UNSUPERVISED LEARNING MODELS FOR UNLABELED GENOMIC, TRANSCRIPTOMIC & PROTEOMIC DATA

EDITED BY: Jianing Xi and Zhenhua Yu

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-967-9
DOI 10.3389/978-2-88971-967-9

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

UNSUPERVISED LEARNING MODELS FOR UNLABELED GENOMIC, TRANSCRIPTOMIC & PROTEOMIC DATA

Topic Editors:
Jianing Xi, Northwestern Polytechnical University, China
Zhenhua Yu, Ningxia University, China

Citation: Xi, J., Yu, Z., eds. (2021). Unsupervised Learning Models for Unlabeled Genomic, Transcriptomic & Proteomic Data. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-967-9

Table of Contents

frontiers in Genetics

EDITORIAL
published: 11 November 2021
doi: 10.3389/fgene.2021.781698

Editorial: Unsupervised Learning Models for Unlabeled Genomic, Transcriptomic & Proteomic Data

Jianing Xi[1]* *and Zhenhua Yu*[2]*

[1]School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China, [2]School of Information Engineering, Ningxia University, Yinchuan, China

Keywords: unsupervised learning, unlabeled data, OMICS data, genome, transcriptome, proteome

Editorial on the Research Topic

Unsupervised Learning Models for Unlabeled Genomic, Transcriptomic & Proteomic Data

UNSUPERVISED LEARNING MODELS FOR UNLABELED GENOMIC, TRANSCRIPTOMIC AND PROTEOMIC DATA

For unveiling the underlying biological mechanisms, the data of genomics, transcriptomics, proteomics, and other types of omics can offer informative cues for the understanding of underlying biological mechanisms (Muers, 2011). Since manual analysis of the huge amounts of these biological data is impractical, computational efforts of bioinformatics has been introduced as the key of unveiling the biological knowledge in omics data (Manzoni et al., 2018). A promising opportunity for omics data analysis is the recent developments in Artificial Intelligence (AI), which empowers bioinformatics research. Inspired by the advanced AI technology (Huang and Xi, 2020), a considerable number of effective and powerful intelligence approaches have been erupting in the bioinformatics research of omics data (Lightbody et al., 2019).

Nevertheless, it should be noted that, the paradigm of supervised learning framework are widely utilized in most of the recent emerging bioinformatics approaches (Min et al., 2017). Despite the achievements yielded by the existing omics data analysis, one of the main shortcomings is that these previously published approaches restrict annotated labels in the omic data as training set (Yu et al., 2019). In consideration of the massive amount of omic data involved in bioinformatics researches, there are extensively manual efforts required from experts, when such amounts of data are annotated with labels (Xi et al., 2021). Consequently, in omics data, a crucial bottleneck in bioinformatics research of omic data is the insufficiency of annotated labels (Yu et al., 2020).

For circumventing the shortage of manual annotations in omics data, a promising solution is to analyze the unlabeled omic data rather than labeled data, which can save considerable costs of annotation (Xi et al., 2020b). Instead of the widely used paradigm of supervised learning, introducing the paradigm of unsupervised learning can open a new window of omic research, demonstrating great potential for unlabeled omic data analysis Xi et al. (2020b). In comparison to the paradigm of supervised learning, unsupervised learning methods may throw light on the unlabeled omic data analysis, which can overcome the issue of high cost of annotated labels in omic data, and promote the research of omic data free from manual labels (Xi et al., 2020a).

This Research Topic focuses on the recent advanced approaches in the methodology of unsupervised learning and their applications on unlabeled omics data. A total of 9 articles

OPEN ACCESS

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

***Correspondence:**
Jianing Xi
xjn@nwpu.edu.cn
Zhenhua Yu
zhyu@nxu.edu.cn

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

Received: *23 September 2021*
Accepted: *25 October 2021*
Published: *11 November 2021*

Citation:
Xi J and Yu Z (2021) Editorial: Unsupervised Learning Models for Unlabeled Genomic, Transcriptomic & Proteomic Data. Front. Genet. 12:781698. doi: 10.3389/fgene.2021.781698

related to unsupervised learning developments on the analysis of genomic data, transcriptomic data, proteomic data, and multi-omic data are included.

For genomic data analysis, three unsupervised learning approaches were published in the Research Topic, unveiling the aspects disease gene selection and copy number variation detection. Specifically, Xie et al. proposes a standard deviation and cosine similarity based unsupervised feature selection algorithms, which is capable of conducting gene selection for stable biomarkers of disease such as cancer through genomic data (Xie J. et al.). At the same time, Fan et al. proposes a hierarchical clustering based framework to predict the disease genes from stage-specific gene regulatory networks (Fan et al.). Furthermore, Xie et al. proposes a local density and minimum distance based density peak clustering method called dpCNV, for detecting relative large range copy number variation from DNA sequencing data (Xie K. et al.). These advanced approaches mainly cover the methodology of feature selection, hierarchical clustering, and density peak estimation, expanding the frontiers of genomic researches.

For transcriptomic data analysis, there are two papers contributing to RNA data research as the roles of bioinformatics tools. One research in this Research Topic is focusing on in single-cell RNA sequencing (Yu et al., 2021), which aims to overcome the zero-inflated data caused by dropout events (Zhao et al.), where Zhao et al. proposes a dimensionality reduction approach on single-cell RNA sequencing data, which is based on a hierarchical autoencoder consisting of a deep count autoencoder for denoising and a graph autoencoder for dimensional reducing. Meanwhile, for long intergenic non-coding RNA (lincRNA) analysis, Lin and Ma proposes a non-negative matrix factorization approach with co-regularization to predict disease-lincRNA associations (Lin and Ma), which integrates four types of information associated to lincRNA. Generally, the two researches are concentrating on the advanced frontiers of either AI technology research or transcriptomic research.

For proteomic data analysis, there are two articles offering the unsupervised learning methods on two aspects. One aspect is to detect overlapping structures in protein functional modules from proteomic data of protein-protein interactions, where Wang et al. proposes a neighboring local clustering coefficient based overlapping community detection algorithm to mine functional modules in these interactions (Wang Y. et al.). Another aspect is to measure the similarity of proteins, where Zhang et al. further incorporates structural information of Gene Ontology (GO) graph to compensate the consideration of only information content of GO terms, and calculates the similarity of proteins through graph embedding methods (Zhang et al.). These protein interaction graph based approaches in the Research Topic also illustrate the frontiers of proteomic research.

For multi-omic data analysis, this Research Topic also collected two studies which include more than one type of omic data. Detailly, Wang et al. proposes a joint matrix tri-factorization framework for discovering complex biological processes (CBPs) of multi-omics molecules regulation, which reflect the activities of various molecules in living organisms (Wang B. et al.). Moreover, in the prediction of cancer subtypes, to effectively utilize rich heterogeneous information in the multiple view fusion graph of multiple omics data, Liu et al. proposes a multi-smooth representation fusion based multi-view spectral clustering method, which consists of graph construction, graph fusion, and spectral clustering for clustering of cancer subtypes from multi-omic data (Liu et al.). These works also show the frontiers of multi-omic research.

In brief, This collection of contributions in the Research Topic provide a window into the frontiers of unsupervised learning models for unlabeled genomic, transcriptomic and proteomic data. Given the remarkable success of unsupervised learning application in bioinformatics problems, we hope that these approaches can throw light on the problem of data annotation cost, extending the frontiers of bioinformatics research of omic data.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work is partially by National Natural Science Foundation of China (Grant Nos. 61901322 and 61901238), and partially by China Postdoctoral Science Foundation (No. 2020M673494).

ACKNOWLEDGMENTS

We would like to thank Dr. Robin Ferdous for his helpful suggestions on organizing this research topic.

REFERENCES

Huang, Q., Xi, J., and Xi, J. (2020). Editorial: Advanced Computer Methods and Programs in Biomedicine. *Math. Biosciences Eng.* 17, 1940–1943. doi:10.3934/mbe.2020102

Lightbody, G., Haberland, V., Browne, F., Taggart, L., Zheng, H., Parkes, E., et al. (2019). Review of Applications of High-Throughput Sequencing in Personalized Medicine: Barriers and Facilitators of Future Progress in Research and Clinical Application. *Brief. Bioinformatics* 20, 1795–1811. doi:10.1093/bib/bby051

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., et al. (2018). Genome, Transcriptome and Proteome: The Rise of Omics Data and Their Integration in Biomedical Sciences. *Brief. Bioinformatics* 19, 286–302. doi:10.1093/bib/bbw114

Min, S., Lee, B., and Yoon, S. (2017). Deep Learning in Bioinformatics. *Brief Bioinform* 18, 851–869. doi:10.1093/bib/bbw068

Muers, M. (2011). Transcriptome to Proteome and Back to Genome. *Nat. Rev. Genet.* 12, 518. doi:10.1038/nrg3037

Xi, J., Li, A., and Wang, M. (2020a). HetRCNA: A Novel Method to Identify Recurrent Copy Number Alternations from Heterogeneous Tumor Samples Based on Matrix Decomposition Framework. *Ieee/acm Trans. Comput. Biol. Bioinf.* 17, 422–434. doi:10.1109/TCBB.2018.2846599

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. (2020b). Inferring Subgroup-Specific Driver Genes from Heterogeneous Cancer Samples via Subspace Learning with Subgroup Indication. *Bioinformatics* 36, 1855–1863. doi:10.1093/bioinformatics/btz793

Xi, J., Ye, L., Huang, Q., and Li, X. (2021). "Tolerating Data Missing in Breast Cancer Diagnosis from Clinical Ultrasound Reports via Knowledge Graph Inference," in KDD'21 Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore (New York, NY, USA: ACM), 3756–3764. doi:10.1145/3447548.3467106

Yu, Z., Du, F., Sun, X., and Li, A. (2019). SCSsim: an Integrated Tool for Simulating Single-Cell Genome Sequencing Data. *Bioinformatics* 36, 1281–1282. doi:10.1093/bioinformatics/btz713

Yu, Z., Du, F., Ban, R., and Zhang, Y. (2020). SimuSCoP: Reliably Simulate Illumina Sequencing Data Based on Position and Context Dependent Profiles. *BMC bioinformatics* 21, 1–18. doi:10.1186/s12859-020-03665-5

Yu, Z., Liu, H., Du, F., and Tang, X. (2021). GRMT: Generative Reconstruction of Mutation Tree from Scratch Using Single-Cell Sequencing Data. *Front. Genet.* 12, 970. doi:10.3389/fgene.2021.692964

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

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

frontiers in Genetics

METHODS
published: 12 January 2021
doi: 10.3389/fgene.2020.622234

Predicting lincRNA-Disease Association in Heterogeneous Networks Using Co-regularized Non-negative Matrix Factorization

Yong Lin[1]* *and Xiaoke Ma*[2]*

[1] School of Physics and Electronic Information Engineering, Ningxia Normal University, Guyuan, China, [2] School of Computer Science and Technology, Xidian University, Xi'an, China

Long intergenic non-coding ribonucleic acids (lincRNAs) are critical regulators for many complex diseases, and identification of disease-lincRNA association is both costly and time-consuming. Therefore, it is necessary to design computational approaches to predict the disease-lincRNA associations that shed light on the mechanisms of diseases. In this study, we develop a co-regularized non-negative matrix factorization (aka *Cr-NMF*) to identify potential disease-lincRNA associations by integrating the gene expression of lincRNAs, genetic interaction network for mRNA genes, gene-lincRNA associations, and disease-gene associations. The Cr-NMF algorithm factorizes the disease-lincRNA associations, while the other associations/interactions are integrated using regularization. Furthermore, the regularization does not only preserve the topological structure of the lincRNA co-expression network, but also maintains the links "lincRNA → gene → disease." Experimental results demonstrate that the proposed algorithm outperforms state-of-the-art methods in terms of accuracy on predicting the disease-lincRNA associations. The model and algorithm provide an effective way to explore disease-lncRNA associations.

Keywords: disease-lincRNA association, non-negative matrix factorization, heterogeneous network, regularization, network analysis

OPEN ACCESS

Edited by:
Jianing Xi,
Northwestern Polytechnical University, China

Reviewed by:
Peng Gao,
Children's Hospital of Philadelphia, United States
Zhong-Yuan Zhang,
Central University of Finance and Economics, China
Wanxin Tang,
Sichuan University, China

***Correspondence:**
Yong Lin
linyong@nxu.edu.cn
Xiaoke Ma
xkma@xidian.edu.cn

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

Received: *28 October 2020*
Accepted: *03 December 2020*
Published: *12 January 2021*

Citation:
Lin Y and Ma X (2021) Predicting lincRNA-Disease Association in Heterogeneous Networks Using Co-regularized Non-negative Matrix Factorization. Front. Genet. 11:622234. doi: 10.3389/fgene.2020.622234

1. INTRODUCTION

Long intergenic non-coding RNAs (lincRNAs) are transcripts whose lengths are greater than 200 nucleotides with little or no protein coding potential (Kapranov et al., 2007; Mercer et al., 2009; Wang and Chang, 2011). In the traditional view, lncRNAs are considered as "junk RNAs" because they do not code protein sequences. However, it has been proven that many lncRNAs are dysregulated in human cancers and implicated in disease progression through modulating apoptosis, increasing cellular oncogenic potential, or inhibiting tumor growth (Wilusz et al., 2009; Taftet al., 2010).

With the advent of the next generation sequencing (NGS) techniques, a large number of lincRNAs have been identified (Guttman et al., 2009, 2010; Wang et al., 2009; Popadin et al., 2013), providing a great opportunity to investigate the functions of lncRNAs. Unfortunately, very few lincRNAs have been depicted with explicit molecular mechanisms in cancers through biological experiments or computational approaches (Guo et al., 2013; Zhao et al., 2016; Tang et al., 2017).

Thus, discovering lincRNA patterns that are associated with cancers is urgently needed as it sheds light on the underlying mechanism of diseases.

Therefore, great efforts have been devoted to investigating the functions or patterns of lincRNAs by analyzing omics data, such as DNA sequences, expression profiles, and genomic annotations. For instance, Liao et al. (2011) constructed a co-expression network for protein-coding genes and lincRNAs, and predicted the functions of lincRNAs via analyzing the constructed co-expression network. However, it has been criticized because of the fact that the gene expression profile cannot fully characterize the connections between genes and lincRNAs. To overcome this problem, Guo et al. (2013) developed a global prediction algorithm to infer probable functions of lincRNAs at a large scale by integrating gene expression, a protein-protein interaction (PPI) network, and DNA sequences. Ma et al. (2017a) designed a pipeline to discover disease related lincRNA modules across various clinical stages of cancers, rather than predicting the functions of lincRNAs. Ning et al. (2016) extracted the disease associated with SNPs within human lincRNAs.

Despite numerous research contributions to extract various patterns of lincRNAs, few efforts have been devoted to analyzing lincRNA-disease associations, which can be used to predict implicated diseases. The available methods to predict lincRNA-disease associations can be categorized into two classes: biological experiments-based methods and computational based approaches. The biological experiment-based methods have been criticized because they are time-consuming and costly. Computational based approaches are thus an alternative which can provide critical clues for biologists in revealing the mechanisms of diseases.

However, it is non-trivial to design effective and efficient algorithms to predict the lincRNA-disease associations largely due to two reasons. First, to infer the lincRNA-disease associations, large-scale known association data is a prerequisite. Second, diseases, such as cancers, are complex and difficult to characterize. Thus, it is wise to predict the lincRNA-disease associations by integrating omics data with an immediate purpose to improve the accuracy of prediction. Regarding the first concern, as more experimentally validated lincRNA-disease associations accumulate, researchers have summarized these associations as lincRNA-disease database, such as LncRNADisease (Chen et al., 2012) and Lnc2Meth (Zhi et al., 2018). These known associations provide a great opportunity to infer the lincRNA-disease associations.

Regarding the second concern, many algorithms have been developed to address this issue. For example, Yang et al. (2014) predicted the lncRNA-disease associations by constructing two biological networks, such as lncRNA-implicated disease network and disease network. Then, a propagation algorithm is applied to extract similar lncRNAs and diseases from those constructed networks. To integrate the expression profile, Chen et al. (2012) designed the Laplacian regularized least squares for lncRNA-disease associations, where the tissue expression profiles of intergenic lncRNA (lincRNA) from the Human BodyMap LincRNA project (Cabili et al., 2011). Zhang et al. (2017) proposed a label propagation algorithm to predict lncRNA-disease associations by integrating multiple heterogeneous networks. Fu et al. (2018) developed a matrix factorization-based model to predict disease-lncRNA associations, where multiple data matrices from various heterogeneous sources are factorized into low-rank matrices. Lan et al. (2017) designed a web server for the prediction of the lncRNA-disease. These algorithms achieve promising performance in inferring lncRNA-disease associations.

However, all of these studies solely focus on ranking lncRNA-disease associations via integrating the additional features of lncRNA genes and diseases, which cannot make use of the known prior knowledge to further improve the performance of algorithms. The latent features facilitate the identification of biological patterns, such as copy number and driver genes (Xi et al., 2020a,b). Actually, compared to the lincRNAs, knowledge of protein-coding genes is more redundant. How do you effectively incorporate the prior information into algorithms in order to perform a particular function and/or to infer a disease in the biological systems? For instance, Liao et al. (2011) made use of the gene-lncRNA relation to predict the functions of lncRNAs, implying that integration of omic data is promising for improving the performance of algorithms. Recently, Biswas et al. (2015) designed the *iNMF* algorithm by integrating expression profiles of protein-coding and lncRNA genes, lncRNA-disease and gene-disease associations, and gene genetic interaction networks to predict the diseases of lncRNAs. The experimental results demonstrate that it is wise to integrate omics data to infer lncRNA-disease associations a major motivation for this study.

iNMF jointly factorizes expression profiles of lncRNA and protein-coding genes. However, the method ignores the fact that lncRNAs execute their functions via interactions between them. Thus, we develop a novel algorithm, named co-regularized NMF (Cr-NMF), to predict lincRNA-disease associations via the heterogeneous network with multiple types of association, including lincRNA co-expression, lincRNA-disease, gene-disease, gene genetic and lincRNA-gene associations (As shown in **Figure 1**). The Cr-NMF algorithm decomposes the lincRNA-disease associations into the feature and coefficient matrices; the latent features for lincRNAs regularize the topological structure of lincRNA co-expression network. Furthermore, we also expect that the factorization reflects paths from *lincRNA* $\rightarrow$ *gene* $\rightarrow$ *disease*, which is also represented by regularization. Compared to state-of-the-art algorithms, the proposed algorithm is more accurate in the lincRNA-disease prediction. The proposed model and method provide an effective strategy to predict lncRNA-disease associations.

The rest of this study is organized as follows. Section 2 presents the details of the proposed algorithm. Then, in section 3, we set up experiments to validate the performance of Cr-NMF. Finally, conclusions are drawn in section 4.

2. ALGORITHM

The algorithm consists of two major components: the objective function construction and optimization rules, as shown in

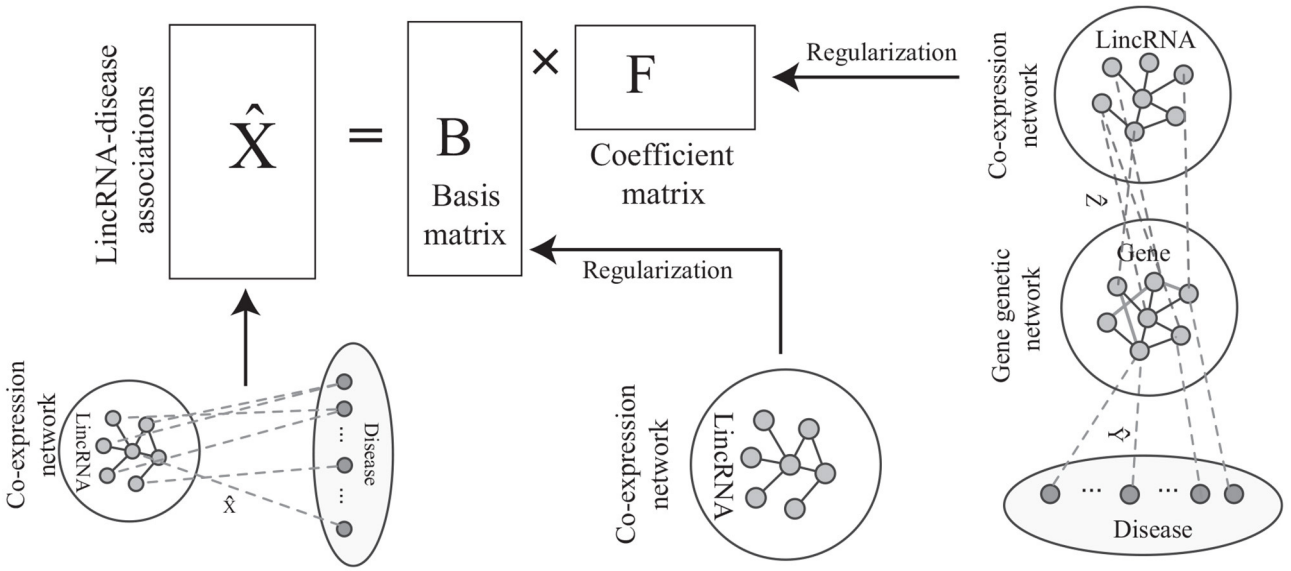

FIGURE 1 | Overview of the Cr-NMF algorithm. It factorizes the known disease-lincRNA associations, regularizing the other associations/interactions. Specifically, the basis matrix is the latent representation of lincRNAs, which preserves the topological structure of lincRNA co-expression networks. The coefficient matrix regularizes the links "lincRNA → gene → disease".

TABLE 1 | Notations.

Notation	Definition and description
n_g, n_d, n_l	Number of genes, diseases, and lincRNAs
$G^{[g]}$	Gene genetic interaction network
$G^{[l]}$	lincRNA co-expression network
$\widehat{X}$	known lincRNA-disease associations
$\widehat{Y}$	known gene-disease associations
$\widehat{Z}$	genes-lincRNAs associations
$W^{[g]}, W^{[l]}$	weighted adjacency matrix for $G^{[g]}$ and $G^{[l]}$
$w_{ij}^{[g]}$	the element at i-th row jth column in matrix $W^{[g]}$
D	the degree diagonal matrix, i.e., $D = diag(d_1, \ldots, d_n)$
$\overline{W}^{[g]}$	normalized $G^{[g]}$, i.e., $\overline{W}^{[g]} = D^{-1/2} W^{[d]} D^{-1/2}$
$W^{'}$	transpose of matrix W
$w_{i.}$	the i-th row of matrix W
$w_{.j}$	the j-th column of matrix W
$\|W\|_F$	Frobenius norm of matrix W
$Tr(W)$	the Tr of matrix W, i.e., $Tr(W) = \sum_i w_{ii}$

Figure 1. The procedure and analysis of the proposed algorithm are addressed in this section.

2.1. Notations

Before presenting the detailed description of the proposed algorithm, let us introduce some terminologies that are widely used in the sections that follow.

The notations for the algorithm are summarized in **Table 1**. Let n_g be the number of genes, n_d be the number of diseases, n_l be the number of lincRNAs. The lincRNA co-expression network is denoted by $G^{[l]} = (V^{[l]}, E^{[l]})$, where $V^{[l]}$ is the set of lincRNAs and $E^{[l]}$ is the interaction sets based on lincRNA co-expression coefficients. The adjacency matrix for $G^{[l]}$ is denoted by matrix $W^{[l]}$, where $w_{ij}^{[l]}$ is the weight on edge (i, j) in $G^{[g]}$. Because $G^{[l]}$ is undirected, $W^{[l]}$ is symmetric. The degree of the i-th lincRNA in $G^{[l]}$ is defined as the sum of weights on the edges connecting to it, i.e., $d_i = \sum_i w_{ij}^{[l]}$. The degree matrix of $G^{[l]}$ is the diagonal one with degree sequence, i.e., $D^{[l]} = diag(d_i^{[l]}, \ldots, d_{n_l}^{[l]})$. Given network $G^{[l]}$, we construct a normalized Laplacian matrix $L^{[l]} = I - (D^{[l]})^{-1/2} W^{[l]} (D^{[l]})^{-1/2}$. Analogously, we construct the normalized Laplacian matrix for $G^{[g]}$ as $L^{[g]} = I - (D^{[g]})^{-1/2} W^{[g]} (D^{[g]})^{-1/2}$.

The known lincRNA-disease associations are represented by $\widehat{X}$, where the row represents a lincRNA and column denotes a disease. The known gene-disease associations are denoted by $\widehat{Y}$, where rows correspond to genes and columns denote diseases. Thegene-lincRNA associations $\widehat{Z}$ are constructed based expression data, where the rows correspond to genes, columns to lincRNAs, and $z_{ij} = 1$ if the i-th gene and j-th lincRNA are associated with at least one disease, 0 otherwise.

2.2. Objective Function

NMF aims at learning the representation parts of the original data (Lee and Seung, 1999) by approximating the target matrix into the product of two low-ranking matrices. Specifically, given matrix W, NMF decomposes W into two non-negative matrices $B_{(m+n)\times k}$ and $F_{(m+n)\times k}$ such that

$$W \approx BF', s.t. B \geq 0, F \geq 0, \quad (1)$$

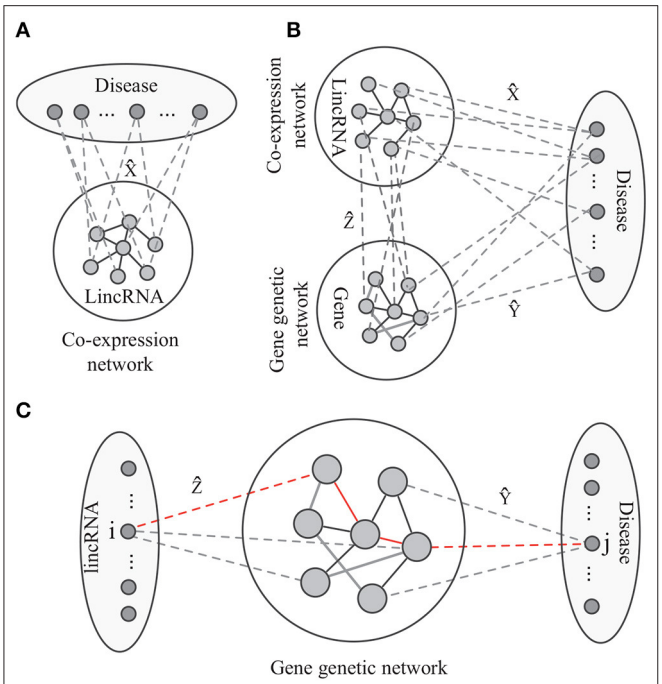

FIGURE 2 | The paradiagrams for disease-lincRNA association prediction. **(A,B)** The label propagation model predicts the disease-lincRNA associations by each single disease based on the lincRNA co-expression network, and **(C)** the proposed algorithm integrates lincRNA co-expression network, gene genetic networks and various associations. The dashed lines show disease-lincRNA association (matrix $\widehat{X}$), disease-gene association (matrix $\widehat{Y}$), and lincRNA-gene associations (matrix $\widehat{Z}$).

where B is the basis matrix and F is the feature matrix. NMF has been widely applied for graph analysis (Ma et al., 2018a), link prediction (Ma et al., 2017b, 2018b), bioinformatics (Chen and Zhang, 2016; Ma et al., 2016, 2018c).

As shown in **Figure 2A**, the label propagation-based model has been widely studied and successfully applied to predict phenotype-gene associations (Hwang and Kuang, 2010; Vanunu et al., 2010; Hwang et al., 2011). The model aims at identifying the disease-lincRNA associations X under some constraints. Thus, the objective function of label propagation model is defined as

$$O_{lp} = \theta Tr(X'L^{[l]}X) + (1-\theta)\|X - \widehat{X}\|_F^2, \tag{2}$$

where $\theta \in (0,1)$ is a parameter to balance the contributions of the two terms, $Tr(\cdot)$ is the Tr function and $\|\cdot\|_F$ is the Frobenius norm. To further improve the performance of label propagation model, Petegrosso et al. (2017) proposed transfer learning-based label propagation model to integrate omics data to predict phenome-genome association.

Given the disease-lincRNA associations $\widehat{X}$, Cr-NMF first factorizes $\widehat{X}$ into the product of matrix B and F, i.e.,

$$\widehat{X} = BF, \quad s.t. \quad B \geq 0, F \geq 0, \tag{3}$$

where $B \in R^{n_l \times r}$ is the basis matrix, $F \in R^{r \times n_d}$ is the feature matrix, r is the number of latent variables (usually, $r \ll \min\{n_l, n_d\}$). By casting Equation (3) as an optimization form, we obtain the following objective function as

$$O_{NMF} = \frac{1}{2}\|\widehat{X} - BF\|_F^2, \quad s.t. \quad B \geq 0, F \geq 0. \tag{4}$$

On the one hand, matrix B is considered to be the representations of lincRNAs in the latent space, where each row $b_{i.}$ is interpreted as latent representation of the i-th lincRNA. We expect the latent representations in matrix B preserve the local topological structure of lincRNAs $G^{[l]}$. Specifically, if a pair of lincRNAs are close in terms of the latent representation, they are well connected in $G^{[l]}$ and vice versa. Cai et al. (2010) demonstrated that

$$\begin{aligned} O_{G^{[l]}} &= \frac{1}{2}\sum_i\sum_j \|b_{i.} - b_{j.}\|^2 w_{ij}^{[l]} \\ &= Tr(B'D^{[l]}B) - Tr(B'W^{[l]}B) \\ &= Tr(B'L^{[l]}B). \end{aligned} \tag{5}$$

On the other hand, the disease-lincRNA associations are also related to the topological structure of the gene interaction network, lincRNA-gene association (**Figure 2B**), and the disease-gene associations. The association between the i-th lincRNA and the j-th disease follows the pattern *lincRNA* $\rightarrow$ *gene* $\rightarrow$ *gene network* $\rightarrow$ *disease*. For example, in **Figure 2C**, the i-th lincRNA and j-th disease are connected by the red path. There is a good biological interpretation for this pattern: the lincRNAs transduce signal to the target genes. The dysfunctional signal possibly leading to an abnormal response via interaction among genes, resulting in diseases. Thus, the disease-lincRNA association w_{ij} can be defined as a product of weights on all the paths connecting the i-th lincRNA and j-th disease, i.e.,

$$x_{ij} = \sum_k \widehat{z}_{ik} w_{ij}^{[g]} \widehat{y}_{kj}. \tag{6}$$

The underlying assumption for Equation (6) is that the more paths connecting a lincRNA and disease, the more likely it is to be a true association. Transforming Equation (6) into matrix form, we obtain

$$X = \widehat{Z}W^{[g]}\widehat{Y}. \tag{7}$$

Transforming Equation (7) into an optimization problem, we obtain

$$O_{G^{[g]}} = \frac{1}{2}\|X - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2. \tag{8}$$

Because we use NMF to approximate X, Equation (8) is re-written as

$$O_{G^{[g]}} = \frac{1}{2}\|BF - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2. \tag{9}$$

Combining Equations (4,5), and (9), the objective function of the proposed algorithm is defined as

$$O = O_{NMF} + \alpha O_{G^{[l]}} + \beta O_{G^{[g]}}, \tag{10}$$

where parameter α, β control the contributions of two terms $O_{G^{[l]}}$ and $O_{G^{[g]}}$. The disease-lincRNA prediction problem is transformed into an optimization problem as

$$\begin{aligned} \min_{B,F} \quad & \frac{1}{2}\|\widehat{X} - BF\|^2 + \alpha Tr(B'L^{[l]}B) \\ & + \frac{\beta}{2}\|BF - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2 \\ s.t. \quad & B \geq 0, F \geq 0. \end{aligned} \tag{11}$$

In the next subsection, we address how to optimize the problem in Equation (11).

2.3. Optimization

An iterative two-step strategy is adopted because direct optimization to Equation (11) is difficult, where we optimize matrices B and F by fixing parameters. At each iteration, either matrix B or F is optimized first, whereas the other is fixed. Iteration is repeated until the algorithm converges or the maximum number of iterations is reached.

Let the objective function of Equation (11), i.e.,

$$\begin{aligned} \mathcal{L} = & \frac{1}{2}\|\widehat{X} - BF\|^2 + \alpha Tr(B'L^{[l]}B) \\ & + \frac{\beta}{2}\|BF - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2. \end{aligned} \tag{12}$$

We handle the non-negative constraints for matrices B and F using the Larange method. Specifically, let ϕ_{ij} and ψ_{ij} be the Larange multiplier for the constraints b_{ij} and f_{ij}, respectively. Considering $\Phi = [\phi_{ij}]$, $\Psi = [\psi_{ij}]$, the Larange $\mathcal{L}$ of Equation (12) can be formulated as

$$\begin{aligned} \mathcal{L} = & \frac{1}{2}\|\widehat{X} - BF\|^2 + \alpha Tr(B'L^{[l]}B) \\ & + \frac{\beta}{2}\|BF - \widehat{Z}W^{[g]}\widehat{Y}\|_F^2 + \Phi B + \Psi F. \end{aligned} \tag{13}$$

The partial derivatives of $\mathcal{L}$ with respect to basis matrix B and feature matrix F are calculated as

$$\frac{\partial \mathcal{L}}{\partial B} = (1+\beta)BFF' - \widehat{X}F' + 2\alpha L^{[l]}B - \widehat{Z}W^{[g]}\widehat{Y}F' + \Phi, \tag{14}$$

and

$$\frac{\partial \mathcal{L}}{\partial F} = B'\widehat{X} - B'BF + \beta B'BF - B'\widehat{Z}W^{[g]}\widehat{Y} + \Psi. \tag{15}$$

According to the Karush-Kuhn-Tucker conditions $\phi_{ij}b_{ij} = 0$ and $\psi_{ij}f_{ij} = 0$, we obtain the updated rules

$$B = \frac{\widehat{X}F' + \widehat{Z}W^{[g]}\widehat{Y}F'}{(1+\beta)BFF' + 2\alpha L^{[l]}B}B, \tag{16}$$

and

$$F = \frac{B'BF + B'\widehat{Z}W^{[g]}\widehat{Y}}{B'\widehat{X} + \beta B'BF}F. \tag{17}$$

The Cr-NMF algorithm is presented in Algorithm 1.

Algorithm 1: The Cr-NMF algorithm

Input:

$G^{[l]}$: Co-expression network for lincRNAs;
$M^{[g]}$: Expression profile for genes;
$M^{[l]}$: Expression profile for lincRNAs;
$\widehat{X}$: Known disease-lincRNA associations;
$\widehat{Y}$: Known disease-gene associations;
α, β: Parameters control relevant importance.

Output:

X: Predicted disease-lincRNA associations.

Step 1: Data Processing

1: Construct co-expression network $G^{[l]}$ for lincRNAs using expression profile $M^{[l]}$;
2: Construct gene-lincRNA associations $\widehat{Z}$ using $M^{[l]}$ and $M^{[g]}$;
3: Construct Laplacian matrix $L^{[g]}$ for $G^{[g]}$;
4: Construct Laplacian matrix $L^{[l]}$ for $G^{[l]}$;

Step 2: Matrix Factorization

5: Make initial matrices B and F;
6: Update matrix B according to Equation (16);
7: Update matrix F according to Equation (17);
8: Goto Step 5 until the algorithm is convergent;

Step 3: Predict disease-lincRNA associations

9: Predict disease-lincRNA association as $X = BF$;
10: **return** X

2.4. Algorithm Analysis

The complexity of algorithm is investigated. On the space complexity of algorithm, the space for the gene genetic network is $O(n_g^2)$. The space for lincRNA co-expression network is $O(n_l^2)$. The space of disease-lincRNA association, disease-gene associations, and gene-lincRNA association is $O(n_d n_l)$, $O(n_d n_g)$, and $O(n_g n_l)$, respectively. The space of basis matrix B and feature matrix F is $O((n_l+n_d)r)$, where r is the number of latent variables. Thus, the total space of Cr-NMF is $O(n_l^2 + n_g^2 + n_d n_l + n_d n_g + n_g n_l + (n_l + n_d)r)$. Because $n_d \ll n_g$ and $n_l \ll n_g$, the total space of the proposed method is $O(n_g^2)$.

The running time of the proposed algorithm depends on the updating rules in Equations (16) and (17). Thus, the time complexity of Cr-NMF is the same as that of NMF, i.e., $O(tkn^2)$, where t is the number of iteration (Lin, 2007). Thus, the overall running time for RNMF-MM is $O(tkn^2) + O(n^2) = O(tkn^2)$, indicating that the proposed algorithm is also efficient in comparison with the NMF algorithm.

3. RESULTS

In this section, we validate the performance of the proposed algorithm. The data, parameter selection as well as the performance of algorithms are addressed in turns.

3.1. Data

The lincRNAs are downloaded from the Human BodyMap project, which provides a catalog of lincRNAs from RNA-seq data across 22 tissues (Cabili et al., 2011). The catalog contains transcript expression profile across the tissues using the Cufflinks (Trapnell et al., 2010).

The association dataset of lincRNAs and diseases are extracted from the LncRNADisease database (Chen et al., 2012) in January 2015. There are 1564 lincRNAs and their associations with 1641 diseases. We employ the OMIM API function call (Hamosh et al., 2005) to retrieve closely matched phenotype IDs, resulting in a set of 684 OMIM phenotypes (mainly disease) associated with lincRNAs. All the diseases without matching any valid OMIM phenotype ID are removed. Finally, we obtain the lincRNA-disease association among 562 lincRNAs and 645 OMIM diseases.

The mRNA-disease associations are downloaded from DisGeNET software (Bauer-Mehren et al., 2010), where 16,666 mRNA genes are associated with 13,135 diseases. Similar to the lincRNA-disease associations, we use the OMIM function call to map disease names to matched phenotype IDs, and only these diseases with at least one lincRNAs are selected. Finally, 180,266 gene-disease associations are obtained among 645 OMIM diseases and 13,425 coding-genes.

The gene genetic interaction network is extracted from Lin et al. (2010), where 4,836,794 interactions among coding-genes. Only these genes associated with at least one disease are retained, resulting 3,264,923 interactions among 13,425 genes.

In this study, we want to make use the connections between lincRNAs and coding-genes. Based on Biswas et al. (2015), we construct the lincRNA-gene association network from diseases. Specifically, if the i-th lincRNA is connected to the j-th coding-gene if and only if both of them are associated with at least a disease. Based on this strategy, there are 1,775,375 edges among 562 lincRNAs and 13,425 coding-genes.

3.2. Settings

To fully validate the performance of the proposed algorithm, we select five well-known algorithms for a comparative comparison: NMF (Lee and Seung, 1999), non-smooth NMF (nsNMF) (Pascual-Marqui et al., 2001), integrated NMF (iNMF) (Biswas et al., 2015), Label Propagation (LP) (Hwang et al., 2011), and Random Walk (RW) (Li and Patra, 2010). All these algorithms can be categorized into two classes: matrix decomposition based and topological structure based methods. The matrix decomposition-based algorithms include NMF, nsNMF, and iNMF, while the topological structure-based methods are LP and RW.

To evaluate the performance of these algorithms, three measures, including mean absolute error (MAE), Accuracy and root mean squared error (RMSE), are employed to quantify the accuracy of algorithms. They are defined as Herlocker et al. (2004):

$$MAE(\widehat{X}, X) = \frac{1}{|\tau|} \sum_{(i,j)\in\tau} |\widehat{x}_{ij} - x_{ij}|, \tag{18}$$

$$Accuracy(\widehat{X}, X) = 1 - MAE(\widehat{X}, X), \tag{19}$$

$$RMSE(\widehat{X}, X) = \sqrt{\frac{1}{|\tau|} \sum_{(i,j)\in\tau} (|\widehat{x}_{ij} - x_{ij}|)^2}, \tag{20}$$

$$RSS(\widehat{X}, X) = \sqrt{\sum_{i,j} (|\widehat{x}_{ij} - x_{ij}|)^2}, \tag{21}$$

where $\widehat{X}$ and X are the observed association matrix and the predicted association matrix, respectively. τ is the set of lincRNA-disease association for prediction, i.e., τ is considered as the test set.

3.3. Parameter Selection

Three parameters are involved in the proposed algorithm, where parameter α determines the relevant importance of lincRNA co-expression networks, parameter β controls the relevant importance of the gene genetic network, and parameter k is the number of features for the basis and coefficient matrices. Similar to Ref., we set $\alpha = \beta$ by assuming that the lincRNA co-expression network and gene genetic network are equally important in discovering the lincRNA-disease associations.

We first investigate how parameter k determines the performance of the proposed algorithm. **Figure 3A** illustrates how RSS changes from 3 to 54 with a gap 3. From **Figure 3A**, we conclude that as k increases from 3 to 33, RSS dramatically decreases, which implies that the accuracy of the proposed algorithm increases. As k increases from 34 to 54, RSS increases. There is a good reason why this occurs. When k is small, the number of the latent features is insufficient to characterize the lincRNA-disease associations. When k is large, the number of the latent features is redundant. $k = 33$ reaches a good balance between them since RSS reaches the minimum. In the experiment, we set $k = 33$.

We then investigate how parameter α and β affect the performance of the Cr-NMF algorithm. **Figure 4** shows that how MAE and RMSE change as $\alpha \in \{0.001, 0.01, 0.1, 1, 10, 100\}$. It is shown that the proposed algorithm achieves the best performance when $\alpha = 1$. Furthermore, the proposed algorithm is robust since the perturbation of performance is subtle if $\alpha \in [10, 100]$, indicating that Cr-NMF is not sensitive to parameter α and β. Even though MAE and RMSE decrease when $\alpha \in [10, 100]$, the change is subtle.

Finally, we check the convergence of the proposed algorithm. **Figure 3B** shows how RSS changes as the number of iterations increases. It is easy to assert that, when the number of iterations reaches 60, the algorithm converges because *RSS* does not change dramatically any more. Thus, the number of iterations is set as 60 in the experiments. The result demonstrates that the proposed algorithm is efficient.

3.4. Performance of Various Algorithms on Predicting lincRNA-Disease Associations

By setting $\alpha(\beta) = 10$, $k = 33$, and the number of iterations as 60, we apply Cr-NMF to the omic data to predict the lincRNA-disease associations. To quantify the performance of various algorithms, the accuracy in Equation (19) is adopted, where it is also used in Biswas et al. (2015). Because all of these compared algorithms have a factor of randomness, we get rid of randomness of algorithms by running each algorithm 50 times, and the mean of accuracy is used to quantify the performance of algorithms.

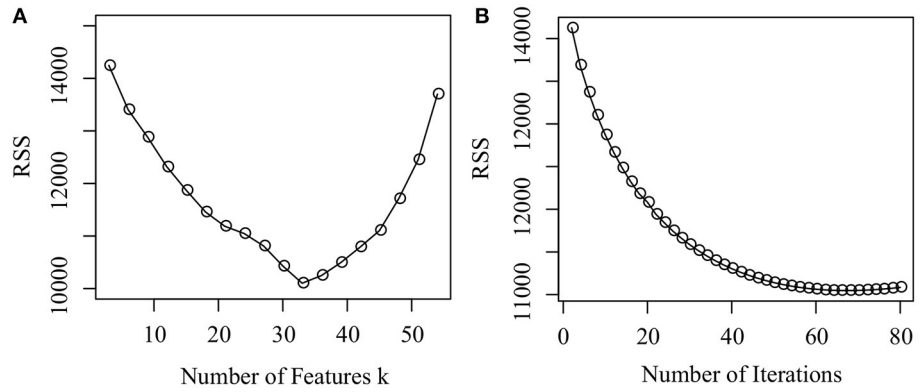

FIGURE 3 | Parameter selection and convergence analysis. **(A)** How the RSS changes as the number of features changes from 3 to 54, and **(B)** How the RSS changes as the number of iterations increases from 1 to 100.

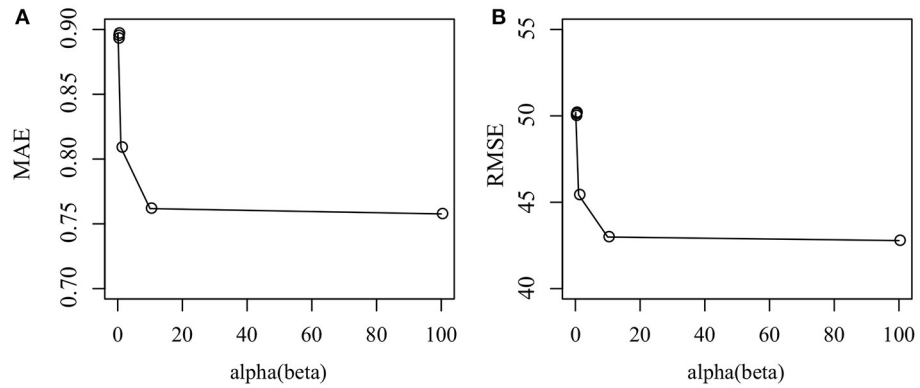

FIGURE 4 | How parameter affects the performance of the proposed algorithm in terms of various measurements: **(A)** MAE, and **(B)** RSME.

The leave-one-out cross validation (LOOCV) is adopted to measure the accuracy of each algorithm. Specifically, for each disease, we remove all the associations between the disease and lincRNA genes. The accuracy of various algorithms is depicted in **Figure 5A**. It is easy to draw conclusions such as: (1) the Cr-NMF algorithm achieves the best performance in LOOCV, followed by the iNMF algorithm. In detail, the accuracy of Cr-NMF is 0.823 ± 0.009, which is 1.9% higher than the iNMF algorithm on predicting disease-lincRNA associations. (2) Both Cr-NMF and iNMF algorithms outperform the rest of the methods, implying the integration of omic data is promising on predicting disease-lincRNA associations. Moreover, (3) The random walk and label propagation algorithms are worst in terms of accuracy. There are two reasons why the proposed algorithm outperforms the other approaches. First, the Cr-NMF algorithm directly factorize associations between diseases and lincRNAs, which captures the latent features to characterize the disease-lincRNA associations. Second, the factorization preserves the paths from "disease $\rightharpoonup$ lincRNA $\rightarrow$ protein-coding gene," which more precisely infers disease-lincRNA associations. The RW and LP algorithms are much worse than the others, implying that the topological structure is insufficient to characterize the relations between diseases and lincRNAs.

In order to further validate the performance of the proposed algorithm, we take the disease-lincRNA associations before 2015 January as training set, and set the data between 2015 and 2017 July as testing set, as shown in **Figure 5B**. It is easy to assert that the proposed algorithm is best, followed by iNMF. Specifically, the accuracy of algorithms is 0.647 (Cr-NMF), 0.594 (iNMF), 0.587 (nsNMF), 0.598 (sNMF), 0.575 (LP), 0.412 (RW). Careful comparison between **Figures 5A,B** indicates that the accuracy of various algorithms in the external validation decreases dramatically. However, the relative performance of these algorithms is similar. The results demonstrate that the proposed algorithm is promising in predicting disease-lincRNA associations.

4. CONCLUSION

LncRNAs are critical regulators in human diseases and disorder pathways. Thus, it is necessary to understand the associations

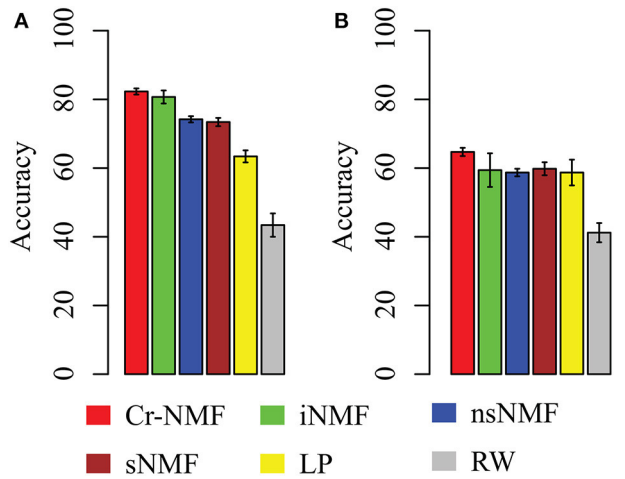

FIGURE 5 | The accuracy of various algorithms on predicting disease-lincRNA associations in terms of various strategies: **(A)** leave-one-out cross validation (LOOCV), and **(B)** external validation, where Y-axis denotes mean accuracy and error bar represents standard deviation.

between lncRNAs and diseases since these relations shed light on revealing the mechanisms of complex diseases. Compared to the protein-coding genes, a very little is known about the associations of lncRNAs and diseases. The next generation of sequencing technique discovers novel lncRNAs at an unprecedent speed. Therefore, there is a critical need to develop sophisticated computational tools to predict the relations between lncRNAs and diseases.

In this study, we proposed an NMF-based algorithm to predict lincRNA-disease associations by integrating multiple types of interaction data, such as co-expression interactions between lincRNAs, disease-lincRNA associations, disease-gene associations, gene genetic interactions, and lincRNA-gene links. There are two advantages of the proposed algorithm. First, it is able to explain each of the associated lincRNA as well as disease in a latent feature space. Second, the proposed algorithm takes the path from lincRNA to disease, i.e., "disease ⇀ lincRNA → protein-coding gene," which improves the accuracy of the prediction. The results demonstrate that the propose method outperforms state-of-the-art algorithms in terms of accuracy.

There are some limits in the proposed algorithm. First, there are two parameters involved in the methods and we solve this issue by a step search strategy in the experiments. A better and faster way to accomplish this needs to be developed. Particularly, how to infer the values of parameters by making use of the biological knowledge in diseases is ideal. Second, even though the proposed algorithm integrates omics data, incorporating additional data, such as disease networks, mutation data in genes would obtain even more meaningful results. In a future study, we will address these issues.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: TCGA.

AUTHOR CONTRIBUTIONS

YL and XM constructed the original idea and designed the experiments. XM wrote the manuscript. YL proofread the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the NSFC (Grant No. 61562070), Scientific Research Projects of Colleges and Universities in Ningxia (NGY2018-136), Major Scientific Research Projects in Ningxia (2019BDE03015), and the Ningxia Science and Technology Leading Talent Project (201601).

REFERENCES

Bauer-Mehren, A., Furlong, L. I., and Sanz, F. (2010). Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Bioinformatics* 26, 2924–2926. doi: 10.1038/msb.2009.47

Biswas, A., King, M., Kim, D.-C., Ding, C. H. Q., Zhang, B., and Wu, X. (2015). Inferring disease associations of the long non-coding RNAs through non-negative matrix factorization. *Netw. Model. Anal. Health Inform. Bioinform.* 4:9. doi: 10.1007/s13721-015-0081-6

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. 25, 1915–1927. doi: 10.1101/gad.17446611

Cai, D., He, X., Han, J., and Huang, T. S. (2010). Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1548–1560. doi: 10.1109/TPAMI.2010.231

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

Chen, J., and Zhang, S. (2016). Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* 32, 1724–1732. doi: 10.1093/bioinformatics/btw059

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

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., et al. (2013). Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Res.* 41:e35. doi: 10.1093/nar/gks967

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D, et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227. doi: 10.1038/nature 07672

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., et al. (2010). *Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510. doi: 10.1038/nbt.1633

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of

human genes and genetic disorders. *Nucleic Acids Res.* 33(Suppl_1), D514–D517. doi: 10.1093/nar/gki033

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inform. Syst.* 22, 5–53. doi: 10.1145/963770.963772

Hwang, T., and Kuang, R. (2010). "A heterogeneous label propagation algorithm for disease gene discovery," in *Proceedings of the 2010 SIAM International Conference on Data Mining* (Siam, OH: SIAM), 583–594.

Hwang, T., Zhang, W., Xie, M., Liu, J., and Kuang, R. (2011). Inferring disease and gene set associations with rank coherence in networks. *Bioinformatics* 27, 2692–2699. doi: 10.1093/bioinformatics/btr463

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488. doi: 10.1126/science.1138341

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, 458–460. doi: 10.1093/bioinformatics/btw639

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

Li, Y., and Patra, J. C. (2010). Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 26, 1219–1224. doi: 10.1093/bioinformatics/btq108

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., et al. (2011). Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic Acids Res.* 39, 3864–3878. doi: 10.1093/nar/gkq1348

Lin, A., Wang, R. T., Ahn, S., Park, C. C., and Smith, D. J. (2010). A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome Res.* 20, 1122–1132. doi: 10.1101/gr.104216.109

Lin, C. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* 19, 2756–2779. doi: 10.1162/neco.2007.19.10.2756

Ma, X., Dong, D., and Wang, Q. (2018a). Community detection in multi-layer networks using joint nonnegative matrix factorization. *IEEE Trans. Knowl. Data Eng.* 31, 273–286. doi: 10.1109/TKDE.2018.2832205

Ma, X., Sun, P., and Qin, G. (2017b). Nonnegative matrix factorization algorithms for link prediction in temporal networks using graph communicability. *Pattern Recogn.* 71, 361–374. doi: 10.1016/j.patcog.2017.06.025

Ma, X., Sun, P., and Wang, Y. (2018b). Graph regularized nonnegative matrix factorization for temporal link prediction in dynamic networks. *Phys. A Stat. Mech. Appl.* 496, 121–136. doi: 10.1016/j.physa.2017.12.092

Ma, X., Sun, P., and Zhang, Z. (2018c). An integrative framework for protein interaction network and methylation data to discover epigenetic modules. *IEEE ACM Trans. Comput. Biol. Bioinform.* 16, 1855–1866. doi: 10.1109/TCBB.2018.2831666

Ma, X., Tang, W., Wang, P., Guo, X., and Gao, L. (2016). Extracting stage-specific and dynamic modules through analyzing multiple networks associated with cancer progression. *IEEE ACM Trans. Comput. Biol. Bioinform.* 15, 647–658. doi: 10.1109/TCBB.2016.2625791

Ma, X., Yu, L., Wang, P., and Yang, X. (2017a). Discovering DNA methylation patterns for long non-coding rnas associated with cancer subtypes. *Comput. Biol. Chem.* 69, 164–170. doi: 10.1016/j.compbiolchem.2017.03.014

Mercer, R. T., Dinger, M. E., and Mattick, J. M. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159. doi: 10.1038/nrg2521

Ning, S., Yue, M., Wang, P., Liu, Y., Zhi, H., Zhang, Y., et al. (2016). Lincsnp 2.0: an updated database for linking disease-associated snps to human long non-coding RNAs and their TFBSs. *Nucleic Acids Res.* 45, D74–D78. doi: 10.1093/nar/gkw945

Pascual-Marqui, R.D., Pascual-Montano, A.D., Kochi, K.,and Carazo, J.M. (2001). Smoothly distributed fuzzy c-means: a new self-organizing map. *Pattern Recogn.* 34, 2395–2402. doi: 10.1016/S0031-3203(00)00167-9

Petegrosso, R., Park, S., Hwang, T. H., and Kuang, R. (2017). Transfer learning across ontologies for phenome–genome association prediction. *Bioinformatics* 33, 529–536. doi: 10.1093/bioinformatics/btw649

Popadin, K., Gutierrez-Arcelus, M., Dermitzakis, E. T., and Antonarakis, S. E. (2013). Genetic and epigenetic regulation of human lincRNA gene expression. *Am. J. Hum. Genet.* 93, 1015–1026. doi: 10.1016/j.ajhg.2013.10.022

Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., and Mattick, J. S. (2010). Non-coding RNAs: regulators of disease. *J. Pathol.* 220, 126–139. doi: 10.1002/path.2638

Tang, W., Zhang, D. Z. and Ma, X. (2017). RNA-sequencing reveals genome-wide long non-coding RNAs profiling associated with early development of diabetic nephropathy. *Oncotarget* 8:105832. doi: 10.18632/oncotarget.22405

Trapnell, C., William, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* 6:e1000641. doi: 10.1371/journal.pcbi.1000641

Wang, K., and Chang, H. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914. doi: 10.1016/j.molcel.2011.08.018

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–1504. doi: 10.1101/gad.1800909

Xi, J., Li, A., and Wang, M. (2020a). Hetrcna: a novel method to identify recurrent copy number alternations from heterogeneous tumor samples based on matrix decomposition framework. *IEEE ACM Trans. Comput. Biol. Bioinform.* 17, 422–434. doi: 10.1109/TCBB.2018.2846599

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. (2020b). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 36, 1855–1863. doi: 10.1093/bioinformatics/btz793

Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F., et al. (2014). A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS ONE* 9:e87797. doi: 10.1371/journal.pone.0087797

Zhang, J., Zhang, Z., Chen, Z., and Deng, L. (2017). Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. *IEEE ACM Trans. Comput. Biol. Bioinform.* 16, 396–406. doi: 10.1109/TCBB.2017.2701379

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., et al. (2016). Noncode 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* 44, D203–D208. doi: 10.1093/nar/gkv1252

Zhi, H., Li, X., Wang, P., Gao, Y., Gao, B., Zhou, D., et al. (2018). Lnc2meth: a manually curated database of regulatory relationships between long non-coding RNAs and DNA methylation associated with human disease. *Nucleic Acids Res.* 46, D133–D138. doi: 10.1093/nar/gkx985

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

frontiers in Genetics

METHODS
published: 13 January 2021
doi: 10.3389/fgene.2020.632311

A Density Peak-Based Method to Detect Copy Number Variations From Next-Generation Sequencing Data

Kun Xie[1†], *Ye Tian*[1,2†] *and Xiguo Yuan*[1,2*]

[1] The School of Computer Science and Technology, Xidian University, Xi'an, China, [2] Xi'an Key Laboratory of Computational Bioinformatics, The School of Computer Science and Technology, Xidian University, Xi'an, China

Copy number variation (CNV) is a common type of structural variations in human genome and confers biological meanings to human complex diseases. Detection of CNVs is an important step for a systematic analysis of CNVs in medical research of complex diseases. The recent development of next-generation sequencing (NGS) platforms provides unprecedented opportunities for the detection of CNVs at a base-level resolution. However, due to the intrinsic characteristics behind NGS data, accurate detection of CNVs is still a challenging task. In this article, we propose a new density peak-based method, called dpCNV, for the detection of CNVs from NGS data. The algorithm of dpCNV is designed based on density peak clustering algorithm. It extracts two features, i.e., local density and minimum distance, from sequencing read depth (RD) profile and generates a two-dimensional data. Based on the generated data, a two-dimensional null distribution is constructed to test the significance of each genome bin and then the significant genome bins are declared as CNVs. We test the performance of the dpCNV method on a number of simulated datasets and make comparison with several existing methods. The experimental results demonstrate that our proposed method outperforms others in terms of sensitivity and F1-score. We further apply it to a set of real sequencing samples and the results demonstrate the validity of dpCNV. Therefore, we expect that dpCNV can be used as a supplementary to existing methods and may become a routine tool in the field of genome mutation analysis.

Keywords: copy number variations, next-generation sequencing data, density peak, null distribution, read depth

OPEN ACCESS

Edited by:
Zhenhua Yu,
Ningxia University, China

Reviewed by:
Liu Guangming,
Xi'an University of Technology, China
Xinhui Yu,
China University of Mining and Technology, China

***Correspondence:**
Xiguo Yuan
xiguoyuan@mail.xidian.edu.cn

[†] These authors have contributed equally to this work

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

Received: *23 November 2020*
Accepted: *21 December 2020*
Published: *13 January 2021*

Citation:
Xie K, Tian Y and Yuan X (2021) A Density Peak-Based Method to Detect Copy Number Variations From Next-Generation Sequencing Data. Front. Genet. 11:632311. doi: 10.3389/fgene.2020.632311

INTRODUCTION

Copy number variation (CNV) is an important category of DNA structural variations, including amplifications or losses of DNA fragments with a length of more than 1 kilo base-pairs (bp) (Freeman et al., 2006; Yuan et al., 2012b). The mutation rate of CNV loci is much higher than that of single nucleotide polymorphisms (SNP) across the whole genome. CNV is one of the important pathogenic factors affecting human complex diseases (Shlien and Malkin, 2009; Fridley et al., 2012; Xi et al., 2020a,b). Therefore, it is necessary and meaningful to analyze CNVs when studying and treating complex diseases especially human cancers. Generally, the mechanisms for the formation of CNVs can be classified into two categories: DNA recombination and DNA error replication (Martin et al., 2019). In each category of the mechanisms, CNVs are usually presented in either amplification or deletion states. The major step of CNV analysis in samples obtained from human

cancers is to identify which genome regions are CNVs and determine the corresponding states (i.e., either amplification or deletion). Therefore, it is required to develop statistically computational methods to analyze the data generated by different sequencing technologies.

There are three primary types of technologies that can produce data sets for the detection of CNVs: array comparative genomic hybridization (aCGH), SNP array, and next-generation sequencing (NGS) technologies. Currently, various computational methods have already been developed for analyzing each type of the data sets. For example, aiming at aCGH data, classic methods include fastRPCA (Nowak et al., 2011), PLA (Zhou et al., 2014), WaveDec (Cai et al., 2018), and graCNV (Auer et al., 2007). Meanwhile, aiming at SNP array data, famous methods include GISTIC (Beroukhim et al., 2007), STAC (Diskin et al., 2006), SAIC (Yuan et al., 2012b), and AISAIC (Zhang et al., 2014). In comparison with these two types of data, NGS data is at the highest resolution and is used widely for the detection of CNVs in recent years. Due to the inherent characteristics behind NGS data, the CNV detection methods using NGS data can be classified into four categories (Zhao et al., 2013): pair-end mapping, split-read, *de no* assembly, and read depth (RD) based approaches. The intention of the pair-end mapping-based approach is that it determines CNVs according to the difference of the length between the two ends of paired reads mapped to the reference and the insert fragment, while the split-read based approach determines CNVs by splitting the sequence and observing the distance of the split reads mapped to the reference sequence. *De no* assembly approach is usually used to find out novel inserted sequences (Yuan et al., 2019b). These three categories of approaches are appropriate for the detection of CNVs with a limited size, since the pair-end mapping and split-read based approaches are subject to the length of inserted fragments and the *de no* assembly method is subject to the cost of computation time. Nevertheless, CNVs are usually ranging at a large scope of interval in size, and can be up to more than tens of M base-pairs. Relative to the above three categories, the RD based approach is more versatile in detecting CNVs with any sizes. The major principle of this approach is to determine CNVs according to the variance of RDs across the genome to be analyzed.

The RD based approach is generally implemented through the following four steps (Duan et al., 2014; Yuan et al., 2019a): (1) mapping sequencing reads to a reference genome and extracting a read count profile, (2) dividing the genome into non-overlapping bins and calculating a RD value for each bin based on the read count profile, (3) making normalization and correction to the RD values, and (4) analyzing the corrected RD values to declare CNVs. The theoretical assumption underlying the RD based approach is that the RD value of one bin or one region is roughly related to its corresponding copy number, i.e., the larger the RD value, the larger the copy number, and vice versa. Therefore, the key point here is how to design an appropriate scheme to reasonably analyze the RD values. The currently popular methods for detecting CNVs using RD values include but are not limited to: RDXplorer (Yoon et al., 2009), CNVnator (Abyzov et al., 2011), GROM-RD (Smith et al., 2015), XCAVATOR (Magi et al., 2017), Control-FREEC (Boeva et al., 2012), CNVkit (Talevich et al., 2016), CNAseg (Ivakhno et al., 2010), CopywriteR (Kuilman et al., 2015), SeqCNV (Chen et al., 2017), CloneCNA (Yu et al., 2016), iCopyDAV (Dharanipragada et al., 2018), DeAnnCNV (Zhang et al., 2015), CNV_IFTV (Yuan et al., 2019c), CONDEL (Yuan et al., 2020), and CNV-LOF (Yuan et al., 2019a). Each of these methods has its own characteristics and advantages. For example, Control-FREEC makes the best use of GC-content to normalize the read count profile so as to find out CNV regions, and iCopyDAV chooses an appropriate bin size and uses thresholds for RD values to declare CNVs. Although much effectiveness has been achieved by these methods, some factors such as low-level tumor purity (i.e., the fraction of tumor cells in the sequencing sample), limited coverage depth and GC-content bias still pose a big challenge to the detection of CNVs with small amplitudes. Therefore, it would be necessary and meaningful to seek for new methods that can grasp the essential characteristics of sequencing data associated with CNVs.

Given the above, we summarize several aspects that should be considered to improve the detection of CNVs. In the first place, it is necessary to make a smooth or segmentation to the observed RD profile, so that adjacent bins with similar amplitudes can be merged into the same region and the bins showing a local mutation state cannot be masked. In the second place, it is meaningful to extract effective features from sequencing data that can make an accurate distinguishing between mutated and normal genome regions. In the last place, it is necessary to design a reasonable model for displaying the extracted features and perform a suitable analysis of the features to determine CNVs.

With a careful consideration of the problems described above, in this article, we propose a new method, called dpCNV, for the detection of CNVs from NGS data. The motivation and underlying idea of dpCNV could be demonstrated as below. It considers the inherent correlations among adjacent positions on the genome, and thus analyzes CNVs based on the unit of genome segments rather than individual bins. These segments can be produced by performing a segmentation process on the RD profile. It carefully takes into account that CNV regions usually accounts for a small fraction of the whole genome and many CNVs just display a "local" outlier state, and thus extracts two related features (i.e., local density and minimum distance) from the RD profile based on the density peak algorithm (Rodriguez and Laio, 2014). Finally, dpCNV analyzes the two feature values for each segment through multivariate Gaussian distribution and calculates the corresponding p-value to declare whether it is a CNV. We perform a large number of simulation experiments to test the dpCNV method and make comparisons with several existing methods. The experimental results demonstrate the merit of the proposed method. Moreover, we apply it to analyze a set of real sequencing samples and prove its validity.

