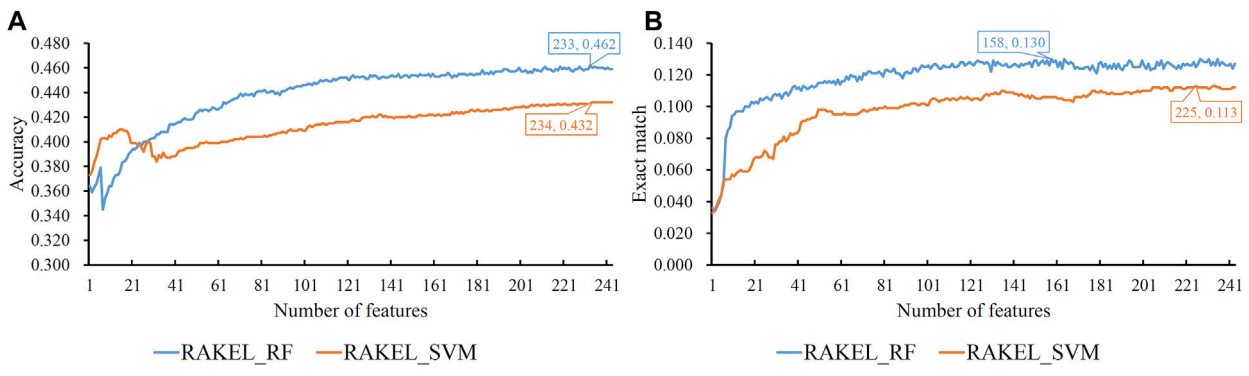


FIGURE 6 | IFS curves on linkage features using different classification methods. **(A)** Accuracy is set to the *Y*-axis **(B)** Exact match is set to the *Y*-axis. RAKEL_RF/RAKEL_SVM indicates that RAKEL with RF/SVM as the base classifier is used to construct the multi-label classifiers.

on S_{te}. Their performance is listed in **Tables 2**, **3**. For the three classifiers selected by accuracy, the optimum classifiers with RF or SVM yielded the accuracies of 0.536 and 0.537 (**Table 2**), respectively, which were slightly lower than those on S_{tr}. The accuracy of the efficient classifier with RF produced the accuracy of 0.530 (**Table 2**), same as that on S_{tr}. These results indicated that the generalization of these classifiers was quite good. As for the three classifiers selected by exact match, they provided exact match values of 0.171, 0.157 and 0.159 (**Table 3**), respectively. They were lower than those on S_{tr}. However, the decrease was in an acceptable range. Thus, the generalization of these classifiers was also satisfied.

3.5 Comparison With Other Classifiers

In this study, we adopted a novel set of features to represent each mouse protein and constructed some multi-label classifiers to predict their functions. This section adopted some classic features to construct the classifiers and make some comparisons.

Two types of embedding features were used in this study. They were derived from the protein functional domain and PPI network. For the protein functional domain, the classic usage of encoding proteins was the one-hot scheme. In detail, a protein was encoded into a binary vector under such scheme. Each domain was used as a dimension, and the component was set to one if the protein had the corresponding domain annotation; otherwise, the component was set to zero. Here, 16797 domains were involved, inducing a 16797-D binary vector for each mouse protein. For an easy description, these features were called as domain features in this study. As for the PPI network, such information can be directly used by selecting all linkages between a protein and all proteins in the network and collecting them in a vector to encode the protein. Accordingly, each mouse protein was represented by a 20684-D vector, as 20684 proteins were found in the PPI network. These features were called as linkage features. Each mouse protein was represented by domain features or linkage features or both of them, inducing three representations of proteins. We investigated the performance of classifiers on each protein representation.