The remainder of this article is organized as follows. Section "Materials and Methods" demonstrates the workflow of dpCNV and the related principles. In section "Results," simulation studies are designed to evaluate the performance of the proposed method and its peer methods, as well as validations by applying it to a set of real sequencing samples. Section "Conclusion" discusses the proposed method and summarizes an outline of future work.

MATERIALS AND METHODS

Workflow of dpCNV

The workflow of the dpCNV method is demonstrated in **Figure 1**. The dpCNV method works by starting from an input of a sequenced tumor sample and a reference genome. The sequenced tumor sample is aligned to the reference genome by using the commonly used alignment tool BWA (Li and Durbin, 2009), and then a read count profile is extracted from the alignment result by using SAMtools (Li et al., 2009). With the read count profile, a RD profile is produced with a pre-defined bin size, such as 1000 base pairs (bp), which is moderate in the detection of CNVs (Yuan et al., 2020).

Based on the RD profile, the dpCNV method performs CNV analysis via the following four steps. (I) It implements a segmentation process on the RD profile to generate small genome segments, each of which usually include a set of adjacent and correlated bins. Here, the segmentation is carried out by using the Fused-Lasso algorithm (Tibshirani and Wang, 2008). (II) It extracts two features as the statistic and calculates the corresponding values via density peak algorithm. (III) It establishes a two-dimensional null distribution via multivariate Gaussian distribution and tests significance for each segment. (IV) It declares CNVs via a threshold of significance level and determines CNV statuses (i.e., amplification or deletion) via a RD cutoff.

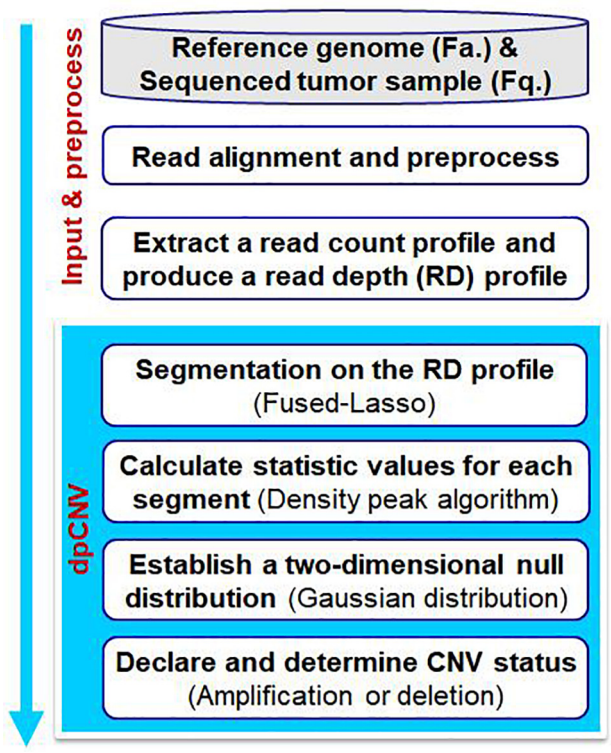

FIGURE 1 | Workflow of the dpCNV method to detect copy number variations from tumor samples with next-generation sequencing data.

Segmentation on the RD Profile

With the RD profile, a GC-content bias correction process is carried out through a similar approach with the works (Abyzov et al., 2011; Yuan et al., 2019a), and then a segmentation process is implemented on the corrected RD profile. The purpose of the segmentation is to divide the whole RD profile into a set of small segments, each of which is composed by adjacent bins, and is to provide a segment-based unit for the detection of CNVs rather than a bin-based unit. Theoretically, the segment-based unit can help to increase the independence of elements in significance testing, so that a reasonable evaluation of *p*-values can be expected to be achieved (Yuan et al., 2012b). Nevertheless, the bin-based unit may result in a conservativeness of *p*-value evaluation since adjacent bins are usually correlated (Yuan et al., 2019c).

There are various existing approaches that can carry out segmentation on the RD profile. Here we choose the Fused-Lasso algorithm for this task (Tibshirani and Wang, 2008). In comparison with other segmentation algorithms, the Fused-Lasso algorithm performs better in smoothing adjacent bins with highly similar RD values while remaining local fluctuations among the resulted segments (Tibshirani and Wang, 2008). For convenience, the resulted segments are denoted by:

$$S = \{s_1, s_2, s_3, \ldots, s_n\} \tag{1}$$

where n denotes the total number of segments that have been achieved. The following steps of analyzing CNVs are based on the set of S.

Calculation of Statistic Values for Each Segment

With the segment-based RD profile S, we adopt the density-based peak algorithm to extract two features as the statistic for each segment: local density (ρ) and minimum distance (δ), and to calculate their corresponding values. With the consideration of that regions with changed copy numbers are inherently different from those of normal copy numbers and only account for a small part of the whole genome, we transfer the problem of detecting CNVs to the issue of identifying outliers from the set of segments with features of ρ and δ. Accordingly, each segment can be regarded as an object or a point in the two dimensional space of ρ and δ. In the following text, we make a detailed description to these two features and the calculation approach.

Before describing the two features ρ and δ, we introduce the Euclidean distance between any two objects (segments) s_i and s_j. Given the total number of segments of n, an Euclidean distance matrix $M_{n\times n}$ can be obtained, where each element (d_{ij}) can be calculated by the Euclidean distance formula:

$$d_{ij} = \sqrt{(\rho_i - \rho_j)^2 + (\delta_i - \delta_j)^2} \tag{2}$$

where ρ_i and δ_i represent the feature values of object s_i, and the same to ρ_j and δ_j. With the Euclidean distance matrix $M_{n\times n}$, an adjustable distance threshold γ is introduced according to the theorem of the density peak algorithm (Rodriguez and Laio, 2014). This threshold can be explained as a radius of each object

s_i and is used to calculate how many objects are adjacent to the object s_i within the distance of γ. Then, the concept of local density ρ for each object is produced.

Definition 1

The local density ρ_i of the object s_i is defined as the number of objects adjacent to the object s_i with the radius γ, and can be calculated by using Eq. 3:

$$\rho_i = \sum_{j \neq i}^{n} \chi(d_{ij} - \gamma) \tag{3}$$

where $\chi(x) = 1$ if $x < 0$, and otherwise, $\chi(x) = 0$.

Definition 2

The minimum distance δ_i of the object s_i is defined as the minimum value among the distances between the object s_i and those objects with higher density than s_i, and can be expressed as Eq. 4:

$$\delta_i = \min_{j:\rho_i < \rho_j} (d_{ij}). \tag{4}$$

For the object s_i with the highest density, the value δ_i is defined as the maximum distance between the object and the rest of objects in the set S, and can be expressed as Eq. 5:

$$\delta_i = \max_j(d_{ij}) \quad \text{if} \quad \rho_i \geq \rho_{j \neq i}. \tag{5}$$

For a clear understanding of local density and minimum distance, we use an example to describe the distribution of a set of objects with respect to the values of the two features, as shown in **Figure 2**. For the example, we can see that the objects at the abnormal area (outliers) are near to the left and bottom side of the distribution. From the basic idea of density peak algorithm, outliers usually have a larger minimum distance and a smaller local density than those of other objects. Here, the abnormal area denotes the place of outlier objects, and normal area denotes the cluster of most objects. More details about the density peak algorithm is referred to Rodriguez and Laio (2014).

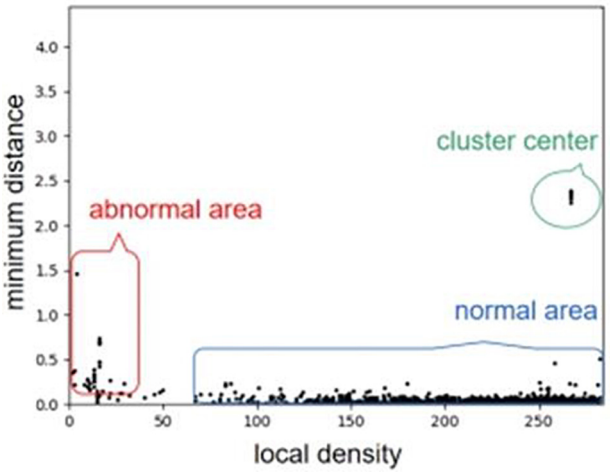

FIGURE 2 | An example of describing the distribution of a set of objects with respect to the values of the two features. We can note that the objects at the abnormal area (outliers) are near to the left and bottom side of the distribution.

Establish of a Two-Dimensional Null Distribution

With the statistic values in a two-dimensional space [i.e., local density (ρ) and minimum distance (δ)], the task now is how to design an appropriate model to test the significance of them. Since the values of the two features are usually at different scopes, it is not appropriate to combine them as a single feature value for the analysis. Therefore, it would be reasonable to design a model that can analyze the statistic values in a two-dimensional space. To mirror this, we establish a multivariate (i.e., two-dimension) Gaussian distribution as the null distribution based on the observed statistic values, and then evaluate a p-value for each of them. The multivariate Gaussian distribution is expressed as Eq. 6:

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)\,|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{6}$$

where μ is a two-dimensional vector, representing the mean values of local density and minimum distance, i.e., $\mu = [\overline{\rho}, \overline{\delta}]$, and Σ represents the covariance matrix of the two features.

The reason about why to choose a multivariate Gaussian distribution as the null distribution can be explained as below. Assuming that there are no CNVs in the segment-based RD profile S, and then the mean RD value should be around the sequencing coverage depth of the whole genome and the variance is primarily contributed by random artifacts such as sequencing and mapping errors. From this viewpoint, the RD values can be approximately modeled by a Gaussian distribution (Yuan et al., 2020). Theoretically, with a Gaussian distributed object, the deduced local density (ρ) and minimum distance (δ) would also follow Gaussian distribution, respectively. Therefore, the joint of the two features can be approximately modeled by a two-dimensional Gaussian distribution. For a clear understanding of this, we depict an example using a simulated dataset to show the distribution of the statistic values (ρ, δ) in **Figure 3**.

Declaration and Determination of CNV Statuses

Based on the two-dimensional null distribution above, the p-value (p_i) for each object (segment) s_i can be calculated. We define a commonly used significance level α as the cutoff for declaring CNVs, i.e., if p_i is less than α, then the object s_i will be declared as a CNV status; otherwise, it is regarded as a normal status. According to our experience and a large number of simulation experiments, we find that the value of α is appropriate to be assigned with 0.005.

With the abnormal objects, we further deduce their types (i.e., amplification or deletion) of CNV according to their RD values. Here, we use the average RD value of the objects in the cluster center (shown in **Figure 3**) as the baseline (r_b) of normal copy number. This is consistent with that the objects in the cluster center are regarded as normal objects according to the density

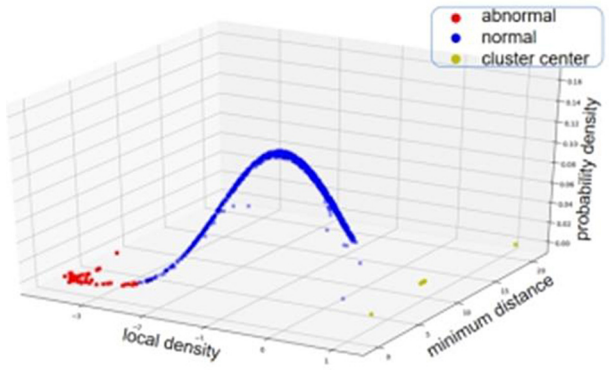

FIGURE 3 | An example of showing the two-dimensional Gaussian distribution of the statistic values (i.e., local density and minimum distance) based on simulation data. The blue points represent the segments with normal copy numbers while the red points represent the segments with abnormal copy numbers.

peak algorithm. Subsequently, for each abnormal object, if its RD value is larger than r_b, then it is regarded as an amplification event, otherwise, it is regarded as a deletion event.

RESULTS

The dpCNV software is implemented in Python language, and the code is publicly available at https://github.com/BDanalysis/dpCNV/. In order to demonstrate the performance and usefulness of our proposed method, we first conduct a number of simulation experiments and make comparisons with several existing methods in terms of precision, sensitivity and F1-score (the harmonic mean of sensitivity and precision). Then, we apply the proposed method to a set of real sequencing samples, which have been obtained from the European Genome-phenome Archive (EGA) databases.[1] To assure a fair comparison between dpCNV and other methods, we use the default parameter values in the implementation of the compared methods.

Simulation Studies

Simulation studies are usually regarded as an appropriate and feasible way to assess the performance of existing and newly developed methods (Yuan et al., 2012a, 2017, 2018). This is because that the ground truth CNVs embedded in the simulated data sets could be used for an exact calculation of sensitivity and precision for the methods. Currently, there are many methods for simulating NGS data have been proposed. Here, we use one of our previously developed simulation methods, IntSIM (Yuan et al., 2017), for the simulation of NGS data with ground truth CNVs. Two primarily factors (i.e., tumor purity and depth of coverage) have been considered in the simulation process. Specifically, six scenarios have been simulated by setting different values of tumor purity (0.2, 0.3, and 0.4) and coverage depth (4× and 6×), and in each scenario 50 replicated samples have been produced.

With these simulated data sets, the dpCNV method and four peer methods (including FREEC, GROM-RD, CNVnator, and CNV_IFTV) are performed. Their results and comparisons are depicted in **Figure 4**. Here, the precision is calculated as the ratio of the number of correctly detected CNVs to the number of all declared CNVs, while the sensitivity is calculated by the ratio of the number of correctly detected CNVs to the total number of ground truth CNVs. From the **Figure 4**, one could observe that the performances of most methods are improving along with the increasing of tumor purity and coverage depth. Comparatively, the dpCNV method is superior in terms of the trade-off (F1-score) between precision and sensitivity in each of the simulation scenarios. With respect to sensitivity, dpCNV ranks first in all the simulation scenarios, followed by FREEC or CNV_IFTV. With respect to precision, GROM-RD and CNVnator display larger values than other methods.

The fact that dpCNV is superior to other methods under this study is due to the following reasons. Firstly, the relationship between adjacent bins has been taken into account by performing a segmentation process. In this process, most noised data points can be smoothed, and some local variations can be remained. In addition, two meaningful features (i.e., local density and minimum distance) are extracted from the segmented data based on a density peak algorithm. Secondly, a two-dimensional null distribution has been established for testing the significance of each genome segment. This can help to relieve the conservativeness of p-value assessment and provide a meaningful null hypothesis testing.

Real Data Applications

To further validate the performance of dpCNV, we apply it to three whole-genome sequencing data (EGAD00001000144_LC, EGAR00001004802_2053_1, and EGAR00001004836_2561_1) obtained from the EGA project. These samples include a lung cancer sample and two ovarian cancer samples. Besides, we also perform three peer methods (FREEC, CNVnator, CNV_IFTV) on these samples for comparisons. Since real sequencing data usually have no ground truth CNVs, it is difficult for us to exactly calculate the sensitivity and precision for the methods. Nevertheless, we analyze the overlapping results among the compared methods to observe the consistence between their results, as shown in **Figure 5**. We can note that CNVnator gets the largest number of overlaps with other methods, followed by dpCNV and FREEC. However, the total number of detected CNVs detected by CNVnator is also the largest. This means that it is not appropriate to determine which method is superior just according to the number of overlapped CNVs. Nevertheless, we adopt the overlapping density score (ODS) proposed in our previous work (Yuan et al., 2020) to evaluate the methods. The ODS is calculated by using Eq. 7. The comparative result is shown in **Table 1**, from which we can notice that dpCNV achieves the highest ODS in the analysis of two ovarian tumor samples and FREEC gets the highest ODS in the analysis of the lung tumor sample:

$$\mathrm{ODS} = m_{cnv} \cdot m'_{cnv} \tag{7}$$

[1]https://www.ebi.ac.uk/ega/

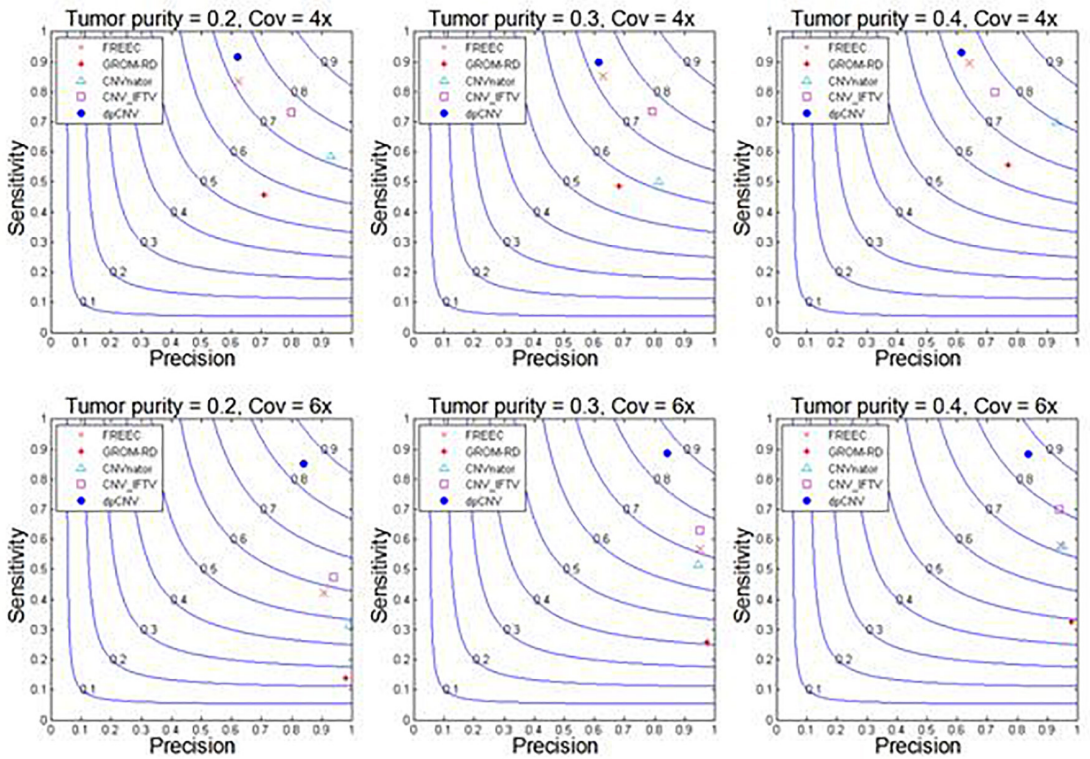

FIGURE 4 | Performance comparisons between our proposed method and the four peer methods in terms of sensitivity, precision, and F1-score (colored curves) on simulation data.

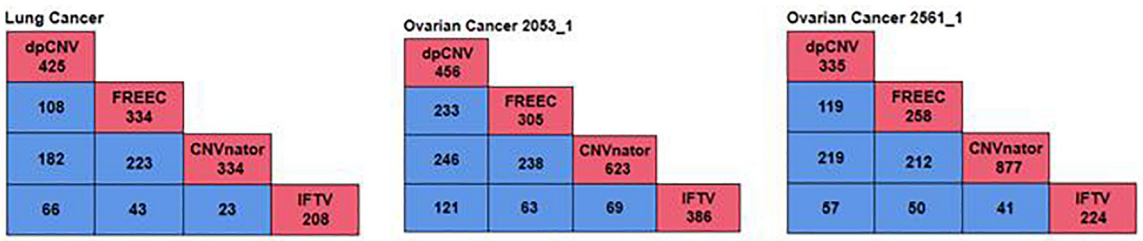

FIGURE 5 | The overlapping results of four methods on the three samples. The red boxes represent the total number of CNVs detected by each method, while the blue boxes denote the number of overlapping CNVs detected by any two methods.

TABLE 1 | Comparison of ODS between dpCNV and three peer methods on real samples.

Sample	dpCNV	FREEC	CNV_IFTV	CNVnator
EGAD00001000144_LC	99.4	**114.02**	47.96	19.06
EGAR00001004802_2053_1	**155.25**	152.89	35.75	44.2
EGAR00001004836_2561_1	**263.16**	192.79	57.04	114.7
Average	**172.6**	153.23	46.92	59.32

Bold value denotes the largest values in each line.

where m_{cnv} denotes the total overlapped CNVs divided by the number of compared methods and m'_{cnv} denotes the total overlapped CNV divided by the number of CNVs detected by itself.

An overview of the numbers of CNVs detected by the four methods are shown in **Figure 6**, where we could clearly take an overview of distribution on 22 autosomes of results called by dpCNV, FREEC, CNVnator, and IFTV, respectively. Each circus diagram is composed of two parts, the upper part consists of four arcs corresponding to the four detection methods and the lower part consists of 22 arcs corresponding to the 22 autosomes. In the lung cancer diagram, dpCNV obtains the largest number of CNVs while CNVnator obtains the smallest number of CNVs. In the diagrams of the two ovarian cancer samples, CNVnator gets the largest number of CNVs while FREEC and dpCNV get relatively fewer CNVs.

In addition, based on the COSMIC (catalog of somatic mutations in cancer) database, we analyze the CNVs detected

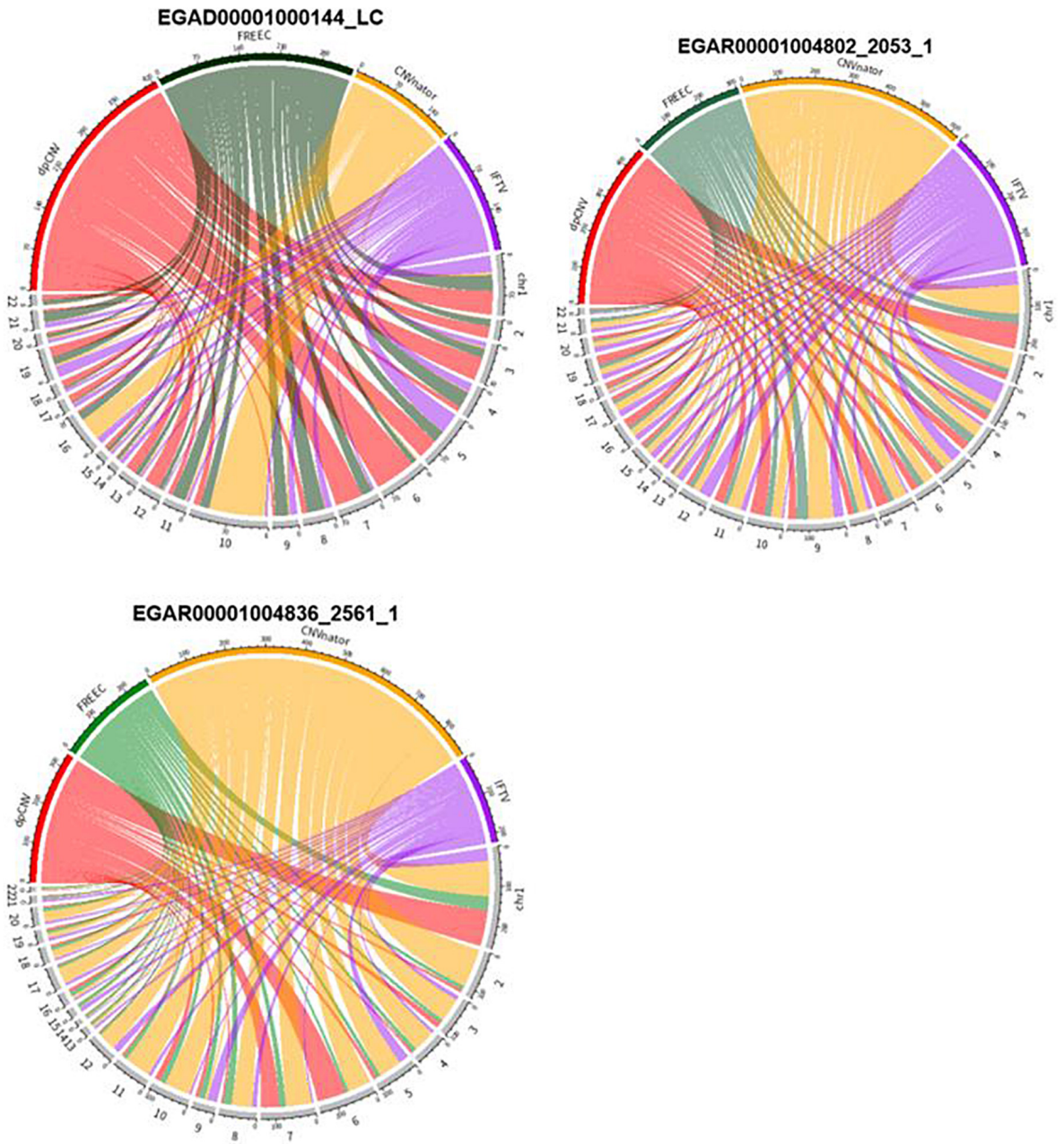

FIGURE 6 | The circus diagram on three real samples. The upper part consisting of four arcs indicates the four methods, while the lower part consisting of 22 arcs denotes 22 autosomes. The length of each arc in upper part represents the total number of detected CNVs.

by our proposed method on three whole genome sequencing data from biological meanings. For example, 425 CNVs detected by dpCNV from the lung cancer sample are compared to the COSMIC database. There are 151 cytobands and 405 genes in the comparative result. We may notice that many cytobands contain a lot of meaningful genes. For example, the cytoband 11p15.5 contains IFITM1 (Sakamoto et al., 2020) and IFITM3 (Infusini et al., 2015). Many of genes are confirmed to be tumor driver genes and closely related to non-small cell lung cancer, such as C3orf21 (Yang et al., 2017), ZNF454 (Zhu et al., 2020), and C10orf137 (Zheng et al., 2013). For the two ovarian cancer samples, dpCNV gets 225 cytobands and 128 cytobands, 285 genes and 529 genes overlapped with the COSMIC database, respectively, in which there are many important tumor driver genes corresponding to ovarian cancer, such as PUM1 (Guan et al., 2018), GOLPH3L (Feng et al., 2015), PIWIL4 (Guo et al., 2009), and KNDC1 (Yu et al., 2020).

CONCLUSION

Accurate detection of CNVs is a crucial step for a comprehensive analysis of genomic mutations in the study of genome evaluation and human complex diseases. In this article, a new method named dpCNV is proposed for the detection of CNVs from NGS data. The central point of dpCNV is that it extracts two meaningful features based on the density peak algorithm and establishes a two-dimensional null distribution to test the significance of genome segments. dpCNV is different from traditional methods and have some new characteristics: (1) it considers the intrinsic correlations among genome bins, and adopts Fused-Lasso segmentation algorithm to smooth the noise data between adjacent bins; (2) it carefully takes into account that CNV regions usually accounts for a small fraction of the whole genome and many CNVs just display a "local" outlier state, and thus extracts two related features (i.e., local density and minimum distance) from the RD profile based on the density peak algorithm; (3) it analyzes the two feature values for each segment through multivariate Gaussian distribution and calculates the corresponding *p*-value to declare whether it is a CNV.

The performance of dpCNV is assessed and validated through simulation studies and applications to a set of real sequencing samples. In simulation experiments, dpCNV outperforms four peer methods (FREEC, GROM-RD, CNVnator, and CNV_IFTV) in terms of sensitivity and F1-score. In real sample experiments, dpCNV is performed on three whole genome sequencing samples including a lung cancer sample and two ovarian samples, and is compared with three peer methods (FREEC, CNVnator, and CNV_IFTV). Here, we have not make comparison with GROM-RD since it has not obtained results from these real sequencing samples. In this comparison, we make an evaluation of the four methods by using ODS. The result indicates that dpCNV obtains a better performance than other methods. In addition, we demonstrate the biological meanings of the detected CNVs by referring the COSMIC database.

With regard to the future work, we plan to make a further improvement to the current version of the dpCNV method from the following aspects. In the first place, we will design a strategy to predict tumor purity and integrate it to the detection of CNVs. In the second place, we intend to predict absolute copy numbers for each CNV region, since absolute copy numbers might provide much information of the study of chromosome instability. In the third place, we intend to combine the detection of CNVs with other types of genomic mutations into a pipeline analysis, which will help to improve the efficiency of genomic mutation analysis. In the last palace, it is necessary to explore the detection of CNVs by using mRNA sequencing data. Generally, RD values obtained from the sequencing data on DNA are closely related with copy numbers. A high expression of mRNAs might be associated with a large copy number. Therefore, using mRNA sequencing data may facilitate the detection of CNVs in tumor genomes.

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/s.

AUTHOR CONTRIBUTIONS

KX and YT participated in the study and design of algorithms and experiments, and participated in writing the manuscript. XY directed the whole work, conceived of the study and help, and edited the manuscript. YT participated in the analysis of the performance of the proposed method. All authors read the final manuscript and agreed the submission.

FUNDING

This work was supported by the Natural Science Foundation of China under grant no. 61571341.

REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110

Auer, H., Newsom, D. L., Nowak, N. J., McHugh, K. M., Singh, S., Yu, C. Y., et al. (2007). Gene-resolution analysis of DNA copy number variation using oligonucleotide expression microarrays. *BMC Genom.* 8:111. doi: 10.1186/1471-2164-8-111

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. U S A.* 104, 20007–20012. doi: 10.1073/pnas.0710052104

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., et al. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 28, 423–425. doi: 10.1093/bioinformatics/btr670

Cai, H., Chen, P., Chen, J., Cai, J., Song, Y., and Han, G. (2018). WaveDec: a wavelet approach to identify both shared and individual patterns of copy-number variations. *IEEE Trans. Biomed. Eng.* 65, 353–364. doi: 10.1109/tbme.2017.2769677

Chen, Y., Zhao, L., Wang, Y., Cao, M., Gelowani, V., Xu, M. C., et al. (2017). SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data. *BMC Bioinform.* 18:147. doi: 10.1186/s12859-017-1566-3

Dharanipragada, P., Vogeti, S., and Parekh, N. (2018). iCopyDAV: integrated platform for copy number variations-Detection, annotation and visualization. *PLoS One* 13:e0195334. doi: 10.1371/journal.pone.0195334

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeckert, C. J., et al. (2006). STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res.* 16, 1149–1158. doi: 10.1101/gr.5076506

Duan, J., Deng, H. W., and Wang, Y. P. (2014). Common copy number variation detection from multiple sequenced samples. *IEEE Trans. Biomed. Eng.* 61, 928–937. doi: 10.1109/tbme.2013.2292588

Feng, Y. L., He, F., Wu, H. N., Huang, H., Zhang, L., Han, X., et al. (2015). GOLPH3L is a novel prognostic biomarker for epithelial ovarian Cancer. *J. Cancer* 6, 893–900. doi: 10.7150/jca.11865

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., et al. (2006). Copy number variation: new insights in genome diversity. *Genome Res.* 16, 949–961. doi: 10.1101/gr.3677206

Fridley, B. L., Chalise, P., Tsai, Y. Y., Sun, Z., Vierkant, R. A., Larson, M. C., et al. (2012). Germline copy number variation and ovarian cancer survival. *Front. Genet.* 3:142. doi: 10.3389/fgene.2012.00142

Guan, X., Chen, S., Liu, Y., Wang, L. L., Zhao, Y., and Zong, Z. H. (2018). PUM1 promotes ovarian cancer proliferation, migration and invasion. *Biochem. Biophys. Res. Commun.* 497, 313–318. doi: 10.1016/j.bbrc.2018.02.078

Guo, L. M., Liu, M., Li, X., and Tang, H. (2009). The expression and functional research of PIWIL4 in human ovarian Cancer. *Prog. Biochem. Biophys.* 36, 353–357. doi: 10.3724/sp.j.1206.2008.00478

Infusini, G., Smith, J. M., Yuan, H., Pizzolla, A., Ng, W. C., Londrigan, S. L., et al. (2015). Respiratory DC Use IFITM3 to avoid direct viral infection and safeguard virus-specific CD8+ T cell priming. *PLoS One* 10:e0143539. doi: 10.1371/journal.pone.0143539

Ivakhno, S., Royce, T., Cox, A. J., Evers, D. J., Cheetham, R. K., and Tavare, S. (2010). CNAseg–a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* 26, 3051–3058. doi: 10.1093/bioinformatics/btq587

Kuilman, T., Velds, A., Kemper, K., Ranzani, M., Bombardelli, L., Hoogstraat, M., et al. (2015). CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol.* 16:49.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Magi, A., Pippucci, T., and Sidore, C. (2017). XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. *BMC Genom.* 18:747. doi: 10.1186/s12864-017-4137-0

Martin, J., Tammimies, K., Karlsson, R., Lu, Y., Larsson, H., Lichtenstein, P., et al. (2019). Copy number variation and neuropsychiatric problems in females and males in the general population. *Am. J. Med. Genet. Part B, Neuropsychiatric Genet.* 180, 341–350. doi: 10.1002/ajmg.b.32685

Nowak, G., Hastie, T., Pollack, J. R., and Tibshirani, R. (2011). A fused lasso latent feature model for analyzing multi-sample aCGH data. *Biostatistics* 12, 776–791. doi: 10.1093/biostatistics/kxr012

Rodriguez, A., and Laio, A. (2014). Machine learning. clustering by fast search and find of density peaks. *Science* 344, 1492–1496. doi: 10.1126/science.1242072

Sakamoto, S., Inoue, H., Kohda, Y., Ohba, S. I., Mizutani, T., and Kawada, M. (2020). Interferon-Induced transmembrane protein 1 (IFITM1) promotes distant metastasis of small cell lung Cancer. *Int. J. Mol. Sci.* 21:4934. doi: 10.3390/ijms21144934

Shlien, A., and Malkin, D. (2009). Copy number variations and cancer. *Genome Med.* 1:62.

Smith, S. D., Kawash, J. K., and Grigoriev, A. (2015). GROM-RD: resolving genomic biases to improve read depth detection of copy number variants. *PeerJ* 3:e836. doi: 10.7717/peerj.836

Talevich, E., Shain, A. H., Botton, T., and Bastian, B. C. (2016). CNVkit: genome-wide copy number detection and visualization from targeted dna sequencing. *PLoS Comput. Biol.* 12:e1004873. doi: 10.1371/journal.pcbi.1004873

Tibshirani, R., and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* 9, 18–29. doi: 10.1093/biostatistics/kxm013

Xi, J., Li, A., and Wang, M. (2020a). HetRCNA: a novel method to identify recurrent copy number alternations from heterogeneous tumor samples based on matrix decomposition framework. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 422–434. doi: 10.1109/tcbb.2018.2846599

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. (2020b). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 36, 1855–1863.

Yang, L., Wang, Y., Fang, M., Deng, D., and Zhang, Y. (2017). C3orf21 ablation promotes the proliferation of lung adenocarcinoma, and its mutation at the rs2131877 locus may serve as a susceptibility marker. *Oncotarget* 8, 33422–33431. doi: 10.18632/oncotarget.16798

Yoon, S. T., Xuan, Z. Y., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109

Yu, S. Q., Shen, J. Y., Fei, J., Zhu, X. Q., Yin, M. C., and Zhou, J. W. (2020). KNDC1 is a predictive marker of malignant transformation in borderline ovarian tumors. *OncoTargets Therapy* 13, 709–718. doi: 10.2147/ott.s223304

Yu, Z., Li, A., and Wang, M. (2016). CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data. *BMC Bioinform.* 17:310. doi: 10.1186/s12859-016-1174-7

Yuan, X., Bai, J., Zhang, J., Yang, L., Duan, J., Li, Y., et al. (2020). CONDEL: detecting copy number variation and genotyping deletion zygosity from single tumor samples using sequence data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1141–1153.

Yuan, X., Gao, M., Bai, J., and Duan, J. (2018). SVSR: a program to simulate structural variations and generate sequencing reads for multiple platforms. *IEEE/ACM Trans. Comput. Biol. Bioinform.* Online ahead of print.

Yuan, X., Li, J., Bai, J., and Xi, J. (2019a). A local outlier factor-based detection of copy number variations from NGS data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* Online ahead of print.

Yuan, X., Xu, X., Zhao, H., and Duan, J. (2019b). ERINS: novel sequence insertion detection by constructing an extended reference. *IEEE/ACM Trans. Comput. Biol. Bioinform.* Online ahead of print.

Yuan, X., Yu, J., Xi, J., Yang, L., Shang, J., Li, Z., et al. (2019c). CNV_IFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* Online ahead of print.

Yuan, X., Miller, D. J., Zhang, J., Herrington, D., and Wang, Y. (2012a). An overview of population genetic data simulation. *J. Comput. Biol.* 19, 42–54. doi: 10.1089/cmb.2010.0188

Yuan, X., Yu, G., Hou, X., Shih, Ie, M., Clarke, R., et al. (2012b). Genome-wide identification of significant aberrations in cancer genome. *BMC Genomics* 13:342. doi: 10.1186/1471-2164-13-342

Yuan, X., Zhang, J., and Yang, L. (2017). IntSIM: an integrated simulator of next-generation sequencing data. *IEEE Trans. Biomed. Eng.* 64, 441–451. doi: 10.1109/tbme.2016.2560939

Zhang, B., Hou, X., Yuan, X., Shih, Ie, M., Zhang, Z., et al. (2014). AISAIC: a software suite for accurate identification of significant aberrations in cancers. *Bioinformatics* 30, 431–433. doi: 10.1093/bioinformatics/btt693

Zhang, Y., Yu, Z., Ban, R., Zhang, H., Iqbal, F., Zhao, A., et al. (2015). DeAnnCNV: a tool for online detection and annotation of copy number variations from whole-exome sequencing data. *Nucleic Acids Res.* 43, W289–W294.

Zhao, M., Wang, Q. G., Wang, Q., Jia, P. L., and Zhao, Z. M. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinform.* 14:S1. doi: 10.1186/1471-2105-14-S11-S1

Zheng, C. X., Gu, Z. H., Han, B., Zhang, R. X., Pan, C. M., Xiang, Y., et al. (2013). Whole-exome sequencing to identify novel somatic mutations in squamous cell lung cancers. *Int. J. Oncol.* 43, 755–764. doi: 10.3892/ijo.2013.1991

Zhou, X., Liu, J., Wan, X., and Yu, W. (2014). Piecewise-constant and low-rank approximation for identification of recurrent copy number variations. *Bioinformatics* 30, 1943–1949. doi: 10.1093/bioinformatics/btu131

Zhu, Q. Q., Wang, J., Zhang, Q. J., Wang, F. X., Fang, L. H., Song, B., et al. (2020). Methylation-driven genes PMPCAP1, SOWAHC and ZNF454 as potential prognostic biomarkers in lung squamous cell carcinoma. *Mol. Med. Rep.* 21, 1285–1295.

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

frontiers in Genetics

METHODS
published: 23 April 2021
doi: 10.3389/fgene.2021.665416

CBP-JMF: An Improved Joint Matrix Tri-Factorization Method for Characterizing Complex Biological Processes of Diseases

Bingbo Wang[1], Xiujuan Ma[1], Minghui Xie[1], Yue Wu[1], Yajun Wang[2], Ran Duan[1], Chenxing Zhang[1], Liang Yu[1], Xingli Guo[1] and Lin Gao[1]**

[1] School of Computer Science and Technology, Xidian University, Xi'an, China, [2] School of Humanities and Foreign Languages, Xi'an University of Technology, Xi'an, China

OPEN ACCESS

Edited by:
Jianing Xi,
Northwestern Polytechnical University, China

Reviewed by:
Hao Wu,
Shandong University, China
Feng Li,
Qufu Normal University, China
Kai Shi,
Guilin University of Technology, China

***Correspondence:**
Bingbo Wang
bingbowang@xidian.edu.cn
Lin Gao
lgao@mail.xidian.edu.cn

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

Received: 08 February 2021
Accepted: 01 March 2021
Published: 23 April 2021

Citation:
Wang B, Ma X, Xie M, Wu Y, Wang Y, Duan R, Zhang C, Yu L, Guo X and Gao L (2021) CBP-JMF: An Improved Joint Matrix Tri-Factorization Method for Characterizing Complex Biological Processes of Diseases. Front. Genet. 12:665416. doi: 10.3389/fgene.2021.665416

Multi-omics molecules regulate complex biological processes (CBPs), which reflect the activities of various molecules in living organisms. Meanwhile, the applications to represent disease subtypes and cell types have created an urgent need for sample grouping and associated CBP-inferring tools. In this paper, we present CBP-JMF, a practical tool primarily for discovering CBPs, which underlie sample groups as disease subtypes in applications. Differently from existing methods, CBP-JMF is based on a joint non-negative matrix tri-factorization framework and is implemented in Python. As a pragmatic application, we apply CBP-JMF to identify CBPs for four subtypes of breast cancer. The result shows significant overlapping between genes extracted from CBPs and known subtype pathways. We verify the effectiveness of our tool in detecting CBPs that interpret subtypes of disease.

Keywords: non-negative matrix factorization, complex biological processes, multi-dimensional genomic data, disease, subtype

INTRODUCTION

Complex biological processes (CBPs) are the coordinated effect of multiple molecules, which result in some functional pathways and the vital processes occurring in living organisms. In addition, the vast amounts of multi-omics data, such as genomics, epigenomics, transcriptomics, proteomics, and metabolomics, can be integrated to understand systems biology accurately (Suravajhala et al., 2016). Hasin et al. (2017) pointed out that a deeper and better understanding of important biological processes and modules can be obtained through multi-omics studies. However, practical tools are still missing to integrate diverse multi-omics data at different biological levels and reveal the CBPs and other problems like the causes of diseases.

Non-negative matrix factorization (NMF) (Lee and Seung, 1999) is a powerful tool for dimension reduction and feature extraction. It has been increasingly applied to diverse fields, including bioinformatics (*e.g.*, high-dimensional genomic data analysis). For example, Brunet et al. (2004) applied NMF and consensus clustering to the gene expression data of leukemia to discover metagenes and molecular patterns. Xi et al. (2018) detected driver genes from pan-cancer data based on another matrix decomposition framework called matrix tri-factorization. Up to

now, several variants of NMF have been proposed, including tri-factorization NMF (Ding et al., 2006), graph-regularized NMF (Cai et al., 2011), joint NMF (Zhang et al., 2012), iNMF (Yang and Michailidis, 2016), *etc*. (more details are in **Supplementary Note 1** of the **Supplementary Materials**). In 2012, jNMF (Zhang et al., 2012) was proposed to identify multi-omics modules by integrating cancer's DNA methylation data, gene expression data, and miRNA expression data. Chen and Zhang (2018) applied joint matrix tri-factorization to discover two-level modular organization from matched genes and miRNA expression data, gene expression data, and drug response data.

Omics data across the same samples contain signal values from expression counts, methylation levels, and protein concentrations, which control biological systems, resulting in so-called multi-dimensional genomic (MG) data. The natural representation of these diverse MG data is a series of matrices with measured values in rows and individual samples in columns. Recently, there are integrative analysis tools based on NMF technique that reveal low-dimensional structure patterns. The low-dimensional structure patterns reflect CBPs and sample groups while preserving as much information as possible from high-dimensional MG data (Stein-O'Brien et al., 2018).

In general, most particular matrix factorization techniques are being developed to enhance their applicability to specific biological problems. Meanwhile, the applications to represent disease subtypes (Biton et al., 2014) and cell types (Fan et al., 2016) have created an urgent need for sample grouping and associated CBP-inferring tools. Moreover, cancer and other complex diseases are heterogeneous, *i.e*., there are various subgroups for a cancer or a complex disease. The study of the heterogeneity of cancer and complex diseases will help us understand the disease further and provide better opportunities to disease treatment (Xi et al., 2020). To address this issue, we extend traditional jNMF and develop CBP-JMF, an improved joint matrix tri-factorization framework for characterizing CBPs that represent sample groups, and implement a Python package. This package takes labeled samples as the prior information and integrates MG data (*e.g*., copy number variation, gene expression, microRNA expression, and/or molecule interaction network) to identify the underlying CBPs which characterize the specific functional properties of each group. CBP-JMF can be used to mark unlabeled samples with groups of known labels. For ease of use, CBP-JMF can recommend reasonable parameter settings for users. CBPs found by CBP-JMF are connected network markers, and they are distinguished between sample groups. These markers usually have specific biological functions and play important roles in phenotypes. As an example, CBPs for subtypes of breast cancer are obtained by CBP-JMF, but they may not have been collected in any reference database yet.

The rest of this paper is organized as follows. Section "Framework of CBP-JMF" deals with the problem formulation of CBP-JMF and the implementation of it. Then, Section "Results" exemplifies our approach by applying CBP-JMF to identify CBPs for different subtypes of breast cancers and compares the results of classifying unlabeled samples with CBP-JMF and its several variants. Finally, Section "Discussion" discusses our results and lists our expectations of our method and the limitations of it. Section "Conclusions" highlights our method.

FRAMEWORK OF CBP-JMF

Problem Definition

Given a non-negative matrix $\mathbf{X} \in \mathbf{R}^{m\times n}$, it can be factorized into three non-negative matrix factors based on matrix tri-factorization: $\mathbf{X} \approx \mathbf{USV}$, where $\mathbf{U} \in \mathbf{R}^{m\times k}$, $\mathbf{S} \in \mathbf{R}^{k\times k}$, and $\mathbf{V} \in \mathbf{R}^{k\times n}$. Factored matrix $\mathbf{S}$ cannot only absorb scale difference between $\mathbf{U}$ and $\mathbf{V}$ but also indicates relationships between the identified k modules.

In CBP-JMF, given a MG dataset composed of P omics, it can be presented by multiple matrices $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(P)}$, as illustrated in **Figure 1**. For each matrix, the rows indicate molecules like genes, and the columns indicate samples; the values in it are related to the meaning of omics. If $\mathbf{X}^{(p)}$ $(p \in [1, P])$ is a matrix of gene expression data, $\mathbf{X}_{ij}^{(p)}$ represents the expression value of the gene in the i-th row on the j-th sample. Basically, each non-negative matrix $\mathbf{X}^{(p)} \in \mathbf{R}^{m\times n}$, $p = 1, 2, ..., P$ is factorized into three non-negative matrix factors based on matrix tri-factorization: $\mathbf{X}^{(p)} \approx \mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}$, where molecular coefficient matrix (MCM) $\mathbf{U}^{(p)} \in \mathbf{R}^{m\times k}$ and sample basis matrix (SBM) $\mathbf{V} \in \mathbf{R}^{k\times n}$ are the pattern indicator matrices of k CBPs and k sample groups, respectively. Scale absorbing matrix (SAM) $\mathbf{S}^{(p)} \in \mathbf{R}^{k\times k}$ explores the relationships between them. Furthermore, MCM describes the structure pattern between molecules (*e.g*., genes), SBM indicates the structure pattern between samples, and SAM absorbs the difference of scales between MCM and SBM (**Figure 1**). Each column of the MCM infers a latent feature associated with a CBP, and the continuous values in it represent the relative contribution of each molecule in the CBP. Meanwhile, each row of the SBM describes the relative contributions of the samples to a latent feature. The sample groups can be detected by comparing the relative weights in each row of the SBM.

Overall, $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(P)}$ can be jointly factorized into specific $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}$, $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(P)}$, and a common matrix $\mathbf{V}$. $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(P)}$ are across the same samples, and $\mathbf{V}$ reveals consistent sample groups of multi-omics data. In CBP-JMF, $\mathbf{V}$ can be divided into $\mathbf{V}^{L}$ and $\mathbf{V}^{UL}$ according to input data, where L and UL mean "labeled" samples and "unlabeled" samples, respectively.

Objective Function of CBP-JMF

Considering that different datasets may play different roles in data integration, we adopted a method that can learn the weights of different input data through a weighted joint tri-NMF:

$$\begin{aligned} &\min \sum_{p=1}^{P} \pi^{(p)} \left\| \mathbf{X}^{(p)} - \mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V} \right\|_F^2 + \omega \|\Pi\|^2 \\ &s.t.\ \pi^{(p)} > 0, \sum_{p=1}^{P} \pi^{(p)} = 1 \end{aligned} \tag{1}$$

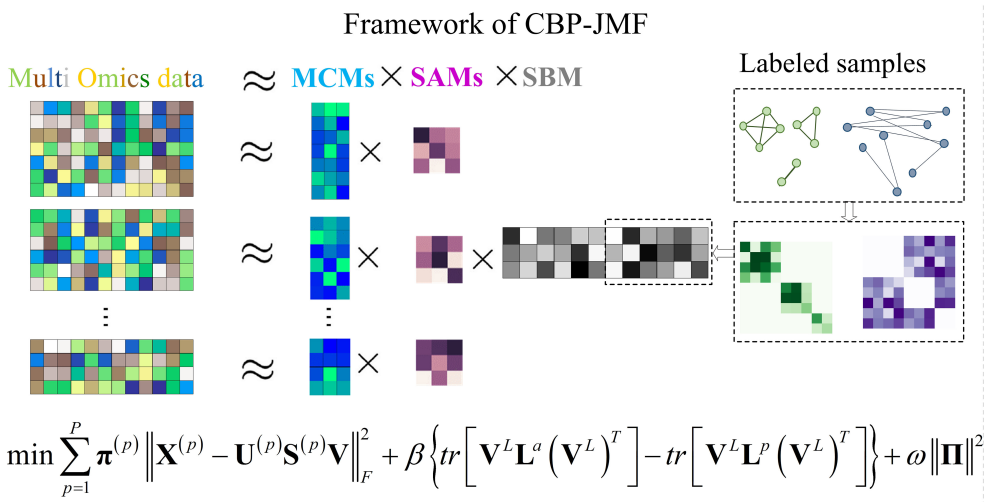

FIGURE 1 | Illustration of the framework and optimization objective function of complex biological processes–joint matrix tri-factorization.

where $\Pi = \left(\pi^{(1)}, \pi^{(2)}, ..., \pi^{(P)}\right)$. CBP-JMF differentiates the importance of datasets by the weight constraint $\|\Pi\|^2$, and $\pi^{(p)}$ will get a weight to represent the contribution of data $\mathbf{X}^{(p)}$ to objective function after optimization. If $\mathbf{X}^{(p)}$ contributes to the optimization of cost function, then it will be given a higher weight $\pi^{(p)}$, or if $\mathbf{X}^{(p)}$ contains lots of noises which hinder the optimization of objective function, it will be given a lower weight $\pi^{(p)}$.

In addition, $\mathbf{V}$ can be divided into labeled $\mathbf{V}^L$ and unlabeled $\mathbf{V}^{UL}$ parts according to the labeled samples and unlabeled samples. In order to learn the correlation between labeled samples, we use a graph Laplacian to represent the distance of labeled sample in latent space (Guan et al., 2015). We use Equations (2) and (3) to denote the distance between labeled samples from the same class and different class in the learned latent space, respectively,

$$\sum_{i=1}^{N^L}\sum_{j=1}^{N^L} \mathbf{W}_{ij}^a \left\| \mathbf{v}_i^L - \mathbf{v}_j^L \right\|_2^2 = tr\left[\mathbf{V}^L \mathbf{L}^a \left(\mathbf{V}^L\right)^T \right] \tag{2}$$

$$\sum_{i=1}^{N^L}\sum_{j=1}^{N^L} \mathbf{W}_{ij}^p \left\| \mathbf{v}_i^L - \mathbf{v}_j^L \right\|_2^2 = tr\left[\mathbf{V}^L \mathbf{L}^p \left(\mathbf{V}^L\right)^T \right] \tag{3}$$

where N^L is the number of labeled samples in $\mathbf{V}$, and $\mathbf{W}^a$ ($\mathbf{W}^{affinity}$) and $\mathbf{W}^p$ ($\mathbf{W}^{penalty}$) are the weighted adjacency matrices (see **Supplementary Note 2** in SM) corresponding to intra-group and inter-group samples respectively. $\mathbf{L}^a$ ($\mathbf{L}^{affinity}$) and $\mathbf{L}^p$ ($\mathbf{L}^{penalty}$) are the Laplacian matrix of $\mathbf{W}^a$ and $\mathbf{W}^p$, respectively, where $\mathbf{L}^a = \mathbf{D}^a - \mathbf{W}^a$, $\mathbf{L}^p = \mathbf{D}^p - \mathbf{W}^p$, $\mathbf{D}^a = \sum_{j=1}^{N^L} \mathbf{W}_{ij}^a$. In machine learning, people try to make samples from the same class near each other in the learned latent space and samples from different class far from each other. This principle can be written as

$$\min\left(tr\left[\mathbf{V}^L \mathbf{L}^a \left(\mathbf{V}^L\right)^T \right] - tr\left[\mathbf{V}^L \mathbf{L}^p \left(\mathbf{V}^L\right)^T \right] \right) \tag{4}$$

Combining weighted joint tri-NMF and the constraints of correlation between labeled samples mentioned above, we give the formulation of the optimization objective function of CBP-JMF as follows (**Figure 1**):

$$\begin{aligned} &\min_{\{\mathbf{U}^{(p)}\}_{p=1}^P, \{\mathbf{S}^{(p)}\}_{p=1}^P, \mathbf{V}} \sum_{p=1}^{P} \pi^{(p)} \left\| \mathbf{X}^{(p)} - \mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V} \right\|_F^2 \\ &+\beta \left\{ tr\left[\mathbf{V}^L \mathbf{L}^a \left(\mathbf{V}^L\right)^T \right] - tr\left[\mathbf{V}^L \mathbf{L}^p \left(\mathbf{V}^L\right)^T \right] \right\} + \omega \|\Pi\|^2 \\ &s.t.\ \forall p, \mathbf{U}_{ij}^{(p)} \geq 0, \mathbf{V}_{ij} \geq 0, \pi^{(p)} \geq 0, \sum_{p=1}^{P} \pi^{(p)} = 1 \end{aligned} \tag{5}$$

Parameters β and ω represent the importance of the graph Laplacian regularization and weight constraint $\|\Pi\|^2$. In total, each $\mathbf{X}^{(p)}$ is factorized into individual molecular matrix $\mathbf{U}^{(p)}$ and scale matrix $\mathbf{S}^{(p)}$ and a common sample matrix $\mathbf{V}$. We allowed all matrices to share the same sample matrix $\mathbf{V}$ for finding common factors in MG data. There is only a part of samples labeled (subtype or subpopulation or subgroup is known as prior information); we incorporate this prior information with graph Laplacian. We can also learn the weights of different input data to conclude the roles that different data matrices play in CBP-JMF.

Optimization and Update Rules of CBP-JMF

To solve the problem of factorization $\mathbf{X} \approx \mathbf{USV}$, we firstly randomly initialize the solution of $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{V}$ and then apply iterative multiplicative updates as the optimization

Algorithm 1 | The CBP-JMF algorithm.

Input:

P data matrices $X^{(1)}, X^{(2)}, ..., X^{(P)}$, parameters $\beta \quad \omega$

Output:

P basis matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}$, P relation matrices $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(P)}$, factor matrices $\mathbf{V}$, weight vector $\Pi = \left(\pi^{(1)}, \pi^{(2)}, ..., \pi^{(P)}\right)$

1: **Begin**
2: Initialize$\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}$, $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(P)}$, $\mathbf{V}$
3: Initialize $\left(\pi^{(1)}, \pi^{(2)}, ..., \pi^{(P)}\right) = \left(\frac{1}{P}, \frac{1}{P}, ..., \frac{1}{P}\right)$
4: **loop**
5: **for** p=1 to P **do**
6: Fix $\mathbf{V}$, update $\mathbf{U^{(p)}}$, $\mathbf{S^{(p)}}$
7: **end for**
8: Fix $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}$, update $\mathbf{V}^{L}$
9: Fix $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(P)}$, update $\mathbf{V}^{UL}$
10: **for** p=1 to P **do**
11: Fix $\mathbf{U}, \mathbf{S}, \mathbf{V}$, compute $c^{(p)} = \left\|\mathbf{X}^{(p)} - \mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}\right\|_F^2$
12: **end for**
13: **Update** Π
14: **break** loop if convergence
15: **End**

approach similar to EM algorithms (Dempster et al., 1977). The optimization procedure of CBP-JMF is as follows.

To clarify the update rules of the objective function of CBP-JMF, we define $O(\mathbf{U}, \mathbf{V}, \mathbf{S}, \Pi) = \sum_{p=1}^{P} \pi^{(p)} \left\|\mathbf{X}^{(p)} - \mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}\right\|_F^2 + \beta\left\{tr\left[\mathbf{V}^L\mathbf{L}^a\left(\mathbf{V}^L\right)^T\right] - tr\left[\mathbf{V}^L\mathbf{L}^p\left(\mathbf{V}^L\right)^T\right]\right\} + \omega\|\Pi\|^2$. Firstly, we fix $\mathbf{V}$ and $\mathbf{S}$ and update $\mathbf{U}$; then, we can get the Lagrange function and let Ψ be the Lagrange multiplier for the constraints $\mathbf{U}_{ij}^{(p)} > 0$.

$$L\left(\mathbf{U}^{(P)}\right) = \mathrm{O}\left(\mathbf{U}^{(P)}\right) + tr\left(\Psi^T\mathbf{U}^{(P)}\right) \tag{6}$$

The partial derivatives of $L\left(\mathbf{U}^{(P)}\right)$ with $\mathbf{U}$ is:

$$\frac{\partial L\left(\mathbf{U}^{(P)}\right)}{\partial \mathbf{U}^{(P)}} = -2\mathbf{X}^{(p)}\mathbf{V}^T\left(\mathbf{S}^{(p)}\right)^T + 2\mathbf{U}^{(P)}\mathbf{S}^{(p)}\mathbf{V}\mathbf{V}^T\left(\mathbf{S}^{(p)}\right)^T + \Psi \tag{7}$$

Based on the KKT conditions $\Psi_{ij}\mathbf{U}_{ij} = 0$, we can get the following update rules:

$$\mathbf{U}^{(P)} \leftarrow \mathbf{U}^{(P)} \circ \frac{\mathbf{X}^{(P)}\mathbf{V}^T\left(\mathbf{S}^{(p)}\right)^T}{\mathbf{U}^{(P)}\mathbf{S}^{(p)}\mathbf{V}\mathbf{V}^T\left(\mathbf{S}^{(p)}\right)^T} \tag{8}$$

Similarly, we can get the update rules for $\mathbf{W}$, $\mathbf{V}^L$, and $\mathbf{V}^{UL}$:

$$\mathbf{S}^{(P)} \leftarrow \mathbf{S}^{(P)} \circ \frac{\left(\mathbf{U}^{(p)}\right)^T\mathbf{X}^{(p)}\mathbf{V}^T}{\left(\mathbf{U}^{(p)}\right)^T\mathbf{U}^{(P)}\mathbf{S}^{(p)}\mathbf{V}\mathbf{V}^T} \tag{9}$$

$$\mathbf{V}^L \leftarrow \mathbf{V}^L \circ \frac{\sum_{p=1}^{P} \pi^{(p)}\left(\left(\mathbf{S}^{(p)}\right)^T\left(\mathbf{U}^{(p)}\right)^T\mathbf{X}^{L(p)}\right) + \beta\mathbf{V}^L\left(\mathbf{D}^p + \mathbf{S}^a\right)}{\sum_{p=1}^{P} \pi^{(p)}\left(\mathbf{S}^{(p)}\right)^T\left(\mathbf{U}^{(p)}\right)^T\mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}^L + \beta\mathbf{V}^L\left(\mathbf{D}^a + \mathbf{S}^p\right)} \tag{10}$$

$$\mathbf{V}^{UL} \leftarrow \mathbf{V}^{UL} \circ \frac{\sum_{p=1}^{P} \pi^{(p)}\left(\left(\mathbf{S}^{(p)}\right)^T\left(\mathbf{U}^{(p)}\right)^T\mathbf{X}^{UL(p)}\right)}{\sum_{p=1}^{P} \pi^{(p)}\left(\mathbf{S}^{(p)}\right)^T\left(\mathbf{U}^{(p)}\right)^T\mathbf{U}^{(p)}\mathbf{S}^{(p)}\mathbf{V}^{UL}} \tag{11}$$

As for updating of π, when $\mathbf{U}$,$\mathbf{V}$, and $\mathbf{S}$ are fixed, minimization of $O(\pi)$ is a convex optimization, and we use convex optimization toolbox to update π.

CBPs Obtained From CBP-JMF

Values in each column of $\mathbf{U}^{(p)}$ represent the relative contribution of each molecule in each module, and values in each row of $\mathbf{V}$ represent the degree of each sample involved in each module. According to the rules of matrix multiplication, the i-th column of basis matrix $\mathbf{U}^{(p)}, p = 1, 2, ..., P$ corresponds to the i-th row of coefficient matrix $\mathbf{V}$, so there is a one-to-one correspondence between subtype and multi-omics module discovered from the columns of $\mathbf{U}^{(p)}$ matrix. Firstly, we need to know the relationship between k modules and subtypes by counting each subtype's value in each module from $\mathbf{V}^{(p)}$ matrix (see **Supplementary Note 3** in **Supplementary Material**).

To select features associated with each module, CBP-JMF calculates the z-scores of each molecule for each column vector of $\mathbf{U}^{(p)}$ as $z = (x - \bar{x})/S_x$, where $\bar{x} = \frac{1}{n}\sum_i x_i$, $S_x^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$. Let $\mathbf{u}_j^{(p)}$ be the j-th column of $\mathbf{U}^{(p)}$ and infer a latent feature associated with j-th CBP. The continuous value $\mathbf{u}_{ij}^{(p)}$ represents the relative contribution of molecule i in the j-th CBP. $\mathbf{u}_{ij}^{(p)}$ can be regarded as x_i, and the length of $\mathbf{u}_j^{(p)}$ can be regarded as n in Equation (12). CBP-JMF calculates a z-score for each value in $\mathbf{u}_j^{(p)}$ and obtains CBP's members through a given cutoff (z-score >2 in our tests). Then, they are mapped to a built-in molecule interaction network (see "Section 'Results'") to extract their connected components as the final CBP.

RESULTS

We applied CBP-JMF to BRCA with multi-omics data. The reason we chose BRCA as example is that breast cancer is a heterogeneous complex disease, and it is the most commonly occurring cancer. BRCA is also a type of cancer that can be divided into smaller groups based on certain characteristics of the cancer cells. Distinct complex biological processes represent different subtypes. Characterizing the processes can provide us comprehensive insights into the mechanisms of how multiple

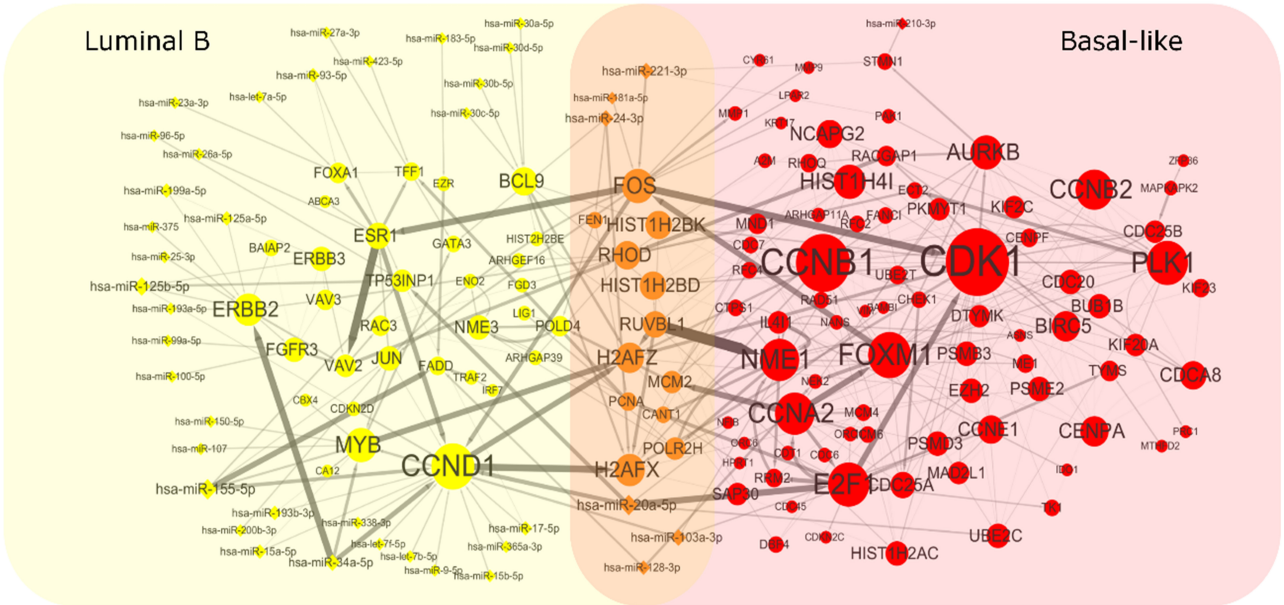

FIGURE 2 | Complex biological processes of luminal B and basal-like subtype. We mapped the genes and miRNAs obtained from luminal B's module and basal-like's module to an integrated gene regulation network. The network was obtained through integrating three databases including Reactom, Kyoto Encyclopedia of Genes and Genomes, and Nci-PID Pathway Interaction Database. The interactions between genes and miRNAs were obtained from miRTarBase. The size of the node is proportional to the size of the degree. The thickness of the edges indicates the strength of the regulatory relationship expressed by the Pearson correlation coefficient between microRNA and gene.