As previously mentioned, proteins were represented by lots of features in each representation. A feature selection procedure was necessary. However, given the large number of features, we first adopted Bortua (Kursa and Rudnicki, 2010; Zhang et al., 2021) to exclude irrelevant features. 37 and 243 features were selected by Bortua for domain and linkage feature representations,

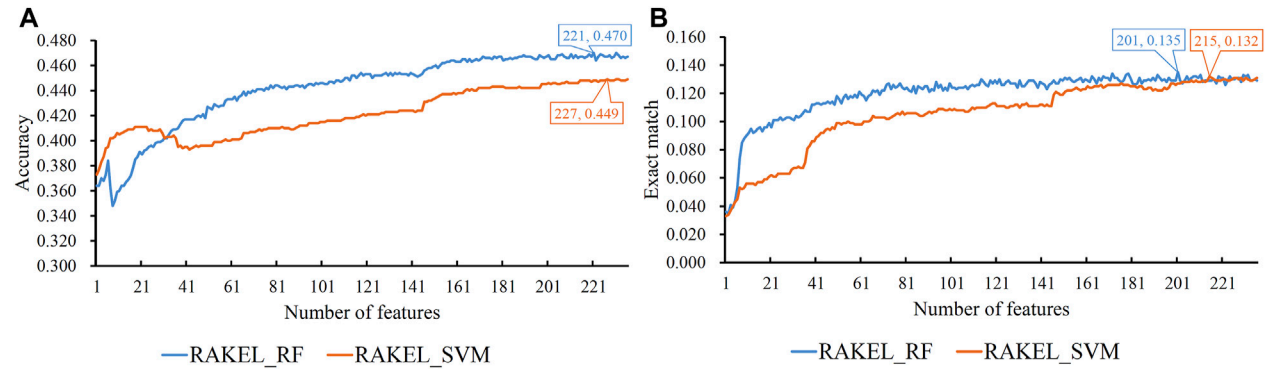


FIGURE 7 | IFS curves on domain and linkage features using different classification methods. **(A)** Accuracy is set to the *Y*-axis. **(B)** Exact match is set to the *Y*-axis. RAKEL_RF/RAKEL_SVM indicates that RAKEL with RF/SVM as the base classifier is used to construct the multi-label classifiers.

respectively. When domain and linkage features were combined together to encode mouse proteins, 236 features were kept by Bortua. Then, these remaining features were evaluated by the mRMR method, resulting in an mRMR feature list for each representation. IFS was used to construct optimum multi-label classifiers for accuracy and exact match. We still used RAKEL to construct the classifiers, and SVM or RF was selected as the base classifier. The IFS results are provided in **Supplementary Tables S3-S5**. Likewise, some IFS curves are plotted in **Figures 5–7**.

The best accuracies for different base classifiers on S_{tr} are listed in **Table 2**, in which those obtained by embedding features are also provided. When the base classifier was RF, the accuracies obtained by domain features, linkage features and both of them were all lower than 0.5, which were much lower than those of the classifiers on embedding features. Furthermore, the base classifier (SVM) yielded similar results (see **Table 2**). As for the exact match, the best values for different base classifiers are listed in **Table 3**, in which those obtained by embedding features are also listed. Evidently, the exact match obtained by embedding features was also higher than that obtained by domain features or linkage features or both of them regardless of the base classifier used (RF or SVM). The improvement was at least 3%. Furthermore, from **Tables 2**, **3**, the classifiers with embedding features also yielded better performance on test dataset S_{te} than those with domain features or linkage features or both of them. All above results indicated that the novel features used in this study were more efficient than the features produced by traditional methods in predicting protein functions. In addition, it can be observed from **Tables 2**, **3** that when domain and linkage features were combined to represent proteins, the classifiers were always better than those only using domain features or linkage features. This fact indicated that combination of the domain and network information of proteins can improve the performance of classifiers. These two types of information can complement each other in predicting functions of proteins.

4 CONCLUSION

In this paper, we proposed some multi-label classifiers to predict the functions of mouse proteins. These classifiers adopted novel features, which were derived from protein functional domains and the PPI network via word embedding and network embedding, respectively. The performance of the classifiers was better than those using features extracted by traditional methods, thereby indicating that the novel features have stronger discriminative power. Therefore, the newly proposed classifiers can be used to predict protein functions, and such novel features can be used to tackle other protein-related problems.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://mips.gsf.de/genre/proj/mfungd.

AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. LC and XP performed the experiments. HL, SZ, and ZL analyzed the results. HL, SZ, and LC wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200, XDA26040304), National Key R&D Program of China (2018YFC0910403), the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.909040/full#supplementary-material

Supplementary Table S1 | mRMR feature list yielded by the mRMR method on training dataset.

Supplementary Table S2 | IFS results with embedding features on training dataset.

Supplementary Table S3 | IFS results with domain features on training dataset.

Supplementary Table S4 | IFS results with linkage features on training dataset.

Supplementary Table S5 | IFS results with domain and linkage features on training dataset.

REFERENCES

Aebersold, R., and Mann, M. (2016). Mass-spectrometric Exploration of Proteome Structure and Function. *Nature* 537, 347–355. doi:10.1038/nature19949

Ashburner, M., and Lewis, S. (2002). On Ontologies for Biologists: the Gene Ontology-Uuntangling the Web. *Novartis Found. Symp.* 247, 66244–52252.

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., et al. (2011). The Quantitative Proteome of a Human Cell Line. *Mol. Syst. Biol.* 7, 549. doi:10.1038/msb.2011.82

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2021). The InterPro Protein Families and Domains Database: 20 Years on. *Nucleic Acids Res.* 49, D344–D354. doi:10.1093/nar/gkaa977

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

Cai, Y.-D., and Chou, K.-C. (2005). Using Functional Domain Composition to Predict Enzyme Family Classes. *J. Proteome Res.* 4, 109–111. doi:10.1021/pr049835p

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., et al. (2003). The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* 13, 662–672. doi:10.1101/gr.461403

Chen, L., Li, Z., Zhang, S., Zhang, Y. H., Huang, T., and Cai, Y. D. (2022). Predicting RNA 5-methylcytosine Sites by Using Essential Sequence Features and Distributions. *Biomed. Res. Int.* 2022, 4035462. doi:10.1155/2022/4035462

Chen, L., Feng, K.-Y., Cai, Y.-D., Chou, K.-C., and Li, H.-P. (2010). Predicting the Network of Substrate-Enzyme-Product Triads by Combining Compound Similarity and Functional Domain Composition. *Bmc Bioinforma.* 11, 293. doi:10.1186/1471-2105-11-293

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., et al. (2017). Identify Key Sequence Features to Improve CRISPR sgRNA Efficacy. *IEEE Access* 5, 26582–26590. doi:10.1109/access.2017.2775703

Chen, W., Chen, L., and Dai, Q. (2021). iMPT-FDNPL: Identification of Membrane Protein Types with Functional Domains and a Natural Language Processing Approach. *Comput. Math. Methods Med.* 2021, 7681497. doi:10.1155/2021/7681497

Chivasa, S., and Slabas, A. R. (2012). Plant extracellularATP Signalling: New Insight from Proteomics. *Mol. Biosyst.* 8, 445–452. doi:10.1039/c1mb05278k

Cho, H., Berger, B., and Peng, J. (2016). Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Syst.* 3, 540–548. doi:10.1016/j.cels.2016.10.017

Church, D. M., Goodstadt, L., Hillier, L. W., Zody, M. C., Goldstein, S., She, X., et al. (2009). Lineage-specific Biology Revealed by a Finished Genome Assembly of the Mouse. *PLoS Biol.* 7, e1000112.doi:10.1371/journal.pbio.1000112

Church, K. W. (2017). Word2Vec. *Nat. Lang. Eng.* 23, 155–162. doi:10.1017/s1351324916000334

Cortes, C., and Vapnik, V. (1995). Support-vector Networks. *Mach. Learn* 20, 273–297. doi:10.1007/bf00994018