TABLE 1 | Enrichment analysis of the extracted module gene across six datasets.

Dataset	Online mendelian inheritance in man	CGC	Virhostome	Kinome	Drug target	BRCA pathway
Total	51	43	947	516	61	102
Overlapped nodes	2	5	13	6	3	6
P-value	0.049	0.0003	0.007	0.008	0.010	0.012

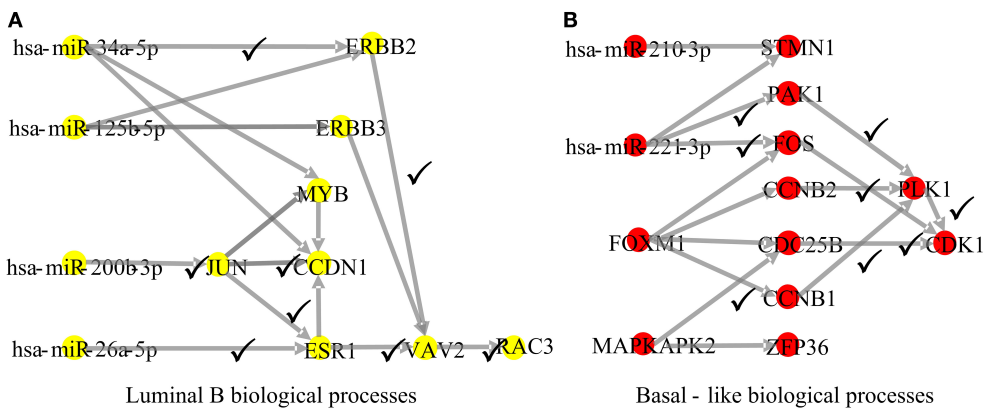

FIGURE 3 | Part of complex biological processes luminal B and basal-like. The edges with checkmarks are the interactions that have been documented. **(A)** Luminal B's biological processes: luminal subtypes are driven by the estrogen/ER pathway. Among all nodes, ERBB2, ERBB3, and ESR1 are involved in the estrogen/ER pathway. **(B)** Basal-like's biological processes: basal-like subtype is driven by the deregulation of various signaling pathways (Notch, MAPK, FoxO signaling pathway, and Wnt/beta-catenin). Among all nodes, MAPKAPK2, CDC25B, CCNB1, CCNB2, PAK1, and STMN1 are known to exist in multiple signaling pathways.

TABLE 2 | Evidences of luminal B's complex biological processes.

Interactions	Literatures	Descriptions
miR-34a->ERBB2	Wang et al., 2017	MiR-34a modulates ErbB2 in breast cancer
ERBB2->VAV2	Wang et al., 2006	ErbB2 colocalizes with Vav2 *via* activation of PI3K
VAV2->RAC3	Rosenberg et al., 2017	Vav2 promotes Rac3 activation at invadopodia
miR-200b->JUN	Jin et al., 2017	MiR-200b upregulates JUN in breast cancer
JUN->CCND1	Cicatiello et al., 2004	CCND1 promoter activation by estrogens in human breast cancer cells is mediated by the recruitment of a c-Jun/c-Fos/estrogen receptor
JUN->ESR1	Stossi et al., 2012	The activation of ESR1 gene locus in a process that was dependent upon activation and recruitment of the c-Jun transcription factor
miR-26a->ESR1	Howard and Yang, 2018	MiR-26a modulates ESR1 in breast cancer
ESR1->VAV2	Grassilli et al., 2014	ESR1 upregulates VAV2 in breast cancer cell lines

TABLE 3 | Evidences of basal-like's complex biological processes.

Interactions	Literatures	Descriptions
CCNB1(CCNB2)->PLK1->CDK1	Li et al., 2019	CCNB1 (CCNB2), PLK1, and CDK1 have interactions in chicken breast muscle
miR221->FOS	Yao et al., 2016	miR221 modulates FOS
miR221->PAK1	Ergun et al., 2015	miR221 modulates PAK1 in breast cancer cell lines
PAK1->PLK1	Maroto et al., 2008	PAK1 regulates PLK1
MAPKAPK2->CDC25B	MAPK signaling pathway	MAPKAPK2 and CDC25B are involved in MAPK signaling pathway
CDC25B->CDK1	Timofeev et al., 2010	Timely assembly of CDK1 required CDC25B

levels of molecules interact with each other and the heterogeneity of breast cancers.

Data

Firstly, we downloaded the Gene Expression (GE) data, miRNA expression (ME) data, and copy number variation (CNV) data across the same set of 738 breast cancer samples from UCSC Xena (Goldman et al., 2018). Secondly, we obtained the sample label information which is classified by PAM50 from The Cancer Genome Atlas Network (Koboldt et al., 2012). Among 738 samples, there are 522 breast cancer samples with labels, including 231 luminal A, 127 luminal B, 98 triple negative/basal-like, 58 HER2-enriched, and eight normal-like. Thirdly, we filtered out some samples, in which more than 90% of the genes have an expression value of zero. For genes and miRNAs, we filtered the genes and miRNAs with an expression value of zero in more than 20% of the samples. Fourthly, we did differential expression analysis for genes using edgeR package (Robinson et al., 2009) in R with P-value < 0.01 and $|\log(\text{fold change})| > 0.5$ to filter out genes which are not associated with breast cancer. Fifthly, we imputed missing miRNA data using knnimpute package in MATLAB. About the CNV data, the GISTIC2 (Mermel et al., 2011) thresholded the estimated values of CNV to -2, -1, 0, 1, and 2, which represent homozygous deletion, single copy deletion, diploid normal copy, low-level copy number amplification, or high-level number amplification. Finally, we obtained the GE data $\mathbf{X}^{(1)} \in \mathbf{R}^{2913\times 725}$ and ME data $\mathbf{X}^{(2)} \in \mathbf{R}^{516\times 725}$. Among 725 samples, 179 samples are marked with subtype labels (80 luminal A, 38 luminal B, 39 basal-like, 22 HER2-enriched) and shared between GE, ME, and CNV datasets. Furthermore, we calculated the Pearson correlation of 179 labeled samples using CNV data to construct $\mathbf{W}^a \in \mathbf{R}^{179\times 179}, \mathbf{W}^p \in \mathbf{R}^{179\times 179}$, and their Laplacian matrices to form the graph Laplacian regularization $tr\left[\mathbf{V}^L\mathbf{L}^a\left(\mathbf{V}^L\right)^T\right] - tr\left[\mathbf{V}^L\mathbf{L}^p\left(\mathbf{V}^L\right)^T\right]$.

Complex Biological Processes for Breast Cancer Subtypes

In our example, we set parameters $k = 4$, β=10, and ω=100, 000. Other parameters and more details can be found in **Supplementary Note 2** of **Supplementary Material**. As a result, we obtained unique matrices $\mathbf{U}^{(1)} \in \mathbf{R}^{2913\times 4}$, $\mathbf{U}^{(2)} \in \mathbf{R}^{516\times 4}$, $\mathbf{S}^{(1)} \in \mathbf{R}^{4\times 4}$, and $\mathbf{S}^{(2)} \in \mathbf{R}^{4\times 4}$ and a common matrix $\mathbf{V} \in \mathbf{R}^{4\times 725}$.

To get heterogeneous CBPs (**Supplementary Table 1**), directed regulatory pathways containing miRNAs and genes, which correspond to each cancer subtype we put subtype-specific multi-omics modules obtained from matrix $\mathbf{U}^{(p)}, p = 1, 2$ onto an integrated gene regulation network from Reactome (Croft et al., 2014), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), and Nci-PID pathway (Schaefer et al., 2009). Then, we add directed regulatory edges from miRNA to the gene supported by miRTarBase (Chou et al., 2018). Finally, we extracted the maximum connected component of the regulation network and showed the discovered characteristic CBPs underlying luminal B and basal-like subtypes in **Figure 2**.

To explore whether the genes in the CBPs of luminal B and basal-like subtype have significant biological importance or

not, we performed an enrichment analysis with all 124 genes from **Figure 2** across six datasets. The datasets are from OMIM (Hamosh et al., 2005), CGC (Futreal et al., 2004), virhostome, kinome (Manning et al., 2002), drug target (Wishart et al., 2008), KEGG pathway of BRCA (Kanehisa and Goto, 2000). Genes associated with breast cancer or breast tissue in the six datasets are selected as the set of enrichment analysis. Genes extracted through CBP-JMF have significant overlapping with known datasets (**Table 1**). Furthermore, for each subtype's CBP, functional enrichment analysis (**Supplementary Figure 4**) shows that four CBPs are mainly enriched in known biological processes and pathways associated with breast cancer, such as cell cycle and various signaling pathways (including p53 signaling pathway and estrogen pathway). However, each CBP also has its specific biological processes and path. This may explain differences between subtypes. As a demonstration, we take the CBPs of luminal B and basal-like as example. Based on the study of the subtypes of BRCA, luminal B is mainly driven by the estrogen/ER pathway (Zhang et al., 2014). In our discovered CBPs, we found several CBPs containing genes like ERBB2, ERBB3, and ESR1 that are related to the estrogen/ER pathway. Besides that, through literature review, miRNAs in luminal B's CBP can regulate the estrogen/ER pathway, such as miR-34a, miR-125b, miR-200b, and so on (**Figure 3**, **Table 2**). In addition, basal-like subtype is mainly driven by the deregulation of various signaling pathways including Notch, MAPK, and wnt/β-catenin signaling pathway (King et al., 2012). In our discovered CBPs, we found genes involved in the above-mentioned pathways, such as MAPKAPK2, CDC25B, PLK1, and so on. Besides that, we also found that miRNAs in CBPs of basal-like, such as miR-221 and miR-210, may regulate the genes above in basal-like subtype (**Figure 3**, **Table 3**). In summary, subtype-specific biological processes can be identified by CBP-JMF, and CBP-JMF can help users discover potential biological targets.

Meanwhile, to classify unlabeled samples into subtypes, CBP-JMF returned predicted labels for unlabeled samples (**Supplementary Note 4** in **Supplementary Material**). **Figure 4** shows the Kaplan–Meier (KM) survival analysis using survival package (Therneau, 2015) on unlabeled samples based on their clinical data in TCGA. We compared our results with other NMF methods (**Supplementary Note 4** of **Supplementary Material**) and found that CBP-JMF achieves more accurate subtype classification results. Unlabeled samples are classified by using GE data and ME data. **Figure 4** indicates that the survival analysis for unlabeled samples has the most significant Cox (Lin and Zelterman, 2002) p-value 0.031 and similar survival curves like the labeled samples. This proves that the CBP-JMF framework is useful for cancer subtyping, as the framework incorporates integration of multi-omics data and samples' prior information.

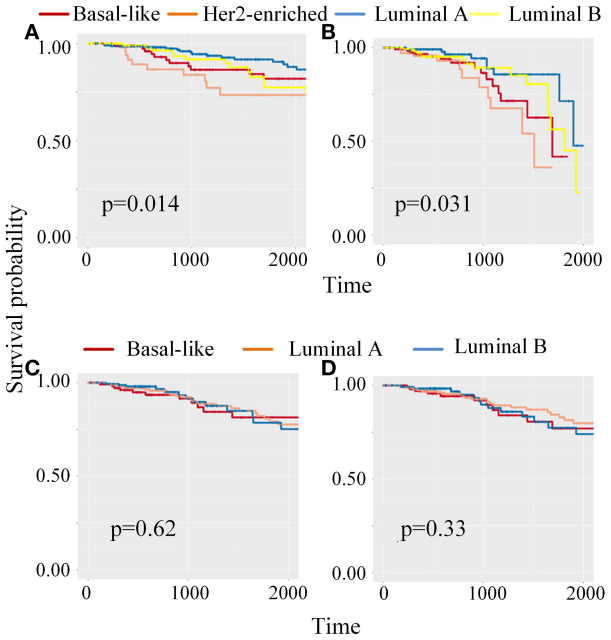

FIGURE 4 | Kaplan–Meier (K–M) survival analysis for patients which are classified using different methods. **(A)** KM survival curve for labeled samples, whose subtypes are known in advance. **(B)** KM survival analysis for unlabeled samples, which are classified using complex biological processes–joint matrix tri-factorization (CBP-JMF) on mRNA expression and miRNA expression data. **(C)** KM survival analysis for unlabeled patients, which are classified using CBP-JMF only on mRNA expression. **(D)** KM survival analysis for unlabeled patients, which are classified on mRNA expression and miRNA expression data without graph embedding regularization.

DISCUSSION

Understanding CBPs is vital to help us further understand the development of disease and intervene in the disease. NMF is an effective tool for dimension reduction and data mining in high-throughput genomic data. In this paper, we proposed CBP-JMF, an improved method of multi-view data analysis. It is designed for heterogeneous biological data based on NMF. Moreover, we created an easy-to-use package in Python. CBP-JMF analyzes multi-dimensional genomic data across the same samples integrally. Our method can discover CBPs that underlie sample groups and classify unlabeled samples through learning the relationship between labeled samples.

We tested this framework on the gene expression data and miRNA expression data of BRCA. CBP-JMF discovered subtype-specific biological processes and classified unlabeled samples into four subtypes. We did survival analysis and function analysis, and the results showed that CBP-JMF has great performance. Furthermore, CBP-JMF is a weighted joint tri-NMF framework in essence. We expect that it can be applied to vast fields including disease subtypes, cell types, and population stratification. Meanwhile, we expect that CBP-JMF can be used to identify hub genes or predict the association between genes or non-coding mRNA and diseases by integrating a variety of data. Though CBP-JMF is efficient to uncover CBPs by integrating multi-omics data, CBP-JMF must integrate different multi-omics data that have the same samples. This weakness limits the use of more types

of data and integrates more information to obtain more significant results.

CONCLUSIONS

In this article, we develop CBP-JMF, a matrix tri-factorization and weighted joint integration tool, for detecting CBPs, which characterize prior disease subtypes and cell groups in Python. We improve its usability by estimating the parameters, such as determining the number of features through consensus clustering. CBP-JMF always gives reference values of all parameters. In applications, CBP-JMF characterizes the CBPs of four subtypes of BRCA based on gene and miRNA expression data from TCGA, and we find the significantly different functional pathways that characterized luminal B and basal-like subtypes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study are publicly available and the addresses for finding them are listed within the article. Prediction results and a reference implementation of CBP-JMF in Python are available at: https://github.com/wangbingbo2019/CBP-JMF.

AUTHOR CONTRIBUTIONS

BW, YWu, and XM conceived and designed the experiments. YWu and MX performed the experiments. XM, RD, CZ, LY, XG, and LG analyzed the data. BW, YWu, XM, and YWa proofread the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61772395 and 61873198, the Fundamental Research Funds for the Central Universities (Nos. JB190306 and ZD2009), and the Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01). We thank LCNBI and ZJLab for the financial support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.665416/full#supplementary-material

REFERENCES

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., et al. (2014). Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.* 9, 1235–1245. doi: 10.1016/j.celrep.2014.10.035

Brunet, J. P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4164–4169. doi: 10.1073/pnas.0308531101

Cai, D., He, X., Han, J., and Huang, T. S. (2011). graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 1548–1560. doi: 10.1109/TPAMI.2010.231

Chen, J., and Zhang, S. (2018). Discovery of two-level modular organization from matched genomic data via joint matrix tri-factorization. *Nucleic Acids Res.* 46, 5967–5976. doi: 10.1093/nar/gky440

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

Cicatiello, L., Addeo, R., Sasso, A., Altucci, L., Petrizzi, V. B., Borgo, R., et al. (2004). Estrogens and Progesterone promote persistent CCND1 gene activation during G1 by inducing transcriptional derepression via c-Jun/c-Fos/estrogen receptor (progesterone receptor) complex assembly to a distal regulatory element and recruitment of Cyclin D1 t. *Mol. Cell. Biol.* 24, 7260–7274. doi: 10.1128/MCB.24.16.7260-7274.2004

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, 472–477. doi: 10.1093/nar/gkt1102

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Ding, C., Li, T., Peng, W., and Park, H. (2006). "Orthogonal nonnegative matrix tri-factorizations for clustering," in *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA), 126–135. doi: 10.1145/1150402.1150420

Ergun, S., Tayeb, T. S., Arslan, A., Temiz, E., Arman, K., Safdar, M., et al. (2015). The investigation of miR-221-3p and PAK1 gene expressions in breast cancer cell lines. *Gene* 555, 377–381. doi: 10.1016/j.gene.2014.11.036

Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 13, 241–244. doi: 10.1038/nmeth.3734

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi: 10.1038/nrc1299

Goldman, M., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., et al. (2018). The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv*, 1–16. doi: 10.1101/326470

Grassilli, S., Brugnoli, F., Lattanzio, R., Rossi, C., Perracchio, L., Mottolese, M., et al. (2014). High nuclear level of Vav1 is a positive prognostic factor in early invasive breast tumors: a role in modulating genes related to the efficiency of metastatic process. *Oncotarget* 5, 4320–4336. doi: 10.18632/oncotarget.2011

Guan, Z., Zhang, L., Peng, J., and Fan, J. (2015). Multi-view concept learning for data representation. *IEEE Trans. Knowl. Data Eng.* 27, 3016–3028. doi: 10.1109/TKDE.2015.2448542

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, 514–517. doi: 10.1093/nar/gki033

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18, 1–15. doi: 10.1186/s13059-017-1215-1

Howard, E. W., and Yang, X. (2018). MicroRNA regulation in estrogen receptor-positive breast cancer and endocrine therapy. *Biol. Proced. Online* 20, 1–19. doi: 10.1186/s12575-018-0082-9

Jin, T., Kim, H. S., Choi, S. K., Hwang, E. H., Woo, J., Ryu, H. S., et al. (2017). microRNA-200c/141 upregulates SerpinB2 to promote breast cancer cell metastasis and reduce patient survival. *Oncotarget* 8, 32769–32782. doi: 10.18632/oncotarget.15680

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of genes and genomes. *Oxford Univ. Press Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

King, T. D., Suto, M. J., and Li, Y. (2012). The wnt/β-catenin signaling pathway: a potential therapeutic target in the treatment of triple negative breast cancer. *J. Cell. Biochem.* 113, 13–18. doi: 10.1002/jcb.23350

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

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

Li, Y., Chen, Y., Jin, W., Fu, S., Li, D., Zhang, Y., et al. (2019). Analyses of microRNA and mRNA expression profiles reveal the crucial interaction networks and pathways for regulation of chicken breast muscle development. *Front. Genet.* 10, 1–15. doi: 10.3389/fgene.2019.00197

Lin, H., and Zelterman, D. (2002). Modeling survival data: extending the cox model. *Technometrics* 44, 85–86. doi: 10.1198/tech.2002.s656

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science 298, 1912–1934. doi: 10.1126/science.1075762

Maroto, B., Ye, M. B., Von Lohneysen, K., Schnelzer, A., and Knaus, U. G. (2008). P21-activated kinase is required for mitotic progression and regulates Plk1. *Oncogene* 27, 4900–4908. doi: 10.1038/onc.2008.131

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12, 1–14. doi: 10.1186/gb-2011-12-4-r41

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rosenberg, B. J., Gil-Henn, H., Mader, C. C., Halo, T., Yin, T., Condeelis, J., et al. (2017). Phosphorylated cortactin recruits Vav2 guanine nucleotide exchange factor to activate Rac3 and promote invadopodial function in invasive breast cancer cells. *Mol. Biol. Cell* 28, 1347–1360. doi: 10.1091/mbc.e16-12-0885

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the pathway interaction database. *Nucleic Acids Res.* 37, 674–679. doi: 10.1093/nar/gkn653

Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., et al. (2018). Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet.* 34, 790–805. doi: 10.1016/j.tig.2018.07.003

Stossi, F., Madak-Erdogan, Z., and Katzenellenbogen, B. S. (2012). Macrophage-elicited loss of estrogen receptor-α in breast cancer cells via involvement of MAPK and c-Jun at the ESR1 genomic locus. *Oncogene* 31, 1825–1834. doi: 10.1038/onc.2011.370

Suravajhala, P., Kogelman, L. J. A., and Kadarmideen, H. N. (2016). Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. *Genet. Sel. Evol.* 48, 1–14. doi: 10.1186/s12711-016-0217-x

Therneau, T. M. (2015). *A Package for Survival Analysis in S. Version 2.38*. Available online at: https://cran.r-project.org/package=survival.

Timofeev, O., Cizmecioglu, O., Settele, F., Kempf, T., and Hoffmann, I. (2010). Cdc25 phosphatases are required for timely assembly of CDK1-cyclin B at the G2/M transition. *J. Biol. Chem.* 285, 16978–16990. doi: 10.1074/jbc.M109.096552

Wang, S. E., Shin, I., Wu, F. Y., Friedman, D. B., and Arteaga, C. L. (2006). HER2/Neu (ErbB2) signaling to Rac1-Pak1 is temporally and spatially modulated by transforming growth factor β. *Cancer Res.* 66, 9591–9600. doi: 10.1158/0008-5472.CAN-06-2071

Wang, Y., Zhang, X., Chao, Z., Kung, H. F., Lin, M. C., Dress, A., et al. (2017). MiR-34a modulates ErbB2 in breast cancer. *Cell Biol. Int.* 41, 93–101. doi: 10.1002/cbin.10700

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, 901–906. doi: 10.1093/nar/gkm958

Xi, J., Li, A., and Wang, M. (2018). A novel unsupervised learning model for detecting driver genes from pan-cancer data through matrix tri-factorization framework with pairwise similarities constraints. *Neurocomputing* 296, 64–73. doi: 10.1016/j.neucom.2018.03.026

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. (2020). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 36, 1855–1863. doi: 10.1093/bioinformatics/btz793

Yang, Z., and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 32, 1–8. doi: 10.1093/bioinformatics/btv544

Yao, M., Gao, W., Yang, J., Liang, X., Luo, J., and Huang, T. (2016). The regulation roles of miR-125b, miR-221 and miR-27b in porcine Salmonella infection signalling pathway. *Biosci. Rep.* 36, 1–11. doi: 10.1042/B.S.R.20160243

Zhang, M. H., Man, H. T., Zhao, X. D, Dong, N., and Ma, S. L. (2014). Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials (review). *Biomed. Rep.* 2, 41–52. doi: 10.3892/br.2013.187

Zhang, S., Liu, C. C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* 40, 9379–9391. doi: 10.1093/nar/gks725

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/fgene.2021.684100

The Unsupervised Feature Selection Algorithms Based on Standard Deviation and Cosine Similarity for Genomic Data Analysis

Juanying Xie[1†], Mingzhao Wang[1,2†], Shengquan Xu[2*], Zhao Huang[1*] and Philip W. Grant[3]*

[1] School of Computer Science, Shaanxi Normal University, Xi'an, China, [2] College of Life Sciences, Shaanxi Normal University, Xi'an, China, [3] Department of Computer Science, Swansea University, Swansea, United Kingdom

OPEN ACCESS

Edited by:
Jianing Xi, Northwestern Polytechnical University, China

Reviewed by:
Quan Zou, University of Electronic Science and Technology of China, China
Fengfeng Zhou, Jilin University, China
Yanchun Zhang, Victoria University, Australia

***Correspondence:**
Juanying Xie
xiejuany@snnu.edu.cn
Shengquan Xu
xushengquan@snnu.edu.cn
Zhao Huang
zhaohuang@snnu.edu.cn

[†]*These authors share first authorship*

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

Received: *22 March 2021*
Accepted: *16 April 2021*
Published: *13 May 2021*

Citation:
Xie J, Wang M, Xu S, Huang Z and Grant PW (2021) The Unsupervised Feature Selection Algorithms Based on Standard Deviation and Cosine Similarity for Genomic Data Analysis. Front. Genet. 12:684100. doi: 10.3389/fgene.2021.684100

To tackle the challenges in genomic data analysis caused by their tens of thousands of dimensions while having a small number of examples and unbalanced examples between classes, the technique of unsupervised feature selection based on standard deviation and cosine similarity is proposed in this paper. We refer to this idea as SCFS (Standard deviation and Cosine similarity based Feature Selection). It defines the discernibility and independence of a feature to value its distinguishable capability between classes and its redundancy to other features, respectively. A 2-dimensional space is constructed using discernibility as x-axis and independence as y-axis to represent all features where the upper right corner features have both comparatively high discernibility and independence. The importance of a feature is defined as the product of its discernibility and its independence (i.e., the area of the rectangular enclosed by the feature's coordinate lines and axes). The upper right corner features are by far the most important, comprising the optimal feature subset. Based on different definitions of independence using cosine similarity, there are three feature selection algorithms derived from SCFS. These are SCEFS (Standard deviation and Exponent Cosine similarity based Feature Selection), SCRFS (Standard deviation and Reciprocal Cosine similarity based Feature Selection) and SCAFS (Standard deviation and Anti-Cosine similarity based Feature Selection), respectively. The KNN and SVM classifiers are built based on the optimal feature subsets detected by these feature selection algorithms, respectively. The experimental results on 18 genomic datasets of cancers demonstrate that the proposed unsupervised feature selection algorithms SCEFS, SCRFS and SCAFS can detect the stable biomarkers with strong classification capability. This shows that the idea proposed in this paper is powerful. The functional analysis of these biomarkers show that the occurrence of the cancer is closely related to the biomarker gene regulation level. This fact will benefit cancer pathology research, drug development, early diagnosis, treatment and prevention.

Keywords: unsupervised feature selection, gene selection, standard deviation, cosine similarity, 2-dimensional space

INTRODUCTION

The rapid development of high-throughput sequencing technology has produced a large amount of genomic data related to protein, gene and life metabolism. It has become a hot spot research field of life medicine to detect biomarkers and undertake related analyses using bioinformatics methods. It is known that the personal medicine program of United States of America and the precision medicine program in China were initiated in 2015 and 2016 respectively (Xie and Fan, 2017). More and more researchers have turned their attention to medical data analysis and to data-driven intelligent medical treatments using artificial intelligence techniques (Orringer et al., 2017; Esteva et al., 2017; Kim et al., 2018; Bychkov et al., 2018).

Cancers have become the main killer of humankind and there are seven persons diagnosed with cancers per minute in China in 2014 (Global Burden of Disease Cancer Collaboration, 2018; Cao and Chen, 2019). According to statistics by the IARC (International Agency for Research on Cancer) from WHO (World Health Organization) and GBD (Global Burden of Disease Cancer Collaboration), cancer cases increased by 28% between 2006 and 2016, and there will be 2.7 million new cancer cases emerging in 2030. Genomics data can reveal cancer related gene expression and regulation. There is a complex regulation network between genes. It has become popular to detect the biomarkers of cancers from the massive genomic data using the feature selection and classification techniques of machine learning (Xie and Gao, 2014; Xie et al., 2016b, 2020a,b; Esteva et al., 2017; Ye et al., 2017; Wang et al., 2017; Dong et al., 2018). The genomic data are usually of very high dimensions and small number of samples, and are always imbalanced, which lead to challenges for the available classification algorithms, especially with regard to the stability and generalization of the available algorithms (Diao and Vidyashankar, 2013). Feature selection algorithms can benefit the classification algorithms' stability and generalization by selecting the key features related to cancers and eliminating the redundant and noisy features simultaneously (Ang et al., 2016; Dashtban and Balafar, 2017; Dong et al., 2018; Xie et al., 2019, 2020a,b).

Feature selection algorithm searches feature subsets from the search space composed of all combinations of features. It is an NP hard problem to detect the optimal feature subset (Fu et al., 1970). The common way is to use heuristics to find it. The feature subset is usually highly relevant to the classification problem and can improve the classification performance of the learning algorithm. Feature selection algorithms can be classified into Filters (Blum and Langley, 1997) or Wrappers (Kohavi and John, 1997) according to whether the feature selection process depends on the later learning algorithms or not. Filters are not dependent on the later learning algorithms while Wrappers are dependent, which lead to the fast efficiency of Filters and the time consuming load of Wrappers. However, wrappers can always detect the feature subset with high performance while with small number of features, but the limitations are that the feature subset can easily fall into overfitting with poor generalization. Therefore the hybrid feature selection algorithms have been studied and become the *ad hoc* research field in recent years (Xie and Wang, 2011; Kabir et al., 2011; Xie and Gao, 2014; Lu et al., 2017; Xie et al., 2019). Furthermore, feature selection algorithms can also be classified as supervised or unsupervised algorithms according to whether the class labels of training data are used or not in the feature selection process. Wrappers are always supervised feature selection algorithms while filters may be supervised, unsupervised or semi-supervised algorithms (Ang et al., 2016). Supervised feature selection algorithms usually realize feature selection by evaluating the correlation between features and class labels, such as mRMR (Minimal redundancy-maximal relevance) proposed by Peng et al. (2005). Supervised feature selection algorithms are always superior to semi-supervised and unsupervised feature selection algorithms in selecting powerful feature subsets due to its using the labels of samples. Semi-supervised feature selection algorithms are always deal with samples some of which having labels while others not, such as LRLS (Label reconstruction based laplacian score) proposed by Wang J. et al. (2013). The situation is that there are amount of data without class labels in the world and it is time-consuming or impossible to get labels for them. Therefore it is very important to study the unsupervised feature selection algorithms. However, the unsupervised feature selection problems are particularly difficult due to the absence of class labels that would guide search for relevant information. Even though, it has attracted many researchers to focus on this field, such as the feature entropy sorting based feature selection algorithm proposed by Dash et al. (1997). It adopted entropy to evaluate the importance of features to realize the unsupervised feature selection. Furthermore, Mitra et al. (2002) proposed the unsupervised feature selection algorithm based on their defined maximum information compression index to eliminate redundant features. Xu et al. (2012) proposed UFS-MI (Unsupervised feature selection approach based on mutual information). He et al. (2006) proposed the unsupervised feature selection algorithm based on manifold learning, and the importance of a feature is evaluated by its power of locality preserving, or, Laplacian Score. Zhao et al. (Zhao and Liu, 2007) proposed SPEC (Spectral analysis based feature selection) algorithm, which studied how to select features according to the structures of the graph induced from a set of pairwise instance similarity and employed the spectrum of the graph to measure feature relevance and elaborate how to realize spectral feature selection. As a result that the features which are consistent with the graph structure would comprise the optimal feature subset. Cai et al. (2010) proposed the MCFS (Multi-Cluster Feature Selection) algorithm, which selected those features to comprise the optimal feature subset such that the multi-cluster structure of the data can be best preserved by solving a sparse eigen-problem and a L1-regularized least squares problem. Hou et al. (2011) proposed a feature selection algorithm via joint embedding learning and sparse regression, which defined the weight using the locally linear approximation to construct graph and unified embedding learning and sparse regression to perform feature selection. Yang et al. (2011) proposed UDFS (Unsupervised discriminative feature selection) algorithm, which obtained the feature subset of the strong discriminant structure by maximizing the local inter-class divergence and minimizing

local intra-class divergence simultaneously while minimizing the L2,1 norm of the coefficient matrix of the linear classifier. Li et al. (2012) proposed the NDFS (Non-negative discriminant feature selection) algorithm, which adopted spectral clustering to learn the cluster labels of the input samples while the feature selection is performed simultaneously. The joint learning of cluster labels and feature selection matrix enabled the NDFS algorithm to detect the most discriminative features. Qian et al. (Qian and Zhai, 2013) proposed an extended unsupervised feature selection algorithm named RUFS (Robust unsupervised feature selection). L2,1 norm minimization method was used in the process of label learning and feature selection to eliminate redundant and noisy features. Xie et al. (2018) proposed a distribution preserving feature selection (DPFS) method for unsupervised feature selection. Those features were selected which can preserve the distribution of the data. Liu et al. (2005) proposed a K-means based feature selection algorithm named as KFS, which performed supervised feature selection on several various clustering results of K-means to get the feature subset. Jiang et al. (2008) presented the CBFS (Clustering-based feature selection) algorithm, which defined the discriminative of each feature based on the difference between different clusters of each feature such that detecting the feature subset. Ling et al. (Ling and Ji, 2007) proposed a clustering ensemble based unsupervised feature selection algorithm by adopting a clustering algorithm to learn data labels and the ReliefF algorithm to perform feature selection. Wang et al. (Wang and Jiang, 2015) proposed unsupervised feature selection algorithm named FSFC (Feature selection method based on feature clustering), which defined the mean-similarity measure for each feature, then group all features into clusters, and select the representative feature from each cluster to comprise the feature subset. Panday et al. (2018) introduced two unsupervised feature selection algorithms by using a cluster-dependent feature-weighting mechanism to reflect the within-cluster degree of relevance of a specific feature. Features with a relatively high weight would comprise the feature subset. Xie et al. (2016a) put forward two unsupervised feature selection algorithms by defining the feature density and feature distance. The denser a feature, the more representative it is, and the more distant of a feature, the less is its redundancy. They adopted the product of the density and the distance of a feature to measure its contribution to the classification. He et al. (2017) proposed the unsupervised feature selection algorithm named DGFS (Decision graph-based feature selection). They defined the local density and the discriminant distance for a feature, and the decision score to evaluate the feature.

To summarize the aforementioned analyses we know that it is very challenging to analyze the genomic data, especially the gene expression data with tens to thousands dimensions while with very small number of samples. The worst thing is that this kind of data are always imbalanced and it is very difficult to get the class labels for the data. Therefore it is very difficult to find a stable and good generalization algorithm for analyzing this kind of genomic data.

To tackle this challenging task, this paper will focus on the feature selection problem for genomic data analysis under an unsupervised learning scenario. It will propose the unsupervised feature selection technique based on the standard deviation and the cosine similarity of variables. We refer to this as SCFS (unsupervised Feature Selection via Standard deviation and Cosine similarity scores of variables), which defines the feature discernibility and feature independence. The standard deviation of a feature is to define its discernibility while the cosine similarity is to define the independence or redundancy of a feature. Three unsupervised feature selection algorithms are derived from SCFS according to the various definitions of feature independence. These three unsupervised feature selection algorithms are SCEFS (Feature Selection via Standard deviation and Cosine similarity with Exponent), SCRFS (Feature Selection via Standard deviation and Cosine similarity with Reciprocal), and SCAFS (Feature Selection via Standard deviation and Anti-Cosine similarity), respectively.

To detect the features with both high discernibility and high independence from the original features easily, we display all features in the two dimensional space with discernibility as x-coordinate and independence as y-coordinate, such that these features centralize in the upper right corner while others in the bottom left corner. These upper right corner features comprise the optimal feature subset. The feature contribution to classification is quantified by the area of the rectangle enclosed by the feature coordinate lines and the coordinate axes, and called the feature score in this paper. Compared to other unsupervised feature selection algorithms, our proposed three unsupervised feature selection algorithms are simple in principles, and with low computational load, and the detected feature subset is sparse while representative.

We test these three unsupervised feature selection algorithms on 18 cancer genomic datasets. The proposed SCEFS, SCRFS and SCAFS can accurately detect the key biomarkers causing cancer diseases. These biomarkers are usually with rich classification information and strong stability. This study provides a base and clue for pathological research, drug development, early diagnosis, treatment and prevention of cancers.

SCFS ALGORITHMS

This section will introduce the proposed unsupervised feature selection algorithms in detail.

Feature Discernibility

Given training dataset $D \in R^{m\times d}$, where m and d are the number of samples and the dimension of the data respectively. The features are represented as $f_1, f_2, \cdots, f_i, \cdots, f_d$, then $D = \left[f_1, f_2, \cdots, f_i, \cdots, f_d\right], f_i \in R^m,\ i = 1, \cdots, d$. The samples are $x_1, x_2, \cdots, x_j, \cdots, x_m$, and $D = \left[x_1; x_2; \cdots; x_j; \cdots; x_m\right]$, $x_j \in R^d, j = 1, \cdots, m$.

Definition 1

Feature discernibility: The discernibility of feature f_i, refers to its distinguishable capability between categories and is denoted by dis_i. The standard deviation of a variable embodies its differences on all samples so the larger the standard deviation, the more differences the variable value has on all samples.

Therefore the standard deviation of a feature can represent its distinguishable capability between categories. The discernibility dis_i of feature f_i is calculated in (1). The larger dis_i, the more distinguishable capability the feature has, so contributes more to the classification.

$$dis_i = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(f_{ji} - \frac{1}{m}\sum_{j=1}^{m} f_{ji}\right)^2} \quad i = 1, 2, \cdots, d; \quad j = 1, 2, \cdots, m \tag{1}$$

where, f_{ji} means the value of sample j on its feature i.

Feature Independence

Feature selection aims to detect the features whose distinguishable capability is strong while the redundancy between them is less. We propose the feature independence definition to measure the redundancy between features. The independence of feature f_i is represented as ind_i, which can be defined using the cosine similarities between features. To represent the redundancy between feature f_i and the other features, we define the cosine similarity matrix $\boldsymbol{C}$ in (2), which quantifies the similarity between feature f_i and other features. We define three types of feature independence in the following definitions (3) - (5).

$$C = (c_{ij})_{d\times d}, \quad i, j = 1, \cdots, d$$
$$c_{ij} = \frac{|f_i \bullet f_j|}{\| f_i \| \times \| f_j \|}. \tag{2}$$

Definition 2

Exponential feature independence: This type of feature independence is defined in (3).

$$ind_i = \begin{cases} \exp\left(\max\limits_{k=1}^{d}(-c_{ik})\right), & i = \arg\max \{dis_j \,|\, j = 1, \cdots, d\}; \\ \exp\left(\min\limits_{k:dis_k > dis_i}(-c_{ik})\right), & \text{otherwise.} \end{cases} \tag{3}$$

Definition 3

Reciprocal feature independence: This type of feature independence is calculated in (4).

$$ind_i = \begin{cases} \max\limits_{k=1}^{d}\left(\frac{1}{c_{ik}}\right), & i = \arg\max \{dis_j \,|\, j = 1, \cdots, d\}; \\ \min\limits_{k:dis_k > dis_i}\left(\frac{1}{c_{ik}}\right), & \text{otherwise.} \end{cases} \tag{4}$$

Definition 4

Anti-similarity feature independence: This kind of feature independence is calculated in (5).

$$ind_i = \begin{cases} \max\limits_{k=1}^{d}(1 - c_{ik}), & i = \arg\max \{dis_j \,|\, j = 1, \cdots, d\}; \\ \min\limits_{k:dis_k > dis_i}(1 - c_{ik}), & \text{otherwise.} \end{cases} \tag{5}$$

The definitions (3)-(5) guarantee that the feature f_i will have the maximal independence as far as possible once it has the maximal discernibility. Otherwise, its independence is quantified using the maximal cosine similarity between it and feature f_k whose discernibility is just higher, such that the independence embodies as low a redundancy as far as possible.

Feature Score

The expected feature subset is the one whose features are strongly related to labels while the redundancy between features is very low (Peng et al., 2005; Ding and Peng, 2005). The discernibility definition (1) in section "Feature Discernibility" shows that the feature with strong distinguishable capability has a large discernibility. The independence definitions in section "Feature Independence" show that a feature with low redundancy has high independence. Therefore the optimal feature subset comprises the features with both high discernibility and high independence. To detect these features with both high discernibility and high independence, we display all features in the 2-dimensional space with discernibility as x-coordinate and independence as y-coordinate such that the upper right corner features are those with both relatively high discernibility and independence. These features comprise the optimal feature subset.

To quantify the contribution of a feature to classification, we introduce the feature score in (6) to measure the significance of the feature. The feature score is defined as the area of the rectangle enclosed by the feature coordinate lines and coordinate axes. From the aforementioned definitions, we know that the features with higher scores have strong discernibility and low redundancy. These features comprise the feature subset, which coincides with the original destination (Fu et al., 1970; Ding and Peng, 2005; Peng et al., 2005) of feature selection.

Definition 5

Feature score: Feature score of f_i is defined as

$$score_i = dis_i \times ind_i \tag{6}$$

Definition (6) guarantees that feature f_i will have a high score when its discernibility and independence are both high implying the feature will benefit classification. Therefore selecting the features with high score as the feature subset satisfies the requirements of the optimal feature subset while guaranteeing the selected features' discernibility is strong and the redundancy is low.

Detailed Steps of SCFS

From the definitions of feature discernibility, feature independence, and feature score, we can display all features

in 2-dimensional space, and select the upper right corner features to comprise the feature subset. Because these upper right corner features are far away from the other features, the feature selection process can be achieved automatically. In addition, three types of independences are used to develop three unsupervised feature selection algorithms named SCEFS, SCRFS, and SCAFS respectively. The pseudo code of our unsupervised feature selection algorithms SCEFS, SCRFS, SCAFS are presented below:

Input

Training data $D \in R^{m\times d}$, where m and d represent the number of samples and features respectively; number of selected features k and the original feature set $\boldsymbol{F}$.

Output

The selected feature subset $\boldsymbol{S}$.

BEGIN

S$\leftarrow \Phi$;

FOR i = 1 to d DO

Calculate the feature discernibility dis_i of f_i using formula (1);

END of FOR

FOR each $f_i \in F$ DO

Calculate the feature independence ind_i of f_i using formula (3), (4) or (5);

Calculate the feature score $score_i$ using formula (6);

END of FOR

Sort features in descending order according to their scores;

Select top k features to comprise the feature subset $\boldsymbol{S}$.

END

A Toy Case Study

In this subsection we will test the correctness of our proposed feature score, arbitrarily choosing SCEFS for illustration. We synthetically generate toy test data using two groups of mean and covariance matrices resulting in two categories of data with normal distributions. There are 20 samples in each category and each sample embodies 100 features.

We adopt a bootstrap approach (Effron and Tibshirani, 1993; Kohavi, 1995) to partition the toy data into training and test subsets so that there are 28 (13 + 15) training samples and 12 (7 + 5) test samples. The feature discernibility, independence and score are calculated by using (1) and (3) and (6) respectively for the training data. All features are represented in 2-dimensional space with discernibility as the x axis and independence as y axis as shown in **Figure 1A**. In **Figure 1B** we display all features in descending order by their scores where the x axis is the number of features and the feature score is the y axis. The circled numbers in **Figure 1** represents the feature ID in the toy data.

The results in **Figure 1** show us that the features with IDs 24, 86, 99, 65, and 4 are the upper right corner features as their feature scores are higher than all others and is the feature subset we are trying to detect. Although features 37 and 42 have comparatively high independence, they do not have comparatively high discernibility; similarly with features 91 and 85, they have sufficiently high discernibility but comparatively low independence, so these four features are not selected for inclusion into the feature subset. The detected features are far away from other features because of their comparatively high scores, which is very clear from **Figure 1B**.

We test the classification capability of the detected features by building SVM classifiers using the SVM tool box LibSVM developed by Professor Lin et al. (Chang and Lin, 2011). The kernel function is a linear function, and the parameters are default except for the penalty factor C = 20. The results of the SVM classifiers achieved 100% accuracy when all the detected features 24, 86, 99, 65, and 4 are in the feature subset, while only 73.15% accuracy with only the top feature 24 in the feature subset, and 95.91% accuracy with the top 3 features 24,86, and 99 included.

Therefore the proposed SCFC method is valid in detecting the sparse and powerful feature subset whose features have comparatively high distinguishable capability and independence between each other so that a powerful classifier can be built using the feature subset.

Complexity Analysis

Assume that there are m samples with d dimensions where it is usual that $d > m$, even $d \gg m$ always holds. The three proposed unsupervised feature selection algorithms SCEFS, SCRFS and SCAFS are all required to calculate the discernibility and independence for each feature. The time complexity of calculating discernibility is $O(dm)$, and for independence is $O(d^2)$, and the time complexity to sort the feature scores is no more than $O(d^2)$. So, from the pseudo code in section "Detailed Steps of SCFS," the total time complexity of all selection algorithms is $O(d^2)$. This is also the time complexity upper bound. The real consuming time may lower than this theoretical analysis by using matrix calculations embedded in MATLAB.

EXPERIMENTS AND ANALYSES

As is well known genomic data analysis is very challenging in bioinformatics, especially gene expression data because this always has tens to thousands of dimensions while having very few samples and the data are always imbalanced. It is very difficult to find stable algorithms with good generalization for analyzing this kind of data.

This subsection will test the power of the unsupervised feature selection algorithms SCEFS, SCRFS, and SCAFS using high dimensional gene expression datasets of cancers. The detailed information of these data sets are shown in **Table 1**. The data sets of Gastric1 (accession: GSE29272), Gastric (accession: GSE37023), Non-small lung cancer (accession: GDS3627) and Prostate2 (accession: GDS2545) are from NCBI Gene Expression Omnibus (GEO) database[1]. The others are from Broad Institute Genome Data Analysis Center[2] and Gene Expression Model Selector[3].

[1]https://www.ncbi.nlm.nih.gov/geo/

[2]http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi

[3]http://www.gems-system.org/

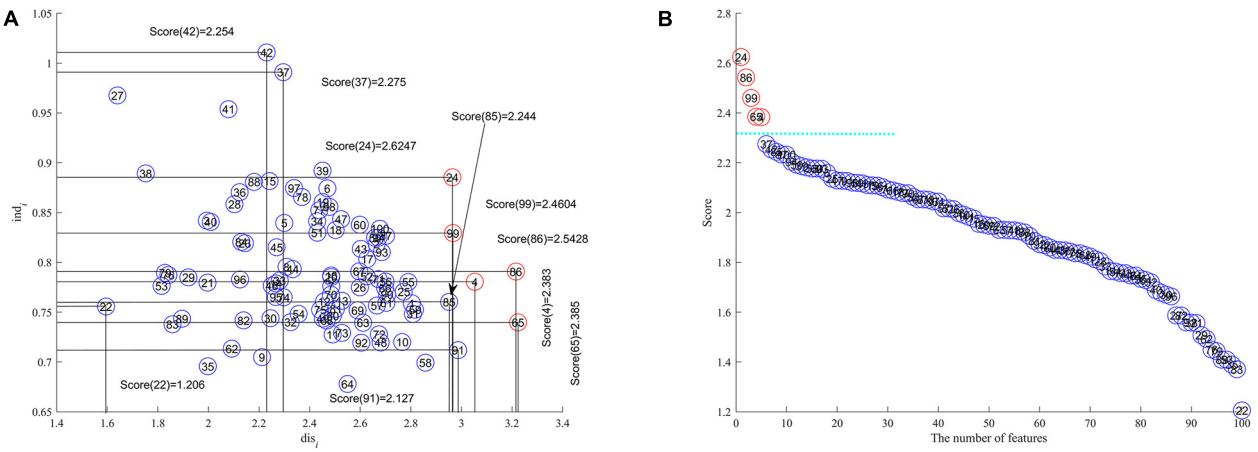

FIGURE 1 | The toy case to test SCFS, **(A)** discernibility and independence are *x*-coordinate and *y*-coordinate respectively, **(B)** the number of features is the *x* axis and feature score the *y* axis respectively.

Experiment Design and Evaluation Metrics

To test the power of our proposed SCEFS, SCRFS and SCAFS in detecting the optimal feature subsets for genomic data, we use them to find the feature subset of the 18 gene expression datasets shown in **Table 1**. Furthermore, we conduct comprehensive comparisons between their performances to that of other unsupervised feature selection algorithms, including EDPFS (unsupervised Feature Selection algorithm based on Exponential Density Peaks) (Xie et al., 2016a), RDPFS (unsupervised Feature Selection algorithm based on the Reciprocal Density Peaks) (Xie et al., 2016a), MCFS (Multi-Cluster Feature Selection) (Cai et al., 2010), Laplacian (Laplacian score for feature selection) (He et al., 2006), UDFS (Unsupervised Discriminative Feature Selection) (Yang et al., 2011), RUFS (Robust Unsupervised Feature Selection) (Qian and Zhai, 2013), NDFS (Non-negative Discriminant Feature Selection) (Li et al., 2012), and DGFS (Decision Graph-based Feature Selection) (He et al., 2017).

TABLE 1 | Descriptions of datasets.

Dataset name	Ng	Ns	Nc	Source
Colon	2000	62	2	Alon et al. (1999)
Leukemia	7129	72	2	Golub et al. (1999)
CNS	7129	90	2	Pomeroy et al. (2002)
CNS2	7129	60	2	Pomeroy et al. (2002)
DLBCL	7129	77	2	Shipp et al. (2002)
Lymphoma	4026	45	2	Alizadeh et al. (2000)
Carcinoma	7457	36	2	Notterman et al. (2001)
SRBCT	2308	83	4	Khan et al. (2001)
ALL1	12625	128	2	Chiaretti et al. (2004)
ALL4	12625	93	2	Chiaretti et al. (2004)
Lung cancer	12600	203	5	Bhattacharjee et al. (2001)
Prostate1	12625	102	2	Singh et al. (2002)
Prostate2	12558	108	3	Chandran et al. (2007)
11_Tumors	12533	174	11	Su et al. (2001)
Leukemia_MLL	12582	72	3	Armstrong et al. (2002)
Gastric	22645	65	2	Wu et al. (2012)
Gastric1	22283	144	2	Wang G. et al. (2013)
Non-small lung cancer	54675	58	2	Kuner et al. (2009)

Note: Ng, Ns and Nc represent the number of features, instances and classes of dataset respectively.

The compared algorithms EDPFS and RDPFS are our previously proposed unsupervised feature selection algorithms, which set the neighbors to be 2% when calculating the density of a feature. The algorithm DGFS set the cutoff distance d_c to the value of 2% of the total number of features, and sorted the feature distances in ascending order using Euclidean distance. The nearest neighbor number K of the compared algorithms MCFS, Laplacian, UDFS, RUFS and NDFS is set to 5. The similarity between features in Laplacian, RUFS and NDFS algorithms are cosine similarity, and the regularization parameter of UDFS and NDFS algorithms are set to 0.1

If there are missing values in the datasets, they are set to the intra-class mean. To avoid the impact from different scales of different features on experimental results due to the large differences among features of genomic data, the maximum and minimum standardization in (7) is used to normalize the data.

$$f'_{i,j} = \frac{f_{i,j} - \min\left(f_{\bullet j}\right)}{\max\left(f_{\bullet j}\right) - \min\left(f_{\bullet j}\right)} \tag{7}$$

where $f_{i,j}$ is the value of sample i on its feature j, $\max\left(f_{\bullet j}\right)$ and $\min\left(f_{\bullet j}\right)$ are the maximum and minimum value of feature j respectively.

Ten-fold cross validation experiments are carried out to test the power of the proposed unsupervised feature selection algorithms. Datasets are partitioned in the following way: the data are first shuffled randomly, and then each type of samples are put into 10 empty sample sets one by one, until each sample is allocated to a subset. Samples are divided into 10 folds evenly while avoiding the case that a fold does not contain samples

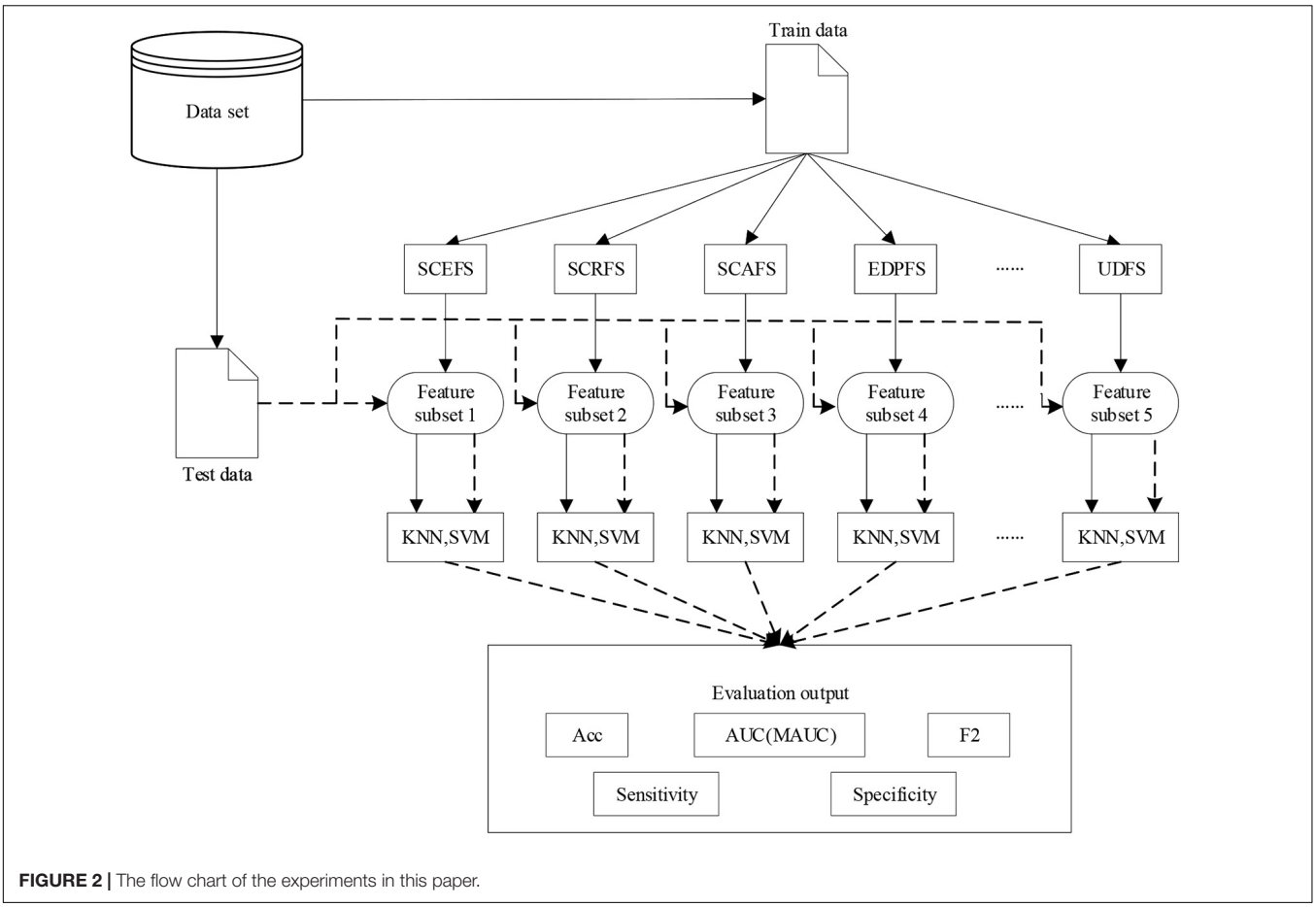

FIGURE 2 | The flow chart of the experiments in this paper.

from some types with small number of samples, especially in the imbalanced datasets. The nine folds comprise the training subset, and the remaining one fold is the test subset. The feature selection algorithms run on the training subset to detect the optimal feature subset, and the test subset is used to evaluate the detected feature subset. This process runs in turn until each fold is used as a test subset. To obtain the statistical experimental results, the above experimental process is run for five times, that is, the 10-fold cross validation experiments are run five times. The performance of a feature selection algorithm is evaluated using the mean classification results of the classifiers built on its selected feature subsets.

The code is implemented in MATLAB R2017b, and the experimental environment is Win10 64bit operating system, 192GB memory, Intel(R) Xeon(R) CPU E5-2666 v3@2.90GHz 2.90GHz (2 processors). The classifier adopts the SVM toolkit LibSVM developed by Lin et al. (Chang and Lin, 2011) and KNN embedded in MATLAB toolbox. The SVM classifier uses a linear kernel function with the penalty factor C = 20 and the default values for other parameters. The KNN classifier uses the nearest neighbor number K = 5. The unsupervised feature selection algorithms are evaluated in terms of the mean classification accuracy (simplified as Acc), AUC (MAUC for multi-class), F2-measure (referred to as F2) (Xie et al., 2019), Sensitivity, and Specificity of 10-fold cross validation experiments of their 5 runs. Where, F2-measure is proposed and defined for analyzing imbalanced data. It avoids the limits of F-measure which focuses on the positive class while ignoring the negative class. It is calculated by:

$$F2{-}measure = 2{*}\frac{precision{*}(\sim precision)}{precision + (\sim precision)} \tag{8}$$

Where, precision and ∼precision are the ratios of the true positive and true negative samples recognized by the classifier to the positive and the negative samples recognized by the classifier, respectively. For multi-class $l\,(l > 2)$ classification problem, we adopt one versus one method to transform the problem to be $l\,(l-1)\,/2$ binary classification problem. The F2 will be calculated using (9), similarly for Sensitivity and Specificity. **Figure 2** shows the flow chart of the whole experiments in this paper.

$$F2 = \frac{4}{l(l-1)} \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} \frac{precision_{ij}{*}(\sim precision)_{ij}}{precision_{ij}+(\sim precision)_{ij}} \tag{9}$$

Performance Comparison

This section will compare the performances of the proposed SCEFS, SCRFS, and SCAFS with other unsupervised feature selection algorithms EDPFS, RDPFS, MCFS, Laplacian, UDFS,

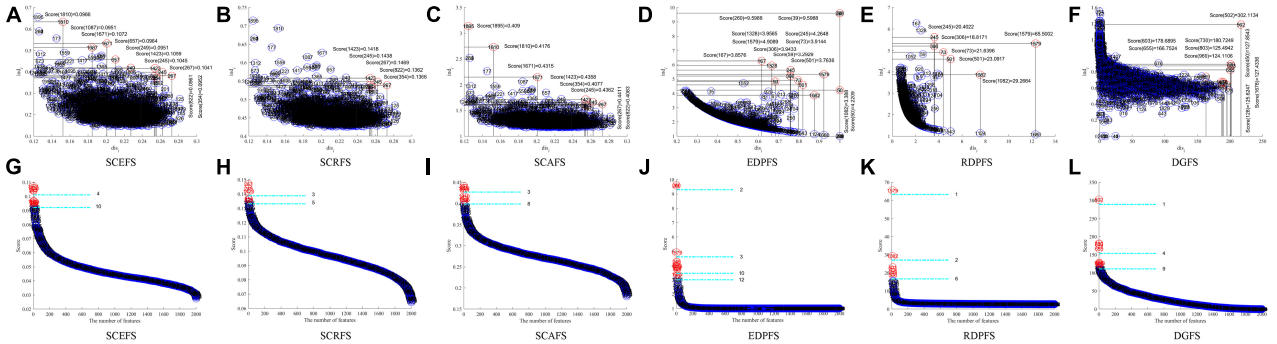

FIGURE 3 | Features displaying in the 2-dimensional space of each algorithm on Colon dataset, **(A)** and **(G)** SCEFS, **(B)** and **(H)** SCRFS, **(C)** and **(I)** SCAFS, **(D)** and **(J)** EDPFS, **(E)** and **(K)** RDPFS, **(F)** and **(L)** DGFS.

TABLE 2 | Performance comparison of KNN and SVM classifiers on Colon dataset by our algorithms and unsupervised feature selection algorithms based on density peaks.

Algorithms	KNN					SVM					Feature numbers
	Acc	AUC	F2	Sen	Spe	Acc	AUC	F2	Sen	Spe	
SCEFS	**0.840**	0.878	**0.817**	0.930	0.673	0.786	0.823	0.691	0.945	0.493	4
	0.824	**0.924**	0.768	**0.940**	0.610	0.795	0.814	0.716	0.940	0.527	10
SCRFS	0.821	0.873	0.809	0.910	0.660	0.757	0.784	0.539	0.970	0.363	3
	0.814	0.897	0.753	0.880	0.617	0.803	0.832	0.734	0.955	0.527	5
SCAFS	0.794	0.855	0.760	0.890	0.617	0.761	0.800	0.594	0.955	0.403	3
	0.834	0.894	0.809	0.920	**0.680**	0.805	0.827	0.745	0.940	0.560	8
EDPFS	0.614	0.779	0.246	0.850	0.180	0.648	0.716	0	**1**	0	2
	0.674	0.799	0.487	0.835	0.380	0.647	0.772	0.016	0.995	0.007	3
	0.811	0.886	0.788	0.890	0.670	**0.819**	0.853	**0.786**	0.935	**0.613**	10
	0.789	0.887	0.736	0.895	0.597	0.812	**0.855**	0.764	0.930	0.603	12
RDPFS	0.647	0.780	0.288	0.900	0.180	0.648	0.776	0	**1**	0	1
	0.647	0.780	0.288	0.900	0.180	0.644	0.776	0	0.995	0	2
	0.740	0.850	0.661	0.825	0.580	0.691	0.842	0.272	0.975	0.163	6
DGFS	0.551	0.698	0.164	0.790	0.11	0.648	0.738	0	**1**	0	1
	0.628	0.803	0.361	0.805	0.297	0.648	0.690	0	**1**	0	4
	0.601	0.763	0.346	0.765	0.303	0.648	0.743	0	**1**	0	9

RUFS, NDFS, and DGFS in selecting feature (gene) subsets on the gene expression datasets of cancers shown in **Table 1**. We first test the correctness of our defined feature score by comparing the proposed SCEFS, SCRFS, and SCAFS to the EDPFS, RDPFS, and DGFS algorithms on classic binary classification data Colon and multiclass classification data Leukemia_MLL. We evaluate the performances of the unsupervised feature selection algorithms in terms of Acc, AUC, F2, Sensitivity and Specificity of the classifier built using the feature subset detected by the algorithms according to feature scores.

Test of Feature Score

This subsection will test the proposed feature score by comparing the proposed SCEFS, SCRFS, and SCAFS with unsupervised feature selection algorithms EDPFS, RDPFS and DGFS. We display the features in 2-dimensional space by using the feature density (in EDPFS, RDPFS and DGFS), feature distance (in EDPFS, RDPFS and DGFS) and feature importance metric γ-score (in EDPFS and RDPFS), or decision graph score γ (in DGFS). It is similar to the proposed SCEFS, SCRFS, and SCAFS to display features in 2-dimensional space using feature independence as y-axis and feature discernibility as x-axis respectively, or display features in feature score descending order in 2-dimensional space using feature score as y-axis and the number of features as x-axis respectively.

Figure 3 shows the Colon cancer data features displayed in 2-dimensioanl space of the aforementioned six unsupervised feature selection algorithms. **Table 2** shows the performances of the six feature selection algorithms in terms of Acc, AUC, F2, Sensitivity, and Specificity of the classifiers built using the detected feature subsets for Colon data. **Figure 4** and **Table 3** are the results of the aforementioned six feature selection algorithms on Leukemia_MLL dataset. The boldface font in **Tables 2**, **3** indicates the best results among the six algorithms.

The experimental results in **Figure 3** show that the proposed unsupervised feature selection algorithms SCEFS, SCRFS and

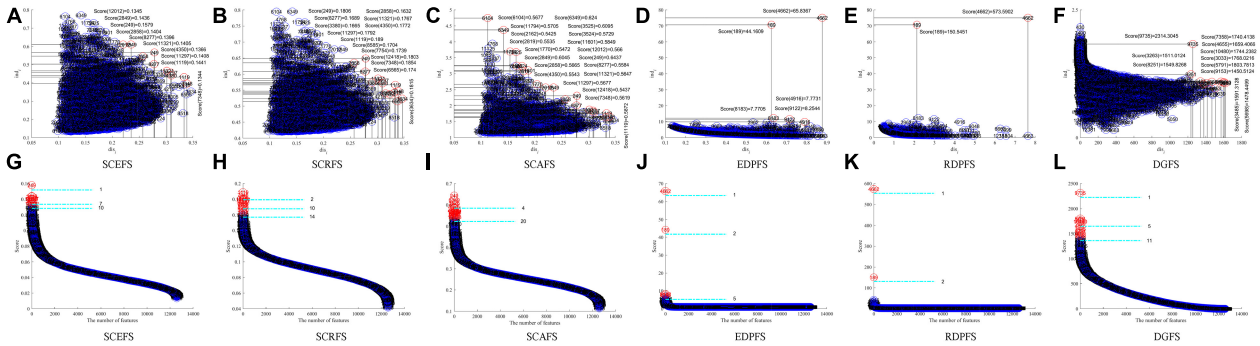

FIGURE 4 | Features displaying in the 2-dimensional space of each algorithm on Leukemia_MLL dataset, **(A)** and **(G)** SCEFS, **(B)** and **(H)** SCRFS, **(C)** and **(I)** SCAFS, **(D)** and **(J)** EDPFS, **(E)** and **(K)** RDPFS, **(F)** and **(L)** DGFS.

TABLE 3 | Performance comparison of KNN and SVM classifiers on Leukemia_MLL dataset by our algorithms and unsupervised feature selection based on density peaks.