Davidi, D., and Milo, R. (2017). Lessons on Enzyme Kinetics from Quantitative Proteomics. *Curr. Opin. Biotechnol.* 46, 81–89. doi:10.1016/j.copbio.2017.02.007

Ding, S., Wang, D., Zhou, X., Chen, L., Feng, K., Xu, X., et al. (2022). Predicting Heart Cell Types by Using Transcriptome Profiles and a Machine Learning Method. *Life* 12, 228. doi:10.3390/life12020228

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data. *Bioinformatics* 28, 3150–3152. doi:10.1093/bioinformatics/bts565

Grover, A., and Leskovec, J. (2016). "node2vec: Scalable Feature Learning for Networks", in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA: ACM), 855–864.

Hanchuan Peng, H. C., Fuhui Long, F. H., and Ding, C. (2005). Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi:10.1109/tpami.2005.159

Hotamisligil, G. S., and Davis, R. J. (2016). Cell Signaling and Stress Responses. *Cold Spring Harb. Perspect. Biol.* 8, a006072. doi:10.1101/cshperspect.a006072

Hu, L., Huang, T., Shi, X., Lu, W.-C., Cai, Y.-D., and Chou, K.-C. (2011). Predicting Functions of Proteins in Mouse Based on Weighted Protein-Protein Interaction Network and Protein Hybrid Properties. *PLoS One* 6, e14556. doi:10.1371/journal.pone.0014556

Huang, G., Chu, C., Huang, T., Kong, X., Zhang, Y., Zhang, N., et al. (2016). Exploring Mouse Protein Function via Multiple Approaches. *PLoS One* 11, e0166580.doi:10.1371/journal.pone.0166580

Kandaswamy, K. K., Chou, K.-C., Martinetz, T., Möller, S., Suganthan, P. N., Sridharan, S., et al. (2011). AFP-pred: A Random Forest Approach for Predicting Antifreeze Proteins from Sequence-Derived Properties. *J. Theor. Biol.* 270, 56–62. doi:10.1016/j.jtbi.2010.10.037

Kohavi, R. "(1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", in Proceedings of the International Joint Conference on Artificial Intelligence (Lawrence Erlbaum Associates), 1137–1145.

Kursa, M., and Rudnicki, W. (2010). Feature Selection with the Boruta Package. *J. Stat. Softw. Articles* 36, 1–13. doi:10.18637/jss.v036.i11

Li, Z., Wang, D., Liao, H., Zhang, S., Guo, W., Chen, L., et al. (2022). Exploring the Genomic Patterns in Human and Mouse Cerebellums via Single-Cell Sequencing and Machine Learning Method. *Front. Genet.* 13, 857851. doi:10.3389/fgene.2022.857851

Liang, H., Chen, L., Zhao, X., and Zhang, X. (2020). Prediction of Drug Side Effects with a Refined Negative Sample Selection Strategy. *Comput. Math. Methods Med.* 2020, 1573543. doi:10.1155/2020/1573543

Liu, H., Hu, B., Chen, L., and Lu, L. (2021). Identifying Protein Subcellular Location with Embedding Features Learned from Networks. *Cp* 18, 646–660. doi:10.2174/1570164617999201124142950

Liu, H., and Setiono, R. (1998). Incremental Feature Selection. *Appl. Intell.* 9, 217–230. doi:10.1023/a:1008363719778

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A Network Integration Approach for Drug-Target Interaction Prediction and Computational Drug Repositioning from Heterogeneous Information. *Nat. Commun.* 8, 573. doi:10.1038/s41467-017-00680-8

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space," in Proceedings Of theInternational Conference on Learning Representations (Arizona, USA): Scottsdale).