Algorithms	KNN					SVM					Features numbers
	Acc	MAUC	F2	Sen	Spe	Acc	MAUC	F2	Sen	Spe	
SCEFS	0.642	0.841	0.539	0.672	0.719	0.624	0.883	0.397	0.564	0.637	1
	0.800	0.945	0.881	0.882	0.946	0.891	0.966	0.900	0.882	0.961	7
	0.803	**0.975**	**0.960**	**0.934**	**0.987**	**0.923**	**0.978**	**0.938**	0.920	**0.973**	10
SCRFS	0.466	0.764	0.424	0.591	0.606	0.410	0.773	0.058	0.128	0.634	2
	0.745	0.919	0.774	0.883	0.789	0.723	0.926	0.639	0.790	0.763	10
	0.721	0.919	0.757	0.901	0.754	0.752	0.949	0.764	0.914	0.769	14
SCAFS	0.719	0.896	0.684	0.798	0.779	0.719	0.918	0.595	0.739	0.780	4
	0.824	0.948	0.895	0.911	0.907	0.875	0.976	0.917	**0.940**	0.927	20
EDPFS	0.416	0.774	0.311	0.561	0.479	0.388	0.730	0	0	0.667	1
	0.477	0.765	0.422	0.644	0.549	0.388	0.713	0	0	0.667	2
	0.565	0.803	0.504	0.670	0.663	0.538	0.795	0.293	0.516	0.631	5
RDPFS	0.416	0.774	0.311	0.561	0.479	0.388	0.730	0	0	0.667	1
	0.477	0.765	0.422	0.644	0.549	0.388	0.713	0	0	0.667	2
DGFS	0.424	0.761	0.350	0.566	0.508	0.412	0.758	0.055	0.106	0.656	1
	0.606	0.846	0.528	0.670	0.702	0.606	0.828	0.285	0.596	0.632	5
	0.670	0.860	0.658	0.794	0.738	0.665	0.868	0.429	0.690	0.672	11

SCAFS can detect two feature subsets of different scales for Colon dataset, while EDPFS, RDPFS and DGFS can detect three or four feature subsets. The number of features in each feature subset detected by our SCEFS, SCRFS and SCAFS ranges from 3 to 10, while EDPFS, RDPFS and DGFS detect from 1 to 12.

As can be seen from the experimental results in **Table 2**, the three proposed unsupervised feature selection algorithms are obviously better than the three compared algorithms EDPFS, RDPFS and DGFS when using KNN classifier. The performance of SCEFS algorithm is the best, and the performance of DGSF algorithm is the worst. However, our previously proposed EDPFS algorithm is better than the proposed SCEFS, SCRFS and SCAFS when using SVM classifier especially when the feature subset size is 10 or 12. The performance of SCEFS, SCRFS and SCAFS is similar, but it is obviously better than RDPFS and DGFS. Although EDPFS, RDPFS and DGFS obtain 100% sensitivity, especially DGFS whose sensitivities are all 100% no matter the feature subset comprise 1, 4 or 9 features, their corresponding F2 and specificity are both 0, which means that all normal people in the test subset are recognized as colon cancer patients using the detected feature subsets.

The results in **Figure 4** show that the six unsupervised feature selection algorithms can detect the 2 or 3 feature subsets of different sizes for Leukemia_MLL dataset. The number of features is from 1 to 20. However, the EDPFS, RDPFS and DGFS algorithms can detect 2 or 3 feature subsets for Leukemia_MLL dataset. The number of features in these feature subsets is from 1 to 11.

As can be seen from results in **Table 3**, the proposed SCEFS can detect the optimal feature subset containing 10 features while having the best performance among the compared 6 unsupervised feature selection algorithms no matter whether using KNN or SVM classifier. It is obvious from the results in **Table 3** that the proposed SCEFS, SCRFS and SCAFS outperformed the unsupervised feature selection algorithms EDPFS, RDPFS and DGFS.

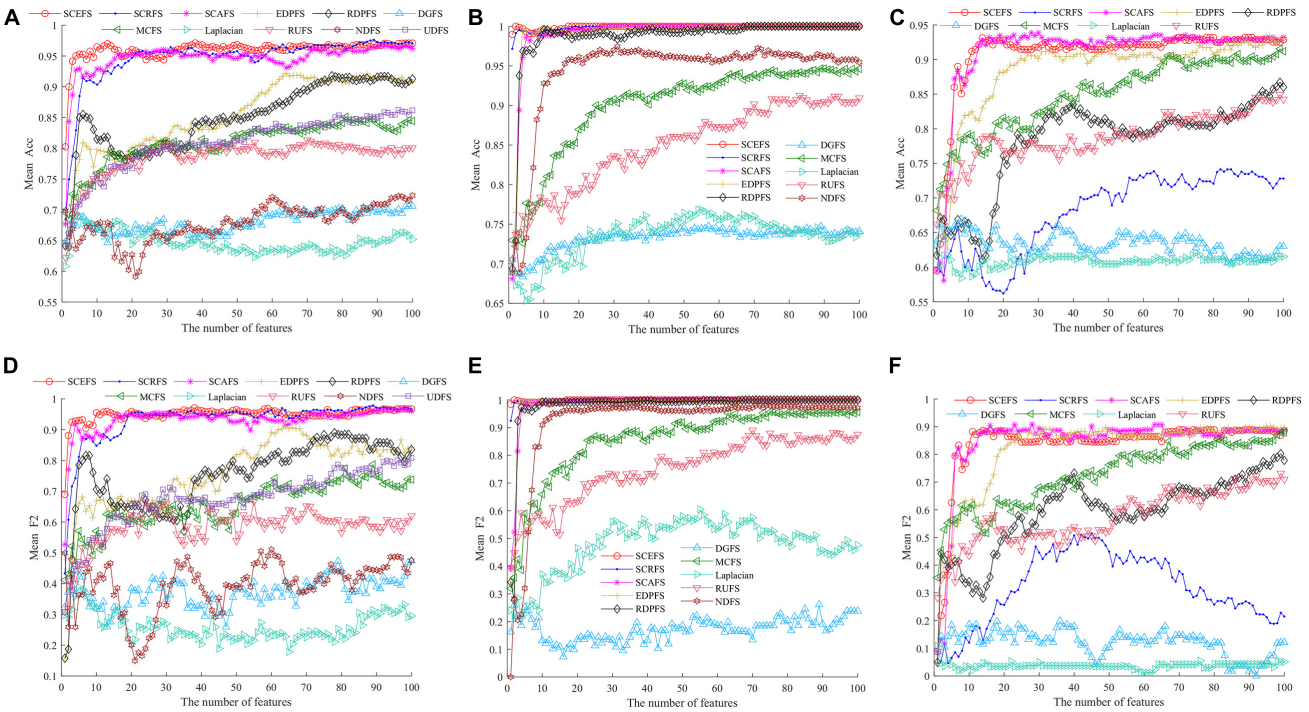

FIGURE 5 | The average accuracy (Acc) and F2 of all algorithms on three datasets using KNN classifier, **(A)** and **(D)** Leukemia, **(B)** and **(E)** ALL1, **(C)** and **(F)** Non-small lung cancer.

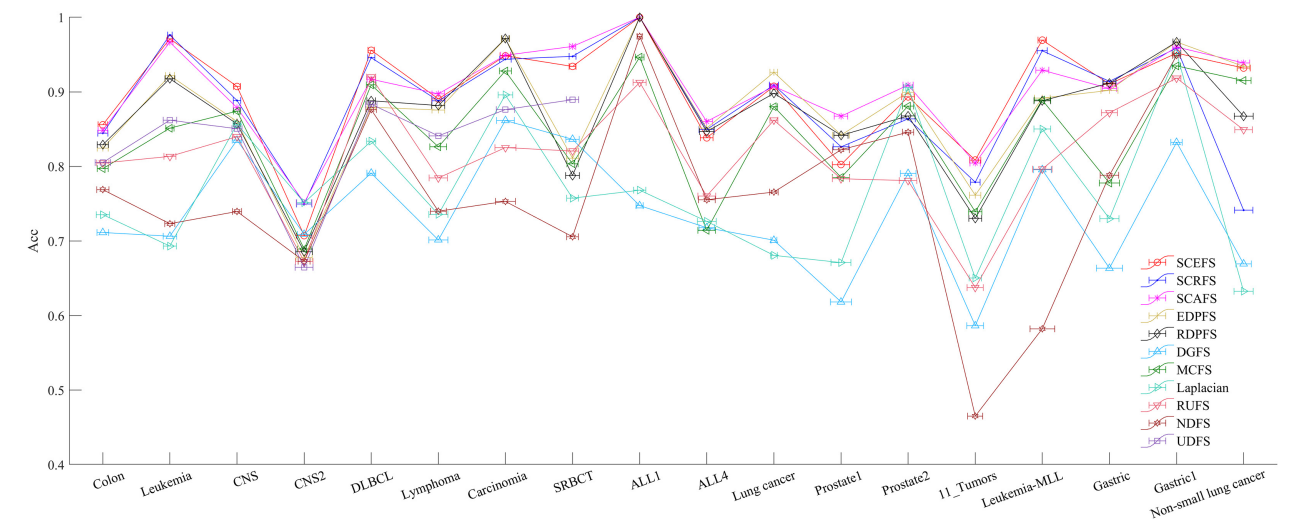

FIGURE 6 | The maximal mean Acc comparison of each algorithm on 18 datasets using KNN classifier.

To summarize the above analyses, we can assert that the proposed three unsupervised feature selection algorithms can detect the feature subset with strong discernibility having low redundancy. The detected feature subset usually comprises of a small number of features, and the classifiers built using the feature subset can obtain a good classification performance especially when the KNN classifier is used. Therefore the proposed SCEFS, SCRFS and SCAFS can realize a dimension reduction for high dimensional data meaning that our proposed feature score is powerful.

Comparison With Other Unsupervised Feature Selection Algorithms

This subsection will compare the performance of our proposed SCEFS, SCRFS and SCAFS to that of the other set of eight unsupervised feature selection algorithms EDPFS, RDPFS,

MCFS, Laplacian, UDFS, RUFS, NDFS and DGFS. We first show, in **Figure 5**, the performance of the above algorithms on three different scales of dimensions of datasets including Leukemia, ALL1 and Non-small lung cancer. Then we compare the performance of the above algorithms on the 18 datasets from **Table 1** in **Figure 6** and **Table 4**, and in **Figure 7** and **Table 5**. The classifier used is KNN due to its simple and good performance in section "Test of Feature Score." These 11 unsupervised feature selection algorithms are evaluated in terms of Acc and F2 of the KNN classifiers built using their detected feature subsets. We assume that the size of the feature subset is up to 100, that is, the feature subset consists of 100 detected features maximally. The NDFS and UDFS are so time consuming that we do not compare the algorithms to UDFS on the datasets with more than 10,000 features, nor for Non-small lung cancer dataset do we compare NDFS to other algorithms.

Figure 5 shows the mean Acc and F2 on Leukemia, ALL 1 and Non-small Lung cancer datasets. **Figure 6** shows the maximal mean Acc of each algorithm of its selecting feature subsets on 18 datasets from **Table 1**. **Figure 7** displays the maximal mean F2 of each algorithm of its selecting feature subsets for 18 datasets from **Table 1**. The horizontal error bar at each data point in **Figures 6**, **7** indicates the standard deviation of the results of 5 runs of 10-fold cross validation experiments and the total error bar length is twice the standard deviation. **Tables 4**, **5** use the triplet of Win/Draw/Loss to evaluate the performance of the three proposed algorithms SCEFS, SCRFS and SCAFS with other unsupervised feature selection algorithms in terms Acc and F2 respectively. For example, for algorithms A and B, the 12/2/4 indicates that algorithm A is superior to algorithm B on 12 datasets, and equal to on 2 datasets, and inferior to on 4 datasets. We make 12/2/4 boldface to indicate that algorithm A defeats algorithm B in performance.

The results in **Figure 5** show that the proposed SCEFS, SCRFS and SCAFS can detect feature subsets with good performance except for SCRFS on Non-small lung cancer dataset. The DGFS and Laplacian are the last two algorithms of the 11 compared unsupervised feature selection algorithms.

The results in **Figures 5A,D** show that the proposed SCEFS, SCRFS and SCAFS are superior to the other eight feature selection algorithms, especially SCEFS that performs best among the 11 feature selection algorithms. It can detect the feature subset containing 13 features which obtaining the Acc of 0.97and F2 of 0.96.

TABLE 4 | The comparison between proposed algorithms and other algorithms in terms of win/draw/loss based on the maximal mean Acc.

Algorithms	SCEFS	SCRFS	SCAFS	EDPFS	RDPFS	DGFS	MCFS	Laplacian	RUFS	NDFS	UDFS
SCEFS	0/18/0	**9/1/8**	7/1/10	**10/1/7**	**12/1/5**	**17/0/1**	**18/0/0**	**16/0/2**	**18/0/0**	**17/0/1**	**18/0/0**
SCRFS	8/1/9	0/18/0	6/1/11	**10/1/7**	**12/1/5**	**18/0/0**	**16/0/2**	**16/0/2**	**17/0/1**	**18/0/0**	**18/0/0**
SCAFS	**10/1/7**	**11/1/6**	0/18/0	**14/1/3**	**14/1/3**	**18/0/0**	**18/0/0**	**17/0/1**	**17/0/1**	**18/0/0**	**18/0/0**

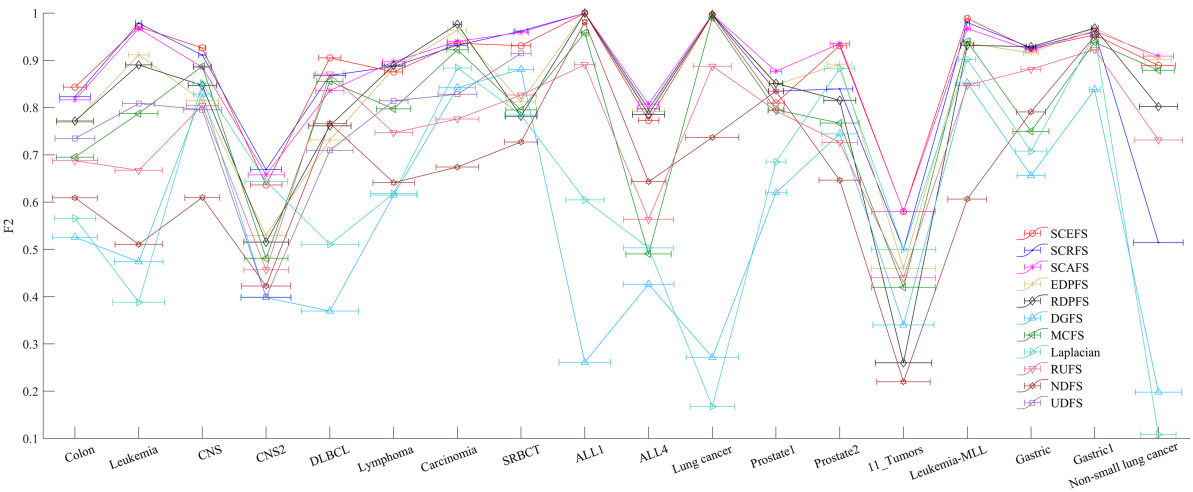

FIGURE 7 | The maximal mean F2 comparison of each algorithm on 18 datasets using KNN classifier.

TABLE 5 | The comparison between proposed algorithms and other compared algorithms in terms of win/draw/loss based on the maximal mean F2.

Algorithms	SCEFS	SCRFS	SCAFS	EDPFS	RDPFS	DGFS	MCFS	Laplacian	RUFS	NDFS	UDFS
SCEFS	0/18/0	**9/1/8**	6/2/10	**10/1/7**	**10/1/7**	**18/0/0**	**17/0/1**	**18/0/0**	**18/0/0**	**17/0/1**	**18/0/0**
SCRFS	8/1/9	0/18/0	8/1/9	**11/1/6**	**10/1/7**	**18/0/0**	**17/0/1**	**16/1/1**	**16/0/2**	**17/0/1**	**18/0/0**
SCAFS	**10/2/6**	**9/1/8**	0/18/0	**14/1/3**	**14/1/3**	**18/0/0**	**16/0/2**	**18/0/0**	**17/0/1**	**18/0/0**	**18/0/0**

TABLE 6 | Runtime of each unsupervised feature selection algorithm on five datasets (in seconds).

Datasets	Algorithms										
	SCEFS	SCRFS	SCAFS	EDPFS	RDPFS	DGFS	MCFS	Laplacian	RUFS	NDFS	UDFS
SRBST	0.335 ± 0.53	0.244 ± 0.17	0.223 ± 0.10	0.590 ± 0.12	0.582 ± 0.13	0.413 ± 0.08	0.956 ± 0.75	0.027 ± 0.02	11.15 ± 5.50	13.56 ± 3.89	37.29 ± 8.92
CNS	1.617 ± 0.36	1.687 ± 0.53	1.611 ± 0.37	6.182 ± 0.95	6.220 ± 0.98	4.789 ± 0.79	2.975 ± 1.76	0.073 ± 0.02	17.85 ± 7.49	240.71 ± 22.15	1003.2 ± 43.63
Prostate2	5.478 ± 1.84	5.480 ± 1.43	5.900 ± 2.07	26.05 ± 9.28	24.19 ± 7.54	16.32 ± 4.76	2.871 ± 0.99	0.174 ± 0.05	43.10 ± 16.77	1851.4 ± 264.9	–
Gastric	12.49 ± 0.63	12.68 ± 0.95	12.63 ± 0.98	51.63 ± 4.26	51.71 ± 4.63	37.85 ± 2.82	2.412 ± 0.34	0.115 ± 0.02	40.71 ± 6.47	6957.0 ± 461.5	–
NLC	96.08 ± 9.59	95.50 ± 10.17	98.18 ± 12.83	756.97 ± 33.47	753.75 ± 40.19	353.86 ± 36.60	8.616 ± 1.59	0.350 ± 0.08	391.18 ± 28.30	–	–

Note: NLC represents the Non-small lung cancer data.

The results in **Figures 5B,E** on ALL1 dataset show that SCEFS and SCRFS algorithms perform very well when the feature subset comprises the top feature, and SCEFS can obtain the maximum Acc and F2 of 1 when selecting the top 2 features. Although SCAFS is not as good as SCEFS and SCRFS, it defeats the other compared feature selection algorithms and converges quickly with increasing features in the feature subset. Its KNN classifier can obtain Acc and F2 higher than 0.95 when there are top 4 features in the feature subset, and get the highest Acc and F2 of 1 when selecting the top 27 features in the feature subset. Our previously proposed EDPFS and RDPFS also perform well on ALL1 dataset, and can detect the feature subset classifying all samples correctly for the test subset.

The results in **Figures 5C,F** on Non-small lung cancer dataset show us that our proposed SCEFS and SCAFS are the top 2 feature selection algorithms among the 11 compared feature selection algorithms, especially SCAFS, which is the best. SCEFS and SCAFS outperform our previously proposed EDPFS. These three are superior to other compared feature selection algorithms. Our proposed SCRFS performs badly on Non-small lung cancer dataset. Its performance is just better than that of the feature selection algorithms DGFS and Laplacian.

The results in **Figure 6** show us that the three proposed unsupervised feature selection algorithms SCEFS, SCRFS and SCAFS can detect the optimal feature subsets with best classification capability on nearly all datasets except for on the Carcinoma, Lung cancer and Gastric1 datasets. Our previously proposed EDPFS or RDPFS performs best on Carcinoma, Lung cancer and Gastric datasets. The performance of DGFS and Laplacian algorithms is poor. The results in **Figure 6** also show us that the error bar of our three proposed algorithms is short on 18 datasets, which indicates that the proposed algorithms are more stable than the other 8 feature selection algorithms in 5 runs of 10-fold cross validation experiments. Therefore the proposed feature selection algorithms can detect the feature subset that has much more stable classification performance than that of other compared feature selection algorithms.

It can be seen from the results in **Table 4** that the proposed SCAFS algorithm is the best, which can select the feature subsets with better classification performance than the algorithms DGFS, MCFS, NDFS and UDSF on 18 genomic data, and is superior to algorithms SCEFS and SCRFS on 10 and 11 data respectively. SCEFS is slightly better than SCRFS, and the former is better than the latter on 9 datasets. Although SCRFS is the worst among the proposed SCEFS, SCRFS, and SCAFS, it is superior to all the other 8 compared unsupervised feature selection algorithms EDPFS, RDPFS, DGFS, MCFS, Laplacian, RUFS, NDFS and UDFS.

The results in **Figure 7** show that the proposed SCEFS, SCRFS and SCAFS perform best on most datasets except for on Carcinoma and Gastric1 datasets in terms of F2 of KNN classifiers built using the selected feature subsets. Our previously proposed RDPFS and EDPFS obtain the best performance on Carcinoma and Gastric1, followed by our proposed SCAFS, SCEFS and SCRFS algorithms. DGFS and Laplacian are the last two unsupervised feature selection algorithms among the overall 11 unsupervised feature selection algorithms. In addition, from the error bar of each algorithm for each dataset, it is clear that the

TABLE 7 | The gene biomarkers of Prostate2 and Non-small lung cancer selected by our algorithms.

Datasets	Algorithms	Gene biomarkers
Prostate2	SCEFS	FOS, DNALI1, VWA5A, BTRC, PMF1-BGLAP, MGAT4C, KAT5, IER2, TRAF6, CYP27A1, CSPG4, MET, TIGR: HG3999-HT4269, LOC100289561, CDKN3, AP2B1, TK2, MSMB, TTPA, YME1L1, B3GALT2
	SCRFS	SEMG1, ALB, TNNT1, CRP, MYL1, CTNNB1, FGB, TNNC1, ACTA1, MYH7, MYLPF, CST4, FGG, HP, APOA1, DDN, MYL3, TPM2, FGA, SEMG2, NEB, SLN, APOC3, PCK1, ENO3, APOC4-APOC2
	SCAFS	CDKN3, FOS, CYP27A1, SSX2B, VWA5A, TTN, TGM4, CCL19, HPGD, CSPG4, AR, MSMB, TNNT1, MYL1, HDAC9, TNNI1, ALOX15B, PMF1-BGLAP, ACTA1, COL2A1, ACTC1, SERPINB5, PEG10, HBB
Non-small lung cancer	SCEFS	KRT5, SPRR1B, DSG3, DSC3, NTS, MAGEA6, MAGEA9B, XIST, SERPINB13, SPRR3, CLCA2, SPRR1A, MAGEA6, MAGEA10-MAGEA5
	SCRFS	GP2, RHOXF1, REG4, ACTN2, NCAN, PRL, REG1B, CYP2F1, FGF3, REG4, RHOXF2B, DEFA5, FRG2EP, GFI1B, BPIFB4, MUC6, EREG
	SCAFS	DSG3, NTS, XIST, SERPINB13, DSC3, SPRR1B, MAGEA9B, CLCA2, LIN28B, MAGEC2, SPRR3

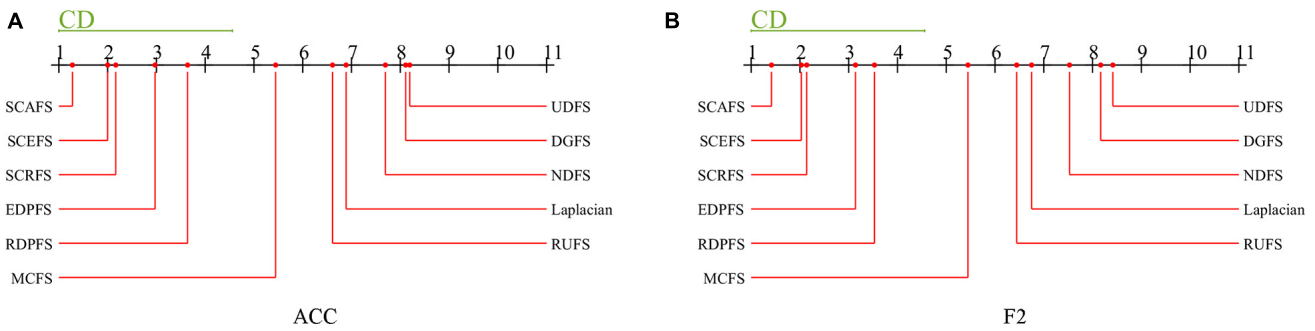

FIGURE 8 | Comparison of unsupervised feature selection algorithms against each other on maximal mean Acc and F2 with Nemenyi's test, **(A)** Acc, **(B)** F2.

proposed SCEFS, SCRFS and SCAFS can select the feature subset with strong stability. Therefore the proposed SCEFS, SCRFS and SCAFS are strong in finding powerful feature subsets.

The results in **Table 5** show us that the proposed SCAFS is the best. It is superior to SCEFS and SCRFS on 10 and 9 datasets respectively, and equal to SCEFS and SCRFS on 2 and 1 datasets respectively. The proposed SCEFS ranks in the second place. Although SCRFS is inferior to SCAFS and SCEFS, it is superior to all the other eight compared unsupervised feature selection algorithms.

Summarizing the above analyses, it can be concluded that the proposed three unsupervised feature selection algorithms SCEFS, SCRFS and SCAFS are superior to our previously proposed EDPFS and RDPFS, and far superior to other compared feature selection algorithms. They can detect the feature subsets with good classification capability and strong stability. The KNN classifier built using the selected feature subsets obtain the expected performance on 18 cancer genomic datasets.

Statistical Significance Test of Algorithms

This subsection will undertake statistical tests on our proposed SCEFS, SCRFS and SCAFS, and the other compared unsupervised feature selection algorithms including EDPFS, RDPFS, DGFS, MCFS, Laplacian, RUFS, NDFS, and UDFS, to judge whether or not the results of our SCEFS, SCRFS and SCAFS are statistically significant. We adopt the Friedman's test to discover the significant difference between the 11 unsupervised feature selection algorithms. If the significant difference has been detected by Friedman's test, then the Nemenyi's test is used as a *post hoc* test to see if there is significant difference between each pair of unsupervised feature selection algorithms. We conduct Friedman's test at α=0.05 using the results of each algorithm in terms of maximal mean Acc and F2 of KNN classifiers built using the selected feature subsets on 18 genomic datasets. If the null hypothesis that "all algorithms have the same performance" does not hold, then we adopt Nemenyi's test to detect the significant difference between each pair of algorithms. We calculate the critical threshold *CD* in (10). If the difference of the mean ranks of a pair algorithm is greater than *CD*, then the null hypothesis that "the two algorithms have the same performance" is rejected, that is, the performances of the two algorithms are significantly different at the confidence degree of $1-\alpha$, that is 0.95; otherwise, the null hypothesis is accepted.

$$CD = q_\alpha \sqrt{\frac{M\,(M+1)}{6N}} \tag{10}$$

In the above M and N are the number of algorithms and datasets respectively, and q_α can be found in textbook. For our Nemenyi's test, $q_\alpha = q_{0.05} = 3.219$, $M = 11$, $N = 18$, so $CD = 3.5587$.

At the statistical significance level of α=0.05, the results of the Friedman's test are here. For maximal mean Acc, $df = 10$, $\chi^2 = 115.76$, $p = 3.652\text{e-}20$; for maximal mean F2, $df = 10$, $\chi^2 = 113.48$, $p = 1.058\text{e-}19$. This Friedman's test shows that p is much less than 0.05 no matter whether for Acc or F2, so we reject the null hypothesis that "all algorithms have the same performance"

at the confidence degree of 0.95 ($= 1 - \alpha$). We can say that there are strong significant differences between these 11 unsupervised feature selection algorithms.

Then as a *post hoc* test, the Nemenyi's test is conducted to detect the significant difference between each pair of algorithms. The Nemenyi's test results are shown in **Figure 8**.

The experimental results in **Figure 8** show us that there is no significant difference between the three proposed unsupervised feature selection algorithms SCAFS, SCEFS, SCRFS in terms of the maximal mean Acc and F2, and there is also no significant difference between our SCAFS, SCEFS, SCRFS and our previously proposed algorithms EDPFS, RDPFS. However, there is significantly different between SCAFS, SCEFS, SCRFS, EDPFS, RDPFS, and MCFS, DGFS, UDFS, NDFS, Laplacian and RUFS algorithms. Our proposed SCAFS, SCEFS, SCRFS are better than the other eight unsupervised feature selection algorithms, especially better than MCFS, DGFS, UDFS, NDFS, Laplacian and RUFS algorithms. Our SCAFS is the best one among the 11 unsupervised feature selection algorithms.

Run Time Comparison

This subsection chooses the five genomic datasets SRBCT, CNS, ProState2, Gastric and Non-Small Lung Cancer with very high dimensionalities to test the time performance of our three unsupervised feature selection algorithms SCAFS, SCEFS, SCRFS, while verifying the correctness of the theoretical time complexity analysis in section "Complexity Analysis." All algorithms are run on the five datasets in 10-fold cross validation experiments for 5 runs. The average run time of each algorithm on five genomic datasets is compared with each other in **Table 6**.

The results in **Table 6** show that the Laplacian algorithm is the fastest one among the 11 unsupervised features selection algorithms on the five genomic datasets. It can complete feature selection in a short time. The proposed SCAFS, SCEFS, SCRFS feature selection algorithms have similar run times. They rank in second place after the Laplacian algorithm on SRBCT and CNS datasets with no more than 10,000 genes, and rank in the third place after Laplacian and MCFS algorithms on ProState2, Gastric and Non-Small Lung Cancer datasets which have more than 10,000 dimensions. They are definitely better than other compared unsupervised feature selection algorithms.

From the above analyses, we can say that although our proposed feature selection algorithms SCAFS, SCEFS, SCRFS are not the most efficient, their time consuming loads are acceptable on high dimensional datasets. They are faster than EDPFS, RDPFS, DGFS, RUFS, NDFS and UDFS algorithms when selecting optimal feature subsets on high dimensional datasets.

The Bioinformatics Interpretation of the Selected Features of Our Algorithms

This subsection will take Prostate2 and Non-small lung cancer datasets as examples to conduct functional analysis on the genes selected by our SCEFS, SCRFS and SCAFS algorithms, and some of which may have known roles in cancer onset and development. **Table 7** summarizes the gene biomarkers of Prostate2 and Non-small lung cancer detected by our SCEFS, SCRFS and SCAFS algorithms.

The literature shows that many genes selected by our three unsupervised feature selection algorithms are associated with the prostate (He et al., 2013; Lu and Chen, 2015; Yu et al., 2015; Fajardo et al., 2016; Sjöblom et al., 2016) and non-small lung cancer (Wang et al., 2004; Monica et al., 2009; Agackiran et al., 2012; Sunaga et al., 2013; Argon et al., 2015; Tantai et al., 2015). For example, the gene MSMB selected by algorithms SCEFS and SCAFS is a key biomarker for prostate cancer (Kim et al., 2015; Sjöblom et al., 2016). The gene of MSMB is located in area 10q11.2 and the protein encoded is a member of the immunoglobulin binding factor family. The protein has inhibin-like activity and is one of the three most common proteins generated by the prostate. Several researches have shown the lower expression of MSMB protein in prostate cancer tissue and the cancer suppressive role in prostate cancer (Abrahamsson et al., 1988; Garde et al., 1999). The genes AR and MET are related to prostate cancer. They are selected by our SCAFS and SCEFS respectively. The gene AR is one of the most important genes in prostate cancer related genes. It has been amply demonstrated that AR gene regulation plays a key role in the survival mechanism of prostate cells (Balk and Knudsen, 2008; Fajardo et al., 2016). The increase of AR expression can reduce the content of prostate specific antigen in serum, and cause benign prostatic hyperplasia, and also has relation with the pathogenesis of prostate cancer. The gene MET participates in the biological processes of endothelial cell morphogenesis, signal transduction, cell surface receptor signaling pathway and cell proliferation. The MET signaling pathway plays an important role in cell migration, apoptosis, proliferation and differentiation, which can promote tumor cells to form more aggressive cell phenotype to avoid immunity and enhance the ability of tumor cells to survive, infiltrate and invade. The genes of KAT5, BTRC, FOS, CTNNB1, TGM4 and SERPINB5 detected by our algorithms have also been shown to be closely related to the occurrence and development of prostate cancer (Cao et al., 2013; He et al., 2013; Bernardo et al., 2015; Lu and Chen, 2015).

The genes DSC3, EREG, KRT5, LIN28B, NTS, XIST and DSG3 etc. selected by our three algorithms are closely connected with development of non-small lung cancer (Wang et al., 2004; Monica et al., 2009; Agackiran et al., 2012; Sunaga et al., 2013; Wen et al., 2014; Argon et al., 2015; Tantai et al., 2015). The gene DSC3 is the component of intercellular desmosome junctions, and involved in the biological processes of cell adhesion, protein stabilization and homophilic cell adhesion via plasma membrane adhesion molecules. Several studies demonstrated that DSC3 was a valuable biomarker for non-small lung cancer from other types of lung cancer (Agackiran et al., 2012; Masai et al., 2013). LIN28B is involved with regulation of transcription with DNA-templated, pre-miRNA processing, miRNA catabolic process and overexpressed in cancer cell lines and primary tumor of human. The gene LIN28B is known to be related to many types of diseases such as obesity, ovarian cancer and colon cancer (Leinonen et al., 2012; Pang et al., 2014; Lu et al., 2016). Recently published research has shown that LIN28B may affect the result of

treatment of non-small lung cancer with radiotherapy, and may be biomarkers for non-small lung cancer (Wen et al., 2014).

Other gene biomarkers such as CDKN3 and SERPINB13 selected in this study may be worth the further prospective studies since they provide the best performance of classification for prostate cancer and non-small lung cancer datasets.

CONCLUSION

This paper presented the unsupervised feature selection algorithms SCEFS, SCRFS, and SCAFS based on feature standard deviation and cosine similarity for tackling the challenges in cancer genomic data analysis. Feature discernibility is proposed and defined using its standard deviation, and also feature independence by cosine similarity. All features are scattered in 2-dimesional space using discernibility as x-axis and independence as y-axis respectively, so that the upper right corner features have both high discernibility and independence, and comprise the optimal feature subset. The feature score is proposed and defined as the area of the rectangle enclosed by the feature coordinate lines and coordinate axes, so as to quantify the contributions of the upper right corner features to classification. The theoretical analysis and the comprehensive experiments on 18 genomic datasets demonstrate that the proposed three unsupervised feature selection algorithms can detect the optimal feature subsets enclosing sparse and strong discernibility while having low redundancy features. The detected features by our proposed feature selection algorithms are most important biomarkers whose regulation levels are closely related to pathogeneses of cancers. This study provides a base for cancer pathological research, drug development, cancer early diagnosis, treatment and prevention.

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/s.

AUTHOR CONTRIBUTIONS

JX made substantial contributions to the conception and revised the work. MW implemented all algorithms and wrote the experimental results. PG read through and revised the manuscript. JX, MW, SX, ZH, and PG discussed and designed this study. All authors read and approved the final manuscript.

FUNDING

This study was supported in part by the National Natural Science Foundation of China under Grant Nos. 61673251, 62076159, 12031010, and 61771297, and is also supported by the National Key Research and Development Program of China under Grant No. 2016YFC0901900, and by the Fundamental Research Funds for the Central Universities under Grant No. GK202105003, and by the Innovation Funds of Graduate Programs at Shaanxi Normal University under Grant Nos. 2015CXS028, 2016CSY009, and 2018TS078 as well.

ACKNOWLEDGMENTS

We are much obliged to those who share the gene expression datasets for us to use in this study.

REFERENCES

Abrahamsson, P. A., Falkmer, H. L. S., and Wadstrom, L. B. (1988). Immunohistochemical distribution of the three predominant secretory proteins in the parenchyma of hyperplastic and neoplastic prostate glands. *Prostate* 12, 39–46. doi: 10.1002/pros.2990120106

Agackiran, Y., Ozcan, A., Akyurek, N., Memis, L., Findik, G., and Kaya, S. (2012). Desmoglein-3 and napsin a double stain, a useful immunohistochemical marker for differentiation of lung squamous cell carcinoma and adenocarcinoma from other subtypes. *Appl. Immunohistochem.* 20, 350–355. doi: 10.1097/PAI.0b013e318245c730

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511. doi: 10.1038/35000501

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.* 96, 6745–6750.

Ang, J. C., Mirzal, A., Haron, H., and Hamed, H. N. A. (2016). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 13, 971–989. doi: 10.1109/TCBB.2015.2478454

Argon, A., Nart, D., and Veral, A. (2015). The value of cytokeratin 5/6, p63 and thyroid transcription factor-1 in adenocarcinoma, squamous cell carcinoma and non-small-cell lung cancer of the lung/akciğerin adenokarsinom, skuamöz hücreli karsinom ve küçük hücreli dışı akciğer kanserlerinde sitokeratin 5/6, p63 ve TTF-1'in değeri. *Turk. J. Pathol.* 31, 81–88. doi: 10.5146/tjpath.2015.01302

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., Boer, M. L., Minden, M. D., et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genet.* 30, 41–47. doi: 10.1038/ng765

Balk, S. P., and Knudsen, K. E. (2008). AR, the cell cycle, and prostate cancer. *Nucl. Recept. Signal.* 6:e001. doi: 10.1621/nrs.06001

Bernardo, M. M., Kaplun, A., Dzinic, S. H., Li, X., Irish, J., Mujagic, A., et al. (2015). Maspin expression in prostate tumor cells averts stemness and stratifies drug sensitivity. *Cancer Res.* 75, 3970–3979. doi: 10.1158/0008-5472.CAN-15-0234

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13790–13795. doi: 10.1073/pnas.191502998

Blum, A. L., and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artif. Intell.* 97, 245–271. doi: 10.1016/S0004-3702(97)00063-5

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P. E., Verrill, C., et al. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci Rep.* 8:3395. doi: 10.1038/s41598-018-21758-3

Cai, D., Zhang, C., and He, X. (2010). "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD International Conference on knowledge Discovery and Data Mining*, (New York, NY: ACM), doi: 10.1145/1835804.1835848

Cao, M., and Chen, W. (2019). Epidemiology of cancer in China and the current status of prevention and control. *Chin. J. Clin. Oncol.* 46, 145–149. doi: 10.3969/j.issn.1000-8179.2019.03.283

Cao, Z., Wang, Y., Liu, Z. Y., Zhang, Z. S., Ren, S. C., Yu, Y. W., et al. (2013). Overexpression of transglutaminase 4 and prostate cancer progression: a potential predictor of less favourable outcomes. *Asian J. Androl.* 15, 742–746. doi: 10.1038/aja.2013.79

Chandran, U. R., Ma, C., Dhir, R., Bisceglia, M., Lyons-Weiler, M., Liang, W., et al. (2007). Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer* 7:64. doi: 10.1186/1471-2407-7-64

Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27. doi: 10.1145/1961189.1961199

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., et al. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* 103, 2771–2778. doi: 10.1182/blood-2003-09-3243

Dash, M., Liu, H., and Yao, J. (1997). "Dimensionality reduction of unsupervised data," in *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, (Danvers, MA: IEEE), doi: 10.1109/TAI.1997.632300

Dashtban, M., and Balafar, M. (2017). Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. *Genomics* 109, 91–107. doi: 10.1016/j.ygeno.2017.01.004

Diao, G., and Vidyashankar, A. N. (2013). Assessing genome-wide statistical significance for large p small n problems. *Genetics* 194, 781–783. doi: 10.1534/genetics.113.150896

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3, 185–205. doi: 10.5555/937976.938050

Dong, X., Han, Y., Zhen, S., and Xu, J. (2018). Actin Gamma 1, a new skin cancer pathogenic gene, identified by the biological feature—based classification. *J. Cell. Biochem.* 119, 1406–1419. doi: 10.1002/jcb.26301

Effron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: CRC Press.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. doi: 10.1038/nature21056

Fajardo, A. M., MacKenzie, D. A., Olguin, S. L., Scariano, L. K., Rabinowitz, L., and Thompson, T. A. (2016). Antioxidants abrogate alpha-tocopherylquinone-mediated down-regulation of the androgen receptor in androgen-responsive prostate cancer cells. *PloS One* 11:e0151525. doi: 10.1371/journal.pone.0151525

Fu, K. S., Min, P. J., and Li, T. J. (1970). Feature Selection in Pattern Recognition. *IEEE Trans. Syst. Sci. Cybern.* 6, 33–39. doi: 10.1109/TSSC.1970.300326

Garde, S. V., Basrur, V. S., Li, L., Finkelman, M. A., Krishan, A., Wellham, L., et al. (1999). Prostate secretory protein (PSP94) suppresses the growth of androgen—independent prostate cancer cell line (PC3) and xenografts by inducing apoptosis. *Prostate* 38, 118–125. doi: 10.1002/(sici)1097-0045(19990201)38:2<118::aid-pros5<3.0.co;2-g

Global Burden of Disease Cancer Collaboration. (2018). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. *JAMA Oncol.* 4, 1553–1568. doi: 10.1001/jamaoncol.2018.2706

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537. doi: 10.1126/science.286.5439.531

He, J., Bi, Y., Ding, L., Li, Z., and Wang, S. (2017). Unsupervised feature selection based on decision graph. *Neural. Comput. Applic.* 28, 3047–3059. doi: 10.1007/s00521-016-2737-2

He, W., Zhang, M. G., Wang, X. J., Zhong, S., Shao, Y., Zhu, Y., et al. (2013). KAT5 and KAT6B are in positive regulation on cell proliferation of prostate cancer through PI3K-AKT signaling. *Int. J. Clin. Exp. Pathol.* 6, 2864–2871.

He, X., Cai, D., and Niyogi, P. (2006). "Laplacian score for feature selection," in *Proceedings of the 18th International Conference on Neural Information Processing Systems*, (Cambridge, MA: MIT Press), doi: 10.5555/2976248.2976312

Hou, C., Nie, F., Yi, D., and Wu, Y. (2011). "Feature selection via joint embedding learning and sparse regression," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, (Catalonia, Spain: AAAI Press).

Jiang, S., Zheng, Q., and Zhang, Q. (2008). Clustering-based feature selection. *Aata Electron. Sinica* 36, 157–160.

Kabir, M. M., Shahjahan, M., and Murase, K. (2011). A new local search based hybrid genetic algorithm for feature selection. *Neurocomputing* 74, 2914–2928. doi: 10.1016/j.neucom.2011.03.034

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7, 673–679. doi: 10.1038/89044

Kim, E. K., Kim, H. E., Han, K., Kang, B. J., Sohn, Y. M., Woo, O. H., et al. (2018). Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. *Sci Rep.* 8:2762. doi: 10.1038/s41598-018-21215-1

Kim, S., Shin, C., and Jee, S. H. (2015). Genetic variants at 1q32. 1, 10q11. 2 and 19q13. 41 are associated with prostate-specific antigen for prostate cancer screening in two Korean population-based cohort studies. *Gene* 556, 199–205. doi: 10.1016/j.gene.2014.11.059

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th international joint conference on Artificial intelligence*, (San Francisco, CA: Morgan Kaufmann Publishers Inc).

Kohavi, R., and John, G. H. (1997). Wrappers for feature subset selection. *Artif. Intell.* 97, 273–324. doi: 10.1016/S0004-3702(97)00043-X

Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C., et al. (2009). Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer* 63, 32–38. doi: 10.1016/j.lungcan.2008.03.033

Leinonen, J. T., Surakka, I., Havulinna, A. S., Kettunen, J., Luoto, R., Salomaa, V., et al. (2012). Association of LIN28B with adult adiposity-related traits in females. *PLoS One* 7:e48785. doi: 10.1371/journal.pone.0048785

Li, Z., Yang, Y., Liu, J., Zhou, X., and Lu, H. (2012). "Unsupervised feature selection using nonnegative spectral analysis," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, (Catalonia, Spain: AAAI Press).

Ling, X., and Ji, G. (2007). A clustering ensemble based unsupervised feature selection approach. *Nanjing Shi Da Xue Bao* 7, 60–63.

Liu, T., Wu, G., and Chen, Z. (2005). An effective unsupervised feature selection method for text clustering. *J. Comp. Res. Develop.* 42, 381–386.

Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., and Gao, Z. (2017). A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing* 256, 56–62. doi: 10.1016/j.neucom.2016.07.080

Lu, L., Katsaros, D., Canuto, E. M., Biglia, N., Risch, H. A., and Yu, H. (2016). LIN-28B/let-7a/IGF-II axis molecular subtypes are associated with epithelial ovarian cancer prognosis. *Gynecol. Oncol.* 141, 121–127. doi: 10.1016/j.ygyno.2015.12.035

Lu, T. L., and Chen, C. M. (2015). Differential requirements for β-catenin in murine prostate cancer originating from basal versus luminal cells. *J. Pathol.* 236, 290–301. doi: 10.1002/path.4521

Masai, K., Tsuta, K., Kawago, M., Tatsumori, T., Kinno, T., Taniyama, T., et al. (2013). Expression of squamous cell carcinoma markers and adenocarcinoma markers in primary pulmonary neuroendocrine carcinomas. *Appl. Immunohistochem.* 21, 292–297. doi: 10.1097/PAI.0b013e31826fd4f3

Mitra, P., Murthy, C., and Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 301–312. doi: 10.1109/34.990133

Monica, V., Ceppi, P., Righi, L., Tavaglione, V., Volante, M., Pelosi, G., et al. (2009). Desmocollin-3: a new marker of squamous differentiation in undifferentiated large-cell carcinoma of the lung. *Mod. Pathol.* 22, 707–717. doi: 10.1038/modpathol.2009.30

Notterman, D. A., Alon, U., Sierk, A. J., and Levine, A. J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res.* 61, 3124–3130.

Orringer, D. A., Pandian, B., Niknafs, Y. S., Hollon, T. C., Boyle, J., Lewis, S., et al. (2017). Rapid intraoperative histology of unprocessed surgical specimens via

fibre-laser-based stimulated Raman scattering microscopy. *Nat. Biomed. Eng.* 1, 1–13. doi: 10.1038/s41551-016-0027

Panday, D., Amorim, R. C., and Lane, P. (2018). Feature weighting as a tool for unsupervised feature selection. *Inf. Process. Lett.* 129, 44–52. doi: 10.1016/j.ipl.2017.09.005

Pang, M., Wu, G., Hou, X., Hou, N., Liang, L., Jia, G., et al. (2014). LIN28B promotes colon cancer migration and recurrence. *PloS One* 9:e109169. doi: 10.1371/journal.pone.0109169

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

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415, 436–442. doi: 10.1038/415436a

Qian, M., and Zhai, C. (2013). ""Robust unsupervised feature selection," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, (Catalonia, Spain: AAAI Press).

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* 8, 68–74. doi: 10.1038/nm0102-68

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 203–209. doi: 10.1016/s1535-6108(02)00030-2

Sjöblom, L., Saramäki, O., Annala, M., Leinonen, K., Nättinen, J., Toloenen, T., et al. (2016). Microseminoprotein-beta expression in different stages of prostate cancer. *PloS One* 11:e0150241. doi: 10.1371/journal.pone.0150241

Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., et al. (2001). Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* 61, 7388–7393.

Sunaga, N., Kaira, K., Lmai, H., Shimizu, K., Nakano, T., Shames, D., et al. (2013). Oncogenic KRAS-induced epiregulin overexpression contributes to aggressive phenotype and is a promising therapeutic target in non-small-cell lung cancer. *Oncogene* 32, 4034–4042. doi: 10.1038/onc.2012.402

Tantai, J., Hu, D., Yang, Y., and Geng, J. (2015). Combined identification of long non-coding RNA XIST and HIF1A-AS1 in serum as an effective screening for non-small cell lung cancer. *Int. J. Clin. Exp. Pathol.* 8, 7887–7895.

Wang, G., Hu, N., Yang, H. H., Wang, L., Su, H., Wang, C., et al. (2013). Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in China. *PloS One* 8:e63826. doi: 10.1371/journal.pone.0063826

Wang, H., Zheng, B., Yoon, S. W., and Ko, H. S. (2017). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur. J. Oper. Res.* 267, 687–699. doi: 10.1016/j.ejor.2017.12.001

Wang, J., Li, Y., and Chen, J. (2013). "Label reconstruction based laplacian score for semi-supervised feature selection," in *2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer*, (Danvers, MA: IEEE), doi: 10.1109/MEC.2013.6885229

Wang, L., and Jiang, S. (2015). Novel feature selection method based on feature clustering. *Appl. Res Comput.* 32, 1305–1308.

Wang, Y., Liang, Z., Yuan, Y., Han, Y., Liu, Y., Liu, N., et al. (2004). Expression of multiple cancer-testis antigen genes in non-small cell lung cancer treated by chemotherapy prior surgery. *Natl. Med. J. China* 84, 464–468.

Wen, J., Liu, H., Wang, Q., Liu, Z., Li, Y., Xiong, H., et al. (2014). Genetic variants of the LIN28B gene predict severe radiation pneumonitis in patients with non-small cell lung cancer treated with definitive radiation therapy. *Eur. J. Cancer* 50, 1706–1716. doi: 10.1016/j.ejca.2014.03.008

Wu, Y., Grabsch, H., Lvanova, T., Tan, L. B., Murray, J., Ooi, C. H., et al. (2012). Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. *Gut* 62, 1100–1111. doi: 10.1136/gutjnl-2011-301373

Xie, J., and Fan, W. (2017). Gene markers identification algorithm for detecting colon cancer patients. *Pattern Recog. Artif. Intell.* 30, 1019–1029. doi: 10.16451/j.cnki.issn1003-6059.201711007

Xie, J., and Gao, H. (2014). The statistical correlation and K-means based distinguishable gene subset selection algorithms. *J Softw.* 25, 2050–2075. doi: 10.13328/j.cnki.jos.004644

Xie, J., Qu, Y., and Wang, M. (2016a). Unsupervised feature selection algorithms based on density peaks. *Journal of Nanjing University* (Natural Sciences). 52, 735–745.

Xie, J., and Wang, C. (2011). Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Syst. Appl.* 38, 5809–5815. doi: 10.1016/j.eswa.2010.10.050

Xie, J., Wang, M., Zhou, Y., Gao, H., and Xu, S. (2019). Differentially expressed gene selection algorithms for unbalanced gene datasets. *Chin. J. Comput.* 42, 1232–1251. doi: 10.11897/SP.J.1016.2019.01232

Xie, J., Wang, M., Zhou, Y., and Li, J. (2016b). "Coordinating discernibility and independence scores of variables in a 2D space for efficient and accurate feature selection," in *12th International Conference on Intelligent Computing*, (Cham, Switzerland: Springer), doi: 10.1007/978-3-319-42297-8_12

Xie, J., Wu, Z., and Zheng, Q. (2020a). An adaptive 2D feature selection algorithm based on information gain and pearson correlation coefficient. *Journal of Shaanxi Normal University (Natural Science Edition)*. 48, 69–81. doi: 10.15983/j.cnki.jsnu.2020.01.019

Xie, J., Zheng, Q., and Ji, X. (2020b). An ensemble feature selection algorithm based on F-score and kernel extreme learning machine. *Journal of Shaanxi Normal University (Natural Science Edition)* 48, 1–8. doi: 10.15983/j.cnki.jsnu.2020.01.001

Xie, T., Ren, P., Zhang, T., and Tang, Y. Y. (2018). Distribution preserving learning for unsupervised feature selection. *Neurocomputing* 289, 231–240. doi: 10.1016/j.neucom.2018.02.032

Xu, J., Zhou, Y., Chen, L., and Xu, B. (2012). An unsupervised feature selection approach based on mutual information. *J. Comput. Res. Develop.* 49, 372–382.

Yang, Y., Shen, H., Ma, Z., Huang, Z., and Zhou, X. (2011). "l2, 1-norm regularized discriminative feature selection for unsupervised learning," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, (Catalonia, Spain: AAAI Press).

Ye, Y., Zhang, R., Zheng, W., Liu, S., and Zhou, F. (2017). RIFS: a randomly restarted incremental feature selection algorithm. *Sci Rep.* 7:13013. doi: 10.1038/s41598-017-13259-6

Yu, Y., Chen, Y., Ding, G., Wang, M., Wu, H., Xu, L., et al. (2015). A novel rabbit anti-hepatocyte growth factor monoclonal neutralizing antibody inhibits tumor growth in prostate cancer cells and mouse xenografts. *Biochem. Biophys. Res. Commun.* 464, 154–160. doi: 10.1016/j.bbrc.2015.06.107

Zhao, Z., and Liu, H. (2007). "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th International Conference on Machine Learning*, (New York, NY: ACM)), doi: 10.1145/1273496.1273641

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

frontiers in Genetics

METHODS
published: 23 June 2021
doi: 10.3389/fgene.2021.689515

Overlapping Structures Detection in Protein-Protein Interaction Networks Using Community Detection Algorithm Based on Neighbor Clustering Coefficient

Yan Wang[1,2], Qiong Chen[1], Lili Yang[1,3], Sen Yang[1], Kai He[1] and Xuping Xie[1]*

[1] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China, [2] School of Artificial Intelligence, Jilin University, Changchun, China, [3] Department of Obstetrics, The First Hospital of Jilin University, Changchun, China

OPEN ACCESS

Edited by:
Jianing Xi, Northwestern Polytechnical University, China

Reviewed by:
Hong-Yu Zhang, Huazhong Agricultural University, China
Jun Meng, Dalian University of Technology, China
Yuxuan Hu, Xidian University, China

***Correspondence:**
Lili Yang
yljlu@jlu.edu.cn

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

Received: *01 April 2021*
Accepted: *31 May 2021*
Published: *23 June 2021*

Citation:
Wang Y, Chen Q, Yang L, Yang S, He K and Xie X (2021) Overlapping Structures Detection in Protein-Protein Interaction Networks Using Community Detection Algorithm Based on Neighbor Clustering Coefficient. Front. Genet. 12:689515. doi: 10.3389/fgene.2021.689515

With the rapid development of bioinformatics, researchers have applied community detection algorithms to detect functional modules in protein-protein interaction (PPI) networks that can predict the function of unknown proteins at the molecular level and further reveal the regularity of cell activity. Clusters in a PPI network may overlap where a protein is involved in multiple functional modules. To identify overlapping structures in protein functional modules, this paper proposes a novel overlapping community detection algorithm based on the neighboring local clustering coefficient (NLC). The contributions of the NLC algorithm are threefold: (i) Combine the edge-based community detection method with local expansion in seed selection and the local clustering coefficient of neighboring nodes to improve the accuracy of seed selection; (ii) A method of measuring the distance between edges is improved to make the result of community division more accurate; (iii) A community optimization strategy for the excessive overlapping nodes makes the overlapping structure more reasonable. The experimental results on standard networks, Lancichinetti-Fortunato-Radicchi (LFR) benchmark networks and PPI networks show that the NLC algorithm can improve the Extended modularity (EQ) value and Normalized Mutual Information (NMI) value of the community division, which verifies that the algorithm can not only detect reasonable communities but also identify overlapping structures in networks.

Keywords: protein-protein interaction network, overlapping structure, clustering coefficient, community detection, central edge

INTRODUCTION

Due to the rapid development of experimental and computing technology, a large number of PPI networks have been mined (Chen et al., 2020). Previous studies have reported that a PPI network can be constructed as a scale-free complex network and satisfies small-world property and high degree of clustering (Ji et al., 2012). Biological functions are performed by many functionally related proteins. Such clustering proteins are called functional module. A module represents a group of

proteins taking part in specific, separable functions such as protein complexes, metabolic pathways or signal transduction systems (Vella et al., 2018). Lots of overlapping structures are shared by the functional modules in PPI networks, indicating some proteins play indispensable roles in different biological processes (Gu et al., 2019). Research on detecting protein functional modules has become one of the most important topics in both life science and computing science since the completion of the Human Genome Project (Ying and Lin, 2020). Detecting overlapping structures in functional modules have good application prospects in protein biological function, disease-causing gene, and drug target prediction.

In recent years, many researchers have designed a large number of algorithms that use the computational methodology to identify overlapping structures in modules. Among myriads of such efforts, network clustering is one of the most popular approaches for analyzing the topological and functional properties of PPI networks (Bhowmick and Seah, 2015). For example, the cluster percolation method (CPM) was the first method to discover overlapping communities. Its main idea was to determine k-connected subgraphs in the network and regard k-connected subgraphs as communities (Palla et al., 2005). By setting different k values, communities with different sizes can be obtained. The clustered communities will overlap, but the division result depends on the value of parameter k. Another common strategy used for community detection was based on edge division. This idea was initially used by Ahn et al., who proposed the classic link clustering community detection algorithm (LC) (Ahn et al., 2010). The LC algorithm first used the classical Jacard distance formula to quantify the distance between edges. The hierarchical clustering method was used to obtain the hierarchical structure of the community, and then the hierarchical structure was cut using the division function of density. Although there were many overlapping structures in the LC algorithm, the division result was quite different from the real community structure. In 2016, based on the density peak clustering algorithm, Huang et al. proposed a novel node distance measurement based on node similarity and the shortest distance between nodes, which could measure the global distance in the network, and applied the density peak clustering algorithm to the community of the detection network structure (Huang et al., 2016). In 2017, Qi et al. (2017). proposed an overlapping community detection algorithm based on the selection of seed nodes (CNS). The two main processes of the CNS algorithm were the selection of the central node and the clustering process. In 2018, Zhang et al. proposed an overlapping community discovery algorithm based on central edge selection (CES) (Zhang et al., 2018). The algorithm introduced the theory of community magnetic interference (CMI), which reduced the probability of the neighboring nodes becoming a central node and made the target central node reliable. Nevertheless, the division result was not sufficiently accurate.

Though the detection of functional modules in PPI networks has aroused widespread attention over the past few years, how to design correct and effective functional module detection methods is still a challenging and important scientific problem in computational biology (Mao and Liu, 2020). One of the main obstacles in community discovery is the accuracy of the division results. To improve the accuracy of community division, this paper proposes an overlapping community detection algorithm based on the neighbor local clustering coefficient (NLC) to select the central edge. The NLC algorithm introduces the clustering coefficient to improve the selection of seeds and optimizes the method of transforming the central node into a central edge set. This actually combines the advantages of the method of selecting seeds based on nodes and those of dividing communities based on edges. In the process of dividing non-central edges, the Jacard distance and the shortest distance between edges are combined to measure the distance between nonadjacent edges. Finally, the community is optimized, and a new pruning method is proposed for excessively overlapping nodes to make the division result more consistent for the real network. The NLC algorithm is applied to networks with real partitions and compared with classic algorithms and recent algorithms in terms of NMI, EQ and coverage rate (CR). The NLC algorithm gives slightly superior results compared to those of other algorithms. The results confirm that the algorithm can be used to find overlapping community structures in complex networks. Then, the algorithm was applied to the PPI networks to determine the overlapping community structures and perform functional enrichment analysis. The results of the enrichment analysis show that we can use the NLC algorithm to predict the function of the proteins in the PPI networks and find the overlapping structures in the protein functional modules.

METHODS

Complex networks are usually represented as graphs with nodes and edges. In a graph $\mathbf{G} = (\mathbf{V},\mathbf{E})$, $\mathbf{V}$ represents a set of nodes and $\mathbf{E}$ represents a set of edges.

Community Detection Algorithm Based on Central Edge Selection (CES)

In 2019, Zhang et al. proposed the CES algorithm based on center-edge selection theory. It is necessary to briefly describe the basic idea of the CES algorithm, which consists of 3 steps: central edge selection, community division, and overlapping node pruning.

In the first step, the community magnetic interference theory (CMI) was used to improve the seed selection. In fact, this theory reduced the influence F of the neighboring nodes of the central node. The definition of the influence F was set as the following formula:

$$F(v) = GF \times \sum_{u \in N(v)} IB(v, u) \tag{1}$$

$$IB(n1, n2) = \frac{D(n1) \times D(n2)}{(1 - sim(n1, n2))^2} \tag{2}$$

where $\mathbf{N}(v) = \{u | u \in \mathbf{V}, (v, u) \in \mathbf{E}\}$, $IB\,(n1, n2)$ is the influence between the node $n1$ and the node $n2$, GF is the coefficient of CMI theory used to revise the value of F, $D(n1)$ represents the degree of node $n1$, $D(n2)$ is the degree of node $n2$, and

$sim(n1, n2) = \frac{|\mathbf{N}(n1) \cap \mathbf{N}(n2)|}{|\mathbf{N}(n1) \cup \mathbf{N}(n2)|}$ represents the similarity between node $n1$ and node $n2$. The Formula (2) is derived from the universal gravitation formula $G = \frac{m_1 \times m_2}{r^2}$.

The second step was to cluster non-central edges to the corresponding communities. This process was mainly divided according to the distance formula between edges. After the completion of edge division, the nodes were divided according to the edge division results.

$$DNC(e_k, \mathbf{CE}_i) = \sum_{e_j \in \mathbf{CE}_i} \frac{ELC(e_k, e_j) \times (\sum_{e_m \in \mathbf{CE}_i} ELC(e_k, e_m) - ELC(e_k, e_j))}{\sum_{e_m \in \mathbf{CE}_i} ELC(e_k, e_m)} \quad (3)$$

where $DNC(e_k, \mathbf{CE}_i)$ represents the distance between edge e_k and central edge set $\mathbf{CE}_i$; e_j and e_m are the edges contained in the central edge set; and $ELC(e_k, e_j)$ represents the similarity between edge e_k and edge e_j, which is defined by the following formula.

$$\begin{aligned} ELC(e_k, e_j) &= ELC(e(a, b), e(c, d)) \\ &= \left| \frac{\mathbf{N}(a) \cap \mathbf{N}(c) + \mathbf{N}(a) \cap \mathbf{N}(d) + \mathbf{N}(b) \cap \mathbf{N}(c) + \mathbf{N}(b) \cap \mathbf{N}(d)}{\mathbf{N}(a) \cup \mathbf{N}(c) + \mathbf{N}(a) \cup \mathbf{N}(d) + \mathbf{N}(b) \cup \mathbf{N}(c) + \mathbf{N}(b) \cup \mathbf{N}(d)} \right| \end{aligned} \quad (4)$$

The last step was pruning overlapping nodes. For all overlapping nodes, the proportion of non-central edges in the connection between the overlapping node and all communities was calculated and compared with a threshold. If the proportion was greater than the threshold, we could determine that the overlapping node belonged to the community.

Limitation of CES

For the CES algorithm, some details need to be optimized. The selection of the seed node is not sufficiently accurate. In the process of clustering non-central edges, CES divides the non-central edges into the central edge set with the minimum distance. When measuring the distance between edges, the CES algorithm can only calculate the distance between edges where the topological distance is less than 3. For instance, in a small benchmark network containing 2 central edge sets and some non-central edges as **Figure 1** shown. According to Formulas (3, 4), $DNC(e(1, 2), \mathbf{CE}_1) = 0$, $DNC(e(1, 2), \mathbf{CE}_2) = 0$. But $DNC(e(1, 2), \mathbf{CE}_1)$ should be smaller than $DNC(e(1, 2), \mathbf{CE}_2)$ because $e(1, 2)$ is closer to $\mathbf{CE}_1$. And $e(1, 2)$ should be divided into community 2. The CES algorithm cannot give a reasonable solution. In the pruning process of overlapping nodes, only the connection between the overlapping nodes and the central node are considered, but most nodes in the network are not connected with the central node. Further research is still important.

Community Detection Algorithm Based on the Neighbor Local Clustering Coefficient (NLC)

To avoid the limitations of CES, we proposed the NLC algorithm. The algorithm combined the seed-based community detection algorithm with the edge-based community detection algorithm, which mainly consisted of four processes: (1) seed selection, (2) transformation from the central node set to the central edge set, (3) expansion of the non-central edge set, and (4) community optimization. The algorithm flow chart is as **Figure 2** shown.

Central Node Selection

The selection of the central node could affect the result of community division. Inspired by the previous method of selecting central nodes, this paper introduced the neighbor local clustering coefficient and proposed a more reasonable selection method of central nodes. The process of selecting seeds is as follows:

1. First, the influence F of each node was calculated. If a node had a greater F than its neighboring nodes, then the node was considered as a candidate central node. The influence F of node u in the network is defined as the following formula.

$$F(u) = \sum_{\substack{v \in N(u) \\ v \neq u}} \frac{D(u) \times (1 + C(u)) \times D(v) \times (1 + C(v))}{(1 - sim(u, v))^2} \quad (5)$$

where $C(u)$ is the local clustering coefficient between the neighbors of node u. The local clustering coefficient quantified the clustering of neighboring nodes to form a cluster (complete graph). The clustering coefficient was defined as the following formula:

$$\begin{cases} C(u) = \frac{2K(u)}{|N(u)| \times (|N(u)| - 1)}, & |\mathbf{N}(u)| \neq 1 \\ C(u) = 0, & |\mathbf{N}(u)| = 1 \end{cases} \quad (6)$$

where $K(u)$ represents the number of connections in the neighbor nodes. As shown in **Figure 1**, the calculation process of $C(12)$ is as follows: $\mathbf{N}(12) = \{8, 11, 13, 14, 15, 16\}$, the connected edges in the neighbors of node 12 are $e(8, 13)$ and $e(11, 16)$, so $K(12) = 2$, $C(12) = \frac{2 \times 2}{6 \times (6-1)} = 0.13$.

2. A good community division results in close connections within the community and sparse connections between communities. Therefore, only when the similarity between the candidate central node and each central node is less than the threshold α, can the candidate central node be added to the central node set **CN**; that is, if $sim(n1, n2) \leq \alpha$ $n1 \in \mathbf{CN}$,

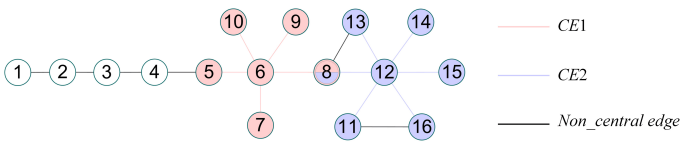

FIGURE 1 | A simple network.