Milo, R. (2013). What Is the Total Number of Protein Molecules Per Cell Volume? A Call to Rethink Some Published Values. *Bioessays* 35, 1050–1055. doi:10.1002/bies.201300066

Mughal, M. J., Mahadevappa, R., and Kwok, H. F. (2019). DNA Replication Licensing Proteins: Saints and Sinners in Cancer. *Seminars Cancer Biol.* 58, 11–21. doi:10.1016/j.semcancer.2018.11.009

Nguyen, T.-T., Huang, J. Z., Wu, Q., Nguyen, T. T., and Li, M. J. (2015). Genome-wide Association Data Classification and SNPs Selection Using Two-Stage Quality-Based Random Forests. *BMC genomics* 16, S5. doi:10.1186/1471-2164-16-s2-s5

Onesime, M., Yang, Z., and Dai, Q. (2021). Genomic Island Prediction via Chi-Square Test and Random Forest Algorithm. *Comput. Math. Methods Med.* 2021, 9969751. doi:10.1155/2021/9969751

Pan, X., Chen, L., Liu, M., Niu, Z., Huang, T., and Cai, Y. D. (2021a). Identifying Protein Subcellular Locations with Embeddings-Based Node2loc. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 19(2):666-675. doi:10.1109/tcbb.2021.3080386

Pan, X., Li, H., Zeng, T., Li, Z., Chen, L., Huang, T., et al. (2021b). Identification of Protein Subcellular Localization with Network and Functional Embeddings. *Front. Genet.* 11, 626500. doi:10.3389/fgene.2020.626500

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: Online Learning of Social Representations", in:Proceedings Of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), 701–710. doi:10.1145/2623330.2623732

Platt, J. (1998). Sequential Minimal Optimizaton: A Fast Algorithm for Training Support Vector Machines,21. *Technical Report MSR-TR*-98-14.

Read, J., Reutemann, P., Pfahringer, B., and Holmes, G. (2016). MEKA: A Multi-label/Multi-Target Extension to WEKA. *J. Mach. Learn. Res.* 17, 1–5.

retmen Kagıalı, Z. C., Şentürk, A., zkan Küçük, N. E., Qureshi, M. H., and zlü, N. (2017). Proteomics in Cell Division. *Proteomics*. 17. 1. doi:10.1002/pmic.201600100

Ruepp, A., Doudieu, O. N., Van Den Oever, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., et al. (2006). The Mouse Functional Genome Database (MfunGD): Functional Annotation of Proteins in the Light of Their Cellular Context. *Nucleic Acids Res.* 34, D568–D571. doi:10.1093/nar/gkj074

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., et al. (2004). The FunCat, a Functional Annotation Scheme for Systematic Classification of Proteins from Whole Genomes. *Nucleic Acids Res.* 32, 5539–5545. doi:10.1093/nar/gkh894

Shen, H.-B., and Chou, K.-C. (2008). PseAAC: a Flexible Web Server for Generating Various Kinds of Protein Pseudo Amino Acid Composition. *Anal. Biochem.* 373, 386–388. doi:10.1016/j.ab.2007.10.012

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.* 43, D447–D452. doi:10.1093/nar/gku1003

Tang, S., and Chen, L. (2022). iATC-NFMLP: Identifying Classes of Anatomical Therapeutic Chemicals Based on Drug Networks, Fingerprints and Multilayer Perceptron. *Curr. Bioinforma.* 36(11):3568-3569. doi:10.2174/1574893617666220318093000

Tsoumakas, G., and Vlahavas, I. (2007). *Random K-Labelsets: An Ensemble Method for Multilabel Classification*. in: 18th European Conference on Machine Learning, 406–417.