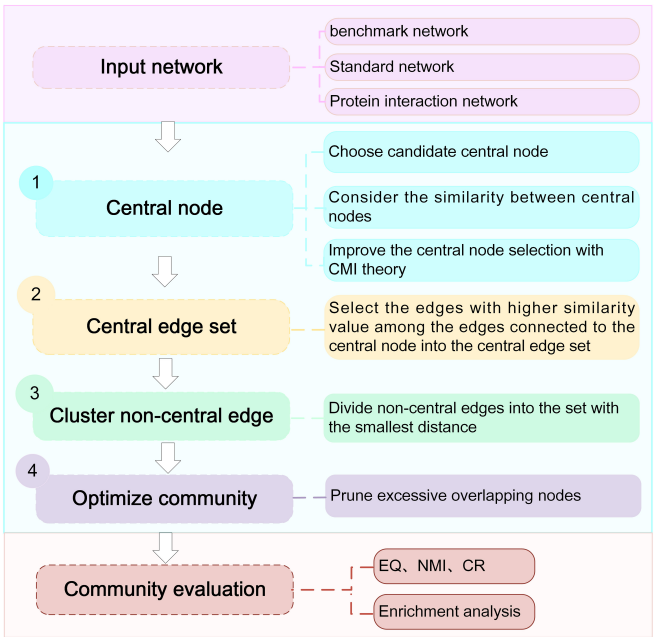

FIGURE 2 | Workflow of the NLC algorithm.

then $\mathbf{CN} = \mathbf{CN} \cup \{\mathbf{n2}\}$, where $n1$ is a central node and $n2$ is a candidate central node.

3. The CMI theory in the CES algorithm was used to revise the weights of the neighbors of the central node, which reduced the possibility of the neighbors becoming the central node. We confirmed that the CMI theory could improve the selection of seed nodes.

The Transformation of the Central Node Set to the Central Edge Set

After selecting the central node, we chose edges that connect to central node and the similarity between two vertices was greater than the average similarity. As shown in Formulas (7, 8).

$$\mathbf{CE} = \{(u, v) | u \in \mathbf{CN} \; and \; \mathrm{sim}(u,v) > \mathrm{ave\_sim}(u)\} \tag{7}$$

$$\mathrm{ave\_}sim(u) = \frac{1}{|\mathbf{N}(u)|} \sum_{v \in N(u)} sim(u, v) \tag{8}$$

where *ave_sim* represents the average similarity between node u and its neighboring nodes, **CE** represents the central edge set, and **CN** represents the central node set. It can be concluded that each central node corresponds to a central edge set. The remaining edges are called non-central edges.

The procedure of central node selection can be described as follows:

Algorithm 1. Central node selection procedure.

Input:
Network: **G** = (**V,E**).
Output:
Central Node Set: **CN**

1. Calculate nodes similarity matrix and clustering coefficient matrix
2. For each $u \in V$do
3. $\mathrm{F}(u) = \sum_{\substack{v \in \mathbf{N}(u) \\ v \neq u}} \frac{D(u) \times (1 + C(u)) \times D(v) \times (1 + C(v))}{(1 - sim(u,v))^2}$
4. End for
5. For each $n \in \mathbf{V}$ do
6. If $F(n) \geq F(\mathbf{N}(n))$ and $sim(n, v) \leq \alpha$, $v \in \mathbf{CN}$ then
7. $\mathbf{CN} = \mathbf{CN} \cup n$
8. End if
9. For each $v \in \mathbf{CN}$
10. $F(v) = GF \times \sum_{u \in N(v)} IB(v, u)$
11. End for
12. End for

Clustering of Non-central Edge

After the central edge set was obtained, the remaining non-central edges could be clustered. The strategies of clustering non-central edges were as follows:

The distance between edges were calculated according to the Formula (9) and then the distance between the non-central edges and each central edge set was calculated. The non-central edge was clustered into the central edge set with the smallest average distance.

$$Dis(a, b) = Jacard(a, b) \times link(a, b) \tag{9}$$

$$DLS(e, \mathbf{CE}) = \frac{\sum_{v \in \mathbf{CE}} Dis(e, v)}{|\mathbf{CE}|} \tag{10}$$

where $Jacard(a, b)$ represents the *Jacard* distance of edge (a, b), $link(a, b)$ is the topological distance of edge (a, b), and

$DLS(e, \mathbf{CE})$ represents the average distance between edge e and central edge set **CE**. To date, the community clustering of edges had been formed. Our next step was to transform the community division of edges into a community division of nodes. If multiple edges connected to a node belong to multiple communities simultaneously, then this node can be considered as an overlapping node, as shown in node 8 in **Figure 1**.

Community Optimization

There were a large number of overlapping nodes in the edge clustering results obtained in the previous process. Therefore, this paper proposed the following method to optimize these excessive overlapping nodes. We only needed to adjust the overlapping nodes in each community. So, the non-overlapping nodes were regarded as the divided parts, and the community was optimized by continuously reducing unnecessary overlapping nodes. The strategies were as follows, and the specific details were shown in algorithm 2.

The proportion of connection between the overlapping nodes and divided parts in each community was calculated. If the proportion was less than the pruning threshold *prune*, the overlapping node did not belong to the community; that is $\frac{con(n,\mathbf{Non\_\ overlap}_j)}{\sum_{k\in clus(n)} con(n,\mathbf{Non\_\ overlap}_k)} < prune$, where $n \notin j$, $\mathbf{Non\_\ overlap}_j$ represents the set of non-overlapping nodes in community j, $con(n, \mathbf{Non\_\ overlap}_j)$ represents the number of connections between overlapping node n and non-overlapping parts in the community j, and $clus(n)$ represents the community to which overlapping node n belongs. If the connection proportion between the overlapping node and each community was less than the threshold *prune*, the overlapping node was only divided into the community with the largest connection proportion. If the size of the community was less than 3, the community was not pruned.

Algorithm 2. Community optimization procedure.

Input:
Community division with excessively overlapping nodes
Output:
Optimized community results

1. For each $n \in$ **overlapping_ node** do
2. For $j \in clus(n)$ do
3. Calculate $con(n, \mathbf{Non\_\ overlap}_j)$
4. $all\_con(n) += con(n, \mathbf{Non\_\ overlap}_j)$
5. End for
6. For $j \in clus(n)$do
7. If $\frac{con(n,\mathbf{Non_\ overlap}_j)}{all_con(n)} < prune$then
8. $clus\_remove(j, clus(n))$
9. End if
10. End for
11. End for
12. Delete the community whose size is less than 3

Time Complexity Analysis

Assuming that the network contains n nodes and m edges, in the power-law distribution, the degree of each node satisfies the distribution $P(degree = k) \propto \frac{1}{k^\gamma}$, where k represents the degree of the node. When the degree of a node is k, the probability of the node may be $\frac{1}{k^\gamma}$. In 2001, Béla Bollobás et al. proposed that the value of γ in large networks is generally always 3 (Bollobás et al., 2001). Therefore, the probability of existence of a node with degree k is $\frac{1}{k^3}$. The average degree in the network is $D\_ave = 1\times\frac{1}{1^3} + 2\times\frac{1}{2^3} + \cdots + n\times\frac{1}{n^3}$. $\lim_{n\to\infty}(\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2}) = \frac{\pi^2}{6}$ (Dunham, 1999), so the total degree of all nodes is $DN = n \times D\_ave \le \frac{\pi^2}{6} \times n$. In a network, the sum of the degrees of all vertices is equal to twice the number of edges in the graph, that is, $m=\frac{DN}{2} \le \frac{\pi^2}{12} \times n$.

The first step of the NLC algorithm is to select the central nodes. First, we need to calculate the similarity between all nodes in the network, and the time complexity is $O(n^2)$. When choosing a central node, we need to access all nodes to calculate its *F*value and compare it with its neighboring nodes. The time is $O(\sum_{v\in V} D(v)) = DN \le \frac{\pi^2}{6} \times n$, where $D(v)$ is the degree of node v. The second step of the algorithm is to transform the central node set into a central edge set, and the time is $O(\sum_{v\in \mathbf{CN}} E(v)) \le |\mathbf{CN}| \times \frac{\pi^2}{6}$, where **CN** is the central node set and $E(v)$ is the size of the central edges of central node v. In the third step, the distance between edges in the network is calculated, and the time complexity is $O(m^2)$. The process of clustering non-central edges needs to calculate the distance between the non-central edges and each community, and the time complexity is $O(m\times|\mathbf{CN}|)$. Finally, the process of community optimization requires calculating the proportion of non-overlapping parts of all the neighbors of overlapping nodes in each community, and the time requires $\sum_{v\in \mathbf{overlapping\_\ node}} D(v)$. Through the above analysis, after omitting the constant of the highest order position, the time complexity of the NLC algorithm is $0(n^2)$.

RESULTS AND DISCUSSION

Datasets

(1) Standard networks

The standard networks used in this paper were Zachary's karate club (Zachary, 1977), American college football (Girvan and Newman, 2002), and books about US politics (polbooks) (Tang, 2014), which are all networks with standard divisions. The karate network is a social network of friendships between 34 members of a karate club at a US university in the 1970s. Each node represents a student, each edge represents the communication relationship between students, and each community represents a team led by a coach. The football network is a network of American football games between Division IA colleges during the regular Fall 2000 season. Each node represents a player, an edge represents a match between players, and a category represents a collection of teams. The polbooks network is a network of books about US politics

published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent co-purchasing of books by the same buyers. The specific conditions of each network are shown in **Table 1**, where NSC represents the number of standard communities.

(2) Benchmark networks

Compared with real world networks, artificial synthetic networks can more effectively measure the accuracy of detected community divisions because they can predict the real network micro characteristics and community divisions (Ren et al., 2019). This paper used the LRF benchmark network to synthesize the network, which was a benchmark method for testing the performance of the algorithm found in the community (Lancichinetti et al., 2008). LFR networks have multiple parameters to control the structure and scale of the synthesized network. The commonly used parameters in LFR are N (number of nodes), K (average degree, the average degree of most large-scale real social networks is approximately 10), Maxk (maximum degree), Mu (mixing parameter), On (number of overlapping nodes), and Om (number of memberships of the overlapping nodes). In this paper, there were 6 networks used for experiments, including two types of networks with different numbers of nodes and different overlapping ratios. The specific parameters and the generated network information were shown in **Table 2**. The visualization results of standard networks and LFR networks were shown in **Figure 3**. The visualization of the network in this paper was drawn by Cytoscape (Shannon et al., 2003). In the visualized results, different colors represented different communities.

(3) PPI networks

The PPI networks used in the experiment were all downloaded from the database of interacting proteins (DIP) (Salwinski et al., 2004). The reliability of the DIP database is high, because it only stores protein interaction verified by experiments, and provides experimental methods used to identify the interaction. The DIP lists protein pairs that are known to interact with each other and are composed of nodes and edges. The nodes represent proteins, and the edges represent the interactions between proteins. The downloaded network version was 20170205 and the downloaded networks were: *M. musculus*, *H. sapiens*, *D. melanogaster*, and *R. norvegicus*. These networks are real networks without standard communities. The unprocessed PPI networks contain some redundant edges and some small structures, so this noise needed to be processed in data processing: the self-circulating edges in the network and some modules with a small scale were removed. The number of nodes and edges of the PPI networks before and after data preprocessing are shown in **Table 3**.

TABLE 1 | Standard networks.

Standard networks	|E|	|V|	NSC
Karate	78	34	2
Football	612	115	12
Polbooks	105	441	3

TABLE 2 | LFR benchmark networks.

LFR benchmark networks	|V|	Maxk	Mu	K	On	Om	|E|	NSC
LFR1	80	15	0.1	10	4	1	764	7
LFR2	80	15	0.1	10	4	2	740	8
LFR3	80	15	0.1	10	4	3	778	8
LFR4	150	15	0.1	10	8	1	1,418	12
LFR5	150	15	0.1	10	8	2	1,478	14
LFR6	150	15	0.1	10	8	3	1,426	17

Evaluation Metrics

To verify whether the community structure detected by the algorithm was reasonable, the algorithm was compared with the CES algorithm, CNS algorithm, CPM algorithm and LC algorithm. The CES algorithm was an edge partition-based algorithm proposed in 2019, and CNS was an algorithm based on node division proposed in 2017. The CPM algorithm and LC algorithm were relatively classic algorithms in the field of overlapping community discovery. The software CFinder (version 2.0.6) is a free software for finding and visualizing overlapping communities, based on the CPM. The clustering result of LC algorithm was obtained by the linkcommon package which includes tools for generating, visualizing, and analyzing overlapping communities (Kalinka and Tomancak, 2011). These algorithms were compared and analyzed with the standard networks, LFR synthesis networks and PPI networks to evaluate the accuracy of this algorithm. To evaluate the performance of the NLC algorithm, we used the following 5 evaluation indicators.

(1) Extended modularity (EQ)

Since the community structure of the complex network was unknown in advance, a metric was needed to measure the community results detected by different community detection algorithms. In this paper, the extended modularity (EQ) (Shen et al., 2009) evaluated the results of overlapping community detections. The value of EQ can be calculated by Formula (11):

$$EQ = \frac{1}{2|\mathbf{E}|}\sum_{i=1}^{|C|}\sum_{v,w\in C}\frac{1}{Ov Ow}\left(A_{vw} - \frac{D_v D_w}{2|\mathbf{E}|}\right) \quad (11)$$

where $|\mathbf{C}|$ represents the number of communities detected, Ov represents the number of communities to which node v belongs, Ow represents the number of communities to which node w belongs, and Avw varies according to different situations: when the node v is connected to the node w, $Avw = 1$; otherwise, $Avw = 0$. The EQ value is between 0 and 1, and a larger value is better.

(2) Normalized Mutual Information (NMI)

The normalized mutual information (NMI) used in this paper is proposed by Lancichinetti et al. (2009) and widely used in overlapping community evaluations, which is defined as the following formula:

$$NMI(R|P) = 1 - [H(R|P) + H(P|R)]/2 \quad (12)$$

where R is the real community, P is the predicted division result, and $H(R|P)$ is the normalized conditional entropy of R with respect to P. The NMI value is between 0 and 1; the closer the

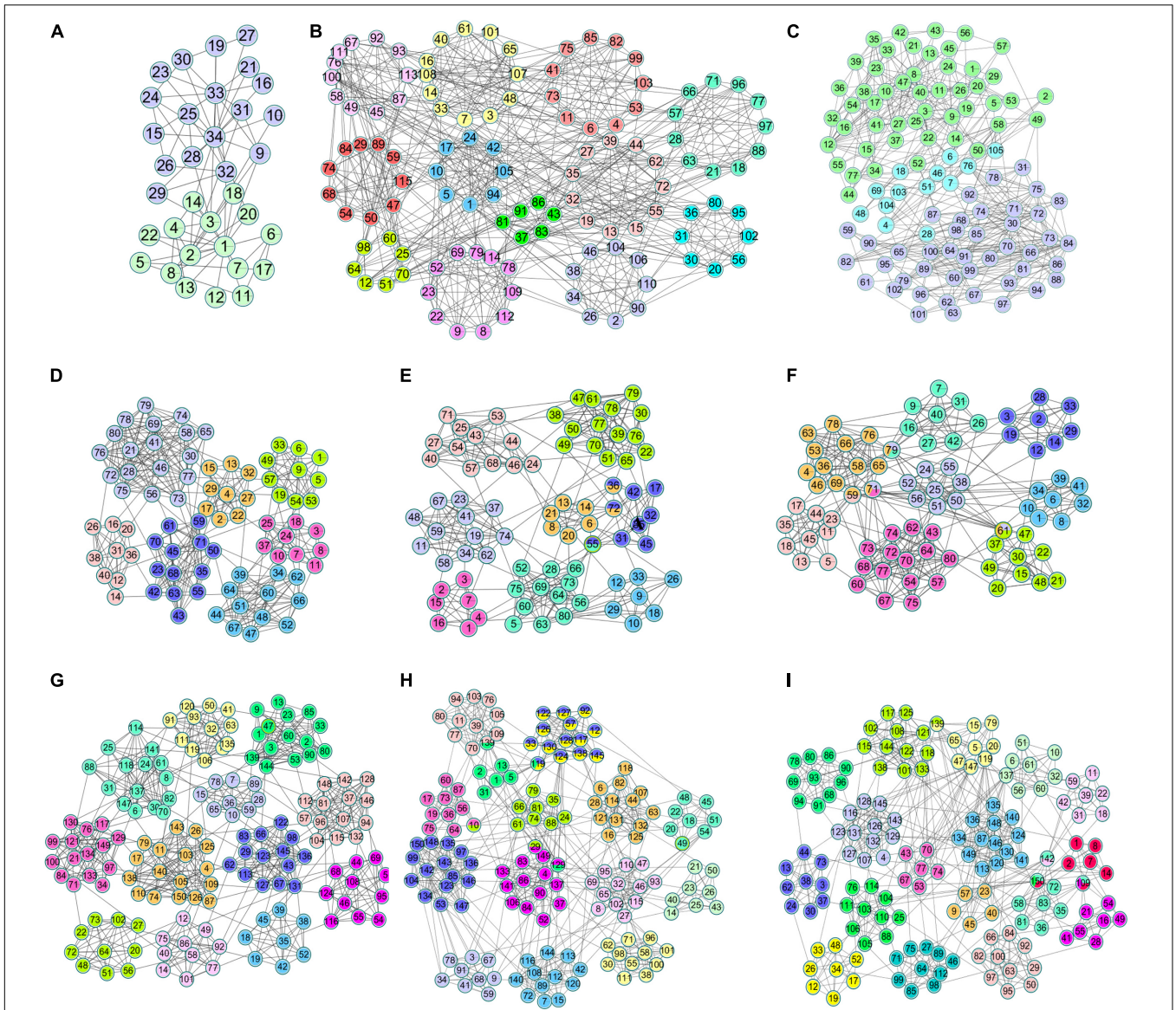

FIGURE 3 | The visualization of standard networks and LFR networks. Standard networks **(A)** Karate network, **(B)** Football network, **(C)** Polbooks network; LFR networks: **(D)** LFR1 network, **(E)** LFR2 network, **(F)** LFR3 network, **(G)** LFR4 network, **(H)** LFR5 network, **(I)** LFR6 network.

value is to 1, the closer it is to the real community. The NMI value is 1 when the result of community division matches the real community completely.

(3) Coverage Rate (CR)

The coverage rate is used to evaluate the coverage of community detection, which is defined as the following formula.

$$CR = \frac{n'}{n} \cdot 100\% \tag{13}$$

where n' represents the number of nodes detected by the community detection algorithm and n represents the total number of nodes in the network.

(4) Number of Normalized Communities (NNC)

The NNC is used to evaluate the difference between the true and predicted values, which is defined as Formula (14):

$$NNC = \boldsymbol{max}(1 - \frac{|NSC - NPC|}{NSC}, 0) \tag{14}$$

where NSC represents the number of standard communities and NPC represents the number of communities predicted by algorithms. The NNC value is between 0 and 1; the closer the value is to 1, the closer it is to the number of standard communities. When the NNC value is 1, the predicted number of communities is consistent with the actual number of communities.

(5) Enrichment analysis

To detect whether the community detected by the algorithm has biological significance, functional enrichment analysis of the protein community is necessary. Enrichment analysis of a gene set refers to comparing the gene set to a database that is classified and annotated according to prior knowledge, using the hypergeometric distribution algorithm to obtain the gene ontology terms with significant enrichment of genes of the gene set. The gene ontology term corresponding to the smallest *p*-value was used as the functional annotation of the protein community. Among these databases, Gene Ontology (GO) (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) are commonly used. The GO annotation contains three indicators: biological process (BP), cellular component (CC) and molecular function (MF). BP describes the biological processes in which proteins are involved. CC describes the location of the proteins in the cell for biological activity. MF describes the biochemical activity of proteins. KEGG provides a complete metabolic pathway, including the metabolism of carbohydrates, nucleosides, and amino acids. and the biodegradation of organic matter. The values of the above four indicators are all expressed by *p*-values, where the closer the *p*-value is to 0, the more significant the biological significance of the divided communities is. During the experiment, the cluster profiler of the R package was used for enrichment analysis (Yu et al., 2012).

Experimental Setup

Parameters in the NLC Algorithm

The algorithm proposed in this paper mainly involves three parameters, which are the community magnetic interference coefficient *GF*, similarity threshold α and pruning coefficient *prune*. The similarity threshold α prevented excessive similarity between two communities. The similarity threshold α was tested between 0 and 1. According to experimental experience, in the karate network $\alpha = 0.14$; in the football network $\alpha = 0.30$; and in the polbooks network $\alpha = 0.10$; the threshold α values of four PPI networks were set to 0.1. The parameter *GF*was used to control the centrality of the neighboring nodes of the central node and was set as $GF = \frac{link\_num}{node\_num}$. The pruning coefficient *prune* reduced the excessive overlapping nodes, and the value was between 0 and 1 for the experiment. The relationship between coefficient *prune*, EQ, NMI and overlapping rate (OR) in the standard networks and PPI networks were shown in **Figures 4**, **5**. The OR was used to describe the proportion of overlapping nodes in the community.

As is demonstrated in **Figure 4**, we can see that different values of parameter *prune* can have various influences on the experiment result. The selection of *prune* was based on the EQ, NMI and *OR*-values. If there is no overlapping structure in the network, we only need to select the corresponding parameters when the EQ and NMI values are relatively good; if there is an overlapping structure in the network, we also need to consider the overlapping structure in the network. Finally, in the three networks of karate, football, and polbooks, *prune* were selected as 0.42, 0.30, 0.31, respectively.

Figure 5 shows the relationship between EQ, OR and *prune* in the PPI networks. The selection of *prune* was based on the value of EQ and OR. In the PPI networks, some proteins have multiple functions and form protein overlapping nodes. Hence, we need to maintain some overlapping structures while the value of EQ is high. In the four PPI networks, the *prune* values were set as 0.32.

The Experimental Results on Networks With Standard Division

Figure 6 shows the clustering results of the karate, football and polbooks networks based on the NLC algorithm. Colors represent

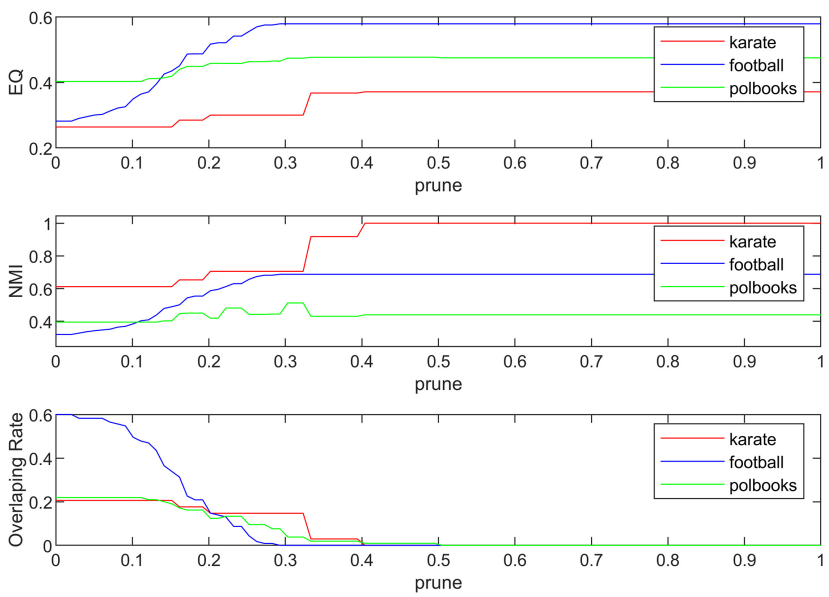

FIGURE 4 | The relationship between EQ, NMI, OR and *prune* in the standard networks.

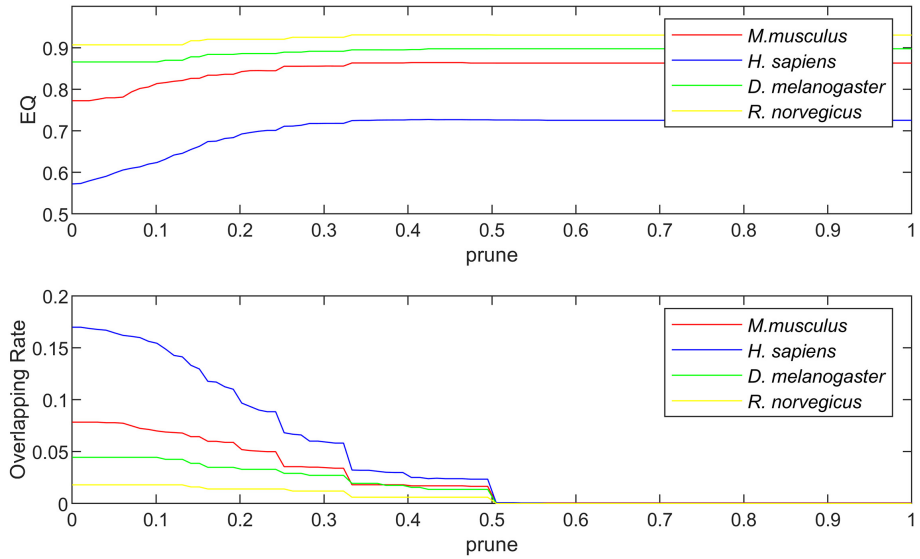

FIGURE 5 | The relationship between EQ, OR and *prune* in the PPI networks.

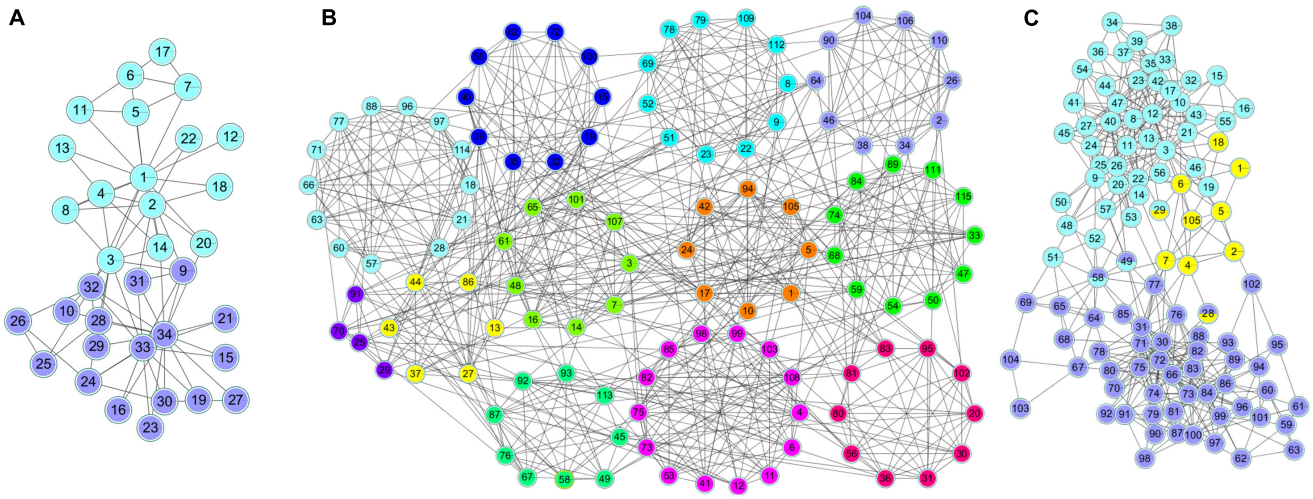

FIGURE 6 | The division results of the NLC algorithm in three standard networks: **(A)** Karate, **(B)** Football, and **(C)** Polbooks network.

communities, and nodes with overlapping colors represent that they can belong to multiple communities.

The NLC algorithm was compared with other four algorithms including CES, CNS, CPM, and LC, by comparing the EQ, NMI, CR and NNC values in networks with standard division: the karate, football, polbooks, and LFR networks. The results were shown in **Figure 7**.

In the CPM algorithm, we set the value of k as 3 in the karate, football and polbooks networks; we set the k as 5 in the LFR networks. In the CES algorithm, we set the coefficient $GF = 4.2\times\frac{link\_num}{node\_num}$ in three standard networks. The parameter in the CNS algorithm was set to 0.4 according to Qi (Qi et al., 2017). In the karate, football, polbooks, LFR 3 and LFR 6 networks, the number of predicted communities (NPC) by the LC algorithm were quite different from the actual number of communities, so the NNC values were 0 in these networks. And the values of NPC obtained by five algorithms were shown in the **Supplementary Materials**. The NLC algorithm could completely pair the karate network and had the best EQ value and NMI value. In the football network, the CPM algorithm had the best *EQ*-value and *NMI*-value but the *NNC*-value was smaller than the NLC. In the polbooks network, the NLC algorithm also had the best EQ and NMI. The NLC algorithm could completely pair LFR synthetic networks (LFR1 and LFR4) without overlapping nodes. In the LFR3 network, the CPM algorithm had the highest *EQ*-value, but the NLC algorithm had the highest *NMI*-value. In the LC algorithm, The NLC algorithm not only had good division results in the LFR network with overlapping structures

but also could be applied in the non-overlapping networks. In general, the NLC algorithm had better division result than the other four algorithms.

The Experimental Results of PPI Networks

The calculation of NMI and NNC requires not only the predicted communities, but also the real communities. Since the real communities in the PPI networks is unknown, the NMI and NNC metrics cannot be calculated. The NLC algorithm was compared with the CES, CNS, CPM, and LC algorithms, by comparing the EQ, CR and NPC values in the four PPI networks: *M. musculus, H. sapiens, D. melanogaster* and *R. norvegicus*. The results were shown in **Figure 8**.

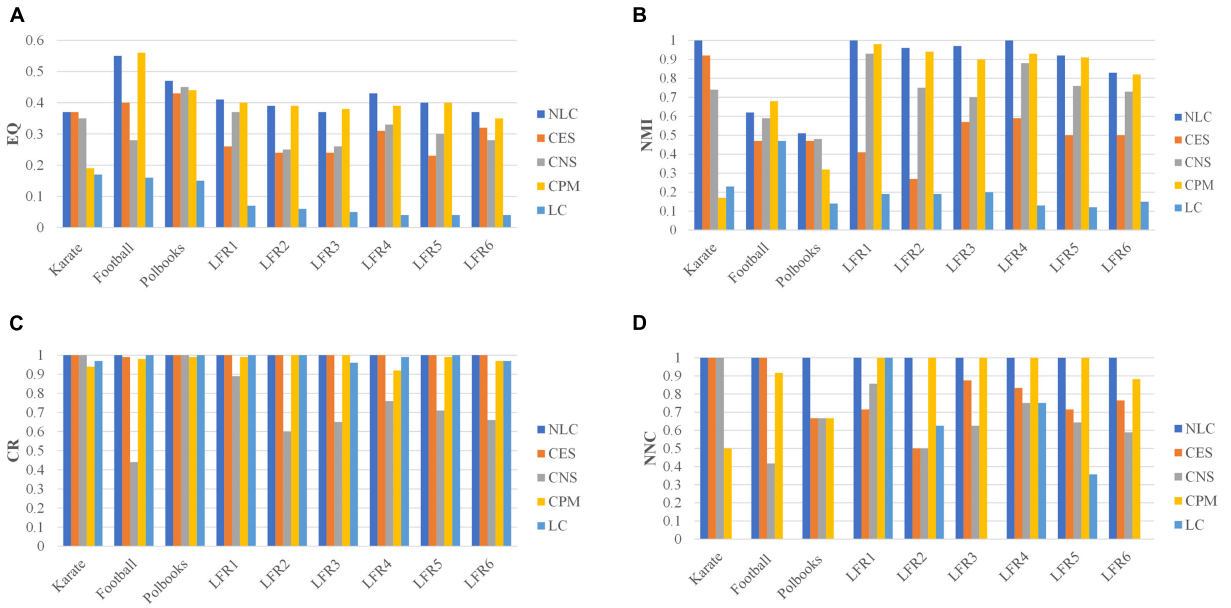

FIGURE 7 | The division results of five algorithms in standard networks and LFR networks including three indicators: **(A)** EQ, **(B)** NMI, **(C)** CR, and **(D)** NNC.

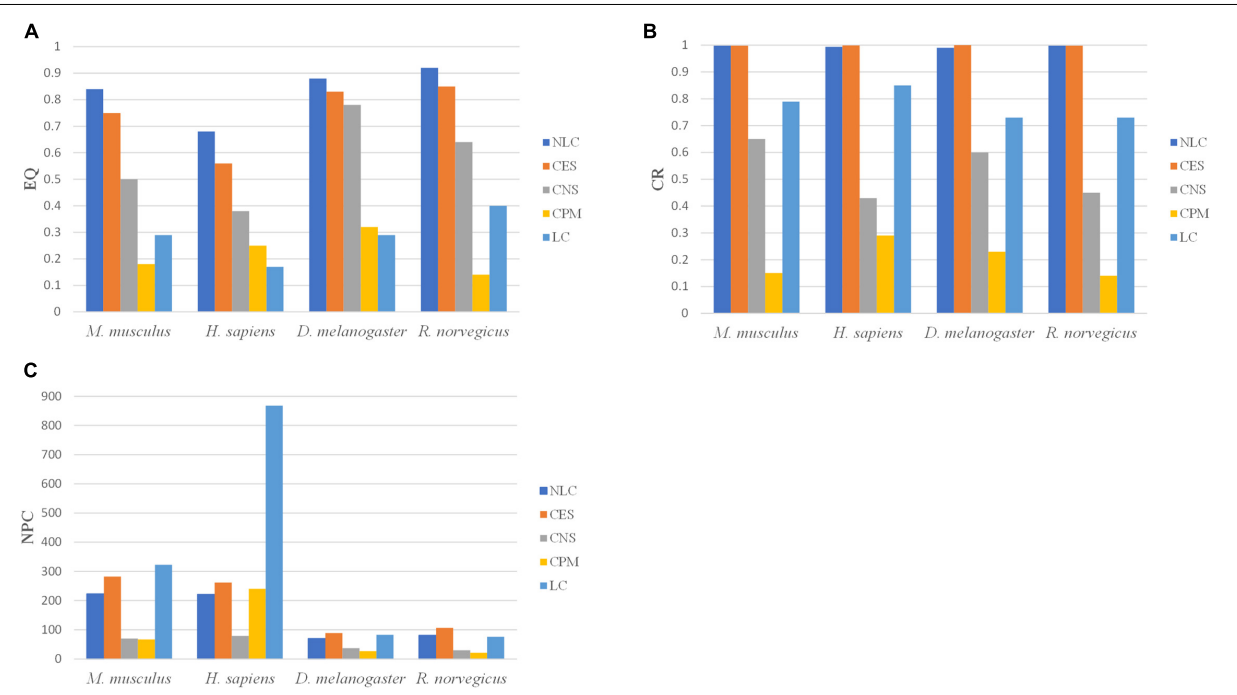

FIGURE 8 | The division results of five algorithms in PPI networks including three indicators: **(A)** EQ, **(B)** CR, and **(C)** NPC.

In the CPM algorithm, the parameter k was set to 3 in four PPI networks. In the CES algorithm, the parameter *GF* in *M. musculus*, *H. sapiens*, *D. melanogaster,* and *R. norvegicus* networks were set to 0.8, 0.3, 0.3, 0.8, respectively. The clustering results of the NLC had higher EQ and CR values in the four PPI networks than the other algorithms. The CPM and LC algorithms had the smallest EQ. The division categories of the algorithm proposed in this paper were always at an intermediate value when compared with other algorithms, indicating that the divided community structure obtained by this algorithm was relatively more reasonable. Moreover, the division effect of the developed algorithm was better than that of the other four algorithms from the perspective of EQ and CR values.

For enrichment analysis, it was necessary to calculate the p-value of the BP, MF, CC categories and KEGG pathways for each protein community, and the smallest p-value is selected as the result of enrichment analysis for a particular protein community. To better reflect the enrichment result of the protein community, communities with more than 2 proteins were left, because the communities with only two proteins are more likely to generate noise on the enrichment results. In our experiment, we set the threshold of the p-value as 0.05. Generally, the gene or protein was considered to be significantly expressed when the $p < 0.05$; otherwise, the community was regarded as an insignificant expression community. In **Figure 9**, a p-value threshold sequence of 1E-12 to 1E-03 was set, and the proportion

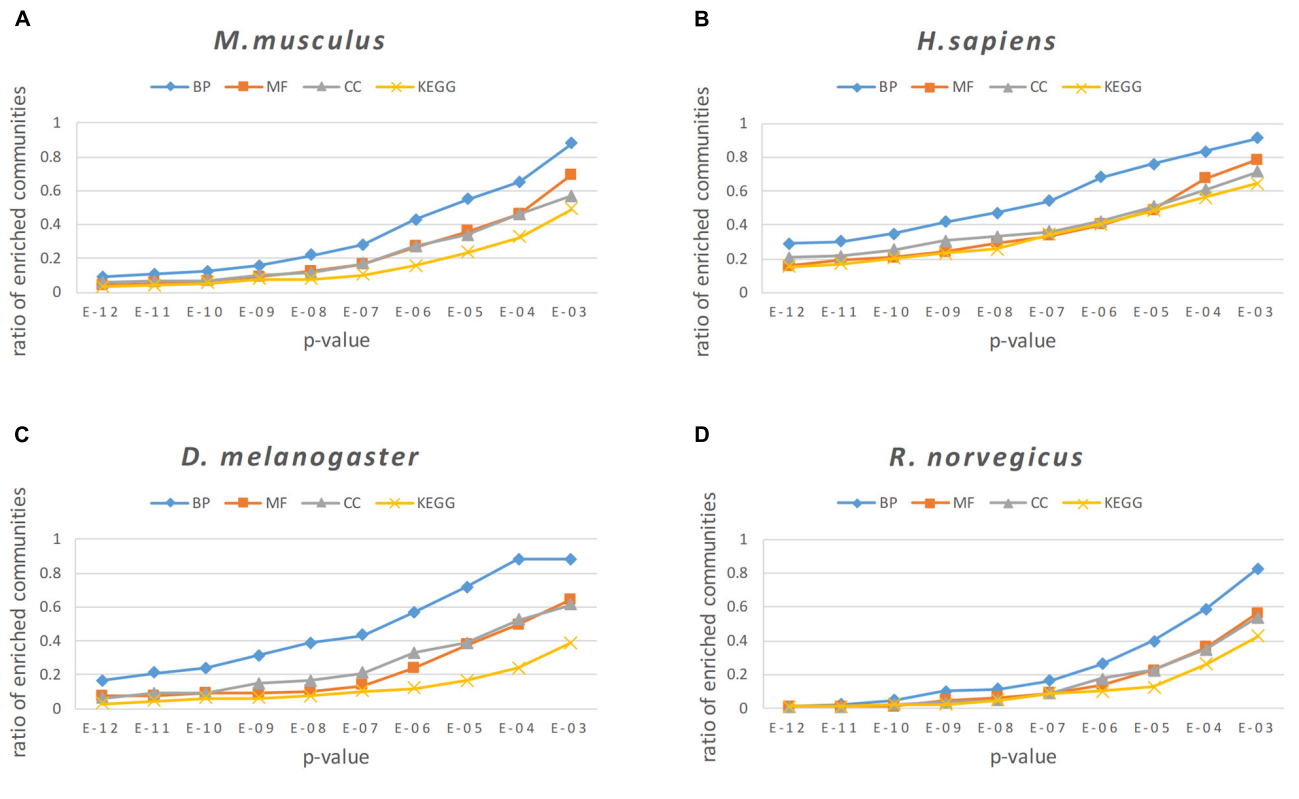

FIGURE 9 | The enrichment analysis results of the NLC in four PPI networks: **(A)** *M. musculus*, **(B)** *H. sapiens*, **(C)** *D. melanogaster,* and **(D)** *R. norvegicus.*

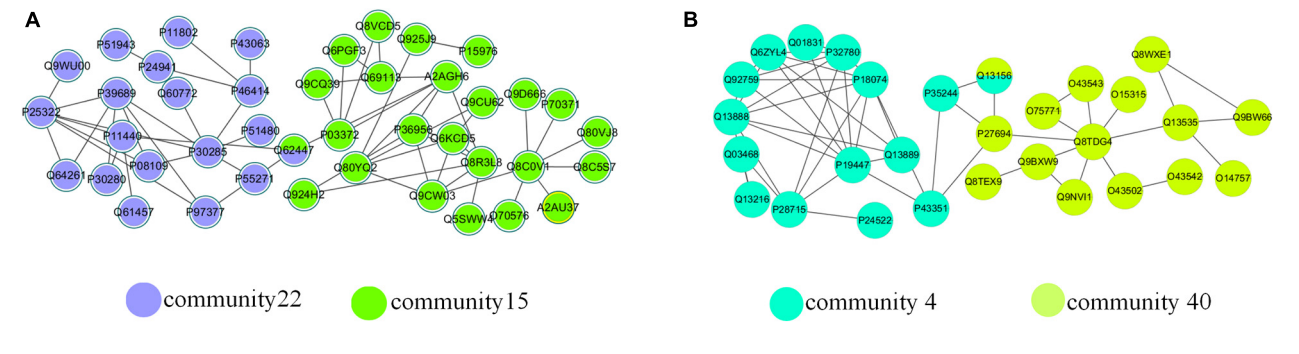

FIGURE 10 | Overlapping structures in **(A)** *M. musculus* network and **(B)** *H. sapiens* network.

TABLE 3 | The PPI networks.

	Before data preprocessing		After data preprocessing	
PPI networks	**|E|**	**|V|**	**|E|**	**|V|**
H. sapiens	7,380	4,670	6,699	4,200
M. musculus	2,597	2,329	2,319	2,006
D. melanogaster	711	626	614	518
R. norvegicus	619	665	497	504

of modules less than or equal to this *p*-value threshold was counted for all protein modules found in the four PPI networks.

From the results of enrichment analysis shown in **Figure 9**, the algorithm proposed in this paper obtains good enrichment results in the BP, MF, and CC classes and KEGG pathways. The BP analytical result was the best in the enrichment analysis, indicating that proteins in the protein community identified by the algorithm in this paper had a high degree of co-participation in biological processes. The BP analytical results show that 97.6% of the communities in the *M. musculus* network had a p-value $\leq 1E-02$, 87.4% communities had a p-value $\leq 1E-03$, and the proportion of communities with a p-value $\leq 1E-02$ in the three networks of *H. sapiens*, *D. melanogaster*, and *R. norvegicus* were 91.4, 88.1, 82.5%, respectively.

There were a large number of overlapping communities in the division results of the NLC algorithm. Taking the *M. musculus* and the *H. sapiens* networks as examples, **Table 5** and **Supplementary Table 1** list the enrichment analysis results of some overlapping communities divided by the NLC algorithm, including the GO ID enriched in the protein community and its functional description, which is the definition of GO terms. **Figure 10** depicts the visual results.

In **Table 4**, the ID is the unique identifier for the GO database or KEGG database. There was an overlapping node Q62447 between communities 15 and 22 divided by the NLC algorithm, and the corresponding protein name is Cyclin-C. Cyclin-C is a component of the mediator complex, which is a coactivator involved in the regulation of gene transcription of almost all RNA polymerase II-dependent genes. Its molecular function is related to the cyclin-dependent protein serine, and there are four biological processes related to Cyclin-C: negative regulation of triglyceride metabolism, positive regulation of RNA polymerase II transcription, protein ubiquitin chemical and RNA polymerase II regulates transcription (Gaudet et al., 2011). Through enrichment analysis, we found that the molecular function of community 15 was the activity of ubiquitin protein ligase, and the cell composition was related to the composition of the mediator complex. The protein Cyclin-C also had the function of community 15. The cellular component of community 22 was a cyclin-dependent protein kinase holoenzyme compound.

TABLE 4 | Enrichment analysis of an overlapping structure in the *M. musculus* network.

Overlapping communities in the *M. musculus* network	**Enrichment analysis**	***p*-value**	**ID**	**Name**
Community 15	BP	1.14E-14	GO:0098813	Nuclear chromosome segregation
	MF	2.26E-09	GO:0061630	Ubiquitin protein ligase activity
	CC	5.59E-18	GO:0016592	Mediator complex.
	KEGG	2.98E-06	mmu04114	Oocyte meiosis
Community 22	BP	8.80e-19	GO:0044843	Cell cycle G1/S phase transition.
	MF	9.69e-24	GO:0016538	Cyclin-dependent protein serine.
	CC	6.31e-21	GO:0000307	Cyclin-dependent protein kinase holoenzyme complex.
	KEGG	1.13e-22	mmu04110	Cell cycle.

TABLE 5 | Enrichment analysis of an overlapping structure in the *H. sapiens* network.

Overlapping communities in the *H. sapiens* network	**Enrichment analysis**	***p*-value**	**ID**	**Name**
Community 4	BP	4.23e-16	GO:0000724	Double-strand break repair via homologous recombination
	MF	1.67e-12	GO:0000400	Four-way junction DNA binding.
	CC	4.18e-15	GO:0033061	DNA recombinase mediator complex.
	KEGG	1.37e-13	hsa03440	Homologous recombination
Community 40	BP	3.38e-31	GO:0006289	Nucleotide-excision repair
	MF	7.17e-17	GO:0008353	RNA polymerase II CTD heptapeptide repeat kinase activity
	CC	2.20e-22	GO:0005675	Transcription factor TFIIH holo complex.
	KEGG	9.96e-31	hsa03420	Nucleotide excision repair

Its molecular function was to regulate the activity of cyclin-dependent protein serine/threonine kinase. The protein Cyclin-C also had the molecular function of community 22.

As we can see from the **Table 5**, the biological function of community 4 is DNA binding, and the biological function of community 40 is the excision of nucleotides. The overlapping node between community 4 and community 40 is Q13156, which corresponds to RPA4. The biological functions of RPA4 are participation in single-stranded DNA binding, DNA replication and repair, double-strand break repair via homology, DNA damage checkpoint (Haring et al., 2010), DNA replication initiation (Keshav et al., 1995), Nucleotide excision repair (Kemp et al., 2010). The overlapping protein Q13156 has both the biological function of communities 4 and 40.

By analyzing the two examples of overlapping communities in the *M. musculus* and *H. sapiens* networks above, we can conclude that the biological function of the overlapping protein is related to the biological function of the community where it is located, so we can use the algorithm proposed to predict the functions of overlapping proteins.

CONCLUSION

This paper proposes an overlapping community detection algorithm based on the neighbor clustering coefficient to select the central edge. First, the node with the largest local influence in the network was found and determined as the central node. The central node was converted into central edge set. Then, the non-central edge was assigned to the community with the smallest distance. Finally, the community was optimized, and the excessively overlapping nodes were pruned according to the pruning strategy. The experimental results of the five algorithms on three types of networks show that the EQ and CR values of the NLC algorithm in this paper were improved and could identify overlapping structures better than the previously established algorithms. Applying the NLC algorithm to PPI networks can help us find overlapping structures in protein functional modules and discover unknown functions of proteins. In future work, we will continue to improve the algorithm so that it can adapt to changes in dynamic networks and further explore the application of the algorithm in biological information.

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/s.

AUTHOR CONTRIBUTIONS

QC collected the data and performed the experiments. YW conceived the project and designed the study. LY, KH, SY, and XX wrote the manuscript. All authors read and approved the final manuscript for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 62072212 and 61902144), the Development Project of Jilin Province of China (Nos. 20200401083GX, 2020C003, 20200403172SF) and Chinese Postdoctoral Science Foundation (No. 801212011421).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.689515/full#supplementary-material

REFERENCES

Ahn, Y.-Y., Bagrow, J. P., and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature* 466, 761–764. doi: 10.1038/nature09182

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

Bhowmick, S. S., and Seah, B. S. (2015). Clustering and summarizing protein-protein interaction networks: a survey. *IEEE Trans. Knowl. Data Eng.* 28, 638–658. doi: 10.1109/tkde.2015.2492559

Bollobás, B. e, Riordan, O., Spencer, J., and Tusnády, G. (2001). The degree sequence of a scale−free random graph process. *Rand. Struct. Algorith.* 18, 279–290. doi: 10.1002/rsa.1009

Chen, Y., Wang, W., Liu, J., Feng, J., and Gong, X. (2020). Protein interface complementarity and gene duplication improve link prediction of protein-protein interaction network. *Front. Genet.* 11:291. doi: 10.3389/fgene.2020.00291

Dunham, W. (1999). *Euler the Master of Us All*. Washingdon, DC: Mathematical Association of America 15:28.

Gaudet, P., Livstone, M. S., Lewis, S. E., and Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefi. Bioinform.* 12, 449–462. doi: 10.1093/bib/bbr042

Girvan, M., and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Gu, L., Han, Y., Wang, C., Chen, W., Jiao, J., and Yuan, X. (2019). Module overlapping structure detection in PPI using an improved link similarity-based Markov clustering algorithm. *Neural Comput. Appl.* 31, 1481–1490. doi: 10.1007/s00521-018-3508-z

Haring, S. J., Humphreys, T. D., and Wold, M. S. (2010). A naturally occurring human RPA subunit homolog does not support DNA replication or cell-cycle progression. *Nucleic Acids Res.* 38, 846–858. doi: 10.1093/nar/gkp1062

Huang, L., Li, Y., Wang, G. S., and Wang, Y. (2016). Community detection method based on vertex distance and clustering of density peaks. *J. Jilin Univ. Eng. Technol. Edn.* 46, 2042–2051.

Ji, J., Zhang, A., Liu, C., Quan, X., and Liu, Z. (2012). Survey: functional module detection from protein-protein interaction networks. *IEEE Trans. Knowl. Data Eng.* 26, 261–277. doi: 10.1109/tkde.2012.225

Kalinka, A. T., and Tomancak, P. (2011). linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics* 27, 2011–2012. doi: 10.1093/bioinformatics/btr311

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kemp, M. G., Mason, A. C., Carreira, A., Reardon, J. T., Haring, S. J., Borgstahl, G. E., et al. (2010). An alternative form of replication protein a expressed in normal human tissues supports DNA repair. *J. Biol. Chem.* 285, 4788–4797. doi: 10.1074/jbc.M109.079418

Keshav, K. F., Chen, C., and Dutta, A. (1995). Rpa4, a homolog of the 34-kilodalton subunit of the replication protein A complex. *Mol. Cell Biol.* 15, 3119–3128. doi: 10.1128/MCB.15.6.3119

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *N. J. Phys.* 11:033015. doi: 10.1088/1367-2630/11/3/033015

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Phys. Rev.* 78, 046110. doi: 10.1103/PhysRevE.78.046110

Mao, Y., and Liu, Y. (2020). Functional module mining in uncertain PPI network based on fuzzy spectral clustering. *J. Comput.* 31, 91–106. doi: 10.3966/199115992020083104008

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818. doi: 10.1038/nature03607

Qi, J., Xun, L., Yi, W., and Information, S. O. (2017). Overlapping community detection algorithm based on selection of seed nodes. *Appl. Res. Comput.* 34, 3534–3537. doi: 10.1016/j.compeleceng.2018.03.012

Ren, H., Xiao, J., Cui, W., and Xu, X. (2019). Construction and applications of benchmark networks for community detection based on null models. *J. Univ. Electr. Sci. Technol. China* 48, 440–448.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32(suppl 1), D449–D451. doi: 10.1093/nar/gkh086

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shen, H., Cheng, X., Cai, K., and Hu, M.-B. (2009). Detect overlapping and hierarchical community structure in networks. *Phys. A Stat. Mech. Appl.* 388, 1706–1712. doi: 10.1016/j.physa.2008.12.021

Tang, X. (2014). *A Network of Books About US Politics Published Around the Time of the 2004*. New York, NY: Springer international publishing. doi: 10.6084/m9.figshare.1149952.v1

Vella, D., Marini, S., Vitali, F., Di Silvestre, D., Mauri, G., and Bellazzi, R. (2018). MTGO: PPI network analysis via topological and functional module identification. *Sci. Rep.* 8:5499. doi: 10.1038/s41598-018-23672-0

Ying, K., and Lin, S. (2020). Maximizing cohesion and separation for detecting protein functional modules in protein-protein interaction networks. *PLoS One* 15:e0240628. doi: 10.1371/journal.pone.0240628

Yu, G., Wang, L., Han, Y., and He, Q. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* A *J. Integr. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* 33, 452–473.

Zhang, F., Ma, A., Wang, Z., Ma, Q., Liu, B., Huang, L., et al. (2018). A central edge selection based overlapping community detection algorithm for the detection of overlapping structures in protein–protein interaction networks. *Molecules* 23, 2633. doi: 10.3390/molecules23102633

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 July 2021
doi: 10.3389/fgene.2021.717557

Prediction of Disease Genes Based on Stage-Specific Gene Regulatory Networks in Breast Cancer

*Linzhuo Fan, Jinhong Hou and Guimin Qin**

School of Computer Science and Technology, Xidian University, Xi'an, China

Breast cancer is one of the most common malignant tumors in women, which seriously endangers women's health. Great advances have been made over the last decades, however, most studies predict driver genes of breast cancer using biological experiments and/or computational methods, regardless of stage information. In this study, we propose a computational framework to predict the disease genes of breast cancer based on stage-specific gene regulatory networks. Firstly, we screen out differentially expressed genes and hypomethylated/hypermethylated genes by comparing tumor samples with corresponding normal samples. Secondly, we construct three stage-specific gene regulatory networks by integrating RNA-seq profiles and TF-target pairs, and apply WGCNA to detect modules from these networks. Subsequently, we perform network topological analysis and gene set enrichment analysis. Finally, the key genes of specific modules for each stage are screened as candidate disease genes. We obtain seven stage-specific modules, and identify 20, 12, and 22 key genes for three stages, respectively. Furthermore, 55%, 83%, and 64% of the genes are associated with breast cancer, for example *E2F2*, *E2F8*, *TPX2*, *BUB1*, and *CKAP2L*. So it may be of great importance for further verification by cancer experts.

Keywords: breast cancer, DNA methylation, differentially expressed genes, stage-specific gene regulatory networks, WGCNA

OPEN ACCESS

Edited by:
Jianing Xi,
Northwestern Polytechnical University, China

Reviewed by:
Wenbin Liu,
Guangzhou University, China
Junyi Li,
Harbin Institute of Technology, China
Henry Han,
Fordham University, United States

***Correspondence:**
Guimin Qin
gmqin@mail.xidian.edu.cn

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

Received: *31 May 2021*
Accepted: *24 June 2021*
Published: *15 July 2021*

Citation:
Fan L, Hou J and Qin G (2021) Prediction of Disease Genes Based on Stage-Specific Gene Regulatory Networks in Breast Cancer. Front. Genet. 12:717557. doi: 10.3389/fgene.2021.717557

INTRODUCTION

Breast cancer is one of the most common malignant tumors in women, and it is the main disease factor that causes cancer deaths in women worldwide. According to statistics (Siegel et al., 2021), breast cancer accounts for 30% of female cancers. In China, breast cancer incidence has two peaks: one is 45–55 years old, and the other is 70–74 years old. From the perspective of age distribution, the incidence of breast cancer gradually increases from the age of 30, and reaches a peak at the age of 55. About 40% of female patients are under 50 (Wild et al., 2020). The symptoms of early breast cancer are unobvious and easy to be overlooked. In the late, cancer cells would metastasize far away, which causes multiple organ diseases, which seriously threatens the lives of patients. However, the current disease genes for breast cancer diagnosis and treatment are far from enough, and it is particularly important to find new candidate disease genes.

Epigenetics is currently a promising field in cancer research. As an important part of epigenetics, DNA methylation has received increasing attention, which is the process of adding methyl groups to DNA molecules and essential for cell development. The functional epigenetic module

(FEM) algorithm (Jiao et al., 2014) has verified the inverse correlation between DNA methylation and gene expression, and a large number of researchers have studied the effect of DNA methylation on breast cancer. Bediaga et al. (2010) analyzed the DNA methylation of cancer-related gene regulatory regions in breast cancer paired samples, and effectively identified 15 individual CpG loci that were differentially methylated in breast cancer tumor subtypes, which provides evidence that DNA methylation profile can predict breast cancer subtypes. Based on DNA methylation in whole blood and specific genes, Tang et al. (2016) studied the level of DNA methylation in the blood of breast cancer patients and healthy controls, and found that epigenome-wide blood DNA of breast cancer patients is hypomethylated, and the frequency of *BRCA1* and *RASSF1A* methylation is higher. Lu et al. (2017) explored the relationship between *RUNX3* gene methylation and breast cancer, and the results showed that the hypermethylation of *RUNX3* plays a significant role in the pathological stage and prognosis of breast cancer, which has great potential as a molecular marker for early diagnosis of breast cancer. De Almeida et al. (2019) analyzed the correlation between genome-wide methylation and gene expression by matching breast cancer DNA methylation with normal tissues in the TCGA, and identified new DNA methylation markers, including *PRAC2*, *TDRD10*, *TMEM132C,* etc., are expected to become diagnostic and prognostic markers of breast cancer.

There are also bioinformatics experts who study breast cancer based on biological molecular networks. Cai et al. (2019) used WCGNA to screen out the gene modules related to the risk of breast cancer metastasis, combined with the PPI network to screen out five key genes related to breast cancer progression and verified them. Lin et al. (2020) constructed a PPI network to screen hub genes, used modular analysis and survival analysis to identify potential target genes and pathways that may affect the occurrence and development of HER-2 positive breast cancer. Tang J. N. et al. (2018) identified five candidate biomarkers by analyzing the co-expression network, and used candidates in the basic and clinical research of breast cancer. Xi et al. (2018a) detected that *TP53* and *PNRM1* driver genes play an important role in breast cancer through matrix tri-factorization framework with pairwise similarity constraints. Guo et al. (2017) explained the mechanism of breast cancer development by identifying key pathways in breast cancer tissue and constructing the network of transcription factors (TFs) and microRNA (miRNA). Qiu et al. (2019) established the gene co-expression network for identifying modules related to breast cancer development, and discovered hub genes that may be used as markers of invasive breast cancer. Xi et al. (2018b) discovered mutated driver genes by using a robust and sparse co-regularized matrix factorization framework with prior information from mRNA expression patterns and interaction network. By combining the subspace learning framework, Xi et al. (2020) proposed the DriverSub algorithm to infer specific driver genes from heterogeneous breast cancer samples.

In this article, we propose a computational framework to predict candidate stage-specific disease genes of breast cancer based on the stage-specific gene regulatory networks. Firstly, we screen out differentially expressed genes and hypermethylated/hypomethylated genes by comparing tumor samples and normal samples. Secondly, we construct and analyze three stage-specific gene regulatory networks by taking stage information into account. Thirdly, we identify stage-specific modules by module division. Finally, we predict candidate stage-specific disease genes.

Our contributions consist of two points:

(1) We integrate stage information and DNA methylation information to construct a stage-specific gene regulatory network for breast cancer, which may help doctors identify patient's disease stage more quickly and design better treatment strategy.
(2) The proposed computational framework is effective in predicting breast cancer related genes, which will help experts to explore the molecular mechanisms of breast cancer.

MATERIALS AND METHODS

Our computational framework for predicting candidate disease genes includes four parts: Stage-specific gene regulatory networks construction, Module division, Topological properties analysis and gene set enrichment analysis, Candidate disease genes prediction (**Figure 1**).

Data Preprocessing

We download breast cancer phenotype data, gene expression profile and DNA methylation data from TCGA (Tomczak et al., 2015) (The Cancer Genome Atlas), which is currently the largest public cancer database, containing nearly 40 common cancer types and tens of thousands of samples. There are 60,484 genes and 1,217 samples in the gene expression profile, and 485,578 CpG sites and 890 samples in the DNA methylation data, respectively. We only retain the sample pairs, i.e., each tumor sample has a corresponding normal sample. Then, we divide the samples according to the stage information, and obtain 29 pairs, 94 pairs, 32 pairs of samples in stage I, stage II, and stage III, respectively. There are only two pairs of samples in stage IV that meet the experimental standards, which is not convincing. Therefore, we exclude samples in stage IV. For the DNA methylation data, we first convert the CpG site into the gene. As there are many CpG sites in a gene, we just use their mean β value to represent the DNA methylation level of the gene. For the gene expression profile, we download normalized FPKM data and filter out 15% genes with missing values. Then we select samples that have both cancer tissue and normal tissue.

The Gene Expression Omnibus (GEO) (Barrett et al., 2005) database includes a large amount of sequencing data and omics data, which is comprehensive and free. We download the GSE15852 and GSE69914 datasets from GEO (Liu et al., 2017). GSE15852 is the raw gene expression data from 43 human breast cancers and their corresponding normal tissues. GSE69914 is DNA methylation profiling of 50 normal samples

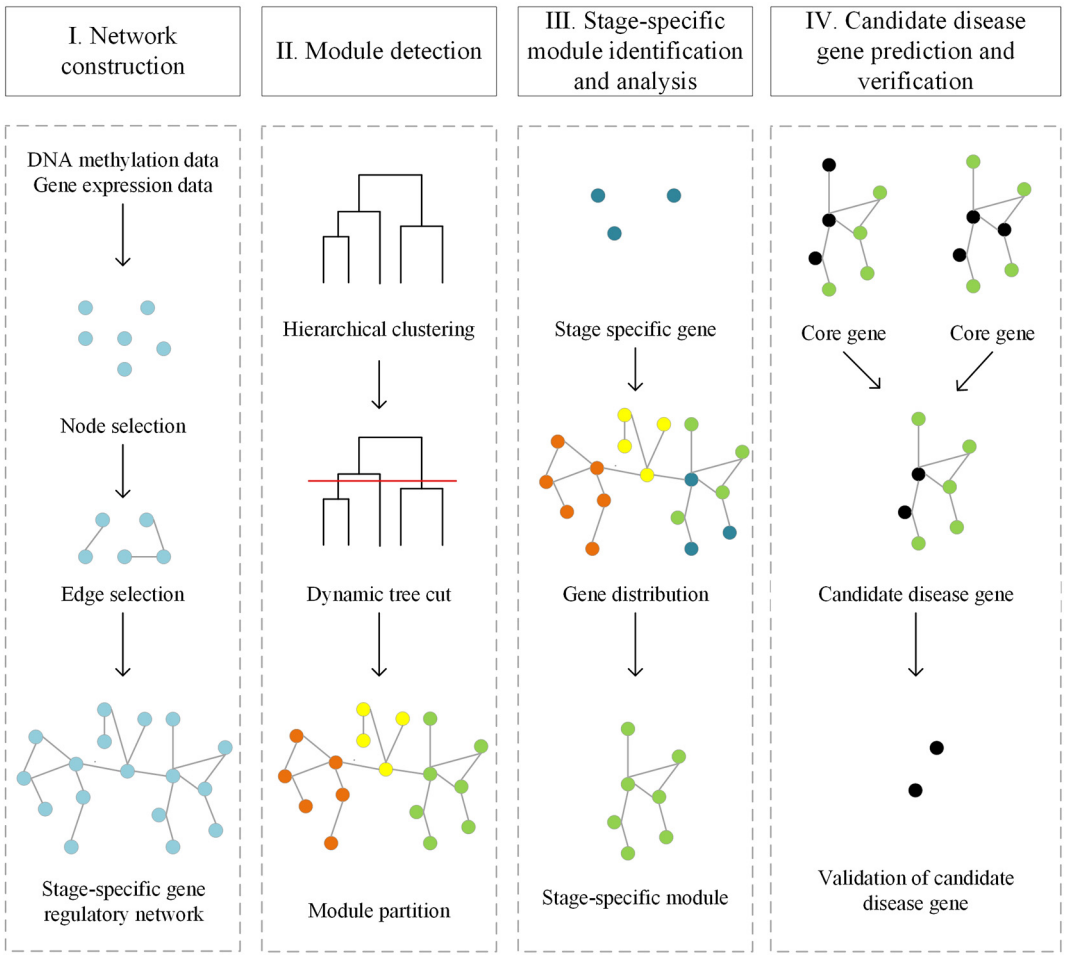

FIGURE 1 | Workflow of the computational framework for predicting disease genes based on stage-specific gene regulatory network.

from healthy women, 42 matched normal-adjacent breast cancer pairs (84 samples), 263 unmatched breast cancers, seven normal samples from BRCA1 carriers and four BRCA1 breast cancers. We only use 42 matched pairs of normal-adjacent breast cancer.

Differentially Expressed Genes and Hypomethylated/Hypermethylated Genes Identification

For the gene expression profile, we use Limma (Ritchie et al., 2015) in the R package to screen the differentially expressed genes, and use *p*-value less than 0.05 and |log FC| less than 0.5 as the threshold. For the DNA methylation data, we define β value greater than 0.8 as hypermethylated genes and β value less than 0.2 as hypomethylated genes. Then we take the intersection of the differentially expressed genes and the hypermethylated/hypomethylated genes and obtain 1,027 genes, 1,012 genes, and 1,220 genes in stage I, stage II, and stage III, respectively. Then we compare the relationship between the DNA methylation profile and gene expression profile, and find that the higher the gene methylation level, the lower the gene expression. And the results are shown in **Figure 2**.

Stage-Specific Gene Regulatory Networks Construction

Gene Regulatory Network database (GRNdb) (Fang et al., 2020) is a gene regulatory network database, which includes a large number of human and mouse transcription factor and target gene pairs. We download the TF-target gene pairs from the GRNdb, and filter out the pairs in which the target genes are differentially expressed genes and hypermethylated/hypomethylated genes (Qin et al., 2019). Then we calculate the Pearson Correlation Coefficient (PCC) for each TF-target gene pair based on their expression level, and the cut-off is set as 0.5 and construct stage-specific gene regulatory networks.

Module Division

We use WGCNA (Langfelder and Horvath, 2008) to divide the stage-specific gene regulatory network into modules. Firstly, we perform hierarchical clustering on the three stage-specific gene regulatory networks to generate a hierarchical clustering

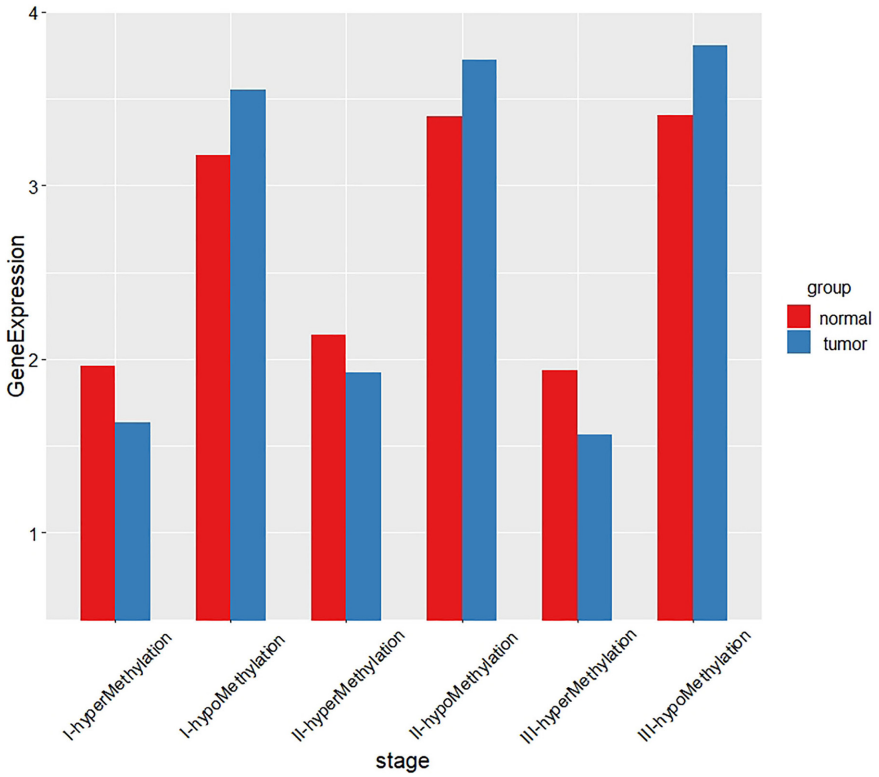

FIGURE 2 | The relationship between DNA methylation and gene expression of each stage.

tree. Then, we use the Dynamic Tree Cut algorithm (Langfelder et al., 2008) to divide the above-generated hierarchical clustering tree and ensure that the number of molecules in each module is at least 30.

Topological Properties Analysis and Gene Set Enrichment Analysis

Hub genes are important for biological processes. We identify and compare hub genes for each gene regulatory network. We perform topological analysis of stage-specific gene regulatory networks using Cytoscape (Shannon et al., 2003), including degree distribution, centrality distribution, and so on. Then, we perform gene set enrichment analysis using Metascape (Zhou et al., 2019).

Candidate Disease Gene Prediction

We filter out candidate disease genes from the above modules and network topological information. Then, we checked them by known disease-related genes from OMIM, COSMIC, and DAVID. Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2005) mainly covers the relationship of genes and diseases, the relationship of genes and phenotypes, and some clinical features. Catalog of Somatic Mutations in Cancer (COSMIC) (Sondka et al., 2018) integrates cancer somatic mutations and provides cancer gene mutation map data information. DAVID (Huang et al., 2009) integrates biological data and analysis tools and provides systematic and comprehensive biological function annotation information for large-scale gene or protein lists. Furthermore, we check the association of the rest of the candidate disease genes and breast cancer in PubMed (Shashikiran, 2016).

RESULTS

Stage-Specific Gene Regulatory Network Construction

We filter out the TF-target gene pairs whose target genes are not differentially expressed genes and hypermethylated/hypomethylated genes, and use the PCC cut-off 0.5 to construct stage-specific gene regulatory networks. There are 1,129, 1,066, and 1,339 nodes and 4,429, 4,879, and 6,461 edges, respectively.

Module Division

We use WGCNA to divide three gene regulatory networks into modules and the results are shown in **Figure 3**. We find that the first-stage network is divided into 11 modules, of which the turquoise module contains up to 270 genes. The number of genes in the remaining modules ranges from 40 to 149. The second-stage network is divided into 10 modules, of which the turquoise module contains 337 genes. The number of genes in the remaining modules ranges

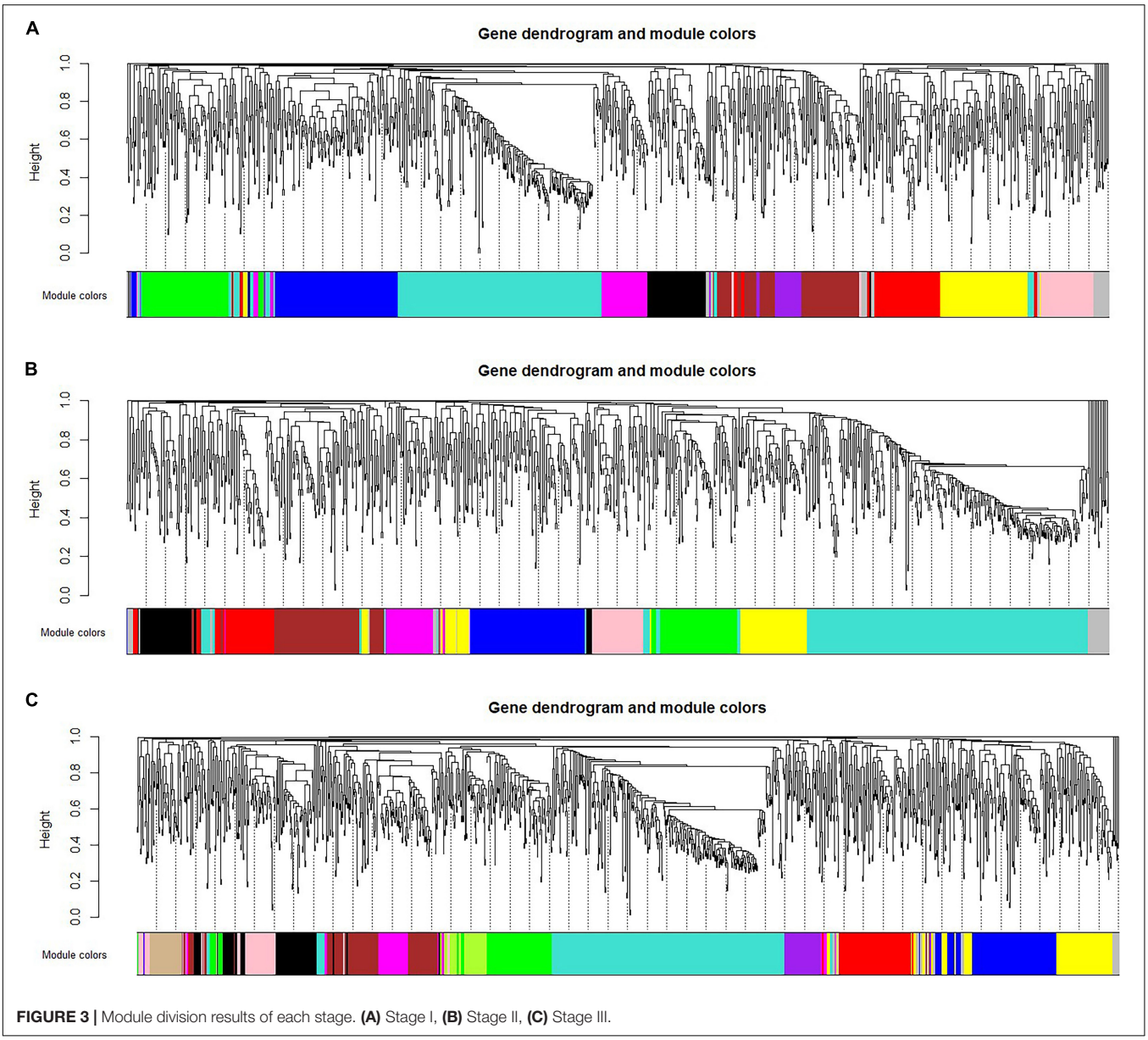

FIGURE 3 | Module division results of each stage. **(A)** Stage I, **(B)** Stage II, **(C)** Stage III.

from 40 to 125. The third-stage network is divided into 13 modules, of which the turquoise module contains 337 genes. The number of genes in the remaining modules ranges from 30 to 142. In particular, the gray modules contain genes that are not classified into any module and discarded. The detailed information of the number of genes in each module is shown in **Table 1**.

We identify differentially expressed genes that only exist in one stage as the stage-specific genes and obtain 92 genes, 60 genes, and 187 genes in stage I, stage II, and stage III, respectively. Then we count the distribution of these genes in each module, as shown in **Table 1**. We find that the specific genes in stage I are mainly distributed in the S1_brown module, S1_turquoise module and S1_blue module, the specific genes in stage II are mainly distributed in the S2_turquoise module, and the specific genes in stage III are mainly distributed in the S3_turquoise module, S3_brown module and S3_green module. Therefore, we regard these seven modules as the specific modules of corresponding stage.

Topological Properties Analysis and Gene Set Enrichment Analysis

We perform network topological analysis for seven specific modules using Cytoscape. For the degree distribution, the degrees of S1_turquoise module, S2_turquoise module, and S3_turquoise module are mainly distributed between 100 and 400, and the degrees of S1_brown module, S1_blue module, S3_brown module, and S3_green module are mainly distributed between 50 and 100, respectively. And the degree distribution of each

module conforms to the power law distribution. The betweenness centrality of most nodes in each module is at a high level. The closeness centrality of most nodes in each module ranges from 0.5 to 0.9. These values indicate that the network corresponding to each module is a dense graph, so the hub genes screened by these three parameters are all core genes.

We use Metascape to perform the joint enrichment analysis on the genes in the seven specific modules, and set *p*-value cut-off 0.01. The joint enrichment results are shown in **Figure 4**. The most significant enrichment item for each module is shown in **Table 2**. According to **Figure 4** and **Table 2**, S1_turquoise, S2_turquoise, and S3_turquoise modules are roughly identical, and these significant pathways are all related to cell transcription and cycle regulation. S3_green, S1_brown, S1_blue, and S3_brown modules are closely related to each other, and these significant pathways are mainly related to gene transcription. In addition, transcription regulation complex (GO:0005667) and chromatin binding (GO:0003682) are the common enrichment items of the seven specific modules. The results show that the stage-specific modules have strong functionality and the genes within the modules are highly correlated.

Candidate Disease Gene Prediction

We predict disease genes based on correlation matrix and network topological properties. Firstly, we calculate the correlation matrix of genes at each specific module, and select genes with correlation cut-off 0.8 and *p*-value cut-off

TABLE 1 | Gene distribution of each module.

Module	Gene count	Specific gene count	Module	Gene count	Specific gene count	Module	Gene count	Specific gene count
S1_black	71	5	S2_black	66	4	S3_black	99	11
S1_blue	149	11	S2_blue	125	5	S3_blue	142	8
S1_brown	120	28	S2_brown	116	3	S3_brown	120	26
S1_green	106	5	S2_green	90	9	S3_green	110	25
S1_grey	42	4	S2_grey	40	3	S3_greenyellow	51	14
S1_magenta	63	4	S2_magenta	58	3	S3_grey	30	2
S1_pink	66	10	S2_pink	58	6	S3_magenta	61	4
S1_purple	40	2	S2_red	68	3	S3_pink	70	14
S1_red	95	8	S2_turquoise	337	21	S3_purple	54	10
S1_turquoise	270	13	S2_yellow	108	2	S3_red	104	15
S1_yellow	107	1				S3_tan	45	2
						S3_turquoise	337	47
						S3_yellow	116	8

S1, S2, and S3 represent stage I, stage II, and stage III, respectively.

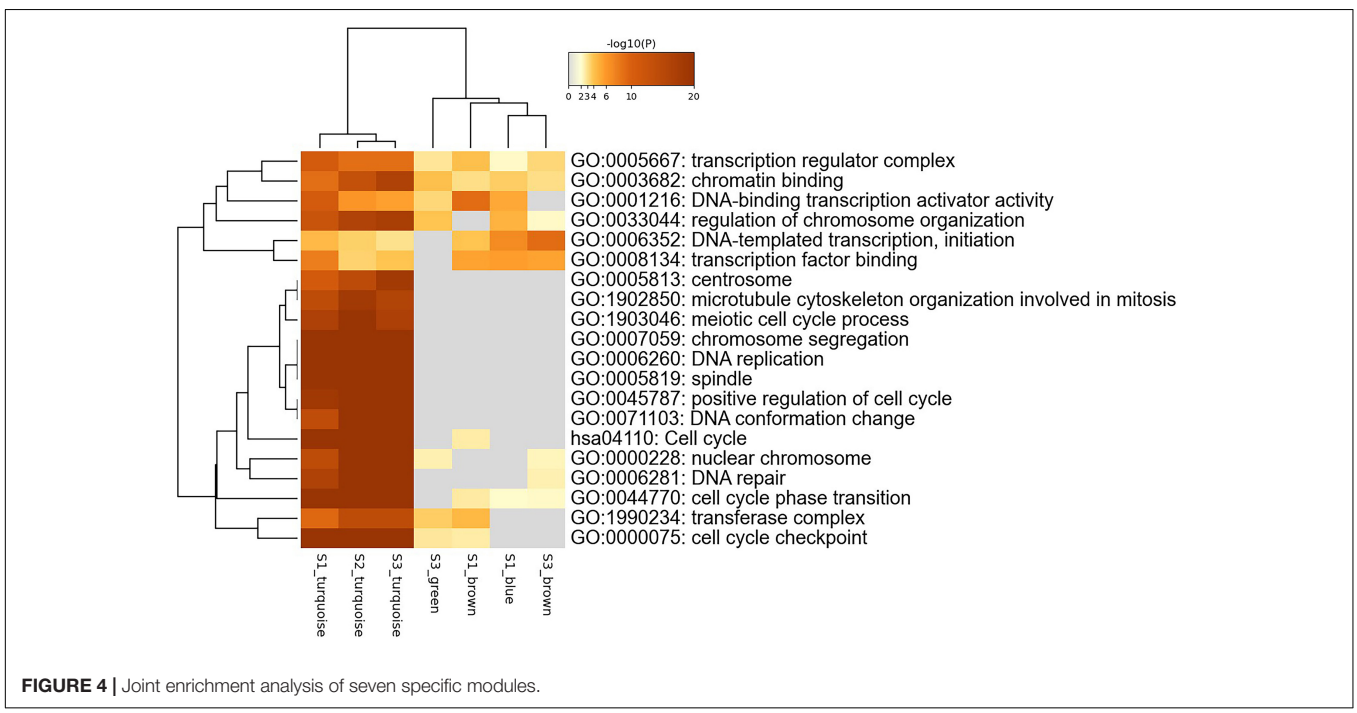

FIGURE 4 | Joint enrichment analysis of seven specific modules.

TABLE 2 | Functional enrichment analysis.

Module	Term	Description	Log10(P)	Count
S1_turquoise	GO:0044770	Cell cycle phase transition	−32.060	52
	GO:0051301	Cell division	−31.343	50
	GO:0006260	DNA replication	−21.835	30
S1_blue	GO:0022411	Cell component disassembly	−8.88	18
	GO:0001046	The core promoter sequence specifically binds to DNA	−8.71	7
	GO:0070897	Transcription pre-priming complex assembly	−3.78	6
S1_brown	GO:0001228	DNA binding transcription activator activity	−9.13	15
	GO:0001227	DNA binding transcription repressor activity	−6.08	10
	GO:0004879	Nuclear receptor activity	−5.54	5
S2_turquoise	GO:0044770	Cell cycle transition	−41.180	66
	GO:0007059	Chromosome segregation	−38.573	50
	GO:0005819	Spindle	−27.162	42
S3_turquoise	GO:0044770	Cell cycle transition	−44.41	69
	GO:0098687	Chromosome region	−38.25	50
	hsa04110	Cell cycle	−28.49	29
S3_brown	GO:0006352	DNA template transcription	−9.100	12
	GO:0001046	The core promoter sequence specifically binds to DNA	−7.678	6
	GO:0034655	Catabolism of nucleobase-containing compounds	−6.932	14
S3_green	GO:0016570	Histone modification	−6.678	12
	GO:0005697	Telomerase holoenzyme complex	−5.816	4
	GO:0034243	Macromolecule methylation	−5.194	5

0.05 as the core genes of each module. Then, we sort the degree distribution, betweenness centrality and closeness centrality of each gene in the seven modules, and select the top 5% as the core gene of each module. The intersection of core genes selected by these two methods are considered as candidate disease genes.

We obtain 20 candidate disease genes in stage I, such as *E2F2*, *E2F8*, *TPX2*, etc., 12 genes in stage II, such as *KPNA2*, *CKAP2L*, *CBX3*, etc., and 22 genes in stage III, such as *RAD21*, *FBXO5*, *CCNE2*, etc. A complete gene list of each stage is shown in **Table 3**. *E2F2, CKAP2L* and *CBX3* are genes shared by three stages. For the remaining candidate genes at different stages, we compare their gene expression data and find that they are indeed different at different stages. And the results are shown in the **Supplementary Figures 1∼2**.

TABLE 3 | Candidate disease genes at each stage.

Stage	Candidate disease genes
Stage I	E2F2*, E2F8#, TPX2*, BUB1*, CKAP2L#, CBX3#, CASC5#, KPNA2#, LMNB1, NEK2*, TTK*, SLC25A36, CREBRF, ZC3H6, PAN2, BTAF1, SLC25A39, DDX49, SLC39A1#, MRPS12
Stage II	E2F2*, E2F8#, TPX2*, KPNA2#, CKAP2L#, CBX3#, DDIAS, BUB1*, CCNE2#, CASC5#, SPDL1, TOP2A*
Stage III	E2F2*, RAD21#, FBXO5#, CCNE2#, CBX3#, STIL#, CKAP2L#, PCNA*, NEK2*, TTK*, CSE1L#, H2AFZ#, NR2F6, TRAPPC6A, IGSF8, FDXR, SLC39A1#, EXOSC5, RBBP5, KDM5B#, H3F3A, CDC42SE1
Common genes	E2F2, CKAP2L, CBX3

**Genes verified by OMIM, COSMIC, DAVID. #Genes verified by PubMed.*

Candidate Disease Gene Verification

In order to determine whether the selected candidate disease genes are effective in the diagnosis and treatment of breast cancer, we use OMIM, COSMIC, and DAVID to verify the candidate genes, and obtain seven genes related to breast cancer. *BUB1* is mitotic checkpoint serine, *E2F2* is a transcription activator, *NEK2* is a serine/threonine-protein kinase, *TPX2* is the target protein for *Xklp2*, *TTK* is essential for spindle establishment and centrosome replication, *PCNA* is the proliferating cell nuclear antigen, and *TOP2A* is DNA topoisomerase 2-alpha. Most of these genes are related to cell proliferation and transcription.

We search the rest candidate disease genes related to the genes in PubMed, and verify whether the genes are related to breast cancer. Kos et al. (2020) found *STIL* is an important prognostic and predictive biomarker for triple-negative breast cancer and HER2-positive breast cancer. At present, there have been studies on pathological assessment of breast cancer based on *STIL*, which is a key step for molecular markers to move toward clinical treatment. Based on the study of differentially expressed hub genes, Qi et al. (2019) proposed that the overexpression of *CCNE2, H2AFZ, TOP2A* is closely related to the diagnosis and poor prognosis of breast cancer. Yuksel et al. (2015) found the overexpression of *CSE1L* has a certain relationship with the distant metastasis of breast cancer and may be a valuable prognostic tool. Tang J. et al. (2018) used WGCNA to construct a co-expression network and found *FBXO5* and *TPX2* are related to the poor prognosis of breast cancer. Liang et al. (2017) found *CBX* family proteins have epigenetic regulatory functions, among which the high expression of *CBX3* is related to the worsening of recurrence-free survival rate of breast cancer patients.

Liu et al. (2018) found *E2Fs* are transcription factors that affect cell proliferation, differentiation and apoptosis, and the high expression of *E2F8* is also related to the deterioration of patients' recurrence-free survival rate, and can be used as a potential target for individualized treatment of breast cancer patients. Zhang et al. (2019) showed that *KDM5B* is up-regulated in breast cancer and many other cancers and its expression is positively correlated with breast cancer metastasis. Duan et al. (2020) and Liu et al. (2020) showed the expression of *KPNA2* and *SLC39A1* in breast cancer tissues is significantly up-regulated, which can regulate the development of breast cancer and provide new targets for breast cancer treatment. *NEK2* is a kind of serine, which plays an important role in mitosis. Cappello et al. (2014) and Chen et al. (2020) have proven *NEK2* is a target for breast cancer. Atienza et al. (2005) has shown through experiments that *RAD21* can enhance the anti-tumor activity of chemotherapeutics by inducing DNA damage and is a new target for cancer drugs. Based on survival analysis and mutation analysis, Fu et al. (2019) found that the high expression of *CKAP2L* and *CASC5* is closely related to the poor prognosis of breast cancer patients. These verified genes are shown in **Table 3**.

In summary, we detect 20, 12, and 22 candidate disease genes for three stages, respectively. Through PubMed search, 11, 10, and 14 genes are verified, respectively. That is 55%, 83%, and 64% of the candidate disease genes are proved to be related to the diagnosis and treatment of breast cancer, respectively, such as *E2F2*, *E2F8*, *TPX2*, *BUB1*, *CKAP2L*, etc. The results show the effectiveness of our computational framework for predicting disease genes.

We also use GSE15852 gene expression profile and GSE69914 DNA methylation profile to verify the validity of the proposed computational framework. Firstly, we screen out 79 differentially expressed genes and hypermethylated/hypomethylated genes. Secondly, we combine with the TF-target gene pairs and construct a gene regulatory network with 195 nodes and 313 edges. Thirdly, we divide the gene regulatory network into four modules: 76 genes in turquoise module, 68 genes in blue module, 18 genes in gray module, and 33 genes in brown module, respectively. In particular, the gray module contains genes that are not classified into any module and discarded. Finally, we screen the candidate disease genes of each module based on correlation matrix and network topological properties, and obtain four genes in turquoise module, four genes in blue module, and two genes in brown module, respectively. In detail, these genes are *H2AFZ, NPM1, MAF, NR3C1, PTGER3, TCF4, IRF1, RARB, CHD2,* and *SMAD4.* Except *PTGER3* and *CHD2*, other genes have been verified. This means that our method is effective, and it may help experts explore breast cancer related genes.

DISCUSSION

At present, the proposed computational framework has only been tested on breast cancer, and satisfactory results have been obtained. In the future, we will try to apply this framework to other types of diseases for discovering more disease-related genes.

CONCLUSION

We propose a computational framework to predict candidate stage-specific disease genes for breast cancer based on stage-specific gene regulatory networks. And we conduct experiments using two breast cancer data sets and find that most predicted genes are related to breast cancer, which shows that our method is effective. We also predict some candidate disease genes that need to be further verified. Nevertheless, our research has some limitations. Our proposed computational framework is based on the public TCGA and GEO datasets, and the noise affects the analysis results. Another limitation is that we should integrate more omics data so that more disease genes may be predicted more accurately.

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

Ethical review and approval was not required for the study on human participants in accordance with the Local Legislation and Institutional Requirements. Written informed consent for participation was not required for this study in accordance with the National Legislation and the Institutional Requirements.

AUTHOR CONTRIBUTIONS

JH and LF conceived and developed the computational framework for predicting disease genes and wrote the manuscript. GQ provided important feedback in the framework process and edited the manuscript. All authors have made significant contributions to the completion and writing of the manuscript and read and approved the final manuscript.

FUNDING

This study was supported by the Natural Science Foundation of Shaanxi Province (No. 2017JM6038) and National Key Research and Development Program of China (2018YFC0116500).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.717557/full#supplementary-material

Supplementary Figures 1∼2 | The gene expression boxplot of different candidate disease genes at different stages.

REFERENCES

Atienza, J. M., Roth, R. B., Rosette, C., Smylie, K. J., Kammerer, S., Rehbock, J., et al. (2005). Suppression of RAD21 gene expression decreases cell growth and enhances cytotoxicity of etoposide and bleomycin in human breast cancer cells. *Mol. Cancer Ther.* 4, 361–368. doi: 10.1158/1535-7163.mct-04-0241

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., et al. (2005). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res.* 33, D562–D566. doi: 10.1093/nar/gki022

Bediaga, N. G., Acha-Sagredo, A., Guerra, I., Viguri, A., Albaina, C., Ruiz Diaz, I., et al. (2010). DNA methylation epigenotypes in breast cancer molecular subtypes. *Breast Cancer Res.* 12:R77. doi: 10.1186/bcr2721

Cai, Y., Mei, J., Xiao, Z., Xu, B. J., Jiang, X. Z., Zhang, Y. J., et al. (2019). Identification of five hub genes as monitoring biomarkers for breast cancer metastasis in silico. *Hereditas* 156:20.

Cappello, P., Blaser, H., Gorrini, C., Lin, D. C., Elia, A. J., Wakeham, A., et al. (2014). Role of Nek2 on centrosome duplication and aneuploidy in breast cancer cells. *Oncogene* 33, 2375–2384. doi: 10.1038/onc.2013.183

Chen, Y. W., Wu, N., Liu, L., Dong, H. Y., and Liu, X. A. (2020). microRNA-128-3p overexpression inhibits breast cancer stem cell characteristics through suppression of Wnt signalling pathway by down-regulating NEK2. *J. Cell. Mol. Med.* 24, 7353–7369. doi: 10.1111/jcmm.15317

De Almeida, B. P., Apolonio, J. D., Binnie, A., and Castelo-Branco, P. (2019). Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. *BMC Cancer* 19:219. doi: 10.1186/s12885-019-5403-0

Duan, M., Hu, F., Li, D., Wu, S., and Peng, N. (2020). Silencing KPNA2 inhibits IL-6-induced breast cancer exacerbation by blocking NF-kappaB signaling and c-Myc nuclear translocation in vitro. *Life Sci.* 253:117736. doi: 10.1016/j.lfs.2020.117736

Fang, L., Li, Y., Ma, L., Xu, Q., Tan, F., Chen, G., et al. (2020). GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic Acids Res.* 49, D97–D103. doi: 10.1093/nar/gkaa995

Fu, Y., Zhou, Q. Z., Zhang, X. L., Wang, Z. Z., and Wang, P. (2019). Identification of hub genes using co-expression network analysis in breast cancer as a tool to predict different stages. *Med. Sci. Monit.* 25, 8873–8890. doi: 10.12659/Msm.919046

Guo, X., Xiao, H., Guo, S., Dong, L., and Chen, J. (2017). Identification of breast cancer mechanism based on weighted gene coexpression network analysis. *Cancer Gene Ther.* 24, 333–341. doi: 10.1038/cgt.2017.23

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514–D517. doi: 10.1093/nar/gki033

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

Jiao, Y., Widschwendter, M., and Teschendorff, A. E. (2014). A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics* 30, 2360–2366. doi: 10.1093/bioinformatics/btu316

Kos, Z., Roblin, E., Kim, R. S., Michiels, S., Gallas, B. D., Chen, W. J., et al. (2020). Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. *NPJ Breast Cancer* 6:17. doi: 10.1038/s41523-020-0156-0

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. *Bioinformatics* 24, 719–720. doi: 10.1093/bioinformatics/btm563

Liang, Y. K., Lin, H. Y., Chen, C. F., and Zeng, D. (2017). Prognostic values of distinct CBX family members in breast cancer. *Oncotarget* 8, 92375–92387. doi: 10.18632/oncotarget.21325

Lin, Y. X., Fu, F. M., Lv, J. X., Wang, M. C., Li, Y., Zhang, J., et al. (2020). Identification of potential key genes for HER-2 positive breast cancer based on bioinformatics analysis. *Medicine* 99:e18445. doi: 10.1097/md.0000000000018445

Liu, J., Li, H., Sun, L., Wang, Z., Xing, C., and Yuan, Y. (2017). Aberrantly methylated-differentially expressed genes and pathways in colorectal cancer. *Cancer Cell Int.* 17:75.

Liu, L., Yang, J., and Wang, C. (2020). Analysis of the prognostic significance of solute carrier (SLC) family 39 genes in breast cancer. *Biosci. Rep.* 40:BSR20200764. doi: 10.1042/BSR20200764

Liu, Z. L., Bi, X. W., Liu, P. P., Lei, D. X., Wang, Y., Li, Z. M., et al. (2018). Expressions and prognostic values of the E2F transcription factors in human breast carcinoma. *Cancer Manag. Res.* 10, 3521–3532. doi: 10.2147/CMAR.S172332

Lu, D. G., Ma, Y. M., Zhu, A. J., and Han, Y. W. (2017). An early biomarker and potential therapeutic target of RUNX 3 hypermethylation in breast cancer, a system review and meta-analysis. *Oncotarget* 8, 22166–22174. doi: 10.18632/oncotarget.13125

Qi, L. N., Zhou, B. T., Chen, J. N., Hu, W. X., Bai, R., Ye, C. Y., et al. (2019). Significant prognostic values of differentially expressed-aberrantly methylated hub genes in breast cancer. *J. Cancer* 10, 6618–6634. doi: 10.7150/jca.33433

Qin, G., Yang, L., Ma, Y., Liu, J., and Huo, Q. (2019). The exploration of disease-specific gene regulatory networks in esophageal carcinoma and stomach adenocarcinoma. *BMC Bioinformatics* 20(Suppl. 22):717. doi: 10.1186/s12859-019-3230-6

Qiu, J., Du, Z., Wang, Y., Zhou, Y., Zhang, Y., Xie, Y., et al. (2019). Weighted gene co-expression network analysis reveals modules and hub genes associated with the development of breast cancer. *Medicine* 98:e14345. doi: 10.1097/MD.0000000000014345

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y. F., 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

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shashikiran, N. D. (2016). Medline, pubmed, and pubmed central ((R)): analogous or dissimilar. *J. Indian Soc. Pedod. Prev. Dent.* 34, 197–198. doi: 10.4103/0970-4388.186748

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer statistics, 2021. *CA Cancer J. Clin.* 71, 7–33. doi: 10.3322/caac.21654

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 18, 696–705. doi: 10.1038/s41568-018-0060-1

Tang, J. N., Kong, D. G., Cui, Q. X., Wang, K., Zhang, D., Gong, Y., et al. (2018). Prognostic genes of breast cancer identified by gene co-expression network analysis. *Front. Oncol.* 8:374. doi: 10.3389/fonc.2018.00374

Tang, J., Kong, D., Cui, Q., Wang, K., Zhang, D., Gong, Y., et al. (2018). Prognostic genes of breast cancer identified by gene co-expression network analysis. *Front. Oncol.* 8:374.

Tang, Q. Q., Cheng, J., Cao, X., Surowy, H., and Burwinkel, B. (2016). Blood-based DNA methylation as biomarker for breast cancer: a systematic review. *Clin. Epigenet.* 8:115.

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Wild, C. P., Weiderpass, E., and Stewart, B. W. (2020). *World Cancer Report: Cancer Research for Cancer Prevention*. Lyon: International Agency for Research on Cancer.

Xi, J., Li, A., and Wang, M. (2018a). A novel unsupervised learning model for detecting driver genes from pan-cancer data through matrix tri-factorization framework with pairwise similarities constraints. *Neurocomputing* 296, 61–73.

Xi, J., Wang, M., and Li, A. (2018b). Discovering mutated driver genes through a robust and sparse co-regularized matrix factorization framework with prior information from mRNA expression patterns and interaction network. *BMC Bioinformatics* 19:214. doi: 10.1186/s12859-018-2218-y

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. (2020). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 36, 1855–1863. doi: 10.1093/bioinformatics/btz793

Yuksel, U. M., Turker, I., Dilek, G., Dogan, L., Gulcelik, M. A., and Oksuzoglu, B. (2015). Does CSE1L overexpression affect distant metastasis development in breast cancer? *Oncol. Res. Treat.* 38, 431–434. doi: 10.1159/000438501

Zhang, Z. G., Zhang, H. S., Sun, H. L., Liu, H. Y., Liu, M. Y., and Zhou, Z. (2019). KDM5B promotes breast cancer cell proliferation and migration via AMPK-mediated lipid metabolism reprogramming. *Exp. Cell Res.* 379, 182–190. doi: 10.1016/j.yexcr.2019.04.006

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10:1523. doi: 10.1038/s41467-019-09234-6

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 August 2021
doi: 10.3389/fgene.2021.733906

SCDRHA: A scRNA-Seq Data Dimensionality Reduction Algorithm Based on Hierarchical Autoencoder

Jianping Zhao[1], Na Wang[1], Haiyun Wang[1], Chunhou Zheng[2] and Yansen Su[2]**

[1] College of Mathematics and System Sciences, Xinjiang University, Ürümqi, China, [2] Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, School of Artificial Intelligence, Anhui University, Hefei, China

OPEN ACCESS

Edited by:
Zhenhua Yu,
Ningxia University, China

Reviewed by:
Rui Fu,
University of Colorado, United States
Jin-Xing Liu,
Qufu Normal University, China
Jing Yang,
North China Electric Power University, China

***Correspondence:**
Haiyun Wang
haiyun_wang_xju@163.com
Yansen Su
suyansen@ahu.edu.cn

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

Received: *30 June 2021*
Accepted: *26 July 2021*
Published: *27 August 2021*

Citation:
Zhao J, Wang N, Wang H, Zheng C and Su Y (2021) SCDRHA: A scRNA-Seq Data Dimensionality Reduction Algorithm Based on Hierarchical Autoencoder. Front. Genet. 12:733906. doi: 10.3389/fgene.2021.733906

Dimensionality reduction of high-dimensional data is crucial for single-cell RNA sequencing (scRNA-seq) visualization and clustering. One prominent challenge in scRNA-seq studies comes from the dropout events, which lead to zero-inflated data. To address this issue, in this paper, we propose a scRNA-seq data dimensionality reduction algorithm based on a hierarchical autoencoder, termed SCDRHA. The proposed SCDRHA consists of two core modules, where the first module is a deep count autoencoder (DCA) that is used to denoise data, and the second module is a graph autoencoder that projects the data into a low-dimensional space. Experimental results demonstrate that SCDRHA has better performance than existing state-of-the-art algorithms on dimension reduction and noise reduction in five real scRNA-seq datasets. Besides, SCDRHA can also dramatically improve the performance of data visualization and cell clustering.

Keywords: scRNA-seq, dimensionality reduction, graph autoencoder, graph attention networks, noise reduction

INTRODUCTION

With the rapid development of single-cell RNA sequencing (scRNA-seq) technology, the research of transcriptomics has changed dramatically (Tang et al., 2013; Xi et al., 2018, 2020). On the one hand, the cell is the unit of an organism, mining data at the single-cell level can help researchers probe the essence and laws of living activities. On the other hand, the scale of scRNA-seq data obtained by researchers is growing, which brings enormous challenges in analysis and computation (Kiselev et al., 2019; Yu et al., 2021). How to transform a high-dimension data into low-dimension embedding while preserving the topological structure of raw data plays an indispensable role in scRNA-seq analysis. Besides, the high noise in scRNA-seq data will make it far too difficult to reduce dimension. One of the most challenging noises is the dropout events, which caused zero inflation in scRNA-seq data (Zhang and Zhang, 2018). The low RNA capture rate leads to the detection failure of an expressed gene resulting in a "false" zero count observation, which is defined as a dropout event. The zero counts consist of "false" zero counts and "true" zero counts, where the true counts represent the lack of expression of a gene in a specific cell, and the false zero counts are dropout

events. A large number of false zero counts will lead to unreliable results of visualization, clustering, and pseudotime inference. Thus, noise reduction is integral for scRNA-seq data analysis as well as dimension reduction.

The new challenges of scRNA-seq data bring new opportunities, these data have spurred the millions of algorithms to derive novel biological insights (Hie et al., 2020; Wang et al., 2021a,b). Because of the high-dimensionality of scRNA-seq, many dimension reduction methods have been proposed for scRNA-seq data. Some of these methods fail to consider zero inflation (dropout) of the scRNA-seq data, including uniform manifold approximation and projection (UMAP) (Becht et al., 2019) and single-cell graph autoencoder (scGAE) (Luo et al., 2021). UMAP is a non-linear dimensionality reduction technique, which is a universal method in high-dimensional gene expression analysis. scGAE is a dimensionality reduction method based on graph autoencoder, which can preserve topological structure in scRNA-seq data. Nevertheless, these methods ignore the impact of dropout events on the output.

On the contrary, many single-cell analysis algorithms take dropout events into account, including zero-inflated factor analysis (ZIFA) (Pierson and Yau, 2015), zero-inflated negative binomial (NB)-based wanted variation Extraction (ZINB-WaVE) (Risso et al., 2018), deep count autoencoder (DCA) (Eraslan et al., 2019), and single-cell model-based deep embedded clustering (scDeepCluster) (Tian et al., 2019). ZIFA focuses on dropout events and assumes the dropout rate for a gene depends on the expression level. However, such a strong assumption lacks flexibility, and it is not quite suitable for real datasets. To solve this challenge, ZINB-WaVE has been proposed, which is general and flexible and uses a zero-inflated negative binomial (ZINB) (Risso et al., 2018) model. Nonetheless, ZIFA and ZINB-WaVE have large computation cost; hence, these methods are not fit for large-scale data. DCA is a deep learning method based on autoencoder in an unsupervised manner, which can be applied to datasets of millions of cells. Different from regular autoencoder, the DCA proposes a ZINB model-based loss function substitute for the conventional mean square error loss function to depict scRNA-seq data better. Based on the framework of DCA, scDeepCluster adds the random Gaussian noise into the encoder to improve the embedded feature representation and executes clustering tasks using deep embedded clustering on latent space. However, both DCA and scDeepCluster are not taking the cell–cell relationships into account.

The recently proposed graph attention network (GAT) (Veličković et al., 2018) is a novel neural network architecture that operates on graph-structured data, which preserves the topological structure in a latent space. In this work, we build the graph autoencoder based on GAT to project the data into a low-dimensional latent expression and maintain the topological structure among cells as possible. Considering the input of the graph autoencoder is single-cell graphs of node matrices and adjacency matrix, the adjacency matrix among cells built by the K-nearest-neighbor (KNN) algorithm is quite considerable for graph autoencoder. Nevertheless, the adjacency matrix will be distorted by the impact of the high sparsity of scRNA-seq data on the KNN algorithm. Therefore, we focus on the impact of dropout events on the output of the KNN algorithm and utilize a scalable denoising method DCA to mitigate zero inflation caused by dropout events. Because the raw data and reconstructed data by DCA have the same dimension, we implement initial dimensionality reduction for the reconstructed data by using principal component analysis (PCA). Based on the latent space constructed by PCA, we build a graph autoencoder to reduce the dimension and get a low dimensional embedding for visualization and clustering. These are the motivations behind our new method SCDRHA. We extensively evaluate our approach with competing methods using five real datasets; the experimental results demonstrate that SCDRHA has better performance than the existing state-of-the-art algorithms on dimension reduction and noise reduction. Besides, SCDRHA can also dramatically improve the performance of data visualization and cell clustering.

MATERIALS AND METHODS

The SCDRHA pipeline for scRNA-seq data analysis consists of two core modules (**Figure 1**). The first model is DCA to alleviate dropout events, which is learned by the ZINB model-based autoencoder. The second model is a graph autoencoder based on GAT, which maps the denoised data by DCA to a low-dimensional latent representation.

Data Preprocessing

To begin, suppose that we have a raw scRNA-seq count matrix C, which is filtered out genes with no count in any cell. C can be represented as a P-by-N dimensional matrix, where P is defined as the total number of genes, N is defined as the total number of cells, and c_{ij} represents the expression value of gene i in cell j.

In this work, we first preprocess the raw scRNA-seq count data, including log transformation and z-score normalization. We have a normalized output X, which is given by

$$X^{'} = log_2(1 + diag(s_j)^{-1}C), \tag{1}$$

$$X = zscore(X^{'}), \tag{2}$$

where s_j is the size factor for every cell j. The advantage of data preprocessing is to preserve the impact of library size differences and transform discrete values to become continuous, allowing for greater flexibility for the subsequent modeling.

Deep Count Autoencoder

To denoise the data after preprocessing and capture the characters of scRNA-seq data, we employ DCA based on the ZINB model, so that we can obtain denoised data, which is beneficial to the stability and accuracy of the subsequent KNN algorithm. Taking the count distribution, overdispersion, and high sparsity of scRNA-seq data into account, DCA applies a ZINB model based on autoencoder to depict the characters of the data, and the loss function of the autoencoder is the likelihood of ZINB distribution.

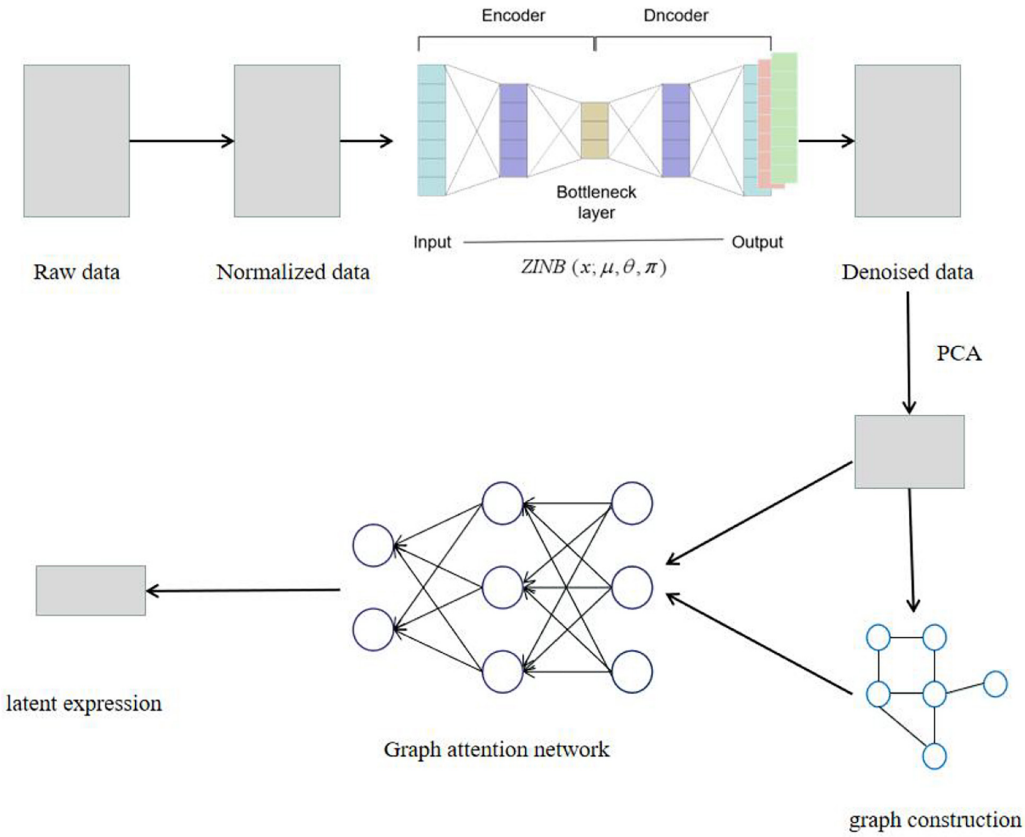

FIGURE 1 | The model architecture of SCDRHA. First, we normalize the raw data. Then, we use deep count autoencoder (DCA) to denoise the data. Ultimately, we use the compressed matrix by PCA and adjacency matrix as the input of graph autoencoder based on graph attention network (GAT) and get a low-dimensional embedding.

The ZINB distribution is a mixture model that consists of two components: a point mass at zero and a negative binomial (NB) component.

$$NB\ (x;\ \mu, \theta) = \frac{\Gamma(x+\theta)}{\Gamma(\theta)\Gamma(x+1)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{x}, \quad (3)$$

$$ZINB\ (x;\ \pi,\ \mu, \theta) = \pi\,\delta_0\ (x) + (1-\pi)\ NB\ (x;\ \mu, \theta). \quad (4)$$

where π, μ, and θ are the parameters of ZINB distribution, which represent the probability of dropout events, mean, and dispersion, respectively. DCA estimates three parameters by using an autoencoder framework; the formulation of the architecture is given below:

$$\begin{aligned} E &= \mathrm{Re}LU(XW_E), \\ B &= \mathrm{Re}LU(EW_B), \\ D &= \mathrm{Re}LU(BW_{\mu}), \\ M &= diag(s_j)\exp(DW_{\mu}), \\ \Pi &= sigmoid(DW_{\pi}), \\ \Theta &= \exp(DW_{\theta}), \end{aligned} \quad (5)$$

where E, B, and D represent the encoder, bottleneck, and decoder layers, respectively. The loss function of DCA is the negative log of the ZINB likelihood:

$$\widehat{\Pi}, \widehat{\mathrm{M}}, \widehat{\Theta} = argmin_{\Pi,\mathrm{M},\Theta} NLL_{ZINB}(X;\ \Pi,\ \mathrm{M},\ \Theta) + \lambda||\Pi||_F^2. \quad (6)$$

where the NLL_{ZINB} function represents the negative log-likelihood of ZINB distribution.

Graph Autoencoder Based on GATs

Graph autoencoder is a very powerful neural network architecture for unsupervised representation learning on graph-structured data. Compared with regular autoencoder, graph autoencoder applies graph neural networks in the encoder, which can better map the graph-structured data. In this work, we construct a graph autoencoder based on GAT to project the high-dimensional data to a low-dimensional latent space. GAT is a novel neural network architecture that extracts the features of the graph and preserves topological structure among cells.

Because the denoised data by DCA have the same dimension as the raw count, we select PCA to embed the gene expression matrix into an intermediate dimension. We select the first F

principal components as the output matrix H of PCA. In this way, it can not only shorten the run time of the subsequent modeling but also enhance the performance of the KNN algorithm to build a more stable and accurate graph.

GAT aims to obtain a power expressive to transform the input feature $H = \{\vec{h}_1, \vec{h}_2, ..., \vec{h}_N\}$ into higher-level feature $H' = \{\vec{h}'_1, \vec{h}'_2, ..., \vec{h}'_N\}$, $\vec{h}_i \in R^F$, and $\vec{h}'_i \in R^{F'}$. GAT learns the final output features of each node by using the information of their neighbor nodes:

$$\vec{h}'_i = \sum_{j \in N_i} \alpha_{ij} W \vec{h}_j, \tag{7}$$

where α_{ij} represents the importance of node *j*'s features to node *i*, W is a shared weight matrix, and $j \in N_i$, N_i is some neighbor of node *i* in the KNN graph. The formula of α_{ij} is given below:

$$\alpha_{ij} = softmax_j(e_{ij}) = \frac{exp\ (e_{ij})}{\sum_{k \in N_i} exp\ (e_{ik})}, \tag{8}$$

where e_{ij} is the attention coefficient, it is defined as:

$$e_{ij} = a(W\vec{h}_i, W\vec{h}_j), \tag{9}$$

where the attention mechanism *a* is a single-layer feedforward neural network. To make coefficients e_{ij} (9) easily compare across different nodes, GAT applies softmax function to normalize them; we can obtain α_{ij} (8). GAT applies the LeakyReLU function as the activation function. After fully expanding out, the coefficients α_{ij} can be expressed as:

$$\alpha_{ij} = \frac{exp\ (LeakyReLU(\vec{a}^T[W\vec{h}_i || W\vec{h}_j]))}{\sum_{k \in N_i} exp\ (LeakyReLU(\vec{a}^T[W\vec{h}_i || W\vec{h}_k]))}, \tag{10}$$

where $\vec{a} \in R^{2F'}$ is a weight vector, and $||$ is the concatenation operation.

Our graph autoencoder has two inputs: compressed expression matrices H by PCA and adjacency matrices A. We apply GAT in the encoder. In our experiments, we encoder the inputs into two latent expressions, and then decode them into the reconstruct expression matrices H' and adjacency matrices A'. The objective of the learning process is to minimize the reconstruction loss:

$$L = \gamma||H - H'||_2^2 + (1 - \gamma)||A - A'||_2^2. \tag{11}$$

TABLE 1 | Basic information about five real single-cell RNA sequencing (scRNA-seq) datasets.

Dataset	Cells	Genes	Clusters	Dropout rate (%)
10X PBMC	4,271	16,499	8	92.24
Mouse ES cell	2,717	24,046	4	65.76
Mouse bladder cell	2,746	19,079	16	94.87
Worm neuron cell	4,186	11,955	10	98.62
Zeisel	3,005	19,972	9	81.21

where γ is a hyperparameter; we set it to be 0.6 in our experiments. It is a hyperparameter, which is used to balance the reconstruction loss of expression matrix and adjacent matrix. Since we mainly use the low-dimensional representation of adjacency matrix for subsequent dimensionality reduction and visualization, we pay more attention to the reconstruction loss of adjacency matrix, and then give more weight to the reconstruction loss of adjacency matrix.

Convergence Analysis

SCDRHA consists of two core modules: DCA and graph autoencoder. How to train these two core modules is also a very important issue, and we give the setting of epochs when training them. Because we refer to the DCA in the process of noise reduction, we use the default value to train the DCA. For graph autoencoder, we first do pretraining, then global training; their epochs are set to 120 and 40, respectively. Because we find that when the number of epochs reaches this number, the value of the loss function of the graph autoencoder changes very little and tends to a more stable state; so, we have reason to think that the optimization objective tends to converge at this time.

TABLE 2 | Average silhouette value under different datasets.

Dataset	PCA	t-SNE	scGAE	SCDRHA
10X PBMC	0.066	0.129	0.112	**0.469**
Mouse ES cell	0.019	0.346	0.337	**0.411**
Mouse bladder cell	0.019	**0.251**	0.032	0.193
Worm neuron cell	−0.143	0.042	−0.026	**0.315**
Zeisel	−0.112	0.113	0.193	**0.317**

Bold values indicate the highest score in the row and the corresponding method has the best performance.

TABLE 3 | Normalized Mutual Information (NMI) score under different datasets.

Dataset	PCA	t-SNE	DCA	scGAE	SCDRHA
10X PBMC	0.320	0.536	0.735	0.650	**0.793**
Mouse ES cell	0.518	0.594	0.856	0.787	**0.951**
Mouse bladder cell	0.522	0.673	0.648	0.664	**0.732**
Worm neuron cell	0.197	0.426	0.467	0.532	**0.752**
Zeisel	0.255	0.469	0.452	0.636	**0.727**

Bold values indicate the highest score in the row and the corresponding method has the best performance.

TABLE 4 | ARI score under different datasets.

Dataset	PCA	t-SNE	DCA	scGAE	SCDRHA
10X PBMC	0.180	0.356	0.723	0.434	**0.781**
Mouse ES cell	0.224	0.594	0.852	0.771	**0.971**
Mouse bladder cell	0.226	0.413	0.529	0.442	**0.550**
Worm neuron cell	0.032	0.290	0.280	0.246	**0.674**
Zeisel	0.129	0.326	0.313	0.502	**0.627**

Bold values indicate the highest score in the row and the corresponding method has the best performance.

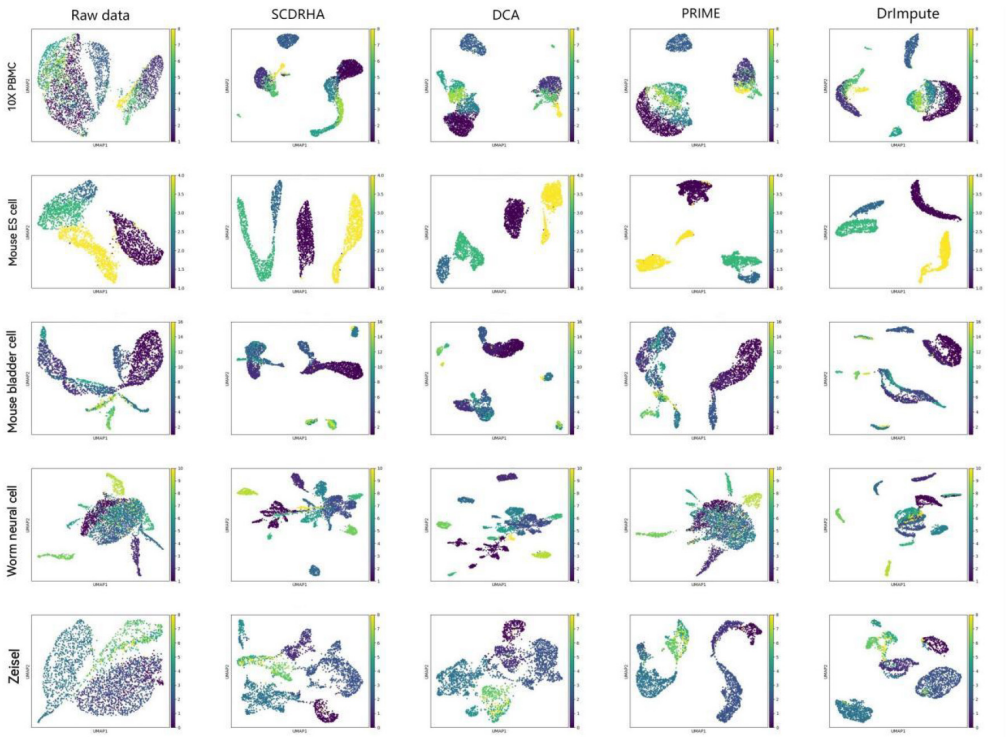

FIGURE 2 | Cell visualization results for all single-cell RNA sequencing (scRNA-seq) datasets. The columns, from left to right, represent the raw data and the data visualization after noise reduction by SCDRHA, DCA, PRIME, and DrImpute. The rows, from left to right, represent visualizations of 10X PBMC, Mouse ES cell, Mouse bladder cell, Worm neuron cell, and Zeisel datasets.

RESULTS

Datasets

To assess the performance of SCDRHA, we focus on relatively large datasets; five real scRNA-seq datasets with known cell types are selected. The basic information about five real datasets is summarized in **Table 1**, and below, we describe these datasets.

(i) The 10X PBMC (Zheng et al., 2017) dataset is provided by the 10X scRNA-seq platform, which is from a healthy human.[1] (ii) The Mouse ES cell (Klein et al., 2015) dataset profiles the transcriptome of the heterogeneous onset of differentiation of mouse embryonic stem cells after Leukemia Inhibitory Factor (LIF) (Klein et al., 2015) withdrawal GSE65525. (iii) The Mouse bladder cell (Han et al., 2018) dataset is from the Mouse Cell Atlas project GSE108097. From the raw count matrix, we select about ∼2,700 cells from bladder tissue. (iv) The Worm neuron cell (Cao et al., 2017) dataset is profiled by single-cell combinatorial indexing RNA sequencing (sci-RNA-seq), which is from the nematode *Caenorhabditis elegans* at the L2 larval stage.[2] (v) The Zeisel et al. (2015) dataset contains 3,005 cells, which are collected from the mouse cortex and hippocampus GSE60361.

The Evaluation of SCDRHA in Dimensionality Reduction

In our experiments, four popular dimension reduction algorithms are used to compare with our algorithm SCDRHA

[1]https://support.10xgenomics.com/single-cell-gene-expression/

[2]http://atlas.gs.washington.edu/worm-rna/docs/

TABLE 5 | NMI score under different datasets.

Dataset	Raw data	DrImpute	PRIME	DCA	SCDRHA
10X PBMC	0.320	0.716	0.682	0.735	**0.793**
Mouse ES cell	0.518	0.609	0.643	0.856	**0.951**
Mouse bladder cell	0.522	0.721	0.693	0.648	**0.732**
Worm neuron cell	0.197	0.665	0.376	0.467	**0.752**
Zeisel	0.255	0.605	0.574	0.452	**0.727**

Bold values indicate the highest score in the row and the corresponding method has the best performance.

TABLE 6 | ARI score under different datasets.

Dataset	Raw data	DrImpute	PRIME	DCA	SCDRHA
10X PBMC	0.180	0.654	0.583	0.732	**0.781**
Mouse ES cell	0.224	0.474	0.497	0.852	**0.971**
Mouse bladder cell	0.226	0.477	0.463	0.529	**0.550**
Worm neuron cell	0.032	0.396	0.215	0.280	**0.674**
Zeisel	0.129	0.465	0.460	0.313	**0.627**

Bold values indicate the highest score in the row and the corresponding method has the best performance.

in five real datasets. These four dimension reduction algorithms include two traditional algorithms (PCA and tSNE) and two novel algorithms for dimensionality reduction of scRNA-seq data (DCA and scGAE).

Firstly, we compare SCDRHA with PCA, t-SNE, and scGAE and use average silhouette value (Rousseeuw, 1987) to evaluate the performance of these methods. It is worth noting that we compress the data into 10 dimensions for comparison, except t-SNE, and do not modify the default parameters in the algorithm. Because the algorithm DCA compresses the data to 32 dimensions by default, it is not selected in this experiment.

As is shown in **Table 2**, only on the Mouse bladder cell dataset, t-distributed stochastic neighbor embedding (t-SNE) performs better than SCDRHA. On the other four datasets, the dimension

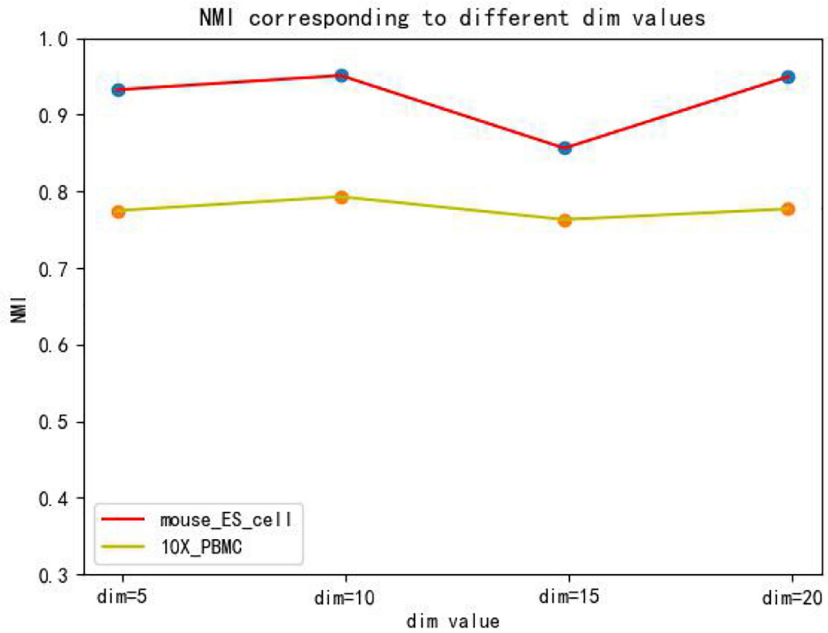

FIGURE 3 | The influence of hidden layer nodes on SCDRHA under Normalized Mutual Information (NMI).

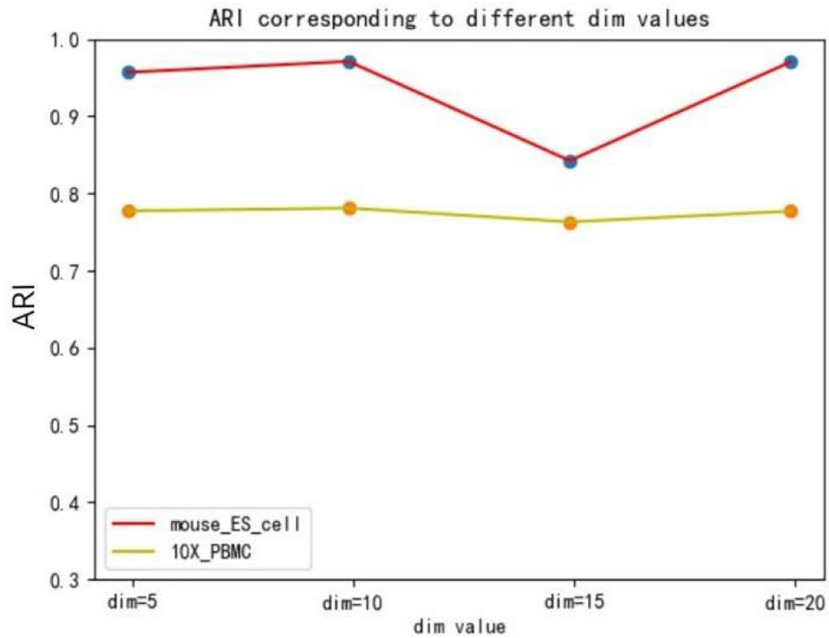

FIGURE 4 | The influence of hidden layer nodes on SCDRHA under adjusted Rand index (ARI).

reduction performance of SCDRHA is obviously better than other methods. The t-SNE is a non-linear dimension reduction algorithm widely used in single-cell dimension reduction and visualization; it can directly project high-dimensional data into two to three dimensions. The Mouse bladder cell dataset has 16 cell clusters; more clusters will distort the computation of average silhouette value.

In order to further test the dimension reduction performance of SCDRHA, we use the embedding expression of different dimensionality reduction methods for clustering analysis. Besides, Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002) and Adjusted Rand index (ARI) (Rand, 1971) are used to evaluate the performance of clustering analysis. To make the results easily comparable across different methods, we employ K-means for clustering analysis and set the parameter K as the real number of clusters in each dataset.

As shown in **Tables 3**, **4**, our experiments illustrate that SCDRHA is superior to other methods in all datasets. It is worth noting that SCDRHA overtakes t-SNE on the Mouse bladder cell dataset, which indicates that denoising single-cell data before dimension reduction can improve the performance of the subsequent analysis.

In a word, our experiments demonstrate that SCDRHA has batter performance in dimension reduction than that other existing methods.

The Evaluation of SCDRHA in Noise Reduction

Since SCDRHA involves the module of noise reduction, we compare SCDRHA with other denoising methods including DCA, PRIME (Jeong and Liu, 2020), and DrImpute (Gong et al., 2018). These methods aim to impute dropout events in scRNA-seq data. At the same time, we also compare the denoised data with the original data.

Visualizing complex, high-dimensional scRNA-seq data in a way that is both easy to understand and faithful to the data is a meaningful task. To further evaluate the SCDRHA comprehensively, we employ UMAP to project the denoised data and the original data into two dimensions for cell visualization. **Figure 2** shows the results of cell visualization for all five scRNA-seq datasets.

We can discover that SCDRHA can clearly divide different types of cells into different clusters. SCDRHA has a better

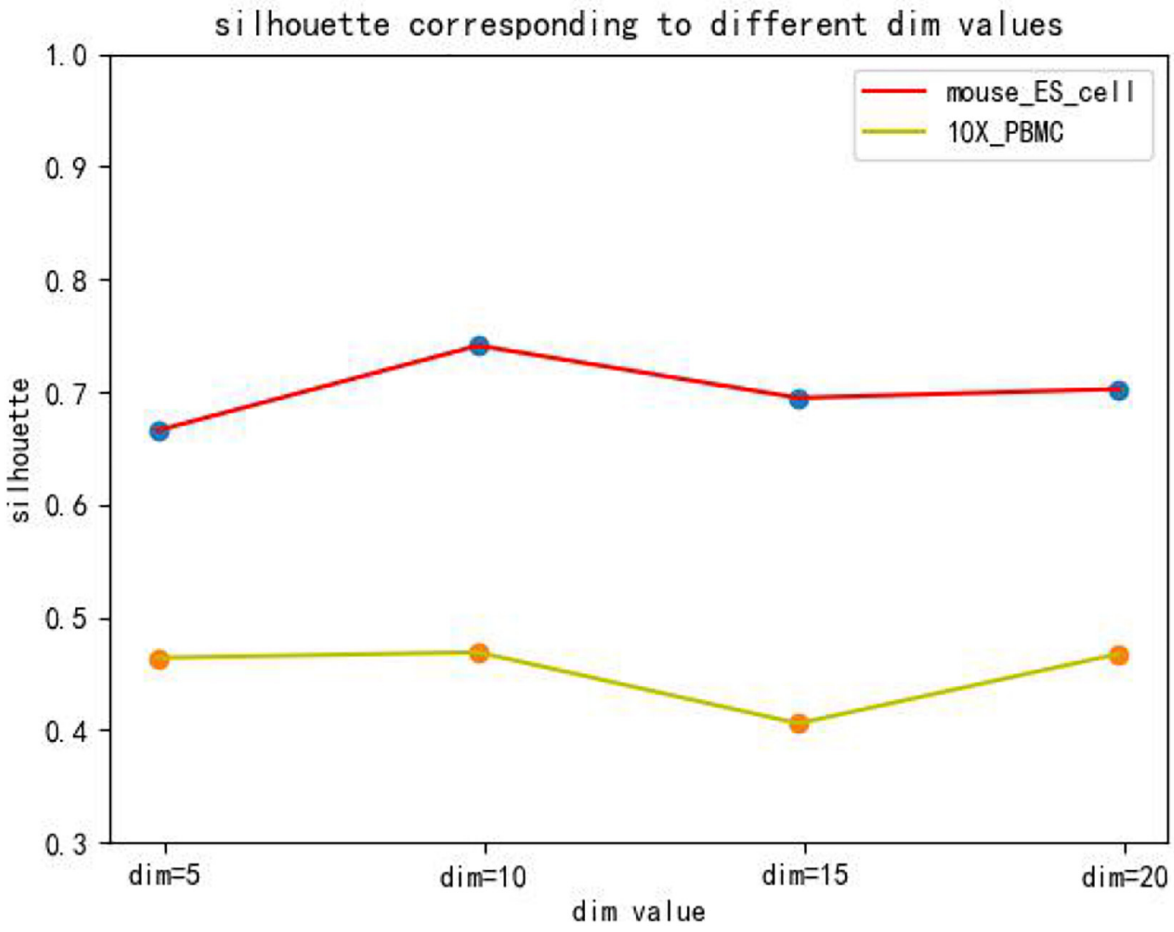

FIGURE 5 | The influence of hidden layer nodes on SCDRHA under Silhouette.

performance on cell visualization than other methods. Comparing raw data with denoised data, we can find that SCDRHA remarkably improves the performance of data visualization. The results demonstrate that SCDRHA has a good ability for noise reduction.

To further evaluate the performance of noise reduction. We also apply K-means for clustering and use the NMI and ARI to assess their ability, thereby testing these methods indirectly. Before clustering analysis, we project the raw data and denoised data into the same dimensions by PCA.

The two metrics (NMI and ARI) of clustering performance are presented in **Tables 5**, **6**. We observe that the clustering results of SCDRHA are better than other algorithms on the five selected datasets. In addition, denoising can significantly enhance the ability of clustering.

Parameter Sensitivity Analysis

The hidden layer nodes of the graph autoencoder are a hyperparameter in SCDRHA, which directly determines the dimension of the final latent expression. To analyze the influence of the hidden layer nodes of graph autoencoder on SCDRHA, we select two datasets (Mouse ES cell and 10X PBMC) as the test datasets. The numbers of hidden layer nodes are set to 5, 10, 15, and 20, respectively. We use the latent space of different dimensions for clustering analysis. Three metrics are applied for analysis. The experimental results are summarized in **Figure 3**.

Figures 3, **4**, **5** show that different values of hidden layer nodes have a slight variation in the dimension reduction and clustering analysis, and when we selected the total number of nodes is 10, the performance under the three indexes is the best. Based on this analysis, the default parameter of the hidden layer nodes in graph autoencoder is set to 10.

Implementation

The SCDRHA is implemented on HP Z840 workstation with 32GB RAM. SCDRHA consists of two portions: one is DCA and the other is graph autoencoder. We refer to the original code of DCA, which is constructed based on TensorFlow 1.15.0[3] and implement DCA using SCANPY 1.7.1, a Python package. We refer to scGAE[4] to build a graph autoencoder that is based on TensorFlow 2.4.1 and Python package spektral 0.6.1. Code and data used in this paper are available at https://github.com/WHY-17/SCDRHA.

Software Package and Setting

When comparing with other methods, we followed the package and instructions provided by the author of each method. We basically use the default parameters of each package, and we used the following packages: (i) PRIME,[5] (ii) DrImpute,[6] (iii) DCA (see text footnote 3), and (iv) scGAE (see text footnote 4).

[3]https://github.com/theislab/dca

[4]https://github.com/ZixiangLuo1161/scGAE

[5]https://github.com/hyundoo/PRIME

[6]https://github.com/gongx030/DrImpute

CONCLUSION

Because of the high dimension of scRNA-seq, many dimension reduction methods have been proposed for scRNA-seq data in recent years. Nevertheless, these dimension reduction methods have some limitations in solving dropout events or maintaining local and global structure in the high-dimensional data. In conclusion, we propose SCDRHA, a scRNA-seq data dimensionality reduction algorithm based on a hierarchical autoencoder. scDeepCluster can learn a latent embedded representation that can denoise the data and preserve the topological structure. SCDRHA denoises the scRNA-seq data to obtain a more stable structure for the subsequent process. To obtain a low-dimension expression and retain the topological structure of single-cell data, we build a graph autoencoder based on GAT. Experimental results demonstrate that SCDRHA has better performance than existing state-of-the-art algorithms on dimension reduction and noise reduction in five real scRNA-seq datasets. Besides, SCDRHA can also dramatically enhance the performance of data visualization and cell clustering. With the rapid development of scRNA-seq technology, the data structure we get is more and more complex. Learning a more flexible and universal distribution to fit the data may be our future research direction.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed for this study. These can be found in the following links: 10X PBMC (https://support.10xgenomics.com/single-cell-gene-expression/), Mouse ES cell (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65525), Mouse bladder cell (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108097), Worm neuron cell (http://atlas.gs.washington.edu/worm-rna/docs/), and Zeisel (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60361).

AUTHOR CONTRIBUTIONS

NW, HW, and JZ constructed the original idea, designed the experiments, and wrote the manuscript. YS and CZ proofread the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the Xinjiang Autonomous Region University Research Program (Nos. XJEDU2019Y002 and XJ2021G023), Key Program of Natural Science Project of Educational Commission of Anhui Province (No. KJ2019A0029), and the National Natural Science Foundation of China (Nos. U19A2064 and 61873001).

REFERENCES

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667. doi: 10.1126/science.aam8940

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* 10, 390–403. doi: 10.1038/s41467-018-07931-2

Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., and Garry, D. J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* 19:220. doi: 10.1186/s12859-018-2226-y

Han, X., Wang, R., Zhou, Y., Fei, L., and Guo, G. (2018). Mapping the mouse cell atlas by microwell-seq. *Cell* 172, 1091–1107.e17. doi: 10.1016/j.cell.2018.02.001

Hie, B., Peters, J., Nyquist, S. K., Shalek, A. K., Berger, B., and Bryson, B. D. (2020). Computational methods for single-cell RNA sequencing. *Soc. Sci. Electr. Publ.* 3, 339–364. doi: 10.1146/annurev-biodatasci-012220-100601

Jeong, H., and Liu, Z. (2020). PRIME: a probabilistic imputation method to reduce dropout effects in single cell RNA sequencing. *Bioinformatics* 36, 4021–4029. doi: 10.1093/bioinformatics/btaa278

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* 20, 273–282. doi: 10.1038/s41576-018-0088-9

Klein, A., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi: 10.1016/j.cell.2015.04.044

Luo, Z., Xu, C., Zhang, Z., and Jin, W. (2021). scGAE: topology-preserving dimensionality reduction for single-cell RNA-seq data using graph autoencoder. *Preprint bioRxiv* [Preprint] doi: 10.1101/2021.02.16.431357

Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* 16, 241–250. doi: 10.1186/s13059-015-0805-z

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* 66, 846–850. doi: 10.1080/01621459.1971.10482356

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9, 284–300. doi: 10.1038/s41467-017-02554-5

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Strehl, A., and Ghosh, J. (2002). Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3, 583–617.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2013). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Tian, T., Wan, J., Song, Q., and Wei, Z. (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nat. Mach. Intell.* 1, 191–198. doi: 10.1038/s42256-019-0037-0

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2018). "Graph attention networks," in *Proceedings of the International Conference on Learning Representations*, Vancouver, BC.

Wang, C., Gao, Y. L., Kong, X. Z., Liu, J. X., and Zheng, C. H. (2021a). Unsupervised cluster analysis and gene marker extraction of scRNA-seq data based on non-negative matrix factorization. *IEEE J. Biomed. Health Inform.* doi: 10.1109/JBHI.2021.3091506 [Epub ahead of print].

Wang, C., Gao, Y. L., Liu, J. X., Kong, X. Z., and Zheng, C. H. (2021b). Single-cell RNA sequencing data clustering by low-rank subspace ensemble framework. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2020.3029187 [Epub ahead of print].

Xi, J., Li, A., and Wang, M. (2018). HetRCNA: a novel method to identify recurrent copy number alternations from heterogeneous tumor samples based on matrix decomposition framework. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 422–434. doi: 10.1109/tcbb.2018.2846599

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., and Huang, Q. (2020). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 36, 1855–1863.

Yu, Z., Liu, H., Du, F., and Tang, X. (2021). GRMT: generative reconstruction of mutation tree from scratch using single-cell sequencing data. *Front. Genet.* 12:692964. doi: 10.3389/fgene.2021.692964

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. doi: 10.1126/science.aaa1934

Zhang, L., and Zhang, S. (2018). Comparison of computational methods for imputing single-cell RNA sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 376–389. doi: 10.1109/TCBB.2018.2848633

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 1–12. doi: 10.1038/ncomms14049

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 September 2021
doi: 10.3389/fgene.2021.718915

Multi-View Spectral Clustering Based on Multi-Smooth Representation Fusion for Cancer Subtype Prediction

Jian Liu[1,2], *Shuguang Ge*[1,2], *Yuhu Cheng*[1,2] *and Xuesong Wang*[1,2]*

[1] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China, [2] Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou, China

It is a vital task to design an integrated machine learning model to discover cancer subtypes and understand the heterogeneity of cancer based on multiple omics data. In recent years, some multi-view clustering algorithms have been proposed and applied to the prediction of cancer subtypes. Among them, the multi-view clustering methods based on graph learning are widely concerned. These multi-view approaches usually have one or more of the following problems. Many multi-view algorithms use the original omics data matrix to construct the similarity matrix and ignore the learning of the similarity matrix. They separate the data clustering process from the graph learning process, resulting in a highly dependent clustering performance on the predefined graph. In the process of graph fusion, these methods simply take the average value of the affinity graph of multiple views to represent the result of the fusion graph, and the rich heterogeneous information is not fully utilized. To solve the above problems, in this paper, a Multi-view Spectral Clustering Based on Multi-smooth Representation Fusion (MRF-MSC) method was proposed. Firstly, MRF-MSC constructs a smooth representation for each data type, which can be viewed as a sample (patient) similarity matrix. The smooth representation can explicitly enhance the grouping effect. Secondly, MRF-MSC integrates the smooth representation of multiple omics data to form a similarity matrix containing all biological data information through graph fusion. In addition, MRF-MSC adaptively gives weight factors to the smooth regularization representation of each omics data by using the self-weighting method. Finally, MRF-MSC imposes constrained Laplacian rank on the fusion similarity matrix to get a better cluster structure. The above problems can be transformed into spectral clustering for solving, and the clustering results can be obtained. MRF-MSC unifies the above process of graph construction, graph fusion and spectral clustering under one framework, which can learn better data representation and high-quality graphs, so as to achieve better clustering effect. In the experiment, MRF-MSC obtained good experimental results on the TCGA cancer data sets.

Keywords: multi-view clustering, cancer subtypes prediction, multi-omics data, spectral clustering, smooth representation, graph fusion

OPEN ACCESS

Edited by:
Zhenhua Yu,
Ningxia University, China

Reviewed by:
Xiaofan Lu,
China Pharmaceutical University, China
Hai-tao Li,
Southeast University, China

****Correspondence:***
Xuesong Wang
wangxuesongcumt@163.com

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

Received: *01 June 2021*
Accepted: *05 August 2021*
Published: *06 September 2021*

Citation:
Liu J, Ge S, Cheng Y and Wang X (2021) Multi-View Spectral Clustering Based on Multi-Smooth Representation Fusion for Cancer Subtype Prediction. Front. Genet. 12:718915. doi: 10.3389/fgene.2021.718915

INTRODUCTION

Cancer is a malignant and heterogeneous disease caused by changes in cellular and molecular expression, epigenetics, transcription, and proteome levels (Burrell et al., 2013). This heterogeneity is reflected in the fact that the same type of cancer will produce subtypes with different representations, which will further affect the clinical treatment plan and prognosis (Bedard et al., 2013). With the development and maturity of the new generation of sequencing technologies, a large number of multi-omics biological data have been collected in some public data sets and are easily accessible to researchers (Schuster, 2008). The Cancer Genome Atlas (TCGA) is a landmark cancer genomics project that stores biological information including mRNA expression data, methylation data, miRNA expression data, and gene mutation data from more than 30 type of cancers and thousands of cancer patients. Therefore, it is particularly important to build a clustering model that makes full use of these biological information to solve the problem of discovering cancer subtypes (Akbani et al., 2014).

In recent years, some effective multi-view clustering methods have been designed and applied to biological data (Shen et al., 2010; Zhang et al., 2012; Mo et al., 2013; Wang et al., 2014; Meng et al., 2016; Ma and Zhang, 2017; Shi et al., 2017; Guo et al., 2019; Yu et al., 2019). In order to achieve the task of clustering, scholars initially focused on feature selecting and feature dimensionality reduction techniques. They all used different strategies to transform or project high-dimensional data into low-dimensional feature space and then realized clustering through K-means. For example, iCluster (Shen et al., 2010) is a Gaussian hidden variable model, and its extended version, iClusterPluse (Mo et al., 2013), is an effective and classical multi-omics data clustering method. It considers that different variable types follow different linear probability relationships, and then constructs a joint sparse model to complete feature selecting and sample clustering tasks. However, iClusterPlus has an obvious drawback: it includes a pre-selecting process for genes that filters out important information, and the clustering results are sensitive to this operation. In order to solve the problem of data preprocessing, many classical dimensionality reduction techniques are applied to the proposed clustering algorithms, e.g., Principal Component Analysis (PCA; Ding and He, 2004), Non-negative Matrix Factorization (NMF; Zhang et al., 2012), etc. Shi et al. (2017) applied the improved PCA to design Pattern Fusion Analysis (PFA) method, which projects each data set into a low-dimensional feature space with local patterns while reducing noise. Then PFA uses the dynamic collimation algorithm to achieve the fusion of feature space.

The above methods only focus on the characteristics of each kind of omics data, without considering the structural characteristics of the data, which can reveal the potential similarity between samples and has great guiding significance for the study of data representation. Considering that the sample (patient) size of the biological data is much smaller than the feature (gene) size, some methods for cancer subtype prediction based on graph learning have been designed. Based on cancer samples, graph learning can quickly construct similar graphs and eventually transform them into spectral clustering problems to achieve clustering. For example, Wang et al. (2014) proposed a widely used clustering algorithm for multi-omics data, named as Similarity Network Fusion (SNF). SNF uses the exponential similarity kernel method to construct a sample similarity network for each data type instead of the dimensionality reduction process, and then uses the nonlinear information fusion technology to integrate these networks into a single similarity network. Inspired by SNF, Ma and Zhang (2017) proposed Affinity Network Fusion (ANF) method, which constructs K-nearest neighbor similar networks of patients for each data type, and then fused these networks based on random step size method. Other algorithms based on graph learning are also very effective in the recognition of cancer subtypes. For example, Yu et al. (2019) proposed Multi-view Clustering using Manifold Optimization (MVCMO), which uses linear search on Stiefel manifold space to solve the spectral clustering optimization problem.

The above methods all use the original omics data matrix to construct the similarity matrix, and fuse the obtained multiple similarity matrices, ignoring the learning of the similarity matrix. In the process of graph fusion, the similarity between sample points is usually different in different views. Some existing algorithms simply take the average value of the affinity graph of multi-omics to represent the result of the fusion graph, and the rich heterogeneous information is not fully utilized. In addition, most of the graph-based multi-view clustering methods separate the data clustering process from the graph learning process, which makes the graph construction independent of the clustering task, leading to the clustering performance highly dependent on the predefined graph. In this paper, we design a Multi-view Spectral Clustering method based on Multi-smooth Representation (MRF-MSC) for the exploration of cancer subtypes. MRF-MSC combines graph learning, graph fusion and spectral clustering into one framework to avoid the above problems. Firstly, MRF-MSC uses the graph regularization method to calculate the smooth representation of each omics data type. The original feature space raw data can be effectively projected into the corresponding sample similarity subspace. The smooth representation can explicitly enhance the grouping effect, that is, it enhances the similarity between samples of the same category and reduces the similarity between samples of different categories (Hu et al., 2014). Secondly, the multi-smooth representation matrices of multi-omics data are integrated to form a fused similarity matrix. Considering that each omics data is of different importance to the prediction of cancer subtypes, MRF-MSC adaptively weights the smooth regularization representation of each omics data by using the self-weighting method in the process of graph fusion. Finally, MRF-MSC optimizes the fused similarity matrix through constrained Laplacian rank to learn a new block diagonal matrix with k connected components (k is the number of classes), which is beneficial for clustering. This problem can be solved by using spectral clustering (Ng et al., 2001). Spectral clustering is a classical data clustering method and widely used in multi-view clustering algorithms (Nie et al., 2016; Kang et al., 2020; Feng et al., 2021; Ge et al., 2021) recently. In order to verify

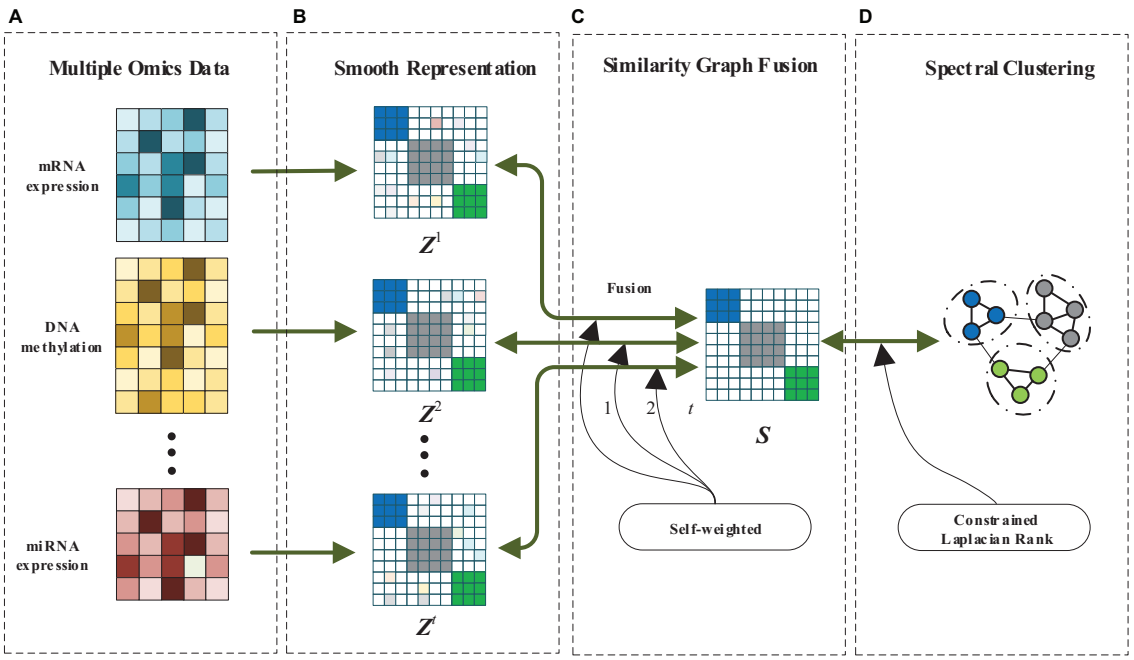

FIGURE 1 | The framework of MRF-MSC. **(A)** Multiple omics data. **(B)** Smooth representation. **(C)** Similarity graph fusion. **(D)** Spectral clustering.

the effectiveness of MRF-MSC, cancer subtypes prediction experiments were carried out on TCGA data sets. The results showed that MRF-MSC was able to obtain more significant clinical differences in cancer typing. In the Breast Invasive Carcinoma (BRCA) analysis, the MRF-MSC results validated previous clinical studies and identified biologically significant cancer subtypes.

MATERIALS AND METHODS

In this paper, we design a MRF-MSC for cancer subtypes prediction. The framework of MRF-MSC as shown in **Figure 1**. Given multi-omics data sets, we first calculate the similarity matrix with smooth representation for each data set to measure the similarity between sample points. Then, the graph fusion and self-weighted methods are used to integrate the multi-smooth representation into a fused similarity matrix. Finally, constrained Laplacian rank and spectral clustering are adopted to optimize the fused similarity matrix, and the clustering results can be obtained.

Smooth Representation of Multi-Omics Data

Given a set of cancer multi-omics data $\boldsymbol{X} = \{\boldsymbol{X}^1, \boldsymbol{X}^2, \cdots, \boldsymbol{X}^t\}$, $\boldsymbol{X}^v \in \mathbb{R}^{m^v \times n}$, where t is the number of data sets, $\boldsymbol{X}^v$ is the v-th omics data, m^v indicates that the v-th dataset has m features, n is the number of samples. In order to obtain the final fused similarity graph, we need to calculate the similarity matrix of each omics data $\boldsymbol{Z}= \{\boldsymbol{Z}^1, \boldsymbol{Z}^2, \cdots, \boldsymbol{Z}^t\}$, $\boldsymbol{Z}^v \in \mathbb{R}^{n \times n}$. This enables the raw omics data to be aggregated into their respective subspaces.

TABLE 1 | Detailed information on five types of cancer multi-omics data sets in Wang et al. (2014).

Cancer type	Number of genes			Number of samples
	mRNA	Methylation	miRNA	
GBM	12,042	1,305	534	215
BRCA	17,814	23,094	354	105
KIRC	17,899	24,960	329	122
LSCC	12,042	23,074	352	106
COAD	17,814	23,088	312	92

Take a single omics data $\boldsymbol{X}^v$ as an example, we introduce a self-representation method to measure the similarity between samples:

$$\boldsymbol{X}^v = \boldsymbol{X}^v \boldsymbol{Z}^v + \boldsymbol{E}^v \tag{1}$$

where $\boldsymbol{Z}^v$ is coefficient matrix which encodes the similarity between the data samples, E^v is error matrix. For Eq. 1, we explicitly strengthen the grouping effect between samples by smooth representation. This can enhance the similarity between samples of the same category and reduce the similarity between samples of different categories. The smooth representation can be roughly written as

$$\min_{\boldsymbol{Z}^v} \left|\left|\boldsymbol{X}^v - \boldsymbol{X}^v \boldsymbol{Z}^v\right|\right|_F^2 + \alpha \Omega\left(\boldsymbol{Z}^v\right) \ s.t. \ \boldsymbol{Z}^v \geq 0 \tag{2}$$

where α is a hyperparameter, Ω is the regularization term of the smooth representation. If two sample points are close to each other in the original feature space, then they should also maintain this property in the new feature space. That is, for samples i

TABLE 2 | Detailed information on five types of cancer multi-omics data sets in Rappoport and Shamir (2018).

Cancer type	Number of genes			Number of samples
	mRNA	Methylation	miRNA	
GBM	12,042	5,000	534	271
BRCA	20,531	5,000	1,046	622
KIRC	20,531	5,000	1,046	181
LSCC	20,531	5,000	1,046	337
COAD	20,531	5,000	705	213

and j, the following rules should be satisfied: $\left|\left|\boldsymbol{x}_i^{\boldsymbol{v}} - \boldsymbol{x}_j^{\boldsymbol{v}}\right|\right|_2 \rightarrow 0 \Rightarrow \left|\left|\boldsymbol{z}_i^{\boldsymbol{v}} - \boldsymbol{z}_j^{\boldsymbol{v}}\right|\right|_2 \rightarrow 0$, where $\boldsymbol{x^v}$ and $\boldsymbol{z^v}$ is the vector of X^v and Z^v, respectively. The smooth representation regularization term in Eq. 2 can be defined as

$$\Omega\left(\boldsymbol{Z}^v\right) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{w}_{ij}^v\left|\left|\boldsymbol{z}_i^v - \boldsymbol{z}_j^v\right|\right|_2 = \mathrm{tr}\left(\boldsymbol{Z}^v\boldsymbol{L}^t\left(\boldsymbol{Z}^v\right)^{\boldsymbol{T}}\right) \quad (3)$$

where tr is the trace of the matrix, and T is the transpose of the matrix, $\boldsymbol{w}_{ij}^v$ is an element in the weight matrix $\boldsymbol{W}^v$ that measures the similarity between sample points. $\boldsymbol{L}^v = \boldsymbol{D}^v - \boldsymbol{W}^v$ is the Laplacian matrix, where $\boldsymbol{D}^v$ is a diagonal degree matrix which diagonal elements satisfy $\boldsymbol{d}_{ii}^v = \sum_{j=1}^{n}\boldsymbol{w}_{ij}^v$. Now, there's a lot of ways to calculate W^v. Here, we construct W^v by using the most common used K-nearest neighbor method. Finally, Eq. 2 can be written as:

$$\min_{\boldsymbol{Z}^v}\left|\left|\boldsymbol{X}^v - \boldsymbol{X}^v\boldsymbol{Z}^v\right|\right|_F^2 + \alpha\mathrm{tr}\left(\boldsymbol{Z}^v\boldsymbol{L}^v\left(\boldsymbol{Z}^v\right)^{\mathrm{T}}\right)\ s.t.\ \mathrm{Z}^v \geq 0 \quad (4)$$

Through Eq. 4, the smooth representation $\boldsymbol{Z}^v$ of each omics data can be obtained.

The Fusion of Multi-Smooth Representations

How to integrate similar graphs in graph learning and make full use of the information of different data sets is the key of multi-view clustering method. After obtaining smooth representations $Z = \{\boldsymbol{Z}^1, \boldsymbol{Z}^2, \cdots, \boldsymbol{Z}^t\}$ of multi-omics data, we want to learn a fused similarity graph $\boldsymbol{S}$ that minimizes the difference between $\boldsymbol{S}$ and $\boldsymbol{Z}^v$. Then the graph fusion process of multi-smooth representations can be denoted as:

$$\min_{\boldsymbol{Z}^v,\boldsymbol{S}}\sum_{v=1}^{t}\left|\left|\boldsymbol{S} - \boldsymbol{Z}^v\right|\right|_F\ s.t.\ \boldsymbol{Z}^v \geq 0 \quad (5)$$

Considering that each omics data is of different importance to the prediction of cancer subtypes, we assign weighting factors $\varepsilon = \{\varepsilon^1, \varepsilon^2, \cdots, \varepsilon^t\}$ to $\boldsymbol{Z} = \{\boldsymbol{Z}^1, \boldsymbol{Z}^2, \cdots, \boldsymbol{Z}^t\}$. ε^v describes the contribution of the v-th smooth representation of each omics data to the graph fusion task. If $\boldsymbol{Z}^v$ is closer to $\boldsymbol{S}$, then its corresponding contribution weight ε^v is larger, which can reduce the impact of poor quality smooth representation on $\boldsymbol{S}$. Here, we adopt the self-weighting method in Nie et al. (2017) to carry out adaptive weighting for the smooth representation. The weighting factor of each smooth representation can be automatically tuned without any additional parameters.

Take the derivative of $\boldsymbol{Z}^v$ in Eq. 5 and set the derivative to zero, we have

$$\sum_{v=1}^{t}\varepsilon^v\frac{\partial\left(\left|\left|\boldsymbol{S} - \boldsymbol{Z}^v\right|\right|_F\right)}{\partial\boldsymbol{Z}^v} = 0 \quad (6)$$

where

$$\varepsilon^v = \frac{1}{2\left(\left|\left|\boldsymbol{S} - \boldsymbol{Z}^v\right|\right|_F\right)} \quad (7)$$

Since ε^v is calculated by $\boldsymbol{Z}^v$, Eq. 6 cannot be solved directly. However, if ε^v is assigned a fixed value as the weighting factor of each smooth representation, then Eq. 6 can be used to solve the following problems:

$$\min_{\boldsymbol{Z}^v,\boldsymbol{S}}\sum_{v=1}^{t}\varepsilon^v\left|\left|\boldsymbol{S} - \boldsymbol{Z}^v\right|\right|_F^2\ s.t.\ \boldsymbol{Z}^v \geq 0 \quad (8)$$

In Eq. 8, since both $\boldsymbol{Z}^v$ and $\boldsymbol{S}$ are goals to be solved, we cannot directly optimize the objective function. We can obtain the objective function of multi-smooth representation fusion by combining Eqs 4, 5 as:

$$\min_{Z^v,S}\sum_{v=1}^{t}\left(\left|\left|\boldsymbol{X}^v - \boldsymbol{X}^v\boldsymbol{Z}^v\right|\right|_F^2 + \alpha\mathrm{tr}\left(\boldsymbol{Z}^v\boldsymbol{L}^v\left(\boldsymbol{Z}^v\right)^{\mathrm{T}}\right) + \beta\varepsilon^v\left|\left|\boldsymbol{S} - \boldsymbol{Z}^v\right|\right|_F^2\right)\ s.t.\ \boldsymbol{Z}^v \geq 0 \quad (9)$$

where β is a hyperparameter.

By solving the above problem, we can learn the smooth representations and fused similarity graph of multi-omics data.

TABLE 3 | Comparison of *P*-values of survival analysis between MRF-MSC and other algorithms on five cancer multi-omics data sets in Wang et al. (2014).

Cancer types	Methods					
	MRF-MSC	iClusterPlus	PFA	SNF	ANF	MVSCO
GBM	**1.71E-5**	2.98E-2	1.82E-4	5.01E-5	5.83E-4	1.42E-3
BRCA	**1.31E-5**	5.52E-2	3.10E-4	6.91E-4	3.62E-4	3.54E-4
KIRC	**1.70E-2**	1.14E-1	7.45E-2	2.90E-2	4.97E-2	1.96E-2
LSCC	**6.58E-4**	5.17E-2	1.13E-2	1.10E-2	8.92E-3	9.13E-3
COAD	**8.24E-4**	4.96E-2	6.71E-2	2.42E-3	9.02E-3	8.51E-3

The best results have been highlighted in bold.

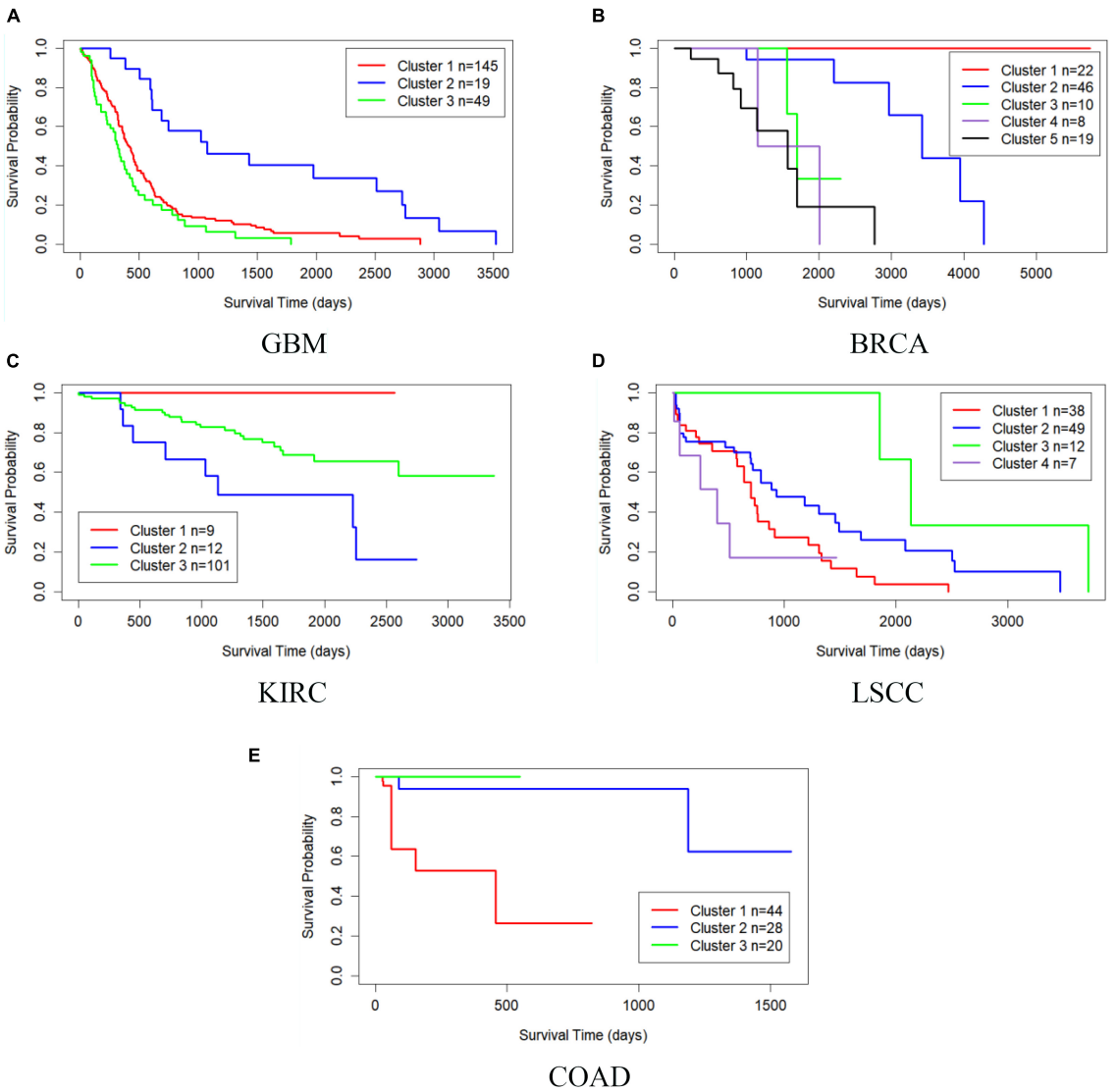

FIGURE 2 | Kaplan–Meier survival curves of MRF-MSC on the cancer types in Wang et al. (2014). **(A)** GBM, **(B)** BRCA, **(C)** KIRC, **(D)** LSCC, and **(E)** COAD.

TABLE 4 | Comparison of *P*-values of survival analysis between MRF-MSC and other algorithms on five cancer multi-omics data sets in Rappoport and Shamir (2018).

Cancer type	Methods					
	MRF-MSC	**iClusterPlus**	**PFA**	**SNF**	**ANF**	**MVSCO**
GBM	**1.43E-6**	3.83E-3	–	7.69E-6	2.17E-1	6.59E-4
	(*k* = 2)	(*k* = 10)		(*k* = 2)	(*k* = 3)	(*k* = 2)
BRCA	**5.25E-13**	1.55E-2	3.54E-9	4.38E-9	2.30E-11	4.26E-12
	(*k* = 4)	(*k* = 4)	(*k* = 3)	(*k* = 3)	(*k* = 5)	(*k* = 3)
KIRC	**7.10E-6**	2.10E-2	2.93E-3	2.53E-2	4.22E-3	2.71E-4
	(*k* = 4)	(*k* = 4)	(*k* = 3)	(*k* = 2)	(*k* = 2)	(*k* = 3)
LSCC	**9.13E-4**	4.63E-3	1.10E-1	9.45E-2	2.19E-2	1.37E-2
	(*k* = 2)	(*k* = 3)	(*k* = 3)	(*k* = 2)	(*k* = 2)	(*k* = 2)
COAD	2.63E-1	7.05E-1	3.21E-1	1.52E-1	**7.68E-2**	1.29E-1
	(*k* = 3)	(*k* = 2)	(*k* = 2)	(*k* = 3)	**(*k* = 3)**	(*k* = 2)

The best results have been highlighted in bold. –Denotes that the algorithm cannot get the clustering result on the data.

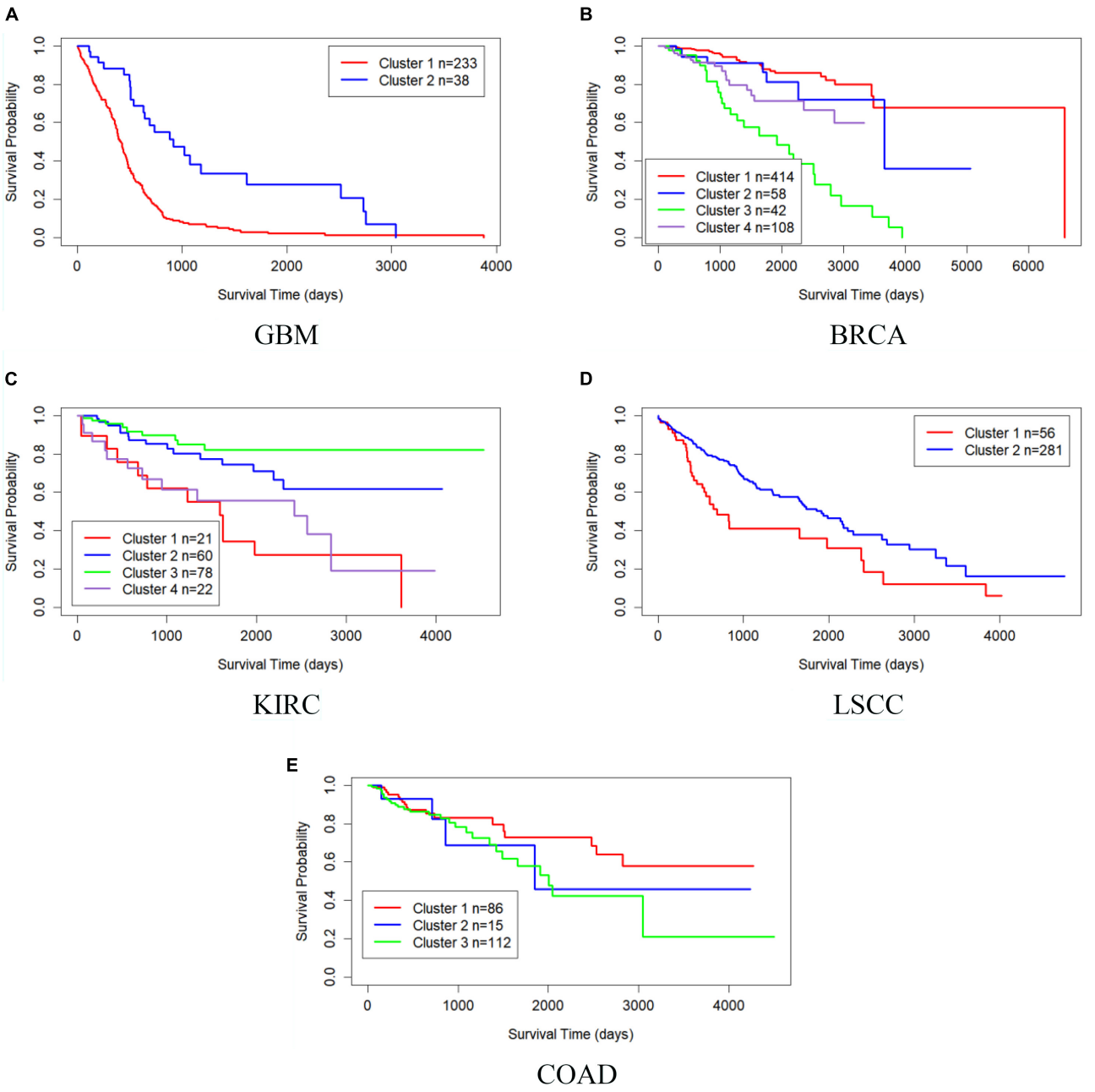

FIGURE 3 | Kaplan–Meier survival curves of MRF-MSC on the cancer types in Rappoport and Shamir (2018). **(A)** GBM, **(B)** BRCA, **(C)** KIRC, **(D)** LSCC, and **(E)** COAD.

In addition, the smooth representation is dynamically weighted during the fusion process, which effectively reduces the influence of the smooth representation of low-quality omics data on the fused similarity graph.

Multi-View Spectral Clustering Based on Multi-Smooth Representation Fusion

After calculating the fused similarity graph $\boldsymbol{S}$, although we can directly cluster $\boldsymbol{S}$ based on spectral clustering, the $\boldsymbol{S}$ obtained by Eq. 9 may not be optimal for the final clustering task. So, we attempt to optimize the clustering structure of S.

Ideally, a graph that is best for clustering tasks should have exactly k connected components, that is, data points are formed into k clusters. This can be done according to the following theorem.

Theorem 1. The number of connected components k of the graph $\boldsymbol{S}$ is equal to the multiplicity of zero eigenvalues of its Laplacian matrix L_S.

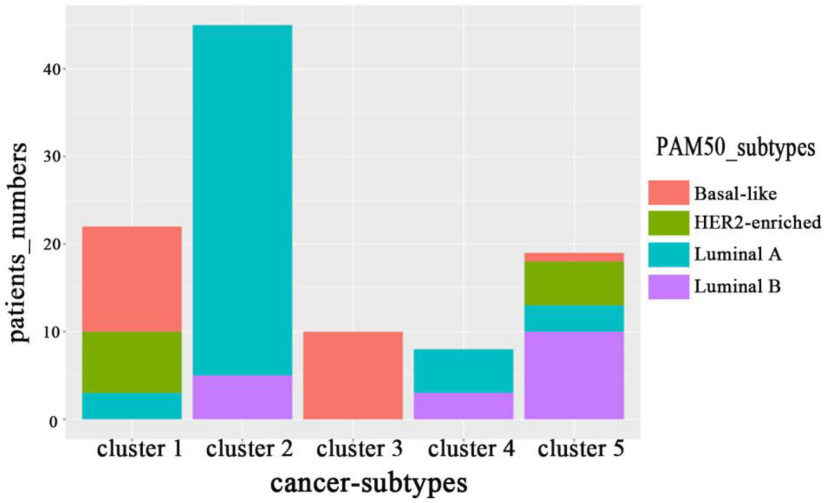

FIGURE 4 | The distribution of subtypes obtained by MRF-MSC on the subtypes: Basal-like, Luminal A, Luminal B and HER2-enriched.

Since the elements in $\boldsymbol{S}$ are non-negative, then $\boldsymbol{L_S}$ is a positive semi-definite matrix. Denote $\sigma_i(\boldsymbol{L_S})$ is the i-th minimum eigenvalue of $\boldsymbol{L_S}$, we can obtain the optimal solution of S through the following constrained Laplacian rank method: $\sum_{i=1}^{k}\boldsymbol{\sigma}_i(\boldsymbol{L_S}) = 0$ and rank $(\boldsymbol{L_S}) = n - k$, where rank (L_S) is the rank of L_S. By Ky Fan's theorem (Fan, 1949), we have

$$\sum_{i=1}^{k}\boldsymbol{\sigma}_i(\boldsymbol{L_S}) = \min_{F,F^{\mathrm{T}}F=I} tr\left(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L_S}\boldsymbol{F}\right) \tag{10}$$

where $\boldsymbol{F}$ is the first k minimum eigenvalues correspond to eigenvectors of $\boldsymbol{L_S}$. The right side of Eq. 10 is the objective function of spectral clustering. Therefore, Eq. 10 establishes the connection between the desired fused graph structure and spectral clustering. The optimization of Eq. 10 results in the fused similarity graph $\boldsymbol{S}$ with exact k connected components.

According to Eqs 9, 10, we combine the smooth representation of multi-omics data, the fusion of multi-smooth representation and multi-view spectral clustering into one framework, and propose the MRF-MSC. The objective function of MRF-MSC can be written as

$$\min_{\boldsymbol{Z}^v,\boldsymbol{S},\boldsymbol{F}}\sum_{v=1}^{t}\left(\left|\left|\boldsymbol{X}^v-\boldsymbol{X}^v\boldsymbol{Z}^v\right|\right|_F^2+\alpha\mathrm{tr}\left(\boldsymbol{Z}^v\boldsymbol{L}^v\left(\boldsymbol{Z}^v\right)^{\mathrm{T}}\right)+\right.$$

$$\left.\beta\varepsilon^v\left|\left|\boldsymbol{S}-\boldsymbol{Z}^v\right|\right|_F^2\right)+\lambda tr\left(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L_S}\boldsymbol{F}\right)\ s.t.\ \boldsymbol{Z}^v\geq 0,\ \boldsymbol{F^TF}=I \tag{11}$$

where α, β, and λ are hyperparameters.

We conclude that MRF-MSC has the following advantages in predicting cancer subtypes using multi-omics data.

(1) The characteristic of biological data is that the sample size is much smaller than the feature size. The smooth representation of the omics data not only retains the characteristic of the original data, but also effectively obtains the similarity between the sample points, which provides a relatively high quality subspace representation for the subsequent graph fusion process.
(2) In general, multi-omics data come from different platforms, which leads to different contribution of each omics data to clustering results. In the process of similar graph fusion, MRF-MSC uses self-weighting to perform multi-smooth representation fusion. In this way, the complementarity of various biological information is realized, the influence of noise data is reduced, and the quality of fused similar graph is improved.
(3) We introduce spectral clustering into MRF-MSC, which can improve the accuracy of the final result. In this joint MRF-MSC framework, the constrained Laplacian rank is used to constrain the structure of the fusion similar graph to obtain a graph structure that is conducive to the clustering task. Moreover, we use the learned graph structure to guide the construction of the graph, so that this mutual learning and iterative method can improve the final clustering result.

Optimization of MRF-MSC

We can optimize $\boldsymbol{Z}^v$, $\boldsymbol{S}$ and $\boldsymbol{F}$ step by step according to Eq. 11 through the idea of iterative optimization.

(1) Fixing $\boldsymbol{S}$ and $\boldsymbol{F}$ to solve $\boldsymbol{Z}^v$

Based on Eq. 11, we can get the objective function Eq. 9 about $\boldsymbol{Z}^v$. It is observed that in Eq. 9, $\boldsymbol{Z}^v$ is independent for each omics data. Therefore, we can update $\boldsymbol{Z}^v$ separately for each omics data. Taking the derivative of $\boldsymbol{Z}^v$ in Eq. 9, we have

$$\left(\left(\boldsymbol{X}^v\right)^{\mathrm{T}}\boldsymbol{X}^v+\beta\varepsilon^v\boldsymbol{I}\right)\boldsymbol{Z}^v+\alpha\boldsymbol{Z}^v\boldsymbol{L}^v=\left(\boldsymbol{X}^v\right)^{\mathrm{T}}\boldsymbol{X}^v+\beta\varepsilon^v\boldsymbol{S} \tag{12}$$

The above equation is a standard Sylvester equation with unique solution. We can easily get the solution result of $\boldsymbol{Z}^v$:

$$\boldsymbol{Z}^v = \left(\left(\boldsymbol{X}^v\right)^{\mathrm{T}} \boldsymbol{X}^v + \beta\varepsilon^v \boldsymbol{I} + \alpha \boldsymbol{L}^v\right)^{-1} \left(\left(\boldsymbol{X}^v\right)^{\mathrm{T}} \boldsymbol{X}^v + \beta\varepsilon^v \boldsymbol{S}\right) \quad (13)$$

(2) Fixing $\boldsymbol{Z}^v$ and $\boldsymbol{F}$ to solve $\boldsymbol{S}$

Based on Eq. 11, we can get the objective function of $\boldsymbol{S}$ as follows:

$$\min_{S} \sum_{v=1}^{t} \beta\varepsilon^v \left|\left|\boldsymbol{S} - \boldsymbol{Z}^v\right|\right|_F^2 + \lambda tr\left(\boldsymbol{F}^{\mathrm{T}} \boldsymbol{L}_S F\right) \quad (14)$$

According to $tr\left(\boldsymbol{F}^{\mathrm{T}} \boldsymbol{L}_S \boldsymbol{F}\right) = \sum_{i,j} \frac{1}{2} \left|\left|f_i - f_j\right|\right|_2^2 s_{ij}$, where s_{ij} is the elements of S, we define $g_{ij} = \left|\left|f_i - f_j\right|\right|_2^2$ and g_i is a vector whose j-th element equal to g_{ij}. So, the Eq. 14 can be calculated by column

$$\min_{s_i} \sum_{v=1}^{t} \beta\varepsilon^v \left|\left|\boldsymbol{s}_i - \boldsymbol{z}_i^v\right|\right|_F^2 + \frac{\lambda}{2} \boldsymbol{g}_i^{\mathrm{T}} \boldsymbol{s}_i \quad (15)$$

Taking the derivative of s_i in Eq. 15, we can obtain the solution of s_i:

$$\boldsymbol{s}_i = \frac{\sum_{v=1}^{t} \varepsilon^v \boldsymbol{z}_i^v - \frac{\lambda \boldsymbol{g}_i}{4\beta}}{\sum_{v=1}^{t} \varepsilon^v} \quad (16)$$

(3) Fixing $\boldsymbol{Z}^v$ and $\boldsymbol{S}$ to solve $\boldsymbol{F}$

Based on Eq. 11, we can get the objective function of F as follows:

$$\min_{F} \lambda tr\left(\boldsymbol{F}^{\mathrm{T}} \boldsymbol{L}_S \boldsymbol{F}\right) \ s.t.\ \boldsymbol{F}^{\mathrm{T}} \boldsymbol{F} = \boldsymbol{I} \quad (17)$$

In the above formula, the optimal solution of F is the k eigenvectors corresponding to the first k minimum eigenvalues. After the iterative optimization, we take each row of the final F as a new representation of each sample, and use the K-means algorithm to calculate the clustering results.

We use pseudo-code to summarize the MRF-MSC solution process in **Algorithm 1**.

Algorithm 1: MRF-MSC algorithm.

Input: cancer multi-omics data $\boldsymbol{X} = \{\boldsymbol{X}^1, \boldsymbol{X}^2, \cdots, \boldsymbol{X}^t\}$, the number of cancer subtypes k, the maximum number of iterations MaxIter, K is the number of neighbors in KNN, hyperparameters α, β and λ.
Output: smooth representation of each omics data $\boldsymbol{Z}^v$, fused similarity graph $\boldsymbol{S}$, eigenvectors $\boldsymbol{F}$.

Initialize $\boldsymbol{S} = \boldsymbol{I}$, $\varepsilon^v = 1/t$.
Repeat
Update $\boldsymbol{Z}^v$ according to Eq. 13,
Set $z_{ij}^v = \max\left(z_{ij}^v, 0\right)$ for every element z_{ij}^v in $\boldsymbol{Z}^v$,
Update $\boldsymbol{S}$ according to Eq. 16,
Update $\boldsymbol{F}$ by optimizing Eq. 17
Update ε^v according to Eq. 7,
Until meeting stop condition
Stop condition: the maximum number of iterations MaxIter is reached or the relative change of $\boldsymbol{S}$ is less than 10^{-3}.

RESULTS AND DISCUSSION

Multi-Omics Data Sets

In order to prove the effectiveness of the MRF-MSC algorithm in cancer subtype prediction, we applied MRF-MSC to the cancer multi-omics data downloaded and preprocessed from TCGA by Wang et al. (2014) and Rappoport and Shamir (2018). We conducted experiments on five cancer types: BRCA, Glioblastoma Multiforme (GBM), Lung Squamous Cell Carcinoma (LSCC), Kidney Renal Clear Cell Carcinoma (KIRC), and Colon Adenocarcinoma (COAD). Each cancer contains three types of cancer expression data from different platforms: mRNA expression, DNA methylation, and miRNA expression. The details on five types of cancer multi-omics data sets in Wang et al. (2014) and Rappoport and Shamir (2018) are shown in **Tables 1**, **2**, respectively. For these cancer types, we also downloaded the patient's clinical information, including all cancer survival data, and BRCA somatic mutation data, copy number data, and clinical data of drug treatments for subsequent analysis and algorithm comparison. The clinical information of BRCA was downloaded from the cBioPortal database.[1]

Evaluation Metrics

We chose the P-value based on the Cox log-rank model in the survival analysis of cancer subtype prediction to measure the MRF-MSC algorithm. For the characteristic that cancer samples have no real labels, it is impossible to use accuracy to evaluate the clustering results. In this case, survival analysis is necessary to verify the degree of difference between cancer subtypes (Mantel, 1966). We established a Cox regression model to obtain the P-value of the log-rank test of survival separation (Goel et al., 2010). If the P-value is smaller, it means that the survival rate between different clusters is more significant. Furthermore, it shows that the greater the difference between clusters, the more likely it is to get potential cancer subtypes with different characteristics.

Comparison Algorithms and Parameter Settings

For comparison, we selected five effective multi-view clustering algorithms in the field of cancer subtype prediction as the comparison algorithm: iClusterPlus, PFA, SNF, ANF, and MVSCO. Their details are as follows.

(1) iClusterPlus (Mo et al., 2013). iClusterPlus considers that different variable types follow different linear probability relationships, and then constructs a joint sparse model to complete the task of sample clustering and feature selection.
(2) PFA (Shi et al., 2017). PFA first uses the method of local information extraction to project each omics data in a low-dimensional space. Then, based on the idea of manifold learning, a dynamic collimation method is constructed to integrate low-dimensional spatial information into the fused feature space. Finally, the K-means method is used to find the label of the sample.

[1] http://www.cbioportal.org/

TABLE 5 | The distribution of clustering results of MRF-MSC on three susceptible genes: TP53, PIK3CA, and ERBB2.

Susceptible genes	Subtypes predicted by MRF-MSC				
	Cluster 1 (22)	**Cluster 2 (46)**	**Cluster 3 (10)**	**Cluster 4 (8)**	**Cluster 5 (19)**
TP53	17	8	9	0	5
PIK3CA	8	23	1	1	4
ERBB2	7	3	0	0	7

The values in this table represent the number of patients counted.

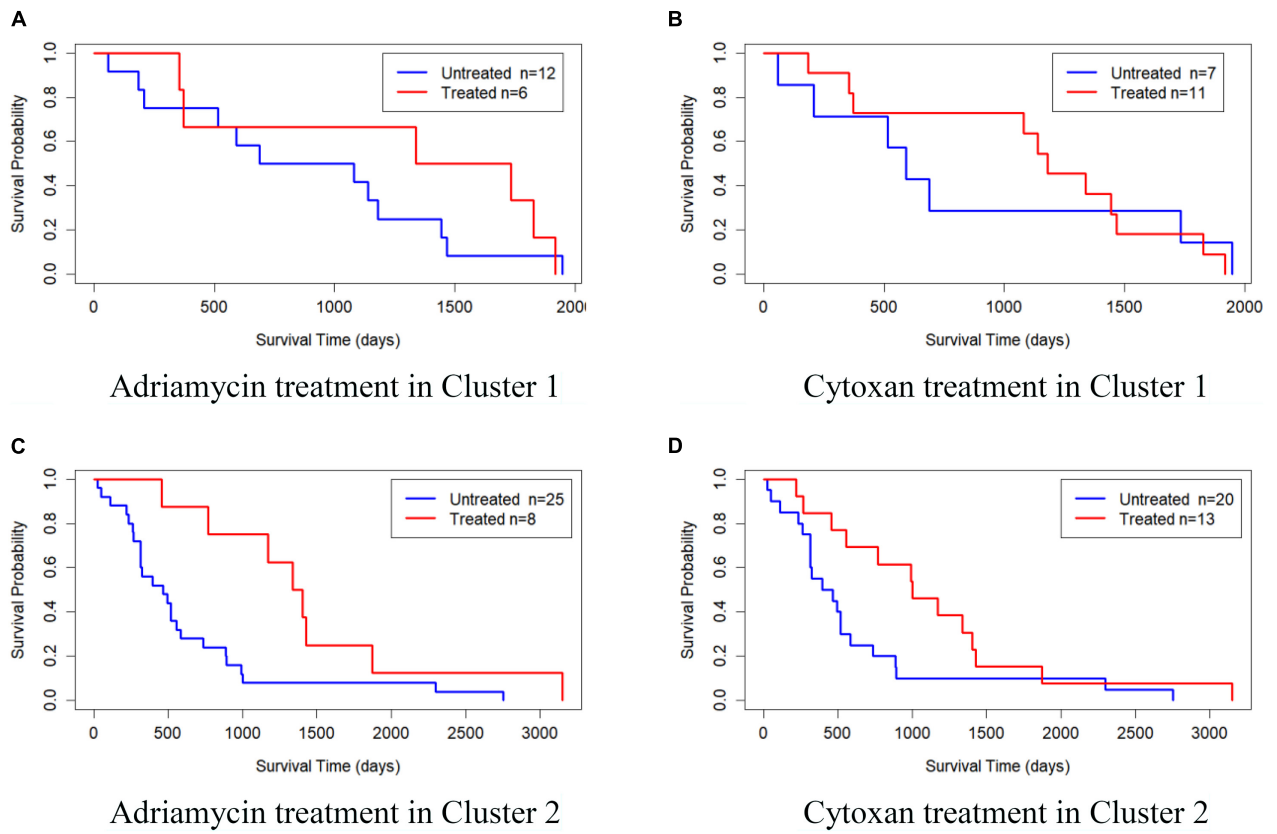

FIGURE 5 | Kaplan–Meier survival curves of drug response in Cluster 1 and Cluster 2. **(A)** Adriamycin treatment in Cluster 1, **(B)** Cytoxan treatment in Cluster 1, **(C)** Adriamycin treatment in Cluster 2, and **(D)** Cytoxan treatment in Cluster 2.

(3) SNF (Wang et al., 2014). SNF first uses the exponential similarity kernel method to define the similarity between the sample points of each omics data. Then, it uses the K-nearest neighbor method and a complete sparse kernel measurement method to obtain the local similarity graph and the global similarity graph of each omics data, respectively. Finally, the information transfer model based on the random walk idea is used to fuse the local information and the global information. Furthermore, spectral clustering method is used to cluster the fused graph.

(4) ANF (Ma and Zhang, 2017). PFA is an improved version of SNF. It constructs a K-nearest neighbor similar network for each omics data, and then merges these networks based on the random step method.

(5) MVSCO (Yu et al., 2019). MVSCO first draws on the method of Zhang et al. (2012) to find the similarity between sample points of each omics data, and then uses the current search method in the Stiefel manifold space to optimize the multi-view spectral clustering problem. Finally, the K-means method is used to predict the label of the sample.

Here, we present the parameter selection range of MRF-MSC algorithm and all comparison algorithms. Three hyperparameter α, β and λ in MRF-MSC are set to $\alpha, \beta, \lambda \in [10^{-6}, 10^{6}]$. iClusterPlus has two penalty parameters α and λ, where α is set to 1 and λ is obtained by automatic learning. In MRF-MSC, SNF, ANF, and MVSCO methods, the number of neighbors of KNN is set to $K \in [5, 50]$. The hyperparameter α in SNF is set to $\alpha \in [0.3, 0.8]$. We used the default parameter to run PFA algorithm.

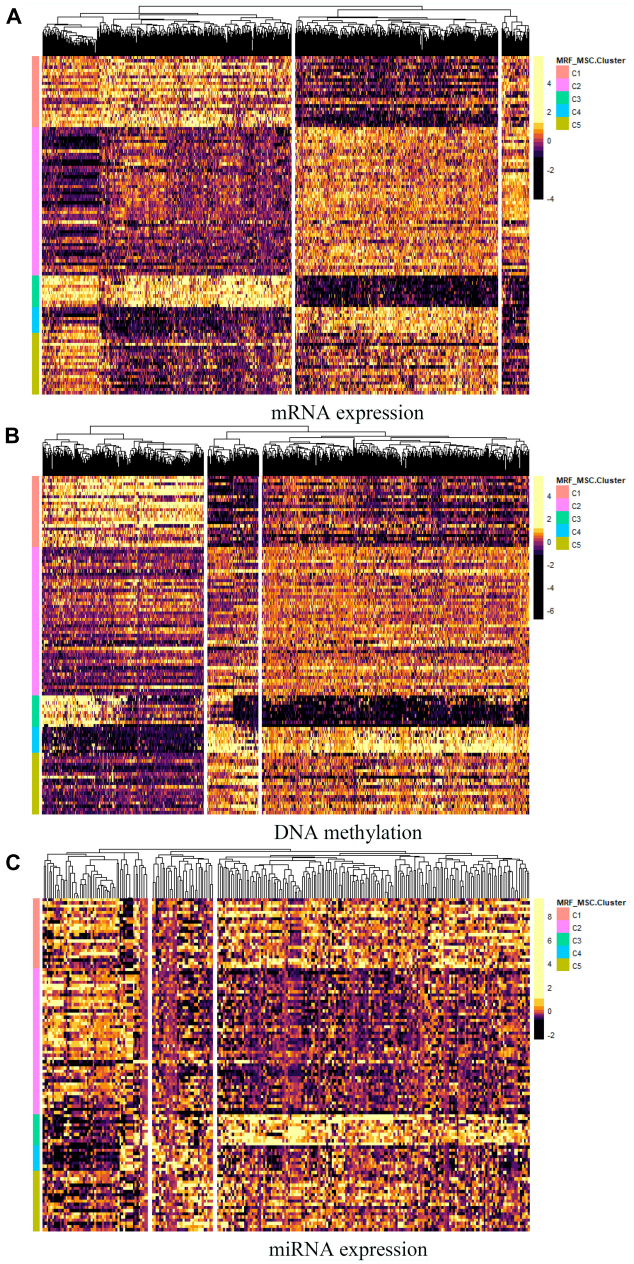

FIGURE 6 | The heatmap of differentially expressed genes in **(A)** mRNA expression, **(B)** DNA methylation, and **(C)** miRNA expression data.

Results on Cancer Multi-Omics Data Sets

Table 3 shows the comparison of *P*-values of survival analysis between MRF-MSC and other algorithms on five cancer multi-omics data sets in Wang et al. (2014), respectively. Since SNF is currently recognized as the most representative cancer subtype prediction algorithm, we used the number of clusters suggested in SNF, that is, GBM is clustered into three categories, BRCA is clustered into five categories, KIRC is clustered into three categories, LSCC is clustered into four categories, and COAD clustered into three categories. Compared with other algorithms, MRF-MSC has the lowest *P*-value on all five types of cancer. **Figure 2** is the Kaplan–Meier survival analysis curve of MRF-MSC on different cancers. Each curve describes the survival time trend of each cancer subtype. The number of samples in each group is also marked in the figure. **Figure 2** shows that MRF-MSC can get significantly different cancer subtypes on all types of cancer.

Table 4 shows the comparison of *P*-values of survival analysis between MRF-MSC and other algorithms on five cancer multi-omics data sets in Rappoport and Shamir (2018), respectively. These cancer data do not have the number of cancer subtypes available for reference. Therefore, we have to determine the number k of these cancer subtypes. iClusterPlus, SNF, and ANF algorithms all have their own way of determining the number of cancer subtypes. For the MRF-MSC, PFA, and MVSCO algorithms, we use Silhouette score (Nguyen et al., 2017) as a reference index for screening the number of cancer subtypes. In the clustering problem, Silhouette analysis is used to study the distance between clusters. Silhouette score measures the closeness of points in the same class compared with points in different classes, which provides a way to evaluate the number of classes. In **Table 4**, the best *P*-value and the corresponding number of clusters k of each algorithm for each cancer type are given. On GBM, BRCA, KIRC, and LSCC data, MRF-MSC algorithm has better experimental results than other algorithms. **Figure 3** is the Kaplan–Meier survival analysis curve of MRF-MSC on different cancers. We can find that MRF-MSC can get significantly different cancer subtypes on all types of cancer. All these results demonstrate the effectiveness of the proposed method in cancer subtype prediction.

Analysis on BRCA Data

Breast Invasive Carcinoma refers to a malignant tumor in which cancer cells have penetrated the basement membrane of breast ducts or lobular alveoli and invaded the interstitium. Many scholars have carried out a series of studies and analyses on the gene level, and have given specific subtypes and treatment programs. Based on the microarray predictive analysis model, Parker et al. (2009) proposed a 50-gene classifier (known as PAM50) to classify BRCA into five subtypes: Basal-like, Luminal A, Luminal B, HER2-enriched, and Normal-like. And each subtype is associated with specific mutant genes. For example, there are a large number of PIK3CA mutations in Luminal A and Luminal B, while Basal-like and HER2-enriched are mainly associated with TP53 mutation and ERBB2 amplification, respectively (Koboldt et al., 2012).

On BRCA data set in Wang et al. (2014), we counted the distribution of clustering results obtained by MRF-MSC on the cancer subtypes: Basal-like, Luminal A, Luminal B, and HER2-enriched in **Figure 4**. Note that, the clinical information for Normal-like cannot be found in Parker et al. (2009). It can be seen from **Figure 4** that Basal-like is mainly distributed in Cluster 1 and Cluster 3, Luminal A is mainly distributed in Cluster 2 and Cluster 4, Luminal B is mainly distributed in Cluster 5, HER2-Enriched is distributed in Cluster 1 and Cluster

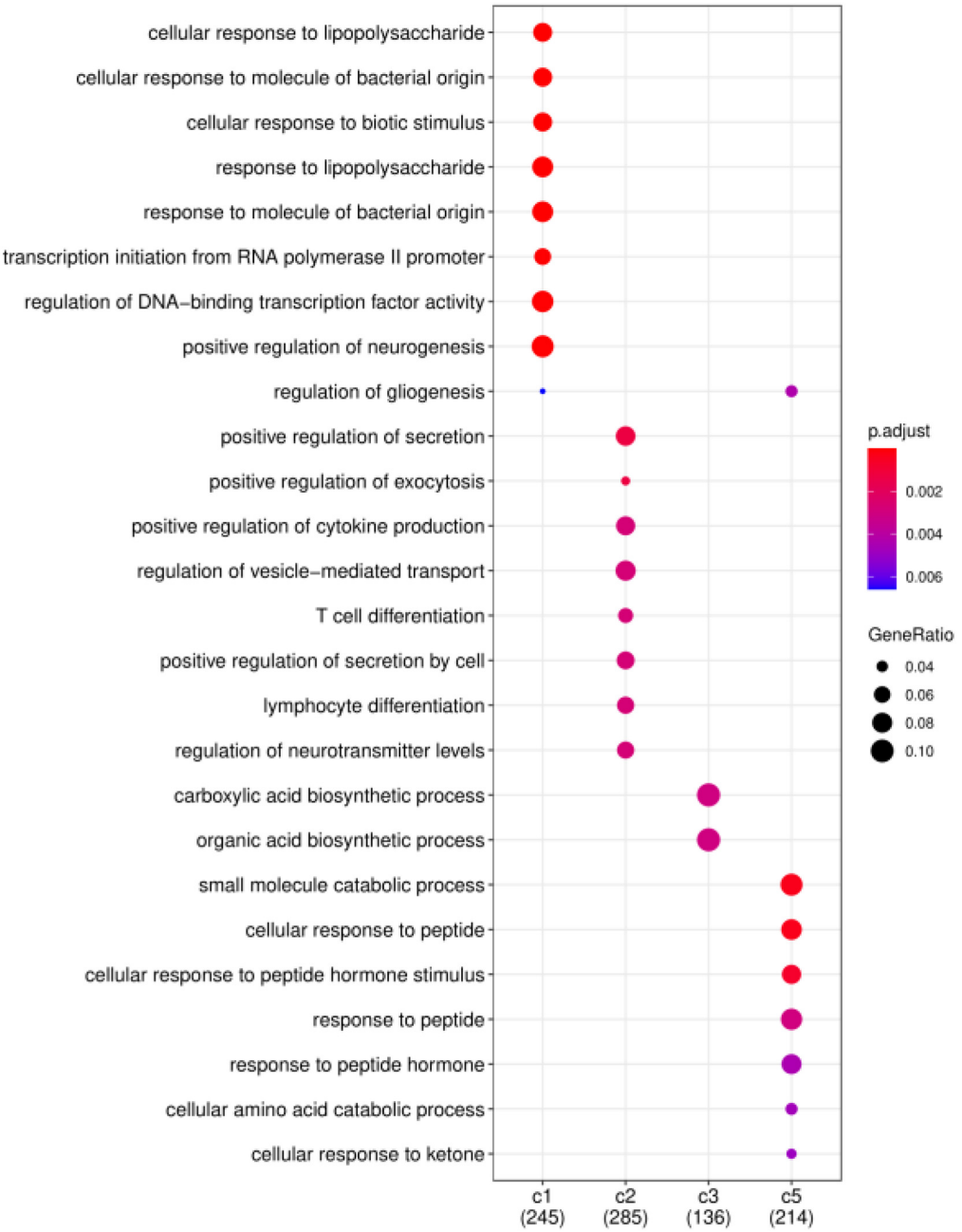

FIGURE 7 | Functional enrichment analysis on BRCA. Significantly different GO biological processes derived from driver genes of different cancer subtypes.

5. This shows that the cancer subtypes obtained by MRF-MSC are related to these known cancer subtypes. Furthermore, we counted the distribution of clustering results of MRF-MSC on three susceptible genes: TP53, PIK3CA, and ERBB2 in **Table 5**. From **Table 5** we can find that there are a large number of TP53 mutations which is in line with the characteristics of the Basal-like subtype. The mutation frequency of PIK3CA in Cluster 2 is much higher than the other clusters show that Cluster 2 is related to the known subtypes: Luminal A and Luminal B. The mutations of ERBB2 are mainly distributed on Cluster 1 and Cluster 5, indicating that HER2-enriched subtype is related to Cluster 1 and Cluster 5. The results in **Figure 4** and **Table 5** are mutually corroborated, proving that MRF-MSC can mine meaningful cancer subtypes.

We also validated the obtained subtypes by comparing the survival of different therapeutic agents in each subtype. We downloaded BRCA drug data from TCGA database and selected Adriamycin and Cytoxan for analysis. Since there are few or no samples in Clusters 3, 4, and 5 for these two drugs, we only established a Cox log-rank model on Cluster 1 and Cluster 2 to

analyze the quality of drug response. **Figure 5** shows the Kaplan–Meier survival curves of drug response in Cluster 1 and Cluster 2. The treated samples and untreated samples are divided into two groups. Clusters 1 and Cluster 2 both responded favorably to Adriamycin and Cytoxan treatment. And the survival of patients with treatment is better than that of patients without treatment. The drug response of Cluster 2 to Adriamycin and Cytoxan (the survival analysis *P*-values of the Cox log-rank model are 9.91×10^{-3} and 4.42×10^{-4}, respectively) is better than that of Cluster 1 (the survival analysis *P*-values of the Cox log-rank model are 0.353 and 0.982, respectively).

Furthermore, differential expressed genes and GO enrichment analysis on BRCA data are performed to compare differences in characteristics between the five clusters obtained by MSR-MSC. For each omics data, we first used Analysis of Variance (ANOVA) method to select the significant differentially expressed genes in five clusters. And the heatmap of differentially expressed genes in mRNA expression, DNA methylation, and miRNA expression data are shown in **Figures 6A–C**, respectively. The specific information of these differentially expressed genes can be found in **Supplementary File 1**. These differentially expressed genes may be closely related to BRCA. For example, the increased expression of GFRA3 (*P*-value = 3.71×10^{-23}) is associated with lymph node metastasis and advanced tumor stage in BRCA (Wu et al., 2013). mir-186 (*P*-value = 7.41×10^{-17}) can regulate the migration and erosion of BRCA by PTTG1 (Li et al., 2013), and mir-197 (*P*-value = 2.71×10^{-17}) targets the tumor-suppressor FUS1 (Du et al., 2009).

Finally, we consider that the driver genes that affect these five clusters should be different. Therefore, based on the DriverNet method (Bashashati et al., 2012), we use BRCA mutation data, copy number data and mRNA expression data to find the driver genes of each cluster. We screened out the unique driver genes of each cluster to construct GO enrichment analysis (Yu et al., 2012). **Figure 7** shows the functional enrichment analysis of four clusters on BRCA. There are too few driver genes in Cluster 4 to form a functional enrichment term. It can be seen that significantly different GO biological processes derived from driver genes of different cancer subtypes (FDR < 0.05). Driver genes in Cluster 1, 2, 3, and 5 are correlated with "cellular response," "positive regulation," "biosynthetic process," and "response to peptide" in GO biological processes, respectively.

CONCLUSION

In the past few decades, many multi-view biological data integration models based on graph learning, matrix decomposition, network fusion, deep learning, nuclear methods and other technologies have been designed and applied to a wide range of bioinformatics topics (Li et al., 2016), such as prediction of drug–target interactions (Liu et al., 2021), identification of cancer driver genes (Bashashati et al., 2012) and genotype-phenotype interactions (Qin et al., 2020). These studies provide meaningful insights into the cause and development of cancer. However, how to effectively mine cancer subtypes with biological characteristics from multi-omics data is still a challenging task for bioinformatics. In this paper, a new cancer subtype prediction method was proposed, named as Multi-View Spectral Clustering Based on Multi-smooth Representation Fusion (MRF-MSC). In order to enable the data samples to retain the original feature space and enhance the grouping effect during data representation, we construct smooth representation for each type of data. Then, based on the method of graph fusion, these smooth representations are integrated into one space, and each smooth representation is given a self-weighted weight to measure their contribution. A fused similarity graph with a consistent structure is obtained through optimization. Finally, constrained Laplacian rank is performed on the fused similarity graph, and the label of the sample is obtained through spectral clustering optimization. We use real cancer data sets to demonstrate the capabilities of MRF-MSC. MRF-MSC can effectively integrate the information of multi-omics data, and is superior to several state-of-the-art integration methods in given evaluation indexes. On BRCA data, through various studies, we have verified that the cancer subtypes predicted by MRF-MSC are significantly different and have biological significance.

In addition, we also admit that MRF-MSC has its shortcomings and limitations. It takes a lot of time to select suitable hyperparameters in the optimization process. Moreover, it is not suitable for binary data (somatic mutation), categorical data (copy number states: loss/normal/gain), and it has no ability to find important genes that affect each subtype. Therefore, we will continue to work hard to improve and expand the capabilities of the MRF-MSC algorithm and explore the heterogeneity of cancer.

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

JL and XW constructed the original idea and designed the experiments. JL and SG wrote the manuscript. JL and YC proofread the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61906198, 61976215, and 61772532) and the Natural Science Foundation of Jiangsu Province (Grant No. BK20190622).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.718915/full#supplementary-material

REFERENCES

Akbani, R., Ng, P. K. S., Werner, H. M. J., Shahmoradgoli, M., Zhang, F., Ju, Z., et al. (2014). A pan-cancer proteomic perspective on the Cancer genome atlas. *Nat. Commun.* 5, 3887–3887. doi: 10.1038/ncomms4887

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, 1–14. doi: 10.1186/gb-2012-13-12-r124

Bedard, P. L., Hansen, A. R., Ratain, M. J., and Siu, L. L. (2013). Tumour heterogeneity in the clinic. *Nature* 501, 355–364. doi: 10.1038/nature12627

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi: 10.1038/nature12625

Ding, C. H. Q., and He, X. (2004). "Cluster structure of K-means clustering via principal component analysis," in *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, (Piscataway, NJ: IEEE) 414–418. doi: 10.1007/978-3-540-24775-3_50

Du, L., Schageman, J. J., Subauste, M. C., Saber, B., Hammond, S. M., Prudkin, L., et al. (2009). miR-93, miR-98, and miR-197 regulate expression of tumor suppressor gene FUS1. *Mol. Cancer Res.* 7, 1234–1243. doi: 10.1158/1541-7786.MCR-08-0507

Fan, K. (1949). On a theorem of weyl concerning eigenvalues of linear transformations: II*. *Proc. Natl. Acad. Sci. U S A.* 36, 31–35. doi: 10.1073/pnas.36.1.31

Feng, J., Jiang, L., Li, S., Tang, J., and Wen, L. (2021). Multi-omics data fusion via a joint kernel learning model for cancer subtype discovery and essential gene identification. *Front. Genet.* 12:647141. doi: 10.3389/fgene.2021.647141

Ge, S., Wang, X., Cheng, Y., and Liu, J. (2021). Cancer subtype recognition based on laplacian rank constrained multiview clustering. *Genes* 12:526. doi: 10.3390/genes12040526

Goel, M. K., Khanna, P., and Kishore, J. (2010). Understanding survival analysis: kaplan-Meier estimate. *Int. J. Ayurveda Res.* 1, 274–278. doi: 10.4103/0974-7788.76794

Guo, Y., Li, H., Cai, M., and Li, L. (2019). Integrative subspace clustering by common and specific decomposition for applications on cancer subtype identification. *BMC Med. Genomics* 12:191. doi: 10.1186/s12920-019-0633-1

Hu, H., Lin, Z., Feng, J., and Zhou, J. (2014). "Smooth representation clustering," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ: IEEE), 3834–3841. doi: 10.1109/CVPR.2014.484

Kang, Z., Shi, G., Huang, S., Chen, W., Pu, X., Zhou, J. T., et al. (2020). Multi-graph fusion for multi-view spectral clustering. *Knowledge Based Systems* 189, 105102. doi: 10.1016/j.knosys.2019.105102

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Li, H., Yin, C., Zhang, B., Sun, Y., Shi, L., Liu, N., et al. (2013). PTTG1 promotes migration and invasion of human non-small cell lung cancer cells and is modulated by miR-186. *Carcinogenesis* 34, 2145–2155. doi: 10.1093/carcin/bgt158

Li, Y., Wu, F.-X., and Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* 19, 325–340. doi: 10.1093/bib/bbw113

Liu, Z., Chen, Q., Lan, W., Pan, H., Hao, X., and Pan, S. (2021). GADTI: graph autoencoder approach for DTI prediction from heterogeneous network. *Front. Genet.* 12:650821. doi: 10.3389/fgene.2021.650821

Ma, T., and Zhang, A. (2017). "Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering," in *Proceedings of 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (Piscataway, NJ: IEEE), 398–403. doi: 10.1109/BIBM.2017.8217682

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Rep.* 50, 163–170.

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2016). MoCluster: identifying joint patterns across multiple omics data sets. *J. Proteome Res.* 15, 755–765. doi: 10.1021/acs.jproteome.5b00824

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U S A.* 110, 4245–4250. doi: 10.1073/pnas.1208949110

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. *Neural Inform. Process. Systems* 14, 849–856.

Nguyen, T., Tagett, R., Diaz, D., and Draghici, S. (2017). A novel approach for data integration and disease subtyping. *Genome Res.* 27, 2025–2039. doi: 10.1101/gr.215129.116

Nie, F., Li, J., and Li, X. (2016). "Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, (Piscataway, NJ: IEEE), 1881–1887.

Nie, F., Li, J., and Li, X. (2017). " Self-weighted multiview clustering with multiple graphs," in *Proceedings of 26th International Joint Conference on Artificial Intelligence*, (Piscataway, NJ: IEEE), 2564–2570. doi: 10.24963/ijcai.2017/357

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27, 1160–1167. doi: 10.1200/JCO.2008.18.1370

Qin, Z., Su, J., Li, M., Yang, Q., Yi, S., Zheng, H., et al. (2020). Clinical and genetic analysis of CHD7 expands the genotype and phenotype of charge syndrome. *Front. Genet.* 11:592. doi: 10.3389/fgene.2020.00592

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* 46, 10546–10562. doi: 10.1093/nar/gky889

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18. doi: 10.1038/nmeth1156

Shen, R., Olshen, A. B., and Ladanyi, M. (2010). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 26, 292–293. doi: 10.1093/bioinformatics/btp659

Shi, Q., Zhang, C., Peng, M., Yu, X., Zeng, T., Liu, J., et al. (2017). Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics* 33, 2706–2714. doi: 10.1093/bioinformatics/btx176

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Wu, Z., Pandey, V., Wu, W. Y., Ye, S., Zhu, T., and Lobie, P. E. (2013). Prognostic significance of the expression of GFRalpha1. GFRalpha3 and Syndecan. *BMC Cancer* 13:34. doi: 10.1186/1471-2407-13-34

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Int. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, Y., Zhang, L.-H., and Zhang, S. (2019). Simultaneous clustering of multiview biomedical data using manifold optimization. *Bioinformatics* 35, 4029–4037. doi: 10.1093/bioinformatics/btz217

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* 40, 9379–9391. doi: 10.1093/nar/gks725

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

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

frontiers in Genetics

METHODS
published: 22 September 2021
doi: 10.3389/fgene.2021.744334

Comparative Analysis of Unsupervised Protein Similarity Prediction Based on Graph Embedding

Yuanyuan Zhang[1,2], Ziqi Wang[1], Shudong Wang[2] and Junliang Shang[3]*

[1] School of Information and Control Engineering, Qingdao University of Technology, Qingdao, China, [2] College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China, [3] School of Information Science and Engineering, Qufu Normal University, Rizhao, China

The study of protein–protein interaction and the determination of protein functions are important parts of proteomics. Computational methods are used to study the similarity between proteins based on Gene Ontology (GO) to explore their functions and possible interactions. GO is a series of standardized terms that describe gene products from molecular functions, biological processes, and cell components. Previous studies on assessing the similarity of GO terms were primarily based on Information Content (IC) between GO terms to measure the similarity of proteins. However, these methods tend to ignore the structural information between GO terms. Therefore, considering the structural information of GO terms, we systematically analyze the performance of the GO graph and GO Annotation (GOA) graph in calculating the similarity of proteins using different graph embedding methods. When applied to the actual Human and Yeast datasets, the feature vectors of GO terms and proteins are learned based on different graph embedding methods. To measure the similarity of the proteins annotated by different GO numbers, we used Dynamic Time Warping (DTW) and cosine to calculate protein similarity in GO graph and GOA graph, respectively. Link prediction experiments were then performed to evaluate the reliability of protein similarity networks constructed by different methods. It is shown that graph embedding methods have obvious advantages over the traditional IC-based methods. We found that random walk graph embedding methods, in particular, showed excellent performance in calculating the similarity of proteins. By comparing link prediction experiment results from GO(DTW) and GOA(cosine) methods, it is shown that GO(DTW) features provide highly effective information for analyzing the similarity among proteins.

Keywords: protein similarity, graph embedding, gene ontology, link prediction, DTW algorithm

OPEN ACCESS

Edited by:
Jianing Xi,
Northwestern Polytechnical University, China

Reviewed by:
Shouheng Tuo,
Xi'an University of Posts and Telecommunications, China
Cheng Liang,
Shandong Normal University, China
Yajun Liu,
Xi'an University of Technology, China

****Correspondence:***
Yuanyuan Zhang
yyzhang1217@163.com

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

Received: *20 July 2021*
Accepted: *25 August 2021*
Published: *22 September 2021*

Citation:
Zhang Y, Wang Z, Wang S and Shang J (2021) Comparative Analysis of Unsupervised Protein Similarity Prediction Based on Graph Embedding. Front. Genet. 12:744334. doi: 10.3389/fgene.2021.744334

INTRODUCTION

Proteomics essentially refers to the study of the characteristics of proteins on a large scale, including the expression level of proteins, the functions of proteins, protein–protein interactions, and so forth. The study of proteome not only provides the material basis for the law of life activities but can also provide the theoretical basis and solutions for elucidating and solving the mechanism of many diseases (Xi et al., 2020a). However, at present, research on the function of proteins is lacking. The functions of proteins encoded by most of the newly discovered genes by genome

sequencing are unknown. For those whose functions are known, their functions have mostly been inferred by methods such as homologous gene function analogy. Therefore, using computational methods to explore the similarity between proteins can effectively improve the efficiency of proteomic studies.

Gene Ontology (GO) (Harris, 2004) describes the function of genes It is a standardized description of the characteristics of genes and gene products, enabling bioinformatics researchers to uniformly summarize, process, interpret, and share the data of genes and gene products. It provides the representation of biological knowledge through structured and controlled terms. GO includes three kinds of ontologies: Biological Processes (BPs), Cell Components (CCs), and Molecular Functions (MFs). The words in the three kinds of ontologies are related to each other and form a Directed Acyclic Graph (DAG), wherein a node denotes a GO term, while an edge denotes a kind of relationship between two GO terms. Therefore, it is of great significance to study the similarity of proteins based on the graph characteristics of GO to explore the function of proteins.

GO has been widely studied in the field of biology (Xi et al., 2020b). GO terms have been used to annotate many biomedical databases [e.g., UniProt database (UniProt Consortium, 2015) and SwissProt database (Amos and Brigitte, 1999)]. The characteristics and structure of GO have made GO terms the basis of functional comparison between gene products (Pesaranghader et al., 2014). GO annotation defines the semantic similarity of genes (proteins) and provides a basis for measuring the functional similarity of proteins. The more information two GO terms share, the more similar they are, and the more the similarity between the proteins annotated by the two GO terms (Hu et al., 2021). In earlier studies, many researchers analyzed protein–protein interaction (PPI) based on GO (Sevilla et al., 2005). Studies on computing protein similarity using GO mainly focus on the IC of GO terms, which is widely used to identify relations between proteins. The uniqueness of GO terms is often evaluated by taking the average of the IC of two terms. The IC of a term depends on the annotating corpus (Sevilla et al., 2005). Three IC-based methods—Resnik's (Resnik, 1999), Rel's (Paul and Meeta, 2008), and Jiang and Conrath's (Jiang and Conrath, 1997)—have been introduced from natural language taxonomies by Lord et al. (2003) to compare genes (proteins). Although the abovementioned methods are used to calculate semantic similarity between two GO terms to achieve good results, they only consider the amount of information of common nodes. They do not consider the information differences between the nodes themselves and ignore the structural information of the terms. The result of term comparison is a rough estimate. For example, in Resnik's method, if the ancestors of two terms are the same, then the similarity of two terms in any layer is not different and cannot be compared. Obviously, this is unreasonable.

This study merged the three categories of ontologies and GO annotations into a large graph called the GO Annotation (GOA) graph. We used three categories of ontologies transformed into a GO graph. Effective graph analysis on GOA and GO graphs can improve our understanding of the structure and node information of GO and proteins. Using the GOA information of the proteins, the similarity among proteins can be calculated, and the relationship between proteins can be predicted. In recent years, graph learning-based analytical methods have made remarkable progress in bioinformatics and other fields (Xi et al., 2021). At present, graph learning-based analytical methods focuses on dynamic graphs. Methods such as SDNE (Wang et al., 2016), DeepWalk (Perozzi et al., 2014), LINE (Tang et al., 2015), Node2vec (Grover and Leskovec, 2016), and SINE (Wang et al., 2020) have been widely used for unsupervised feature learning in the field of data mining and natural language processing. The edge prediction task is applied to the PPI prediction to find new protein interaction relationships. They also provide a basis for calculating protein similarity based on GO, such as GO2vec (Zhong et al., 2019), which used the Node2vec algorithm to compute the functional similarity between proteins.

To explore the performance of graph embedding methods in measuring protein similarity based on GO and GOA, we used four typical graph embedding methods to learn the features of GO terms and proteins. These methods can be divided into two categories. The first category is the random walk method, such as the DeepWalk and Node2Vec methods. The DeepWalk method uses the truncated random walk strategy to obtain the sequence of nodes and point embedding obtained from learning with Word2Vec (Goldberg and Levy, 2014). Node2Vec uses biased random walk to generate a node sequence by balancing the Breadth First Search (BFS) and Depth First Search (DFS) of the graph. The second category is based on deep learning, such as SDNE and LINE methods. SDNE uses an auto-encoder to optimize the first-order and second-order similarity simultaneously, while LINE optimizes the orders of similarity separately. As a result, their learned node embedding can retain the local and global graph structure and is robust to sparse networks. We introduce the overall flowchart of this paper in **Figure 1**, which is divided into two parts. Firstly, in Part A, the features of GO terms are learned based on the GO graph using graph embedding methods. The similarity of proteins is then calculated based on the features of their annotated GO terms by Dynamic Time Warping (DTW) distance (Lou et al., 2016). Secondly, in Part B, the features of proteins are learned based on the GOA graph directly. Then, the cosine similarity of the corresponding features is calculated to measure the similarity of protein. Finally, a link prediction (Li et al., 2018) experiment is performed in the screened-out protein similarity networks, using the area under the curve (AUC) (Lobo, 2010) and area under the precision-recall curve (AUCPR) (Yu and Park, 2014) to evaluate the reliability of the protein network constructed by learned vectors.

MATERIALS AND METHODS

Data Source and Preprocessing

We downloaded GO data in Open Biomedical Ontologies (OBO) format from the GO Consortium Website [1]. The GO protein

[1] http://geneontology.org/page/download-ontology

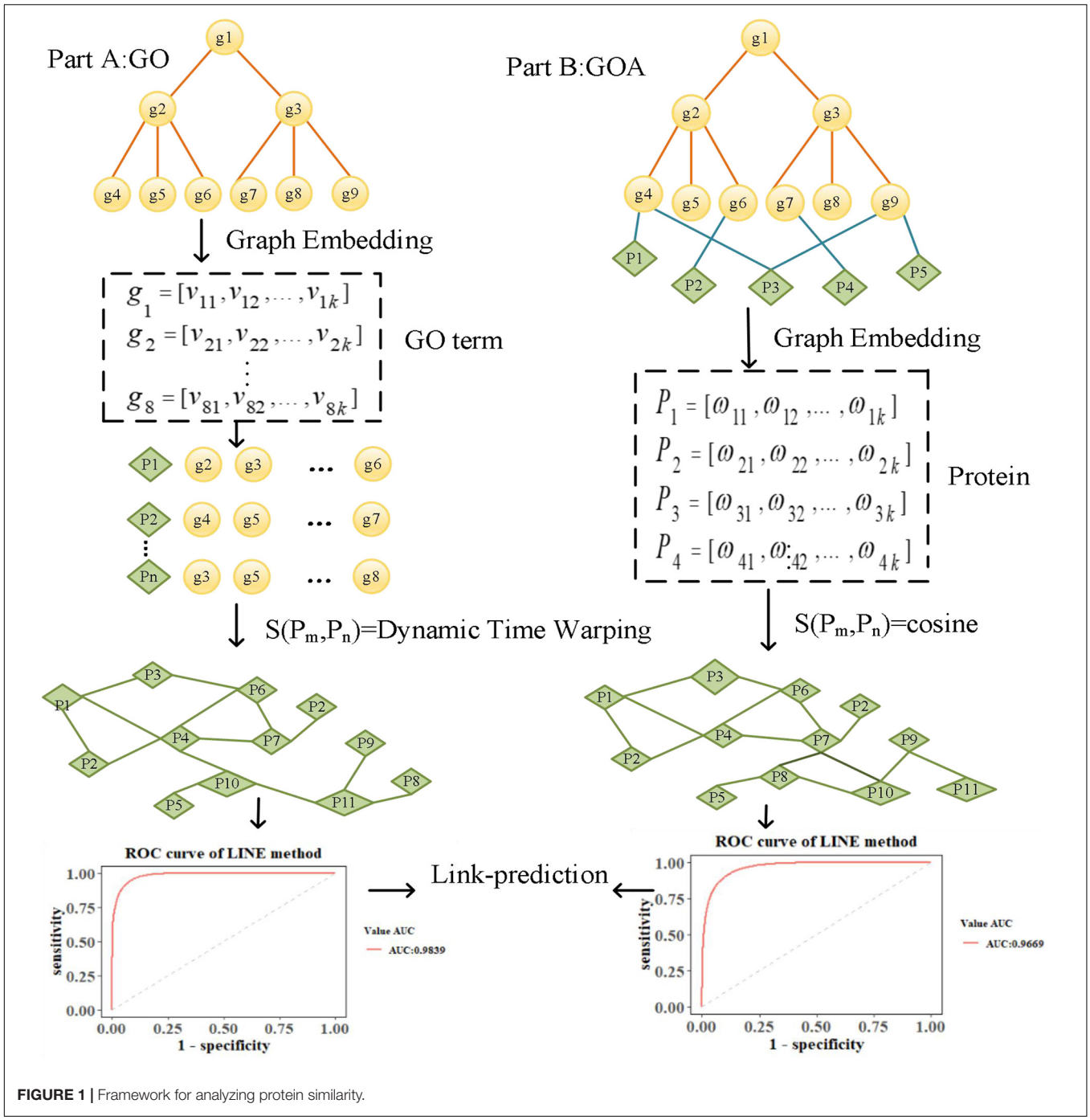

FIGURE 1 | Framework for analyzing protein similarity.

annotations were obtained from the UniProt GOA website [2]. The Yeast dataset contained 2,887 proteins, and the Human dataset contained 9,677 proteins. The GO data were then preprocessed based on the following processes. First, since several GO terms annotate a protein, term–term relations of GO terms and term–protein annotations between GO terms and proteins were combined into a GOA graph. Second, the GO terms were then transformed into an undirected, unweighted GO graph, regardless of the type and direction of the relationship. We summarize the numbers of GO terms and edges in **Table 1**.

Method

Based on different graph embedding methods, the feature of GO terms and proteins was learned into vector representations by fusing GO and GOA graph topologies, respectively. Thus, we could capture the global information based on the graph embedding method, and its learned vectors could calculate

[2]http://www.ebi.ac.uk/GOA

TABLE 1 | Characteristics of GO graphs.

Gene ontology	Term	Edges
BP*	30,705	71,530
CC**	4,380	7,523
MF***	12,127	13,658

*Biological Processes, **Cell Components, and ***Molecular Functions.

the similarity between proteins by the DTW distance and cosine similarity.

Introduction of Different Graph Embedding Methods

In this paper, we used the methods of graph embedding based on random walk and deep learning to learn the features of GO terms and proteins through fusing the topology of GO and GOA graphs, respectively. Random walk-based methods include DeepWalk (Perozzi et al., 2014) and Node2vec (Grover and Leskovec, 2016). The DeepWalk method is divided into two parts: random walk to obtain node sequences and to generate node embedding. Random walk is used to obtain the local information of the node in the graph, and the embedding reflects the local structure of the node in the graph. The path length is controlled by setting the parameter walk-length ($_L$). The more neighborhood nodes (higher-order neighborhood nodes) two nodes have, the more similar they are. **Figure 2A** illustrates the DeepWalk algorithm flow. Node2vec method sets two hyper-parameters p and q to control the random walk and adopts a flexible biased random walk procedure that smoothly combines BFS and DFS to generate node sequences. **Figure 2B** illustrates the Node2vec algorithm flow. Nodes c_i are generated based on the following distribution:

$$P(c_i = \mathrm{x}|\mathrm{c}_{\mathrm{i}-1} = t) = \begin{cases} \frac{\pi_{tx}}{Z} & (if\ (t,x) \in E) \\ 0 & (otherwise) \end{cases} \quad (1)$$

where π_{tx} is the transition probability between nodes t and x, and Z is the normalization constant. According to the node context information, node sequences are generated by setting the sizes of the hyper-parameters p and q to control the random walk strategy. The Skip-gram model is used to obtain the vector representation of the nodes. The random walk graph embedding of nodes reflects the local and global topology information of nodes in the graph.

The second kind of embedding method is SDNE, which proposed a new semi-supervised learning model. Combining the advantages of first-order and second-order estimation, SDNE can capture the global and local structural properties of the graph. The unsupervised part uses a deep auto-encoder to learn the second-order similarity, and the supervised part uses a Laplace feature map to capture the first-order similarity. **Figure 2C** illustrates the SDNE algorithm flow. By inputting the node embedding S_i in the model, where S_i is compressed by the auto-encoder, the feature is then reconstructed. Finally, its loss function is defined as follows:

$$O_2 = \Sigma \left|\left| S_i' - S_i \right|\right|_2^2 \quad (2)$$

LINE is another method based on deep learning, which optimizes the first-order and second-order similarities (**Figure 2D**). The first-order similarity is used to describe the local similarity between pairs of nodes in the graph. The second-order similarity is described as two nodes in the graph not having directly connected edges, but there are common neighbor nodes, which indicate that the two nodes are similar.

Introduction to IC-Based Method

In this paper, we chose two typical IC-based methods to measure the semantic similarity of GO terms, based on Jiang and Conrath (1997) and Rel (Paul and Meeta, 2008). The IC of a term is inversely proportional to the frequency of the term being used to annotate genes in a given corpus, such as the UniProt database. The IC of a GO term g is defined by the negative log-likelihood and is given by

$$IC(g) = -\log p(g) \quad (3)$$

$$p(g) = \frac{freq(g)}{N} \quad (4)$$

where $p(g)$ is the frequency of term g and its offspring in a specific GO annotated corpus. N represents the total number of annotated proteins in the corpus. If there are 50 annotated proteins in a corpus and 10 of them are annotated by term g, the annotation frequency of term g is $p(g) = 0.2$.

Jiang and Conrath and Rel's methods rely on comparing the attributes of terms in GO. Jiang and Conrath's method considered the fact that the semantic similarity between two terms is closely related to the nearest common ancestor corresponding to the two terms. The semantic similarity between two terms is estimated by calculating the amount of IC in the nearest common ancestor. Jiang and Conrath's and Rel's similarities are expressed as follows:

$$sim_{J\&C}(g_1, g_2) = 2^* IC(g_c) - IC(g_1) - IC(g_2) \quad (5)$$

$$sim_{Rel}(g_1, g_2) = \frac{2^* IC(g_c)}{IC(g_1) + IC(g_2)} + (1 - p(g_c)) \quad (6)$$

where g_c is the most informative common ancestor of g_1 and g_2 in the ontology. Given two proteins P_m and P_n annotated with GO terms $G_m = \{g_1, \cdots, g_i\}$ and $G_n = \left\{g_1^{'}, \cdots, g_j^{'}\right\}$, we used the Best Match Average (BMA) method to compute the similarity between two sets of GO terms, which can be expressed as follows:

$$BMA(P_m, P_n) = \frac{1}{2}\left(\frac{1}{n}\sum_{g_m \in G_m} \max_{g_n' \in G_n} sim(g_m, g_n') + \frac{1}{m}\sum_{g_n' \in G_n} \max_{g_m \in G_m} sim(g_m, g_n')\right) \quad (7)$$

where $sim(g_m, g_n')$ is the similarity between term g_m and term g'_n, which could have been calculated using IC-based similarity methods.

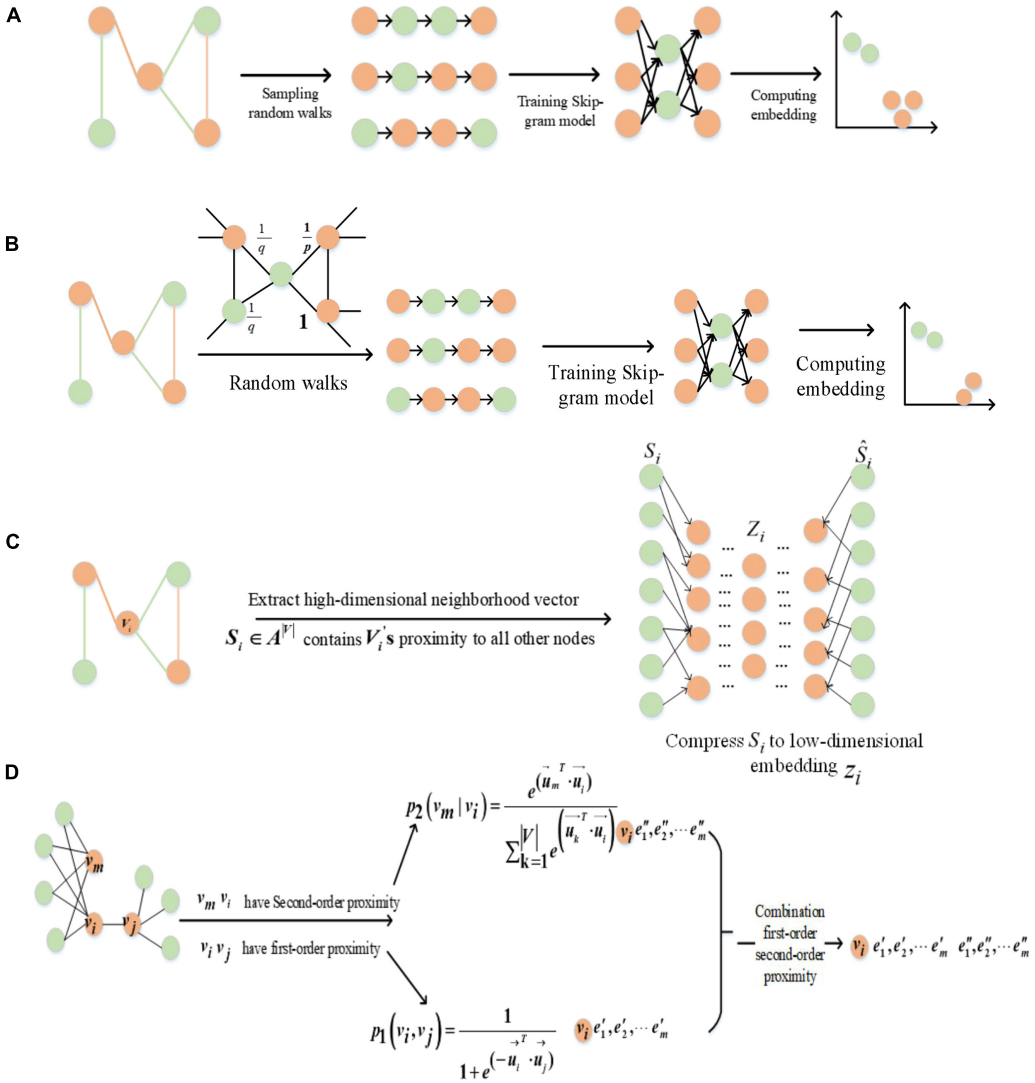

FIGURE 2 | Framework for graph embedding method. **(A)** DeepWalk, **(B)** Node2vec, **(C)** SDNE, and **(D)** LINE.

Protein Similarity Calculation

Each node in the GO graph is represented as a low-dimensional feature vector by considering the topology feature using a graph embedding method. Usually, a protein is annotated by several GO terms. For example, the protein "P03882" is annotated by the GO terms "GO:0004519," "GO:0005739," "GO:0006314," and "GO:0006397." Since a set of GO terms can be represented by its corresponding set of vectors, the similarity between proteins can be calculated based on the similarity of the two sets of GO vectors. Therefore, for any GO term g_i, we use SDNE (Wang et al., 2016), DeepWalk (Perozzi et al., 2014), LINE (Tang et al., 2015), and Node2vec (Grover and Leskovec, 2016) graph embedding methods to learn the low-dimensional feature vector v_i.

We let $G_m = \{g_1, g_2, \cdots, g_m\}$ and $G_n = \{g'_1, g'_2, \cdots, g'_n\}$ denote the sets of GO terms that annotated proteins P_m and P_n; thus, $V_m = \{v_1, v_2, \cdots, v_m\}$ and $V_n = \{v'_1, v'_2, \cdots v'_n\}$ denote the sets of vectors that correspond to $G_m = \{g_1, g_2, \cdots, g_m\}$ and $G_n = \{g'_1, g'_2, \ldots g'_n\}$, respectively. In this paper, we use the idea of DTW to calculate the similarity between two sets of vectors, which is denoted as DTW distance. The smaller the value, the more similar the two proteins. The GO embedding of the two proteins' annotations is concatenated as V_m and V_n, and the lengths are m and n, respectively ($m \neq n$). For constructing the matrix $D_{m\times n}$, the element $D(v_m, v'_n)$ represents the distance between points v_m and v'_n and can be expressed as follows:

$$D(v_m, v'_n) = \min \begin{cases} D(v_{m-1}, v'_n) = Dist(v_{m-1}, v'_n) + d(v_m, v'_n) \\ D(v_m, v'_{n-1}) = Dist(v_m, v'_{n-1}) + d(v_m, v'_n) \\ D(v_{m-1}, v'_{n-1}) = Dist(v_{m-1}, v'_{n-1}) + 2d(v_m, v'_n) \end{cases} \quad (8)$$

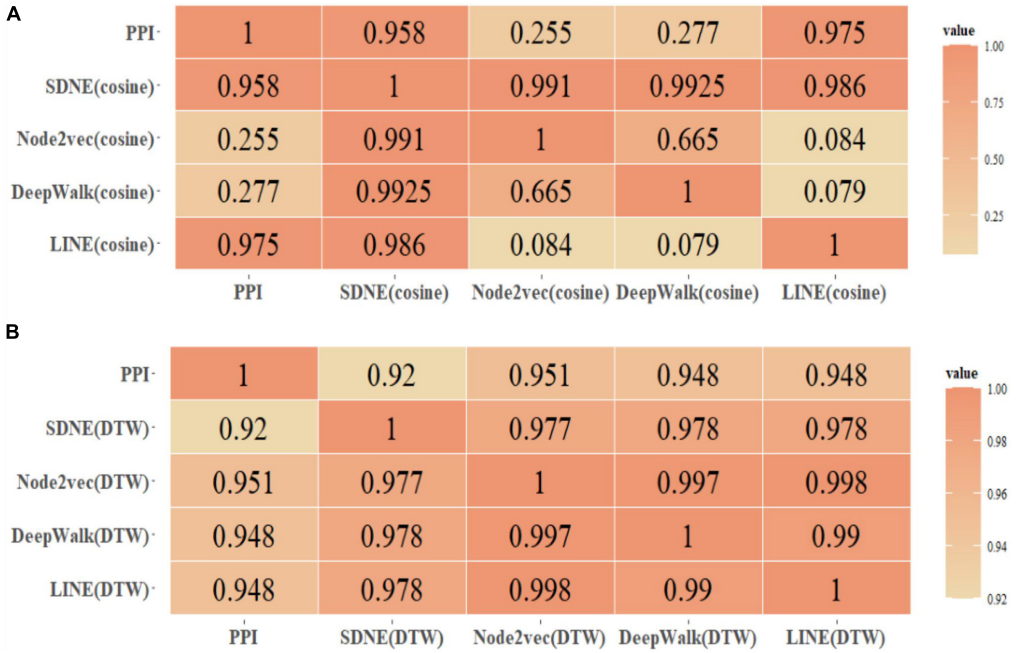

FIGURE 3 | Human protein similarity network ($\tau > 0.4$) and PPI coincidence degree. **(A)** Cosine, **(B)** DTW.

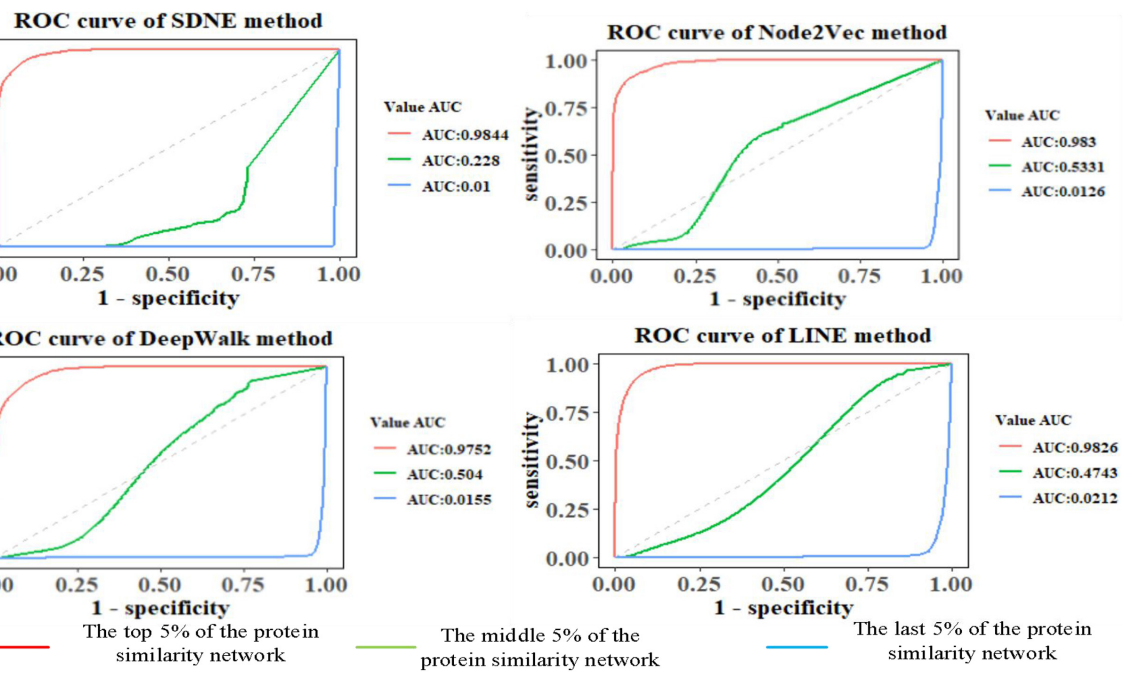

FIGURE 4 | Comparison of prediction results of Human protein similarity networks.

We used the DTW distance method to find a path W through several lattice points in the matrix. The shortest path is the distance between the set of vectors $V_m = \{v_1, v_2, \ldots v_m\}$ and $V_n = \{v'_1, v'_2, \ldots v'_n\}$. We then calculated the distance used to measure the similarity between the two proteins. The process for calculating the DTW distance is presented in **Supplementary Figure 1**.

For any protein P_i, the low-dimensional feature ω_i is directly learned from the GOA graph, which contains the information of term–term and term–protein relations. We use the cosine

distance of the proteins' vector ω to measure the similarity of the proteins. Cosine distance can be expressed as follows:

$$D(P_m, P_n) = \text{cosine}(\omega_m, \omega_n) = \frac{\omega_m \cdot \omega_n}{||\ \omega_m||\ ||\omega_n||} \quad (9)$$

Link Prediction and Evaluation Metrics

When it is difficult to use a unified standard to measure the advantages and disadvantages of a network model, link prediction can be used as a unified comparison method for the similarity nodes in the network. It provides a standard to measure the reliability of the structure of the network. In the comprehensive evaluation, we use two commonly used evaluation indicators, AUC (Lobo, 2010) and AUCPR (Yu and Park, 2014), widely used in dichotomy. Therefore, to evaluate the available networks constructed based on different graph embedding methods in the GO graph and GOA graph, we perform link prediction experiments on the protein similarity network and evaluate the accuracy of the prediction results. For any undirected network $G(V, E)$, we let E be the complete set of $C^2_{|V|}$ node pairs. We first remove 20% of the existing edges E_r in the network. The remaining 80% of the edges E_s are then divided into E_p and E_t, where $E_s = E_p \cup E_t$, $E_P \cap E_t = \emptyset$, and $E = E_r \cup E_s$. Given a link prediction method, each pair of unconnected node pairs v_x and v_y is given a link probability of two nodes. Sorting all the node pairs according to the score value in descending order, we have the top node pair with the highest link probability. The calculation process of the AUC value is presented in **Supplementary Figure 2**. The value of AUCPR is affected by the precision and recall value. For a link prediction experiment, accuracy is defined as the proportion of accurate prediction among the top L prediction edges. If m prediction edges exist, sort the link probability score value in descending order. If m of the top L edges are in the E_t, the precision is defined as follows:

$$Precision = \frac{m}{L} \quad (10)$$

The number of existing edges in the network $M = E - E_r$, where m is the number of edges predicted by the prediction algorithm. The recall index is defined as follows:

$$\text{Recall } = \frac{m}{M} \quad (11)$$

The similarity between nodes is an essential precondition for link prediction, and the more similar the two nodes are, the more likely that a link exists between them. The similarity of network-based structural information definition is called structural similarity. Link prediction accuracy based on structure similarity depends on whether the structure similarity can grasp target structure characteristics. In the link prediction task, there are many methods to calculate the structural similarity between nodes, such as the following:

Common neighbors index

Common Neighbors (CN) (Li et al., 2018) similarity can be called structural equivalence, that is, if two nodes have multiple common neighbors, they are similar. In the link prediction experiment, CN index basic assumption is that if two unconnected nodes have more common neighbors, they are more likely to be connected. For nodes v_x and v_y in the protein similarity network, their neighbors are defined as $\Gamma(x)$ and $\Gamma(y)$, and the similarity of the two nodes is defined as the number of their CN. The index of CN is defined as follows:

$$S_{xy} = \left|\Gamma(x) \cap \Gamma(y)\right| = (A^2) \quad (12)$$

where S represents the similarity matrix and A represents the adjacency matrix of the graph. CN index is based on local information similarity index.

Jaccard index

Based on the common neighbors and considering the influence of the node degree at both ends, the Jaccard (JC) similarity index (Ran et al., 2015) is proposed. JC not only considers the number of two nodes' common neighbors but also considers the number of all their neighbors. JC is defined as follows:

$$S_{xy} = \frac{\left|\Gamma(x) \cap \Gamma(y)\right|}{\left|\Gamma(x) \cup \Gamma(y)\right|} = \frac{(A^2)_{xy}}{\left|\left|\Gamma(x) \cap \Gamma(y)\right|\right|} \quad (13)$$

Resource allocation index

Resource Allocation (RA) (Dianati et al., 2005) index considers the attribute information of the common neighbors of two nodes. In the link prediction process, the common neighbor nodes with higher degrees play a lesser role than those with lower degrees, and the weight of the common neighbor nodes decreases in the form of $1/k$. An example is presented in **Supplementary Figure 3**. RA index (Dianati et al., 2005) is defined as follows:

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{K_z} \quad (14)$$

where K_z is the degree of the common neighbors of nodes v_x and v_y. The calculation process of the RA similarity index is shown in **Supplementary Figure 3**. Assuming that each node's resources are distributed equally to its neighbors, the RA index calculates a node's received resources, which is the similarity between nodes v_x and v_y .

RESULTS

Comparison of Protein Similarity and the Actual PPI Network Coincidence Degree

We downloaded the human yeast protein interaction network from the String database. We then mapped the proteins to the UniProt database, filtered out those proteins that could not be found in the UniProt database, and removed duplicate edges. After filtering, the Yeast dataset consisted of 2,877 proteins with 228,468 interactions, and the Human dataset consisted of 6,882 proteins with 892,054 interactions. Finally, to verify the validity of our calculated protein similarity network, we compared protein similarity and the actual PPI network coincidence degree.

This paper only shows the Human dataset experiment results in **Figure 3**, and the Yeast dataset results are shown in **Supplementary Figures 4**, **5**.

We selected the protein similarity networks ($\tau > 0.4$) and compared them with the PPI dataset downloaded from the

TABLE 2 | AUCPR value of protein similarity prediction in the Human dataset.

Method	The top 5% of the network	The middle 5% of the network	The last 5% of the network
SDNE	0.9105	0.0076	0.0052
Node2vec	**0.9115**	**0.0143**	**0.0055**
DeepWalk	0.8220	0.0127	0.0052
LINE	0.7117	0.0097	0.0052

Bold means the best result in the comparative experiment.

String database to analyze the coincidence degree of the Human and Yeast protein networks. Furthermore, we compared the edge coincidence of the protein similarity network based on different graph embedding methods (as shown in **Figure 3**). The calculation was based on $\frac{E_a \cap E_b}{E_a}(E_a > E_b)$.

By comparing the GO(DTW) and GOA(cosine) methods, it can be seen that the Node2vec graph embedding method performed best in the GO graph. SDNE and LINE methods performed better in the GOA graph, and there was little difference between them in the GOA graph and GO graph. However, Node2vec and DeepWalk performed better in the GO graph. In general, the performance of protein similarity calculation based on different graph embedding methods in the GO graph was better than in the GOA graph. As shown, using graph embedding methods can be effective in calculating protein similarity in GO and GOA graphs. We also proved that using the DTW method to calculate different dimensional protein vector similarities is feasible.

Comparison of Link Prediction Results Based on Different Graph Embedding Methods in GO Graph

The features of GO terms are learned from the GO graph based on different graph embedding methods, and the similarity among proteins is calculated. By selecting the top 5%, middle 5%, and the last 5% of the protein similarity network data, the link prediction is computed for the filtered protein similarity network, and the AUC and AUCPR values are calculated (as shown in **Figure 4** and **Table 2**). This paper only shows the Human dataset experiment result, and the Yeast dataset result is shown in **Supplementary Figure 6** and **Supplementary Table 1**.

We can see that as the similarity of network nodes decreases, the value of AUC decreases. In the top 5% of the protein similarity network, the proteins are more similar, but for AUCPR values, we can see that the performance of the Node2vec method is the best in all the top, middle, and the last 5% of the protein similarity networks. The Node2vec method introduces BFS and DFS into the generation process of the random walk sequence by introducing two parameters p and q. BFS focuses on the adjacent nodes and characterizes a relatively local graph representation; that is, the BFS can explore the local structural properties of the graph, while the DFS can explore the global similarity in context. We found that the AUC value of protein similarity calculated by the graph embedding method decreased gradually with the decrease in the value of the screening protein similarity. Furthermore, it is shown that the edge connection of the protein similarity network calculated by the graph embedding method is reliable.

We also found that the Node2vec graph embedding method performed well in calculating the Yeast protein similarity network (as shown in **Supplementary Figure 6** and **Supplementary Table 1**). Therefore, the GO term vectors fused the local and global information of nodes in the GO graph and contain more information, so the GO(DTW) method performs better in computing protein similarity.

Comparison of Link Prediction Results Based on Different Graph Embedding Methods in the GOA Graph

To reflect the influence of the structure information of the GO annotation on proteins, the features of proteins are learned from the GOA graph based on different graph embedding methods, and the similarity among proteins is calculated (as shown in **Figure 5** and **Table 3**). This paper only shows the Human dataset experiment result, and the Yeast dataset result is presented in **Supplementary Figure 7** and **Supplementary Table 2**.

We screened the top, middle, and last 5% of the protein similar networks and performed the link prediction experiments to observe the values of AUC and AUCPR under different methods. The AUC and AUCPR values decreased gradually with the decrease in the percentage selected. Therefore, it can be seen that the performance of the Node2vec method in the GOA(cosine) method is also better than other graph embedding methods. For the Yeast protein similarity network, we also performed the same experiment and obtained the same experimental conclusions as described above. We found that SDNE graph embedding methods also showed excellent performance in the Yeast dataset (as shown in **Supplementary Table 2**). This is because the SDNE method also defines first-order and second-order similarities. Therefore, calculating the protein similarity network based on these vectors achieved excellent results in the prediction task.

Comparison of Link Prediction Results of Protein Similarity Calculated by IC-Based Method and Based on Graph Embedding Methods

We studied the application of different graph embedding methods to calculate protein similarity in GO and GOA graphs. We screened the top 5% of the protein similarity networks for link prediction analysis (as shown in **Table 4**). Furthermore, we performed an experiment that calculated the density of the protein similarity network based on graph embedding and IC-based methods (as shown in **Table 5**). This paper only presents the Human dataset experiment results, and the Yeast dataset result is presented in **Supplementary Tables 3**, **4**.

The link prediction results from these methods are compared as follows. From **Table 4**, it can be seen that the similarity calculation of proteins based on different graph embedding methods is superior to that of the IC-based methods. We also performed the above experiment for Yeast datasets, and the same conclusion was obtained (as shown in **Supplementary Table 3**). It can be seen that the SDNE and Node2vec graph embedding methods show good performance in the GO graph. Analyzing the

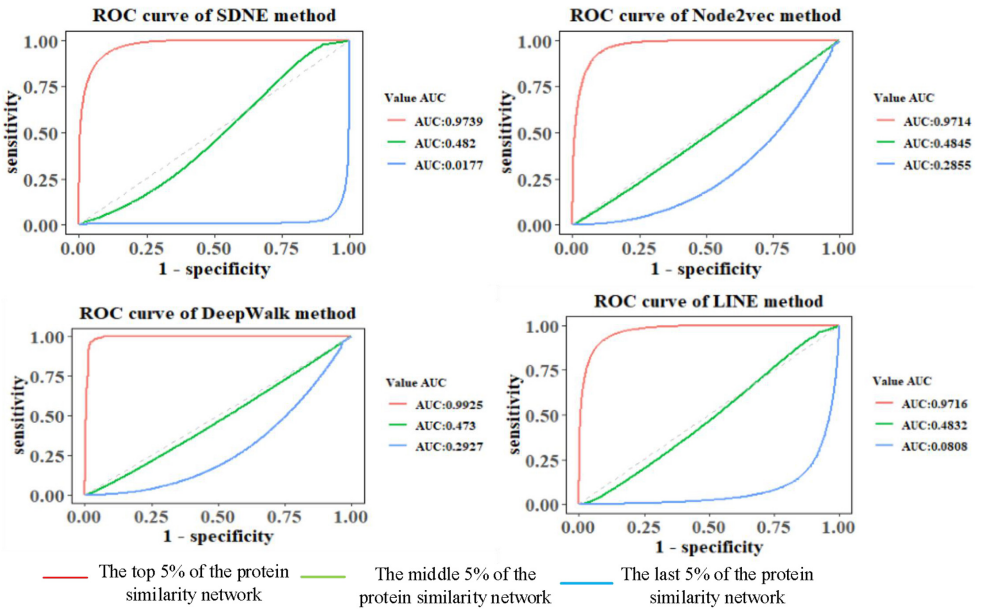

FIGURE 5 | Comparison of prediction results of Human protein similarity networks.

density of the top 5% of the human protein similarity networks, it can be seen that the density of the protein similarity network calculated by the graph embedding method is higher than that calculated by IC-based methods. Therefore, it is shown that the protein similarity network calculated by the IC-based method is sparse, and the similarity of proteins is not as high as that calculated by the graph embedding method. Thus, in the IC-based method, the AUCPR value obtained in link prediction is lower. We also verified this conclusion on the Yeast dataset (as shown in **Supplementary Table 4**).

Based on different graph embedding methods, the features of the GO terms were learned into the vector representations through fusing the topology of the GO graph. Thus, we could capture the global information based on the graph embedding method, and its learned vectors could calculate the similarity between proteins by the DTW distance similarity. As can be seen from the results of the link prediction, the GO(DTW) method performed better than GOA(cosine), and most of the protein similarity networks calculated by the GO(DTW) method are denser than those calculated by the GOA(cosine) method.

TABLE 3 | AUCPR value of Human protein similarity prediction.

Method	The top 5% of the network	The middle 5% of the network	The last 5% of the network
SDNE	0.6578	0.0100	0.0052
Node2vec	**0.8758**	**0.0105**	**0.0069**
DeepWalk	0.8719	0.0094	0.0053
LINE	0.8189	0.0095	0.0053

Bold means the best result in the comparative experiment.

TABLE 4 | AUCPR and AUC values of Human protein similarity prediction (the top 5% of the similarity network).

Method	AUC	AUCPR
SDNE (cosine/DTW)	0.9699/**0.9739**	0.9015/**0.9105**
Node2vec (cosine/DTW)	0.9714/**0.983**	0.8758/**0.9115**
DeepWalk (cosine/DTW)	**0.9925**/0.9752	**0.8719**/0.8220
LINE (cosine/DTW)	**0.9839**/0.9716	**0.8189**/0.7117
Rel.	0.9067	0.1519
Jiang and Conrath	0.8409	0.0669

Bold means the best result in the comparative experiment.

TABLE 5 | Comparison of Human protein similarity network density between different methods.

Method	Nodes	Edges	Density
SDNE (cosine/DTW)	4,797/2,024	1,183,801/713,961	0.1/**0.3**
Node2vec (cosine/DTW)	6,882/2,807	2,841,303/1,183,762	0.12/**0.3**
DeepWalk (cosine/DTW)	6,882/3,079	1,183,876/1,183,707	0.05/**0.2**
LINE (cosine/DTW)	5,586/1,660	1,183,815/206,650	0.07/**0.15**
Rel	5,902	870,987	0.05
Jiang and Conrath	5,883	870,986	0.05

Bold means the best result in the comparative experiment.

TABLE 6 | Prediction results under different similarity indexes (the top 5% of the Human protein similarity network).

Similarity index	CN	JC	RA
SDNE (cosine/DTW)	0.9694/0.981	0.9739/0.9843	**0.9818**/**0.9886**
Node2vec (cosine/DTW)	0.9598/0.9809	0.9714/0.9843	**0.9856**/**0.9886**
DeepWalk (cosine/DTW)	0.9772/0.981	0.9856/0.9842	**0.9885**/**0.9884**
LINE (cosine/DTW)	0.9703/0.9716	0.9716/0.9825	**0.9874**/**0.9853**

Bold means the best result in the comparative experiment.

Similarity Indexes' Results

We performed three different link prediction similarity index experiments on the top 5% of the protein similarity network and found that based on different similarity indexes, the difference in the AUC value is small, which indicates that the calculated protein similarity network structure has improved (as shown in **Table 6**). This paper only presents the Human dataset experiment result, and the Yeast dataset result is presented in **Supplementary Table 5**.

Among the three different similarity evaluation indexes, we found that the AUC value of the RA similarity index based on link prediction is slightly higher than the other two similarity indexes. Furthermore, the results showed that the top 5% of the protein similarity network had higher AUC values in different similarity indexes of link prediction, indicating that the graph embedding method effectively calculated protein similarity. We obtained the same conclusion in the experiment with the Yeast dataset (as shown in **Supplementary Table 5**).

DISCUSSION

Gene Ontology is one of the many biological ontology languages. Its emergence and development reduce the confusion of biological concepts and terms, provide a three-layer (BP, MF, and CC) structure of system definition, and describe the functions of proteins. Therefore, it is important to understand protein function based on GO terms to describe protein similarity.

In this paper, by fusing the GO terms' topology information, we learned the features of GO terms and proteins into vector representations in GO and GOA graph based on different graph embedding methods. Then, the similarity of proteins was calculated based on these vectors using DTW and cosine similarity. Finally, protein similarity networks were screened by selecting different percentages, and a link prediction experiment was used to evaluate the prediction accuracy of different networks. The experimental results indicate that the graph embedding method is better than the IC-based method in protein similarity calculation. Among the two graph embedding methods, the performance of the GO(DTW) method is better than that of the GOA(cosine) method. This is because the GO terms and proteins are treated equally in the GOA graph, and some information may be ignored when learning protein low-dimensional embedding. Therefore, the coincidence degree between the protein similarity network calculated by the GOA(cosine) method and the actual PPI data is not as high as that calculated by the GO(DTW) method. There are potential limitations to our method. First, we transformed directed graphs into undirected graphs, which might result in a loss of structural information. We also treated the GO terms and the proteins equally in the GOA graph, which may ignore some information. Therefore, in our future study, we plan to learn the protein representations in the graph by combining the information in the directed graph and by considering representation learning of heterogeneous graphs that contain GO terms and 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 author/s.

AUTHOR CONTRIBUTIONS

YZ conceived the idea and prepared the experimental data. ZW and YZ debugged the code, conducted the experiments, interpreted the results, and wrote and edited the manuscript. SW and JS advised the study and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61902430, 61873281, and 61972226).

ACKNOWLEDGMENTS

We would like to thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.744334/full#supplementary-material

REFERENCES

Amos, B., and Brigitte, B. (1999). The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* 22, 49–54. doi: 10.1093/nar/22.17.3626

Dianati, M., Shen, X., and Naik, S. (2005). "A new fairness index for radio resource allocation in wireless networks," in *Proceedings of the Wireless Communications & Networking Conference* (New Orleans, LA: IEEE), 785–890.

Goldberg, Y., and Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *OALib J.* 14, 144–156. doi: 10.1017/S1351324916000334

Grover, A., and Leskovec, J. (2016). "node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference*, (New York, NY: ACM), 855–864.

Harris, M. A. (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, 258–261. doi: 10.1093/nar/gkh036

Hu, L., Wang, X., Huang, Y. A., Hu, P., and You, Z. H. (2021). A survey on computational models for predicting protein–protein interactions. *Bioinformatics.* 05, 77–85. doi: 10.1093/bib/bbab036

Jiang, J. J., and Conrath, D. W. (1997). "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of the 10th Research on Computational Linguistics International Conference*, Vol. 11, (Taipei: The

Association for Computational Linguistics and Chinese Language Processing (ACLCLP)), 115–123. doi: 10.1.1.269.3598

Li, S., Huang, J., Zhang, Z., Liu, J., Huang, T., and Chen, H. (2018). Similarity-based future common neighbors model for link prediction in complex networks. *Sci. Rep.* 19, 518–524. doi: 10.1038/s41598-018-35423-2

Lobo, J. M. (2010). AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol.* 17, 145–151. doi: 10.1111/j.1466-8238.2007.00358

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

Lou, Y., Ao, H., and Dong, Y. (2016). "Improvement of dynamic time warping (DTW) Algorithm," in *Proceedings of the 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*, 14 (Guiyang: IEEE), 18–24. doi: 10.1109/DCABES.2015.103

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

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "DeepWalk: online learning of social representations," in *Proceedings of the 2014 ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (New York, NY: ACM), 701–740.

Pesaranghader, A., Rezaei, A., and Davoodi, D. (2014). Gene functional similarity analysis by definition-based semantic similarity measurement of GO terms. *Lecture Notes Bioinformatics.* 12, 203–214. doi: 10.1007/978-3-319-06483-3_18

Ran, S., Ngan, K. N., and Li, S. (2015). "Jaccard index compensation for object segmentation evaluation," in *Proceedings of the 2014 IEEE International Conference on Image Processing*, (Paris: IEEE), 253–259.

Resnik, P. (1999). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* 11, 95–130. doi: 10.1613/jair.514

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J., Martínez-Cruz, L. A., et al. (2005). Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans. Comput. Biol. Bioinformatics.* 24, 330–338. doi: 10.1109/TCBB.2005.50

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "LINE: large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web*, (New York, NY: ACM), 1067–1077.

UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 32, 115–119. doi: 10.1093/nar/gkh131

Wang, D., Peng, C., and Zhu, W. (2016). "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD International Conference on Data Mining*, (New York, NY: ACM), 1225–1234.

Wang, Z., Zhang, Y., Wang, S., and Shang, J. (2020). SINE: second-order information network embedding. *IEEE Access* 1, 98–110. doi: 10.1109/ACCESS.2020.3007886

Xi, J., Li, A., and Wang, M. (2020a). HetRCNA: a novel method to identify recurrent copy number alternations from heterogeneous tumor samples based on matrix decomposition framework. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 422–434. doi: 10.1109/TCBB.2018.2846599

Xi, J., Ye, L., and Huang, Q. (2021). "Tolerating data missing in breast cancer diagnosis from clinical ultrasound reports via knowledge graph inference," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)*, (New York, NY: Association for Computing Machinery), 1–9. doi: 10.1145/3447548.3467106

Xi, J., Yuan, X., Wang, M., Li, A., and Huang, Q. (2020b). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics* 36, 1855–1863.

Yu, W., and Park, T. (2014). AucPR: an AUC-based approach using penalized regression for disease prediction with high-dimensional omics data. *BMC Genomics.* 15:S1. doi: 10.1186/1471-2164-15-S10-S1

Zhong, X., Kaalia, R., and Rajapakse, J. C. (2019). GO2Vec: transforming GO terms and proteins to vector representations via graph embeddings. *BMC Genomics* 20:918. doi: 10.1186/s12864-019-6272-2

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.

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