Wang, Y., Xu, Y., Yang, Z., Liu, X., and Dai, Q. (2021). Using Recursive Feature Selection with Random Forest to Improve Protein Structural Class Prediction for Low-Similarity Sequences. *Comput. Math. Methods Med.* 2021, 5529389. doi:10.1155/2021/5529389

Wu, Z., and Chen, L. (2022). Similarity-based Method with Multiple-Feature Sampling for Predicting Drug Side Effects. *Comput. Math. Methods Med.* 2022, 9547317. doi:10.1155/2022/9547317

Xu, X., Yu, D., Fang, W., Cheng, Y., Qian, Z., Lu, W., et al. (2008). Prediction of Peptidase Category Based on Functional Domain Composition. *J. Proteome Res.* 7, 4521–4524. doi:10.1021/pr800292w

Yang, Y., and Chen, L. (2022). Identification of Drug-Disease Associations by Using Multiple Drug and Disease Networks. *Cbio* 17, 48–59. doi:10.2174/1574893616666210825115406

Yao, S., You, R., Wang, S., Xiong, Y., Huang, X., and Zhu, S. (2021). NetGO 2.0: Improving Large-Scale Protein Function Prediction with Massive Sequence, Text, Domain, Family and Network Information. *Nucleic Acids Res.* 49, W469–w475. doi:10.1093/nar/gkab398

You, R., Yao, S., Xiong, Y., Huang, X., Sun, F., Mamitsuka, H., et al. (2019). NetGO: Improving Large-Scale Protein Function Prediction with Massive Network Information. *Nucleic Acids Res.* 47, W379–w387. doi:10.1093/nar/gkz388

Zhang, C., Lane, L., Omenn, G. S., and Zhang, Y. (2019). Blinded Testing of Function Annotation for uPE1 Proteins by I-TASSER/COFACTOR Pipeline Using the 2018-2019 Additions to neXtProt and the CAFA3 Challenge. *J. Proteome Res.* 18, 4154–4166. doi:10.1021/acs.jproteome.9b00537

Zhang, C., Wei, X., Omenn, G. S., and Zhang, Y. (2018). Structure and Protein Interaction-Based Gene Ontology Annotations Reveal Likely Functions of Uncharacterized Proteins on Human Chromosome 17. *J. Proteome Res.* 17, 4186–4196. doi:10.1021/acs.jproteome.8b00453

Zhang, Y.-H., Zeng, T., Chen, L., Huang, T., and Cai, Y.-D. (2021). Determining Protein-Protein Functional Associations by Functional Rules Based on Gene Ontology and KEGG Pathway. *Biochimica Biophysica Acta (BBA) - Proteins Proteomics* 1869, 140621. doi:10.1016/j.bbapap.2021.140621

Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting Drug Side Effects with Compact Integration of Heterogeneous Networks. *Cbio* 14, 709–720. doi:10.2174/1574893614666190220114644

Zhou, H., Yang, Y., and Shen, H. B. (2017). Hum-mPLoc 3.0: Prediction Enhancement of Human Protein Subcellular Localization through Modeling the Hidden Correlations of Gene Ontology and Functional Domain Features. *Bioinformatics* 33, 843–853. doi:10.1093/bioinformatics/btw723

Zhou, J.-P., Chen, L., Wang, T., and Liu, M. (2020b). iATC-FRAKEL: a Simple Multi-Label Web Server for Recognizing Anatomical Therapeutic Chemical Classes of Drugs with Their Fingerprints Only. *Bioinformatics* 36, 3568–3569. doi:10.1093/bioinformatics/btaa166

Zhou, J. P., Chen, L., and Guo, Z. H. (2020a). iATC-NRAKEL: An Efficient Multi-Label Classifier for Recognizing Anatomical Therapeutic Chemical Classes of Drugs. *Bioinformatics* 36, 1391–1396. doi:10.1093/bioinformatics/btz757

Zhu, Y., Hu, B., Chen, L., and Dai, Q. (2021). iMPTCE-Hnetwork: A Multilabel Classifier for Identifying Metabolic Pathway Types of Chemicals and Enzymes with a Heterogeneous Network. *Comput. Math. Methods Med.* 2021, 6683051. doi:10.1155/2021/6683051

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.

Copyright © 2022 Li, Zhang, Chen, Pan, Li, Huang and Cai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership