Complexity

Biomolecular Networks for Complex Diseases

Guest Editors: Fang X. Wu, Jianxin Wang, Min Li, and Haiying Wang

Biomolecular Networks for Complex Diseases

Complexity

Biomolecular Networks for Complex Diseases

Guest Editors: Kazuo Toda, Jorge L. Zeredo, Sae Uchida, and Vitaly Napadow

Editorial Board

Contents

Hindawi
Complexity
Volume 2018, Article ID 4210160, 3 pages
https://doi.org/10.1155/2018/4210160

Editorial

Biomolecular Networks for Complex Diseases

Fang-Xiang Wu,[1,2] **Jianxin Wang**,[3] **Min Li**,[3] **and Haiying Wang**[4]

[1]*School of Mathematical Sciences, Nankai University, Tianjin 300071, China*
[2]*Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A9*
[3]*School of Information Science and Engineering, Central South University, Changsha, Hunan 410012, China*
[4]*School of Computing and Mathematics, Ulster University, Belfast BT37 0QB, UK*

Correspondence should be addressed to Fang-Xiang Wu; faw341@mail.usask.ca

Received 18 January 2018; Accepted 18 January 2018; Published 15 February 2018

It is widely acknowledged that complex diseases or disorders (e.g., cancer, AIDS, and obesity) stem from the dysfunction of biomolecular networks, not only their isolated components (e.g., genes, proteins, and metabolites). Biomolecular networks typically include gene regulatory networks, protein-protein interaction networks, metabolic networks, and signal transduction networks. With advances in high throughput measurement techniques such as microarray, RNA-seq, ChIP-chip, yeast two-hybrid analysis, and mass spectrometry, large-scale biological data have been and will continuously be produced. Such data contain insightful information for understanding the mechanism of molecular biological systems and have proved useful in diagnosis, treatment, and drug design for complex diseases or disorders. For this special issue, we have invited the researchers to contribute their original studies in modeling/construction, analysis, synthesis, and control of complex disease-related biomolecular networks. This special issue accepted eleven articles for inclusion after rigorous review. We would like to introduce each of them by a short description.

Existing studies have shown that microRNAs (miRNAs) are involved in the development and progression of various complex diseases. Experimental identification of miRNA-disease association is expensive and time-consuming and thus it is appealing to design efficient algorithms to identify novel miRNA-disease association. In the paper "miRNA-Disease Association Prediction with Collaborative Matrix Factorization," Z. Shen et al. developed the computational method of Collaborative Matrix Factorization for miRNA-Disease Association (CMFMDA) prediction to identify potential miRNA-disease associations by integrating miRNA functional similarity, disease semantic similarity, and experimentally verified miRNA-disease associations. Experiments verified that CMFMDA achieved intended purpose and application values with its short consuming-time and high prediction accuracy. In addition, CMFMDA was applied to reveal the potential related miRNAs of Esophageal Neoplasms and Kidney Neoplasms.

The identification of target molecules associated with specific complex diseases is the basis of modern drug discovery and development. The computational methods provide a low-cost and high-efficiency way for predicting drug-target interactions (DTIs) from biomolecular networks. In the paper "SDTRLS: Predicting Drug-Target Interactions for Complex Diseases Based on Chemical Substructures," C. Yan et al. proposed a method (called SDTRLS) for predicting DTIs through RLS-Kron model with chemical substructure similarity fusion and Gaussian Interaction Profile (GIP) kernels. Their computational experiments showed that SDTRLS outperformed the state-of-the-art methods such as SDTNBI.

Vaccines represent one of the most effective interventions to control infectious diseases. Despite the many successes, an effective vaccine against current global pandemics such as HIV, malaria, and tuberculosis is still missing. In the paper "Exploring the Limitations of Peripheral Blood Transcriptional Biomarkers in Predicting Influenza Vaccine Responsiveness," L. Marchetti et al. applied systems biology to vaccinology and employed a recently established algorithm for signature-based clustering of expression profiles, SCUDO,

to provide new insights into why blood-derived transcriptome biomarkers often fail to predict the seroresponse to the influenza virus vaccination. Their analysis revealed that composite measures provided a more accurate assessment of the seroresponse to multicomponent influenza vaccines.

Protein complexes are involved in multiple biological processes, and thus detection of protein complexes is essential to the understanding of complex diseases. In the paper "Predicting Protein Complexes in Weighted Dynamic PPI Networks Based on ICSC," J. Zhao et al. proposed a novel algorithm named improved Cuckoo Search Clustering (ICSC) algorithm for detecting protein complexes in weighted dynamic protein-protein interaction (PPI) networks. The experimental results on both DIP dataset and Krogan dataset demonstrated that ICSC algorithm was more effective in identifying protein complexes than other competing methods.

With the development of gene sequencing technology and other gene detection technologies, huge gene data have been generated. Differentially expressed genes identified from gene expression data play an important role in cancer diagnosis and classification. In the paper "Robust Nonnegative Matrix Factorization via Joint Graph Laplacian and Discriminative Information for Identifying Differentially Expressed Genes," L.-Y. Dai et al. proposed a novel constrained method named robust nonnegative matrix factorization via joint graph Laplacian and discriminative information (GLD-RNMF) for identifying differentially expressed genes, in which manifold learning and the discriminative label information are incorporated into the traditional nonnegative matrix factorization model to train the objective matrix. The experimental results on two publicly available cancer datasets demonstrated that GLD-RNMF was an effective method for identifying differentially expressed genes.

Many of genetic changes represented neutral variations that do not contribute to cancer development which are called passenger mutations. Only a few alterations are causally implicated in the process of oncogenesis which are referred to as driver mutations. Although some large-scale cancer genomics projects have produced different omics data, it is still a major challenge to distinguish pathogenic driver mutations from the so-called random mutated passenger mutations. In the paper "DriverFinder: A Gene Length-Based Network Method to Identify Cancer Driver Genes," P.-J. Wei et al. presented a gene length-based network method, named DriverFinder, to identify driver genes by integrating somatic mutations, copy number variations, gene-gene interaction network, tumor expression, and normal expression data. Their computational experimental results demonstrated the effectiveness of their proposed method.

Although Genome-Wide Association Studies (GWAS) predicted massive genetic variations related to complex traits, they can only explain a small part of the mechanism under the complex diseases known as "missing heritability." In the paper "FAACOSE: A Fast Adaptive Ant Colony Optimization Algorithm for Detecting SNP Epistasis," L. Yuan et al. presented a unified fast framework integrating adaptive ant colony optimization algorithm with multiobjective functions for detecting SNP epistasis in GWAS datasets. Their experimental results from Late-Onset Alzheimer's Disease dataset showed that the proposed method outperformed other methods in epistasis detection and could contribute to the research of mechanism underlying the disease.

Clinical disorders of human brains, such as Alzheimer's disease (AD), schizophrenia (SCZ), and Parkinson's disease (PD), are among the most complex diseases and therapeutically intractable health problems. In recent years, brain regions and their interactions can be modeled as complex brain networks, which describe highly efficient information transmission in a brain. Many brain disorders have been found to be associated with the abnormal topological structures of brain networks. In the paper "Complex Brain Network Analysis and Its Applications to Brain Disorders: A Survey," J. Liu et al. provided a comprehensive overview for complex brain network analysis and its applications to brain disorders.

Colorectal cancer (CRC) ranks 4 in cancer incidences and accounts for approximately 8–10% of cancer-related death and the 5-year survival rate (40–50%) is still not as satisfied as expected. Identifying the "high risk" populations is critical for early diagnosis and improvement of overall survival rate. In the paper "Building Up a Robust Risk Mathematical Platform to Predict Colorectal Cancer," L. Zhang et al. collected relatively complete information of genetic variations and environmental exposure for both CRC patients and cancer-free controls and developed a multimethod ensemble model for CRC-risk prediction by employing such big data to train and test the model. Their results demonstrated that (1) the explored genetic and environmental biomarkers were validated to connect to the CRC by biological function- or population-based evidences, (2) the model could efficiently predict the risk of CRC after parameter optimization by the big CRC-related data, and (3) their innovated heterogeneous ensemble learning model (HELM) and generalized kernel recursive maximum correntropy (GKRMC) algorithm have high prediction power.

Osteoporosis is a type of systemic skeletal disease that is characterized by reduced bone mass and microarchitecture deterioration of bone tissues, thereby leading to the loss of strength and increased risk of fractures. In past decades, a number of genes and SNPs associated with osteoporosis have been found through GWAS method. In the paper "Identifying the Risky SNP of Osteoporosis with ID3-PEP Decision Tree Algorithm," J. Yang et al. proposed a computational method for identifying the suspected risky SNPs of osteoporosis based on the known osteoporosis GWAS-associated SNPs. The experiment result showed that their method was feasible and could provide a more convenient way to identify the suspected risky SNPs associated with osteoporosis.

Major depressive disorder (MDD) is a global mental disorder and has an unfavorable influence on physical and psychological health. Studies demonstrated that MDD is characterized by the alterations in brain functional connections which is also identifiable during the brain's "resting-state." However, the existing approaches to constructing functional connectivity are often biased and as a result the clustering partition of nodes was unclear. In the paper "A Resting-State Brain Functional Network Study in MDD

Based on Minimum Spanning Tree Analysis and the Hierarchical Clustering," X. Li et al. applied minimum spanning tree (MST) analysis and the hierarchical clustering for studying the depression disease. With resting-state electroencephalograms (EEG) from 15 healthy and 23 major depressive subjects, their findings suggested that there was a stronger brain interaction in the MDD group and a left-right functional imbalance in the frontal regions for MDD controls.

In summary, this focus issue has reported the recent progress in the studies of biomolecular networks for complex diseases. We hope that the readers of this focus issue could get some benefits from these newly developed methods.

Fang-Xiang Wu
Jianxin Wang
Min Li
Haiying Wang

Hindawi
Complexity
Volume 2017, Article ID 2713280, 10 pages
https://doi.org/10.1155/2017/2713280

Research Article

SDTRLS: Predicting Drug-Target Interactions for Complex Diseases Based on Chemical Substructures

Cheng Yan,[1,2] Jianxin Wang,[1] Wei Lan,[1] Fang-Xiang Wu,[3] and Yi Pan[4]

[1]*School of Information Science and Engineering, Central South University, Changsha, Hunan 410083, China*
[2]*School of Computer and Information, Qiannan Normal University for Nationalities, Duyun, Guizhou 558000, China*
[3]*Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada*
[4]*Department of Computer Science, Georgia State University, Atlanta, GA 30302, USA*

Correspondence should be addressed to Jianxin Wang; jxwang@mail.csu.edu.cn

Received 8 April 2017; Revised 19 October 2017; Accepted 1 November 2017; Published 3 December 2017

Academic Editor: Daniela Paolotti

It is well known that drug discovery for complex diseases via biological experiments is a time-consuming and expensive process. Alternatively, the computational methods provide a low-cost and high-efficiency way for predicting drug-target interactions (DTIs) from biomolecular networks. However, the current computational methods mainly deal with DTI predictions of known drugs; there are few methods for large-scale prediction of failed drugs and new chemical entities that are currently stored in some biological databases may be effective for other diseases compared with their originally targeted diseases. In this study, we propose a method (called SDTRLS) which predicts DTIs through RLS-Kron model with chemical substructure similarity fusion and Gaussian Interaction Profile (GIP) kernels. SDTRLS can be an effective predictor for targets of old drugs, failed drugs, and new chemical entities from large-scale biomolecular network databases. Our computational experiments show that SDTRLS outperforms the state-of-the-art SDTNBI method; specifically, in the G protein-coupled receptors (GPCRs) external validation, the maximum and the average AUC values of SDTRLS are 0.842 and 0.826, respectively, which are superior to those of SDTNBI, which are 0.797 and 0.766, respectively. This study provides an important basis for new drug development and drug repositioning based on biomolecular networks.

1. Introduction

The identification of target molecules associated with specific diseases is the basis of modern drug discovery and development [1–3]. Therefore, the identification of drug-target interactions (DTIs) is important for drug development. However, it is well known that drug discovering is a cost- and time-consuming process in the field of pharmacology. According to the USA Food and Drug Administration statistical data, the cost of new drug discovery is approximately $1.8 billion and it takes an average of 13 years [4]. Therefore, how to deal with this problem becomes an emerging issue. Over decades, different computational methods and tools [5–13] have been developed to predict large-scale potential DTIs and drug repositing through the unremitting efforts of a large number of researchers and organizations under the development of computing technology.

Meanwhile, many DTI data have been generated with the rapid growth of the public chemical and biological database. For example, PubChem [14] is a freely available chemistry database. There are 7759 drug entities, 4104 target proteins, and 15,199 DTIs present in DrugBank [15] database by now. The freely available online ChEMBL [16] database provides pharmaceutical chemists with a convenient platform for querying target bioactivity data for compounds or targets. In addition, including TTD [17], KEGG [18], SIDER [19], STITCH [20], STRING [21], BindingDB [22], and other various kinds of resources have established the basis for DTI prediction.

Now it is possible for us to quickly and inexpensively identify potential DTIs and repurpose existing drugs [23–27] through the developments of computational methods. These methods are mainly divided into three categories, including basic network-based models, machine learning-based models, and other approaches based on similarity [28].

From the viewpoint of basic network-based model, Cheng et al. [29] developed the method to predict DTIs through network-based inference (NBI). Comparing with drug-based similarity inference (DBSI) and target-based similarity inference (TBSI), NBI is better than them because it is in the full use of the known DTIs. Moreover, node- and edge-weighted NBI was developed via constructing the weight of nodes and edges on drug-target network. Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) was developed by Chen et al., which implemented the random walk on the heterogeneous network (protein-protein similarity network, drug-drug similarity network, and known drug-target interaction networks) [30]. It is an enhanced version of the traditional random walk that improved the predictive performance through making full use of data with the integrated heterogeneous network.

Some machine learning-based approaches were also developed to predict DTIs. Following Bleakley et al. [31] and Mordelet and Vert [32], Bleakley and Yamanishi [33] further proposed the bipartite local model (BLM) to predict DTIs, which used local support vector machine (SVM) classifiers with known DTIs and integrated the chemical structure similarity and protein sequence similarity information. Gaussian Interaction Profile (GIP) kernels on drug-target networks were significant improvements developed by van Laarhoven et al. [34]. In order to solve the problem of negative samples, Lan et al. [35] proposed a prediction method (PUDT) which classified unlabeled samples into the reliable negative examples and likely negative examples based on the similarity of protein structure and achieved good results.

The matrix decomposition technique is also used for predicting DTIs, miRNA-disease associations [36, 37], and so on. It maps the DTI matrix to the low-dimensional matrix to infer the hidden interactions based on the known interactions. Gönen [38] proposed a Bayesian model that combined dimensionality reduction, matrix factorization, and binary classification for predicting DTIs via integrating the drug-drug chemical similarity and protein-protein sequence similarity. Multiple Similarities Collaborative Matrix Factorization (MSCMF) [39] method projected drugs and targets into a common low-rank feature space and significantly improved the results via adjusting the weight of similarity matrix of drugs and of targets. Ezzat et al. [40] developed the Regularized Matrix Factorization method that distinguished from many of nonoccurring edges in the interaction matrix which are actually unknown or hide cases by other similarity information. DrugE-Rank [12] developed a machine learning-based model by combining the advantages of two different types of feature-based and similarity-based methods to improve the prediction performance.

Although the above methods have gained good results in predicting the new DTIs on known drugs, it is also important to predict DTIs of failed drugs and new chemical entities. There are thousands of drugs that are failed in clinical phases and even US National Center for Advancing Translational Sciences is paying US$20 million to research for repurposing 58 failed drugs [41, 42] as the drugs that failed in their initially targeted diseases may be effective in other diseases. Wu et al. [43] proposed an integrated network and chemoinformatics tool for systematic prediction of DTIs and drug repositioning, namely, SDTNBI (substructure-drug-target network-based inference) which predicted new DTIS of failed drugs and new chemical entities by integrating known DTIS and chemical substructure of failed drugs or new chemical entities in a way of resource diffusion. Their study assumed that chemical substructure played key roles in DTIs. This method achieved good prediction results for large-scale failed drugs and new chemical entities based on chemical substructures shared between them and the known drugs.

In this study, we propose a method called SDTRLS (substructure-drug-target Kronecker product kernel regularized least squares) for large-scale DTI prediction and drug repositioning based on the chemical substructures of known drugs, failed drugs, and new chemical entities. Firstly, we compute the substructure similarity and then create a Gaussian Interaction Profile (GIP) kernels for drug entities and target proteins based on known DTIs. The k-nearest neighbor (KNN) was used to compute the initial relational score in the presence of a new chemical entity or failed drug that has no known DTIs. Through similarity network fusion (SNF) technology [44] the similarity of substructure and GIP of drugs are integrated. SNF substantially outperforms single-type data analysis and establishes integrative approaches to predicting DTIs. Finally, the RLS-Kron [34] classifier was used to predict DTIs, which constructs a large kernel that directly relates to the drug-target pairs by combining the similarity kernels of drug entities and target proteins. In order to comprehensively assess the performance of our method, we compare it against current state-of-the-art algorithms with the same data and evaluation criteria. We use the 10-fold cross validation and external validation to show the accuracy and robustness of our method. The computational results show that our proposed SDTRLS is comparable to other five methods in terms of stability. Especially in the G protein-coupled receptors (GPCRs) external validation dataset, the maximum and average AUC values were 0.842 and 0.826, respectively, which are superior to 0.797 and 0.766 from state-of-the-art SDTNBI method. In order to further confirm the prediction ability of STDRLS, we perform an experimental analysis on some prediction results. In summary, we provide a new alternative method for DTI prediction for known drugs, failed drugs, and new chemical entities. It provides the basis for drug discovery, development, and personalized medical treatment in the future.

2. Materials

This study used five internal validation datasets and two external validation datasets. The internal datasets are used to validate the predictions of the new DTIs of known drugs, and the external datasets are used to validate the predictions of all DTIs of new entities and failed drugs. Five internal datasets are G protein-coupled receptors (GPCRs), kinase superfamily (Kinases), ion channels (ICs), nuclear receptors (NRs), and Global. GPCRs and Kinases were downloaded from ChEMBL database. ICs and NRs were collected from the ChEMBL and BindingDB database. The Global is a global network covering genomewide targets where all drugs also

Table 1: Drugs, targets, and DTIs in each dataset.

Datasets	Targets	S_d	S_t	N_{dt}	Sparsity (%)
Internal datasets	GPCRs	4741	97	17,111	3.72
	Kinases	2827	206	13,647	2.34
	ICs	7929	97	8944	1.16
	NRs	5218	35	7366	4.03
	Global	1844	1032	10,185	0.54
External datasets	ExGPCRs	92	46	271	6.4
	ExKinases	188	28	202	3.84

S_d is the number of drugs, S_t is the number of targets, N_{dt} is the known DTIs, and sparsity is the proportion of the of N_{dt} to all possible DTIs in datasets.

come from DrugBank database. Two external datasets were selected from GPCRs and Kinases in DrugBank database, respectively.

The external validation is to predict all DTIs for drugs, so it needs a basic dataset that includes drugs, targets, and known DTIs. GPCRs and Kinases are the basic datasets to ExGPCRs and ExKinases, respectively. The known 17,111 DTIs of GPCRs are the prior knowledge to external validation of ExGPCRs in Table 1.

Table 1 shows that the 92 drugs of ExGPCRs and 4741 of GPCRs are independent of each other. However, the 46 targets of ExGPCRs are the subset of the 92 targets of GPCRs. Furthermore, the relationship of drugs and targets between Kinases and ExKinases is the same as that between ExGPCRs and GPCRs. These datasets can be downloaded from http://lmmd.ecust.edu.cn/methods/sdtnbi/#*. Table 1 contains some statistics of five internal validation datasets and two external datasets.

2.1. Chemical Substructure. In this study, we used seven types of fingerprints to express the chemical substructures of each molecule. All substructure data are generated from PaDEL-Descriptor software, including CDK Fingerprint, CDK Extended Fingerprint, CDK Graph Only Fingerprint, Substructure Fingerprint, Klekota-Roth Fingerprint, MACCS Fingerprint, and PubChem Fingerprint, namely, CDK, CDKExt, Graph, FP4, KR, MACCS, and PubChem, respectively. Each type of substructures of each molecule is represented by a multiple dimensional vector with values of 0 or 1. We only used the substructures that appear in the datasets.

Table 2 contains the overview of the seven substructures of dataset GPCRs, including the dimension of each chemical substructure. The dimensions of substructures were derived from the statistics result of the datasets that include all appearing substructure types.

Table 2: The dimensions on GPCRs.

Chemical substructure types	Dimensions
CDK	1024
CDKExt	1012
FP4	131
Graph	1023
KR	1834
MACCS	153
PubChem	627

3. Methods

3.1. Chemical Substructure Similarity. Let $S = \{s_1, s_2, \ldots, s_K\}$ be a set of all substructures for one type of seven chemical substructures, where K is the dimension of the chemical substructure. For example, the value of K is 1024 in CDK and the value of K is 153 in MACCS. $D = \{d_1, d_2, \ldots, d_m\}$ is the set of drugs, where m is the number of drugs. For one chemical substructure, drug d_i can be represented by a profile (binary vector) of the substructure, that is, $DS(d_i) = \{ds_1(d_i), ds_2(d_i), \ldots, ds_K(d_i)\}$. If drug d_i has s_k, the value of $ds_k(d_i)$ is 1, otherwise 0. For a type of chemical substructure, the substructure similarity $S_{\text{subsim}}(d_i, d_j)$ of drugs d_i and d_j can be computed by the weighted cosine correlation coefficient based on the substructure information [27].

$$S_{\text{subsim}}\left(d_i, d_j\right) = \frac{\sum_{k=1}^{K} w_k ds_k\left(d_i\right) ds_k\left(d_j\right)}{\sqrt{\sum_{k=1}^{K} w_k ds_k^2\left(d_i\right)} \sqrt{\sum_{k=1}^{K} w_k ds_k^2\left(d_j\right)}}, \quad (1)$$

where w_k is the weight of the kth substructure (s_k), which can be calculated by the formula [27]

$$w_k = \exp\left(-\frac{f_k^2}{\delta^2 h^2}\right), \quad (2)$$

where f_k is the frequency of chemical substructure s_k in the whole dataset, δ is the standard deviation of $\{s_k\}_{k=1}^{k=K}$, and h is a parameter (set to be 0.1 in this study). The basic rationale for introducing the weight to compute substructure similarity between drugs and new chemical entities is that substructures with fewer occurrences should occupy a more proportion than substructures which appear frequently.

3.2. Gaussian Interaction Profile Kernel. We denoted that $T = \{t_1, t_2, \ldots, t_n\}$ is the set of n targets. A drug-target network can be represented by a bipartite graph which has an adjacency

matrix $Y \in R^{m*n}$, where the value of y_{ij} is 1 if d_i and t_j have known DTI, otherwise 0. The Gaussian Interaction Profile (GIP) kernel is constructed from the topology information of known DTIs network [10, 34]. The kernel of drugs d_i and d_j can be formulated as

$$K_{\mathrm{GIP},d}\left(d_i, d_j\right) = \exp\left(-\gamma_d \left\|Y\left(d_i\right) - Y\left(d_j\right)\right\|^2\right),$$
$$\gamma_d = \frac{\alpha}{\left(\left(1/N_d\right)\sum_{i=1}^{N_d}\left\|Y\left(d_i\right)\right\|^2\right)}, \tag{3}$$

where $Y(d_i) = \{y_{i1}, y_{i2}, \ldots, y_{in}\}$ is the interaction profile of drug d_i, and α is a parameter that controls the bandwidth; we set the value to be 1 in this study. Similarly, the kernel of targets t_i and t_j can be calculated by (4).

$$K_{\mathrm{GIP},t}\left(t_i, t_j\right) = \exp\left(-\gamma_t \left\|Y\left(t_i\right) - Y\left(t_j\right)\right\|^2\right), \tag{4}$$

$$\gamma_t = \frac{\beta}{\left(\left(1/N_t\right)\sum_{i=1}^{N_t}\left\|Y\left(t_i\right)\right\|^2\right)}, \tag{5}$$

where $Y(t_j) = \{y_{1j}, y_{2j}, \ldots, y_{mj}\}^T$ is the interaction profile of target t_j; we also set the parameter β to be 1.

3.3. Similarity Network Fusion. We have two similarity matrices for drugs (including known drugs, new chemical entities), namely, substructure similarity $S_{\mathrm{subsim}} \in R^{m*m}$ and $K_{\mathrm{GIP},d} \in R^{m*m}$. To construct more comprehensive similarity kernel for drugs, we used the SNF method to fuse two similarity kernels.

Firstly, the row-normalized matrices $P^{(1)}$ and $P^{(2)}$ are calculated from the drug similarity matrices S_{subsim} and $K_{\mathrm{GIP},d}$, respectively. Secondly, according to the K-nearest neighbors (KNN) method, the resultant matrices $S^{(1)}$ and $S^{(2)}$ are obtained from $P^{(1)}$ and $P^{(2)}$ by the following equation[44]:

$$S\left(d_i, d_j\right) = \begin{cases} \dfrac{P\left(d_i, d_j\right)}{\sum_{d_k \in N(d_i)} P\left(d_i, d_k\right)}, & d_j \in N\left(d_i\right), \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $N(d_i)$ is the set of top N similar neighbors of drug d_i. In this study, we set the value of N to be 50. The main idea of SNF is iteratively updating similarity matrices $P^{(1)}$ and $P^{(2)}$ [44].

$$P_{t+1}^{(1)} = S^{(1)} \times P_t^{(2)} \times \left(S^{(1)}\right)^T,$$
$$P_{t+1}^{(2)} = S^{(2)} \times P_t^{(1)} \times \left(S^{(2)}\right)^T, \tag{7}$$

where the parameter t represents the times of iterations, and its value is set to be 20 in this study by considering that the iteration time can not be too long and $\max(\max(\mathrm{abs}(((P_t^{(1)} + P_t^{(2)})/2 - (P_{t-1}^{(1)} + P_{t-1}^{(2)})/2)))) < 10^{-3}$. The initial matrices are defined as $P_{t=1}^{(1)} = P^{(1)}$ and $P_{t=1}^{(2)} = P^{(2)}$. The final similarity matrix $S_{\mathrm{final}} \in R^{m*m}$ of drugs is calculated from the average value of matrices $P_{20}^{(1)}$ and $P_{20}^{(2)}$ ($S_{\mathrm{final}} = (P_{20}^{(1)} + P_{20}^{(2)})/2$).

3.4. Kron_RLS. Kronecker product kernels are used widely in prediction issues of other studies and conditions [45–47]. In this study, we also use a Kronecker product kernel to construct a larger kernel for the drug-target pairs. Then the prediction of DTIs is based on the ranking of the pairs that include known drugs and targets and new entities or failed drugs and targets. The higher rank implies the higher possibility of existing interactions. Based on the kernel of drugs and targets, the Kronecker product kernel of drug-target pairs is constructed as follows [34]:

$$K\left(\left(d_i, t_j\right), \left(d_k, t_l\right)\right) = K_d\left(d_i, d_k\right) K_t\left(t_j, t_l\right), \tag{8}$$

where $K_d(d_i, d_k)$ is the (i,k)th element of the kernel of drugs with S_{final}, while $K_t(t_j, t_l)$ is the (j,l)th element kernel of targets with $K_{\mathrm{GIP},t}$.

According the Kronecker product kernel of formula (8), the predictions of DTIs for all drug-target pairs can be calculated as follows [34]:

$$\mathrm{vec}\left(\widehat{Y}^T\right) = K\left(K + \sigma I\right)^{-1} \mathrm{vec}\left(Y^T\right), \tag{9}$$

where σ is a regularization parameter. The smoother result can be obtained via the higher value σ. We get $\widehat{Y} = Y$ when $\sigma = 0$ which shows no generalization [34]. We also use the eigendecompositions of the kernel matrices according to Laarhoven's study. The eigendecompositions of matrices K_d and K_t are $K_d = \vee_d \wedge_d \vee_d^T$ and $K_t = \vee_t \wedge_t \vee_t^T$, in which $\vee_d$ and $\vee_t$ are the unitary matrices of feature vectors, and $\wedge_d$ and $\wedge_t$ are the diagonal matrices of eigenvalues for drugs and targets, respectively. Since the eigenvalues (vectors) of a Kronecker product are the Kronecker product of eigenvalues (vectors), the Kronecker product kernel of drug-target pairs can be formulated as follows [34]:

$$K = K_d \otimes K_t = \vee \wedge \vee^T, \tag{10}$$

in which

$$\vee = \vee_d \otimes \vee_t,$$
$$\wedge = \wedge_d \otimes \wedge_t. \tag{11}$$

3.5. KNN for New Chemical Entities. New chemical entities or failed drugs have no known associations with targets, which makes it impossible to predict more associations by existing methods. In this study, we used the KNN method to estimate the interaction scores for new chemical entities or failed drugs by the similarity between them and known drugs. For example, we denote a new chemical entity or failed drug as C_{new}, whose interaction score with target t_j can be computed by the formula

$$\mathrm{Score}\left(C_{\mathrm{new}}, t_j\right) = \frac{\sum S_{\mathrm{subsim}}^{(d_i,d_l)} y_{lj}}{\sum S_{\mathrm{subsim}}^{(d_i,d_l)}}, \quad l \in K_{\mathrm{new}}, \tag{12}$$

where $S_{\mathrm{subsim}}^{(d_i,d_l)}$ is the (i,l)th element of chemical substructure similarity matrix $S_{\mathrm{subsim}} \in R^{m*m}$, and y_{lj} is the (l,j)th element of $Y \in R^{m*n}$. K_{new} is the set of top K neighbors according to the S_{subsim} matrix. In this study, we set the value of K to be 4.

Table 3: The performance of 10-fold cross validation on 5 datasets.

		AUC					
Target	FP	DBSI-R	NWNBI	EWNBI*	NBI	SDTNBI*	SDTRLS
GPCRs	CDK	0.896 ± 0.003	0.981 ± 0.001	0.981 ± 0.001	0.980 ± 0.001	0.904 ± 0.003	0.982 ± 0.002
	CDKExt	0.895 ± 0.002	0.981 ± 0.001	0.981 ± 0.001	0.980 ± 0.001	0.901 ± 0.003	0.982 ± 0.002
	FP4	0.896 ± 0.002	0.981 ± 0.001	0.981 ± 0.001	0.980 ± 0.001	0.966 ± 0.002	0.979 ± 0.002
	Graph	0.897 ± 0.002	0.981 ± 0.001	0.981 ± 0.001	0.980 ± 0.001	0.917 ± 0.003	0.980 ± 0.001
	KR	0.909 ± 0.002	0.981 ± 0.001	0.981 ± 0.001	0.980 ± 0.001	0.960 ± 0.002	0.983 ± 0.002
	MACCS	0.881 ± 0.005	0.981 ± 0.001	0.981 ± 0.001	0.980 ± 0.001	0.931 ± 0.002	0.982 ± 0.001
	PubChem	0.895 ± 0.003	0.981 ± 0.001	0.981 ± 0.001	0.980 ± 0.001	0.918 ± 0.003	0.981 ± 0.001
Kinases	CDK	0.886 ± 0.004	0.977 ± 0.002	0.976 ± 0.002	0.976 ± 0.002	0.893 ± 0.004	0.972 ± 0.002
	CDKExt	0.885 ± 0.002	0.977 ± 0.002	0.976 ± 0.002	0.976 ± 0.002	0.892 ± 0.004	0.973 ± 0.002
	FP4	0.880 ± 0.004	0.977 ± 0.002	0.976 ± 0.002	0.976 ± 0.002	0.956 ± 0.003	0.969 ± 0.003
	Graph	0.882 ± 0.003	0.977 ± 0.002	0.976 ± 0.002	0.976 ± 0.002	0.905 ± 0.003	0.971 ± 0.003
	KR	0.901 ± 0.003	0.977 ± 0.002	0.976 ± 0.002	0.976 ± 0.002	0.958 ± 0.003	0.970 ± 0.003
	MACCS	0.880 ± 0.002	0.977 ± 0.002	0.976 ± 0.002	0.976 ± 0.002	0.925 ± 0.003	0.970 ± 0.003
	PubChem	0.877 ± 0.005	0.977 ± 0.002	0.976 ± 0.002	0.976 ± 0.002	0.903 ± 0.003	0.971 ± 0.002
ICs	CDK	0.923 ± 0.002	0.582 ± 0.007	Ø	0.573 ± 0.013	0.932 ± 0.004	0.956 ± 0.005
	CDKExt	0.922 ± 0.003	0.582 ± 0.007	Ø	0.573 ± 0.013	0.931 ± 0.004	0.955 ± 0.005
	FP4	0.916 ± 0.003	0.582 ± 0.007	Ø	0.573 ± 0.013	0.954 ± 0.003	0.943 ± 0.005
	Graph	0.920 ± 0.003	0.582 ± 0.007	Ø	0.573 ± 0.013	0.940 ± 0.003	0.948 ± 0.005
	KR	0.932 ± 0.004	0.582 ± 0.007	Ø	0.573 ± 0.013	0.971 ± 0.002	0.953 ± 0.005
	MACCS	0.919 ± 0.004	0.582 ± 0.007	Ø	0.573 ± 0.013	0.941 ± 0.003	0.950 ± 0.005
	PubChem	0.916 ± 0.003	0.582 ± 0.007	Ø	0.573 ± 0.013	0.937 ± 0.003	0.949 ± 0.005
NRs	CDK	0.860 ± 0.003	0.754 ± 0.013	Ø	0.775 ± 0.014	0.872 ± 0.006	0.916 ± 0.005
	CDKExt	0.859 ± 0.004	0.754 ± 0.013	Ø	0.775 ± 0.014	0.871 ± 0.006	0.916 ± 0.005
	FP4	0.859 ± 0.004	0.754 ± 0.013	Ø	0.775 ± 0.014	0.906 ± 0.005	0.911 ± 0.005
	Graph	0.857 ± 0.005	0.754 ± 0.013	Ø	0.775 ± 0.014	0.878 ± 0.006	0.905 ± 0.006
	KR	0.879 ± 0.008	0.754 ± 0.013	Ø	0.775 ± 0.014	0.932 ± 0.005	0.915 ± 0.005
	MACCS	0.857 ± 0.004	0.754 ± 0.013	Ø	0.775 ± 0.014	0.881 ± 0.006	0.912 ± 0.005
	PubChem	0.855 ± 0.004	0.754 ± 0.013	Ø	0.775 ± 0.014	0.876 ± 0.006	0.912 ± 0.005
Global	CDK	0.849 ± 0.005	0.916 ± 0.006	Ø	0.924 ± 0.005	0.909 ± 0.005	0.936 ± 0.004
	CDKExt	0.849 ± 0.005	0.916 ± 0.006	Ø	0.924 ± 0.005	0.907 ± 0.005	0.936 ± 0.004
	FP4	0.843 ± 0.004	0.916 ± 0.006	Ø	0.924 ± 0.005	0.949 ± 0.004	0.935 ± 0.004
	Graph	0.850 ± 0.005	0.916 ± 0.006	Ø	0.924 ± 0.005	0.921 ± 0.005	0.935 ± 0.004
	KR	0.862 ± 0.002	0.916 ± 0.006	Ø	0.924 ± 0.005	0.947 ± 0.004	0.936 ± 0.004
	MACCS	0.851 ± 0.007	0.916 ± 0.006	Ø	0.924 ± 0.005	0.937 ± 0.004	0.936 ± 0.004
	PubChem	0.848 ± 0.005	0.916 ± 0.006	Ø	0.924 ± 0.005	0.923 ± 0.005	0.936 ± 0.004

Ø represents the fact that we did not compute the prediction performance because of data reason; ∗ stands for the prediction results derived from previous studies.

4. Experiments and Results

4.1. Benchmark Evaluation and Evaluation Indices. In order to demonstrate the performance of our method, we adopt the 10-fold cross validation and external validation. The 10-fold validation was widely used in prediction of DTIs [29, 48, 49] and other interaction prediction in bioinformatics. The main experiment process is that the whole dataset is randomly divided into 10 groups; each group alternates as a testing set, and the rest of the 9 groups alternate as the training set, and this process is repeated 10 times.

Furthermore, the DTIs of new chemical entities and failed drugs are a very important portion in this study. We use two external datasets (ExGPCRs, ExKinases) to evaluate performance of our method by predicting all interactions with them.

We use the AUC (area under the ROC curve) as an evaluation metric for our SDTRLS as for SDTNBI methods, and the values in Tables 3, 5, and 6 are presented in the format

Table 4: The performance of two external validations.

		AUC					
Target	FP	DBSI-R	NWNBI	EWNBI*	NBI	SDTNBI	SDTRLS
ExGPCRs	CDK	0.752	0.756	0.764	0.769	0.753	0.824
	CDKExt	0.751	0.756	0.764	0.769	0.751	0.804
	FP4	0.758	0.756	0.764	0.769	0.784	0.818
	Graph	0.758	0.756	0.764	0.769	0.761	0.842
	KR	0.770	0.756	0.764	0.769	0.797	0.840
	MACCS	0.754	0.756	0.764	0.769	0.758	0.822
	PubChem	0.754	0.756	0.764	0.769	0.759	0.831
ExKinases	CDK	0.851	0.812	0.821	0.828	0.852	0.827
	CDKExt	0.850	0.812	0.821	0.828	0.852	0.834
	FP4	0.850	0.812	0.821	0.828	0.847	0.855
	Graph	0.851	0.812	0.821	0.828	0.852	0.841
	KR	0.851	0.812	0.821	0.828	0.863	0.848
	MACCS	0.851	0.812	0.821	0.828	0.852	0.844
	PubChem	0.850	0.812	0.821	0.828	0.852	0.846

∗ stands for the prediction results derived from previous studies.

of mean ± standard deviation. The larger the AUC value is, the better the prediction is.

4.2. Cross Validation. Table 3 describes the performance evaluation index values of the predicted datasets in the 10-fold cross validation for 5 datasets. SDTRLS's minimum AUC among the seven substructures reaches 0.979 and the average is 0.981, which indicates good prediction results. However, on NRs dataset, the validation results of each substructure are relatively poor and the minimum value is 0.905 based on Graph substructure while the maximum value is 0.916. On Kinases dataset, the verification results are also very stable, with the maximum and minimum values of 0.973 and 0.969, respectively. On ICs dataset, the verification results are not bad, the minimum value of AUC is 0.943 with FP4 substructure, and the maximum value is 0.956 with CDK substructure. Similarly, on Global dataset, the results are stable except for the slightly lower values of 7 substructures, between 0.935 and 0.936. In general, the validation results on GPCR and Kinase datasets are better than the other three datasets. Moreover, the prediction performances of EWNBI, NWNBI, and NBI on Kinases dataset are lightly better than SDTRLS, while SDTRLS has obvious advantage on ICs, NRs, and Global datasets. In addition, because the authors did not provide the data needed for EWNBI method on three datasets (ICs, NRs, and Global) and the prediction results of datasets GPCRs and Kinases are not good, we do not compute the AUC values of EWNBI method on these three datasets. Overall, SDTRLS and SDTNBI provide more stable prediction results on 5 datasets.

4.3. External Validation. Table 4 describes the evaluation results of six methods on two external datasets ExGPCRs and ExKinases; the basic datasets are GPCRs and Kinases, respectively. Overall, external validation results of all prediction methods are worse than 10-fold cross validation results because new chemical entities have no known DTIs. On ExGPCRs dataset, the AUC values of SDTRLS on the 7 substructures are between 0.804 and 0.842. On ExKinases dataset, the AUC values of SDTRLS of the 7 substructures are between 0.827 and 0.855. As can be seen from Table 4, the verification results of all approaches on ExKinases are better than on ExGPCRs. In the validation on ExKinases dataset, there are no obvious differences in AUC values among DBSI-R, SDTNBI, and SDTRLS. On ExKinases, SDTRLS demonstrates its excellent prediction power.

4.4. Comparison with Previous Methods. Since the datasets used in this study are derived from the datasets used in the SDTNBI method, as the state-of-the-art method, its prediction performances are more stable than the other 4 methods. In this study, the comparison is performed in terms of the t-test statistical analyses of SDTRLS and SDTNBI methods, as well as in terms of the parameter-independent AUC value with other 5 methods.

Table 5 shows t-tests results of SDTNBI and SDTRLS on five datasets GPCRs, Kinases, ICs, NRs, and Global, respectively. We can see from Table 5 that the average AUC of our method on each dataset is greater than that of the SDTNBI method, especially in the GPCRs and Kinases datasets, respectively, from 0.928 to 0.981 and from 0.919 to 0.971. Moreover, there were significant differences ($p < 0.05$) in the comparison results of GPCRs, Kinases, and NRs datasets; particularly, the comparison result is more significant ($p < 0.01$) on GPCRs and Kinases datasets. In conclusion, our method is more stable than the SDTNBI method in terms of the 10-fold cross validation.

We also compare the prediction results with other four methods on five datasets GPCRs, Kinases, ICs, NRs, and Global. The four competing methods are NBI, NWNBI, EWNBI, and DBSI-R [29]. NBI applied a mass diffusion-based method to obtain the predicted list by considering

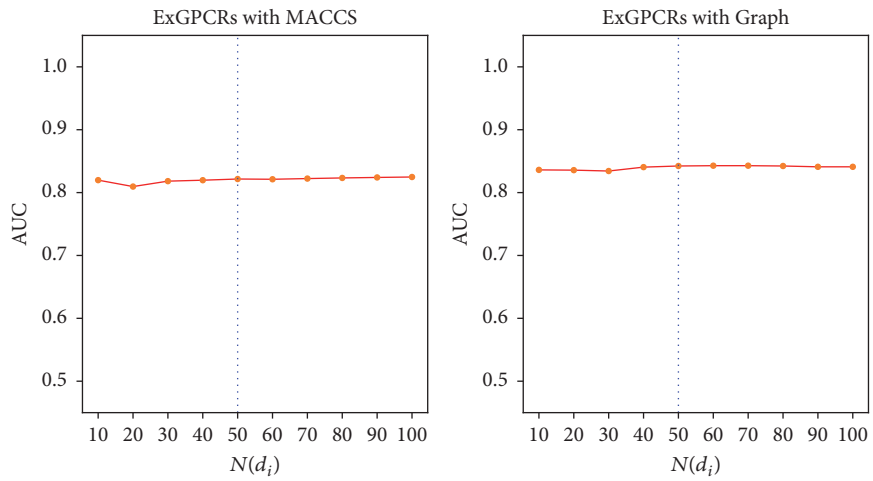


Figure 1: Robustness of SDTRLS with respect to the number of $N(d_i)$: the dotted line is the default value and its prediction performance.

Table 5: The t-tests results of 10-fold cross validations on 5 datasets.

	AUC				
Methods	GPCRs	Kinases	ICs	NRs	Global
SDTNBI	0.928 ± 0.026	0.919 ± 0.028	0.944 ± 0.014	0.888 ± 0.023	0.928 ± 0.017
SDTRLS	0.981 ± 0.001	0.971 ± 0.001	0.951 ± 0.005	0.912 ± 0.004	0.935 ± 0.000
p	0.0002	0.0004	0.248	0.016	0.232

Table 6: The t-tests results of two external validations.

	AUC	
Methods	ExGPCRs	ExKinases
SDTNBI	0.766 ± 0.017	0.853 ± 0.005
SDTRLS	0.825 ± 0.013	0.842 ± 0.009
p	$1.04e-05$	0.019

the bipartite graph. However, in EWNBI method, a DTI network was weighted by the potency of binding affinity or inhibitory activity of the interactions with drugs and targets. The theoretical basis of NWNBI method is that the hub node is more difficult to be influenced. The DBSI method is based on the hypothesis that two similar drugs may have similar targets. Table 3 shows that the SDTRLS method is slightly better than NBI, NWNBI, and EWNBI methods on GPCRs and much better than DBSI-R method. In addition, SDTRLS method is much better than DBSI-R method while being comparable with NBI, NWNBI, and EWNBI methods on Kinases. In general, the SDTRLS approach is comparable to these four methods from the results of the 10-fold cross validation on GPCRs and Kinases datasets.

Table 6 shows results of SDTNBI and SDTRLS on two datasets, ExGPCRs and ExKinases, respectively. From Table 6 we can see that our method greatly outperforms the SDTNBI method, on ExGPCRs in terms of the average AUC and t-test result ($p < 0.01$). In addition, the average AUC of our methods are slightly lower than those of SDTNBI method on ExKinases, which may be due to the sparsity of known DTIs in this dataset.

We compare the prediction result of our method with other four competing methods on the same datasets ExGPCRs and ExKinases. We can see from Table 4 that SDTRLS method outperforms the other four competing methods on ExGPCRs dataset. In addition, SDTRLS method is also comparable with other four competing methods on ExKinases dataset.

4.5. Parameter Analysis for $N(d_i)$ and K. In this section, we analyzed two parameters, including $N(d_i)$ for similarity network fusion and K for new chemical entities. The parameter h was set to be 1 according to previous study [27]. Moreover, GIP is widely used in other studies [10, 34, 37, 50, 51]; we also set the values of both α and β to be 1. All results were validated over external validation of ExGPCRs datasets based on substructures MACCS and Graph. Figure 1 describes that the sensitivity of the prediction performance of SDTRLS with to different numbers $N(d_i)$ of similarity network fusion. SDTRLS had stable prediction performance over a wide range from 10 to 100. The impact of parameter K for new chemical entities on the prediction performance of SDTRLS, in terms of AUC value, is illustrated in Figure 2. SDTRLS was robust to different values of parameter K.

4.6. Case Studies. In order to further confirm the prediction ability of our method, we conduct an experimental analysis

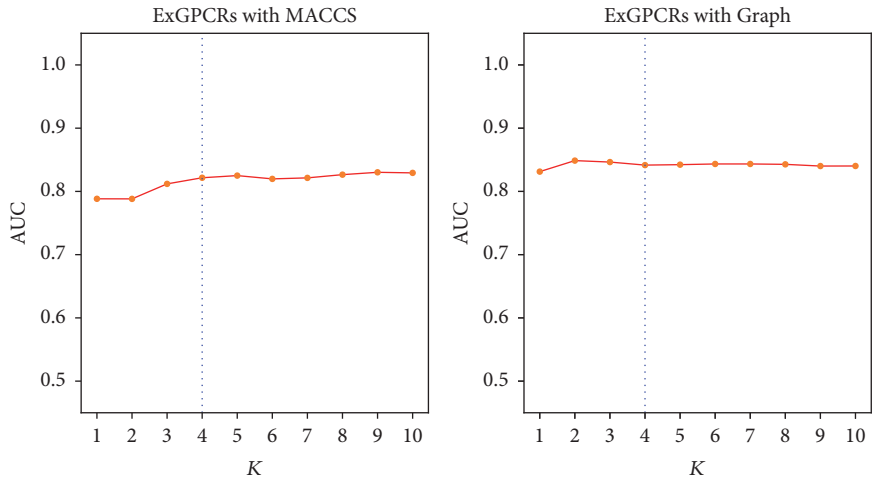


Figure 2: Robustness of SDTRLS with respect to the number of K: the dotted line is the default value and its prediction performance.

on dataset ExGPCRs, and its known DTIs are not used as a priori knowledge when conducting external validation. The selected predictions of drugs are confirmed with DrugBank, ChEMBL, and KEGG databases. Table 7 describes the confirmed result based on ExGPCRs dataset. We select the top five predicted interactions of 5 drugs; the top one predicted interaction of every drug is confirmed by searching databases. Furthermore, 76% of all predicted DTIs (19 out of 25) are also confirmed with three databases; 32% of predicted DTIs (8 out of 25) are simultaneously confirmed with two databases, especially in the predicted result of (DB00209, DB00283, and DB00334); they are all confirmed with the several databases. In addition, we further validate the results marked as unknown in the prediction results; we searched the relevant literature and found the related description. For example, thiethylperazine (DB00372) is an antagonist of human Dopamine D3 (hDRD3_170) according to the description in Petsko and Ringe [52] which shows that the prediction result is meaningful, and other remaining unknown DTIs deserve being validated in the future. In general, it proves that our method is effective in practical applications.

5. Conclusions

The systematic understanding of the interactions between chemical compounds and target proteins is very important for new drug design and development. In the past decades, in order to solve the time-consuming shortcomings of traditional biochemical methods, many computational approaches have been developed to predict DTIs, like machine learning, network inference, and so on. However, these methods mainly focused on new DTIs for known drugs and paid less attention to new chemical entities for DTIs. In addition, their prediction performances are not good enough.

In this study, we have constructed the similarity kernel of approved drugs, failed drugs, and new chemical entities

Table 7: The new confirmation of drug-target interactions based on Graph substructure in the ExGPCRs.

Drug ID	Target ID	Rank	Source
DB00209	hCHRM2_86	1	KEGG
	hCHRM3_98	2	KEGG
	hCHRM1_92	3	DrugBank KEGG
	hCHRM4_87	4	KEGG
	hCHRM5_90	5	KEGG
DB00283	hDRD3_170	1	ChEMBL
	hDRD4_106	2	ChEMBL
	HDRD2_94	3	ChEMBL
	hOPRM1_166	4	ChEMBL
	hOPRK1_173	5	ChEMBL
DB00334	hCHRM2_86	1	DrugBank ChEMBL
	hCHRM3_98	2	DrugBank ChEMBL
	hCHRM1_92	3	DrugBank ChEMBL
	hA1AB_164	4	ChEMBL KEGG
	hA1AD_116	5	ChEMBL KEGG
DB00372	hDRD2_94	1	DrugBank KEGG
	h5HT2A_125	2	Unknown
	h5HT2C_126	3	Unknown
	hDRD3_170	4	Unknown
	h5HT1A_89	5	Unknown
DB00612	hB1AR_88	1	DrugBank KEGG
	hB2AR_84	2	DrugBank
	hB3AR_93	3	Unknown
	hDRD2_94	4	Unknown
	hOPRM1_166	5	Unknown

by weighting the chemical substructures. Then, GIP kernels were calculated from drugs and targets according to the known DTIs. For the new chemical entities or failed drugs, we used the KNN to initialize the DTIs before calculating the GIP kernel. To construct a comprehensive similarity kernel for drugs, SNF method is used to fuse GIP kernel and substructure similarity kernel. Finally, the score of drug-target pairs was predicted by Kron_RLS. We compared the prediction performance with other competing methods via the tenfold cross validation and external validation.

However, there are still some limitations in this study. First, since the target set is specified within the current datasets, it may be not possible to predict the DTIs of the target beyond the datasets. Other similarity information of targets such as the sequence and functional network [53–56] is not used when the similarity kernel of targets is constructed. In addition, the 3D structure of drugs may also need to be considered as important information. It is expected that additional information may improve prediction performance. In the future, more information using other methods such as ClusterViz [57] should be integrated to develop a more efficient prediction method. Nevertheless, this study provides an important basis for new drug development and drug repositioning and also plays an important role in the personalized medical development.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work has been supported in part by the National Natural Science Foundation of China under Grants no. 61772552, no. 61420106009, and no. 61622213.

References

[1] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 673–683, 2004.

[2] N. Novac, "Challenges and opportunities of drug repositioning," *Trends in Pharmacological Sciences*, vol. 34, no. 5, pp. 267–272, 2013.

[3] M. R. Hurle, L. Yang, Q. Xie, D. K. Rajpal, P. Sanseau, and P. Agarwal, "Computational drug repositioning: from data to therapeutics," *Clinical Pharmacology & Therapeutics*, vol. 93, no. 4, pp. 335–341, 2013.

[4] A. L. Hopkins, "Drug discovery: predicting promiscuity," *Nature*, vol. 462, no. 7270, pp. 167-168, 2009.

[5] S. J. Swamidass, "Mining small-molecule screens to repurpose drugs," *Briefings in Bioinformatics*, vol. 12, no. 4, pp. 327–335, 2011.

[6] B. Y. Feng, A. Simeonov, A. Jadhav et al., "A high-throughput screen for aggregation-based inhibition in a large compound library," *Journal of Medicinal Chemistry*, vol. 50, no. 10, pp. 2385–2390, 2007.

[7] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drug-target network," *Nature Biotechnology*, vol. 25, no. 10, pp. 1119–1126, 2007.

[8] S. Zhao and S. Li, "Network-based relating pharmacological and genomic spaces for drug target identification," *PLoS ONE*, vol. 5, no. 7, Article ID e11764, 2010.

[9] S. Alaimo, A. Pulvirenti, R. Giugno, and A. Ferro, "Drug-target interaction prediction through domain-tuned network-based inference," *Bioinformatics*, vol. 29, no. 16, pp. 2004–2008, 2013.

[10] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, and J. Zheng, "Drug-target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2013.

[11] T. van Laarhoven and E. Marchiori, "Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile," *PLoS ONE*, vol. 8, no. 6, Article ID e66952, 2013.

[12] Q. Yuan, J. Gao, D. Wu, S. Zhang, H. Mamitsuka, and S. Zhu, "DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank," *Bioinformatics*, vol. 32, no. 12, pp. i18–i27, 2016.

[13] H. Luo, J. Wang, M. Li et al., "Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm," *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, 2016.

[14] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. Elsevier Science & Technology, 2008.

[15] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, pp. D901–D906, 2008.

[16] A. Gaulton, L. J. Bellis, A. P. Bento et al., "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, no. 1, pp. D1100–D1107, 2012.

[17] C. Qin, C. Zhang, F. Zhu et al., "Therapeutic target database update 2014: a resource for targeted therapeutics," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1118–D1123, 2014.

[18] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[19] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular Systems Biology*, vol. 6, p. 343, 2010.

[20] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild et al., "STITCH 4: integration of protein-chemical interactions with user data," *Nucleic Acids Research*, vol. 42, no. 1, pp. D401–D407, 2014.

[21] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. 1, pp. D808–D815, 2013.

[22] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D198–D201, 2007.

[23] J. T. Dudley, T. Deshpande, and A. J. Butte, "Exploiting drug-disease relationships for computational drug repositioning," *Briefings in Bioinformatics*, vol. 12, no. 4, pp. 303–311, 2011.

[24] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, "Similarity-basedmachine learning methods for predicting drug-target interactions: A brief review," *Briefings in Bioinformatics*, vol. 15, no. 5, pp. 734–747, 2013.

[25] Y. Tabei and Y. Yamanishi, "Scalable prediction of compound-protein interactions using minwise hashing.," *BMC systems biology*, vol. 7, p. S3, 2013.

[26] H. Yabuuchi, S. Niijima, H. Takematsu et al., "Analysis of multiple compound-protein interactions reveals novel bioactive molecules," *Molecular Systems Biology*, vol. 7, article no. 472, 2011.

[27] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, Article ID btq176, pp. i246–i254, 2010.

[28] X. Chen, C. C. Yan, X. Zhang et al., "Drug-target interaction prediction: Databases, web servers and computational models," *Briefings in Bioinformatics*, vol. 17, no. 4, pp. 696–712, 2016.

[29] F. Cheng, C. Liu, J. Jiang et al., "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Computational Biology*, vol. 8, no. 5, Article ID e1002503, 2012.

[30] X. Chen, M.-X. Liu, and G.-Y. Yan, "Drug-target interaction prediction by random walk on the heterogeneous network," *Molecular BioSystems*, vol. 8, no. 7, pp. 1970–1978, 2012.

[31] K. Bleakley, G. Biau, and J.-P. Vert, "Supervised reconstruction of biological networks with local models," *Bioinformatics*, vol. 23, no. 13, pp. i57–i65, 2007.

[32] F. Mordelet and J.-P. Vert, "SIRENE: Supervised inference of regulatory networks," *Bioinformatics*, vol. 24, no. 16, pp. i76–i82, 2008.

[33] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug-target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.

[34] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[35] W. Lan, J. Wang, M. Li et al., "Predicting drug–target interaction using positive-unlabeled learning," *Neurocomputing*, vol. 206, pp. 50–57, 2016.

[36] W. Lan, J. Wang, M. Li, J. Liu, F. X. Wu, and Y. Pan, "Predicting microrna-disease associations based on improved microrna and disease similarities," *IEEEACM Transactions on Computational Biology & Bioinformatics*, vol. PP, no. 99, p. 1, 2016.

[37] C. Yan, J. Wang, P. Ni, W. Lan, F. X. Wu, and Y. Pan, "Dnrlmf-mda:predicting microrna-disease associations based on similarities of micrornas and diseases," *IEEEACM Transactions on Computational Biology & Bioinformatics*, 2017.

[38] M. Gönen, "Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.

[39] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1025–1033, Chicago, Ill, USA, August 2013.

[40] A. Ezzat, P. Zhao, M. Wu, and X. Li, "Drug-target interaction prediction with graph regularized matrix factorization," *IEEEACM Transactions on Computational Biology & Bioinformatics*, vol. 1, p. 1, 2016.

[41] M. Hay, D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal, "Clinical development success rates for investigational drugs," *Nature Biotechnology*, vol. 32, no. 1, pp. 40–51, 2014.

[42] A. Mullard, "Drug repurposing programmes get lift off," *Nature Reviews Drug Discovery*, vol. 11, no. 7, pp. 505-506, 2012.

[43] Z. Wu, F. Cheng, J. Li, W. Li, G. Liu, and Y. Tang, "SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning," *Briefings in Bioinformatics*, pp. 333–347, 2016.

[44] B. Wang, A. M. Mezlini, F. Demir et al., "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014.

[45] J. Basilico and T. Hofmann, "Unifying collaborative and content-based filtering," in *Proceedings of the Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pp. 65–72, July 2004.

[46] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, no. 1, pp. i38–i46, 2005.

[47] M. Hue and J.-P. Vert, "On learning with kernels for unordered pairs," in *Proceedings of the 27th International Conference on Machine Learning, ICML 2010*, pp. 463–470, June 2010.

[48] Z. Xia, L. Wu, X. Zhou, and S. T. Wong, "Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces," *BMC Systems Biology*, vol. 4, no. Suppl 2, p. S6, 2010.

[49] C. Huang, R. Zhang, Z. Chen et al., "Predict potential drug targets from the ion channel proteins based on SVM," *Journal of Theoretical Biology*, vol. 262, no. 4, pp. 750–756, 2010.

[50] Z.-H. You, Z.-A. Huang, Z. Zhu et al., "Pbmda: A novel and effective path-based computational model for mirna-disease association prediction," *PLoS Computational Biology*, vol. 13, no. 3, 2017.

[51] W. Lan, M. Li, K. Zhao et al., "LDAP: a web server for lncRNA-disease association prediction," *Bioinformatics*, vol. 33, no. 3, pp. 458–460, 2016.

[52] G. Petsko and D. Ringe, "ligand-based virtual screening for new antagonists of dopamine receptor D2/D3," *Shanghai Management Science*, 2013.

[53] X. Peng, J. Wang, W. Peng, F. Wu, and Y. Pan, "Protein–protein interactions: detection, reliability assessment and applications," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 798–819, 2016.

[54] B. Zhao, J. Wang, and F. Wu, "Computational Methods to Predict Protein Functions from Protein-Protein Interaction Networks," *Current Protein & Peptide Science*, vol. 18, no. 11, 2017.

[55] B. Zhao, J. Wang, X. Li, and F.-X. Wu, "Essential protein discovery based on a combination of modularity and conservatism," *Methods*, vol. 110, pp. 54–63, 2016.

[56] B. Zhao, J. Wang, M. Li et al., "A New Method for Predicting Protein Functions from Dynamic Weighted Interactome Networks," *IEEE Transactions on NanoBioscience*, vol. 15, no. 2, pp. 133–141, 2016.

[57] J. Wang, J. Zhong, G. Chen, M. Li, F.-X. Wu, and Y. Pan, "ClusterViz: A Cytoscape APP for Cluster Analysis of Biological Network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 815–822, 2015.

Hindawi
Complexity
Volume 2017, Article ID 8362741, 27 pages
https://doi.org/10.1155/2017/8362741

Review Article

Complex Brain Network Analysis and Its Applications to Brain Disorders: A Survey

Jin Liu,[1] Min Li,[1] Yi Pan,[2] Wei Lan,[1] Ruiqing Zheng,[1] Fang-Xiang Wu,[3] and Jianxin Wang[1]

[1]*School of Information Science and Engineering, Central South University, Changsha 410083, China*
[2]*Department of Computer Science, Georgia State University, Atlanta, GA 30302, USA*
[3]*Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N5A9*

Correspondence should be addressed to Jianxin Wang; jxwang@mail.csu.edu.cn

Received 26 April 2017; Revised 18 September 2017; Accepted 27 September 2017; Published 22 October 2017

Academic Editor: Manlio De Domenico

It is well known that most brain disorders are complex diseases, such as Alzheimer's disease (AD) and schizophrenia (SCZ). In general, brain regions and their interactions can be modeled as complex brain network, which describe highly efficient information transmission in a brain. Therefore, complex brain network analysis plays an important role in the study of complex brain diseases. With the development of noninvasive neuroimaging and electrophysiological techniques, experimental data can be produced for constructing complex brain networks. In recent years, researchers have found that brain networks constructed by using neuroimaging data and electrophysiological data have many important topological properties, such as small-world property, modularity, and rich club. More importantly, many brain disorders have been found to be associated with the abnormal topological structures of brain networks. These findings provide not only a new perspective to explore the pathological mechanisms of brain disorders, but also guidance for early diagnosis and treatment of brain disorders. The purpose of this survey is to provide a comprehensive overview for complex brain network analysis and its applications to brain disorders.

1. Introduction

The consensus of the neuroscience community is that a human brain contains about 100 billion (10^{11}) neurons connected by about 100 trillion (10^{14}) synapses [1, 2], which are anatomically organized over multiple space scales and functionally interact over multiple time scales. Therefore, exploring the brain and revealing the neural mechanism of brain activities have been a challenging scientific problem [3–5]. Now it is realized that brain functions are determined not only by a single neuron or a single brain region independently, but also by clusters of neurons, neural circuits within a function block, or a group of interactions between brain regions [6]. A brain can be modeled as a complex network [7–9], which enables highly efficient information transmission. Currently, network neuroscience has become a research hotspot [10–12].

Clinical disorders of human brain networks, such as Alzheimer's disease (AD), schizophrenia (SCZ), and Parkinson's disease (PD), are among the most disabling and therapeutically intractable health problems. Therefore, it is unsurprising that understanding brain network connectivity has long been a central goal of neuroscience and has recently catalyzed an unprecedented era of large-scale initiatives and collaborative projects, such as BRAIN Initiative (http://www.braininitiative.org/) (USA, 2013), Human Brain Project (https://www.humanbrainproject.eu/) (Europe, 2013), Brain/MINDS Project (http://brainminds.jp/) (Japan, 2014), Australian Brain Alliance (http://www.cibf.edu.au/australian-brain-alliance) (Australia, 2016), and China Brain Project (China, 2016) [13]. The goal of these projects is to revolutionize our understanding of the human brain. By accelerating the development and application of innovative technologies, revolutionary new accurate images of the brain can be produced for more accurate understanding of brain functions.

Neuroimaging techniques provide a way for clinicians and researchers to examine the structural and functional changes in the brain disorders *in vivo* [14–26]. Commonly

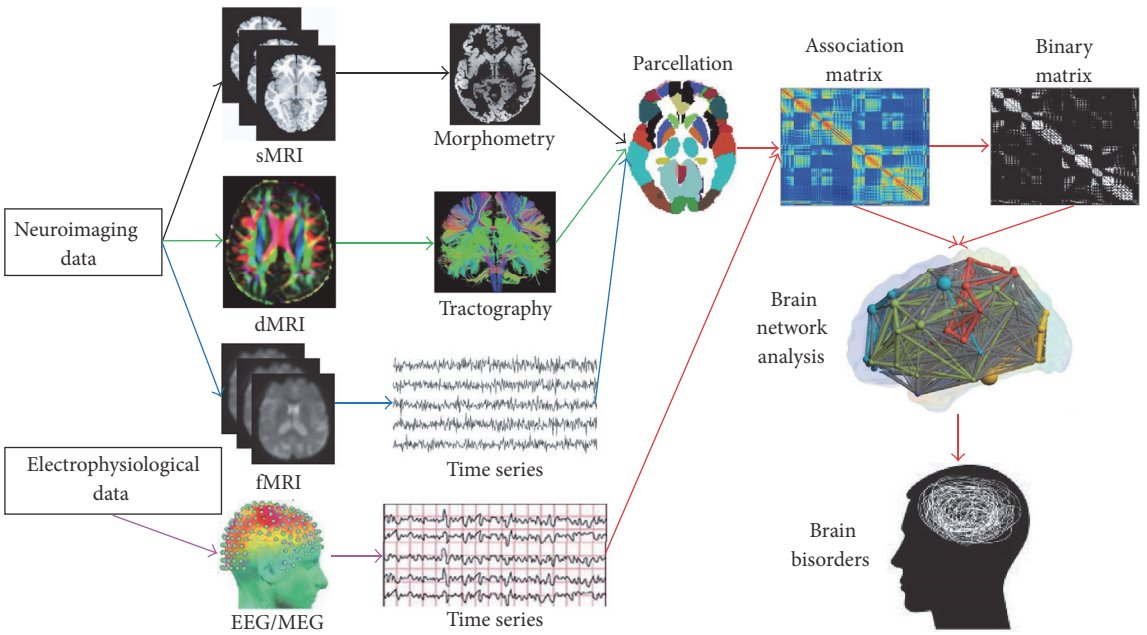


Figure 1: A general perspective for complex brain network analysis methods and its applications in brain disorders.

used modalities include structural magnetic resonance imaging (sMRI, such as T1w MRI) [27], diffusion magnetic resonance imaging (dMRI, such as diffusion tensor imaging (DTI)) [28], and functional magnetic resonance imaging (fMRI, such as rs-fMRI) [29, 30]. Electroencephalography (EEG) [31] and magnetoencephalography (MEG) [32] are noninvasive electrophysiological techniques for recording brain activities. EEG is used to measure voltage sensed by an array of electrodes placed on the scalp. MEG is used to measure the magnetic field outside the head using an array of very sensitive magnetic field detectors (magnetometers). The signals recorded by EEG and MEG directly reflect current flows generated by neurons within a brain. EEG/MEG also have been utilized for the studies of brain disorders [33–36]. Network-based analysis has been widely used in various fields, such as medical image analysis [37–40] and bioinformatics [41–44]. The two techniques (i.e., neuroimaging and magnetoencephalography) have been used to construct brain networks with multiple different scales, which have led to the development of brain network studies [45–48]. The construction of brain networks provides a necessary basis for brain network analysis, which includes global efficiency, local efficiency, modularity, and rich club [49, 50]. A graph can effectively and visually present a brain as a complex network whose topological structures can be quantified [51–53]. Therefore, graph theory has become one of the most important mathematical tools in the field of brain network analysis [54–56].

In this article, we provide a comprehensive review regarding complex brain network analysis and its applications to brain disorders as shown in Figure 1. Firstly, we introduce some basic concepts for constructing the brain networks based on neuroimaging and electrophysiological data, which include structural data and functional data. It is worth mentioning that, in this article, structural data only consider sMRI and dMRI and functional data only consider fMRI and EEG/MEG. After summarizing methods for brain network analysis based on graph theory, we present several applications of brain network analysis in brain disorders. Finally some conclusions are drawn and the directions of future work are pointed out along with brain network analysis.

2. Brain Network Construction

A brain network is typically represented by a graph $G = (V, E)$, where V is the set of vertices (or nodes) and E is the set of edges (or links, also called connections) between pairs of nodes. As nodes and edges are the basic elements of each brain network, the accurate definition of the two elements plays important roles in the brain network analysis [57].

2.1. Nodes. In order to construct a brain network, the first step is to define nodes of the brain network. The nodes should represent different, functionally uniform neurons (which are grouped together to perform the same function) or brain regions. However, since there is no gold standard for brain parcellation, methods for defining nodes of brain networks are varied as follows:

(i) The simplest method is to treat each measurement point as a separate node. This method occurs before the data acquisition. For example, different nodes could correspond to separate voxels in MRI images, different sensors in MEG, or different electrodes in EEG. The advantage of this method is that no additional data processing or assumptions are required to analyze the data at the original resolution or to perform further averaging or aggregating. The weaknesses of this method include the following: (1)

Table 1: Some common atlas in brain network study.

Name	Number of regions	Links	Reference
Brodmann area	104	http://www.fmriconsulting.com/brodmann/Interact.html	[58]
Anatomical Automatic Labeling atlas	116	http://www.cyceron.fr/index.php/fr/plateforme/freeware	[59]
Destrieux atlas	148	https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation	[90, 91]
Desikan-Killiany atlas	68	https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation	[92]
LPBA40	56	http://neuro.imm.dtu.dk/wiki/LPBA40	[93]
Brainnetome atlas	246	http://atlas.brainnetome.org/	[94]
HCP MMP1.0	360	https://balsa.wustl.edu/study/show/RVVG	[69]

there is no guarantee that the measurement points are consistent with the boundaries of functional human cell populations; (2) the boundaries of a specific, functionally specialized human cell population may go beyond the boundaries of a voxel. Thus, this method is often used in EEG/MEG but is rarely used in the other three types of data (i.e., sMRI, dMRI, and fMRI).

(ii) The most common method is to register experimental data to an a priori anatomical parcellation atlas, such as Brodmann area [58] and Anatomical Automatic Labeling (AAL) atlas [59]. For more atlases, please see Table 1. The advantage of this method is that it can easily parcellate a whole brain into many regions (about 10^2) as nodes of a brain network. The weakness of this method is that the resulting regions can show considerable variation in size, which affects any subsequent brain network analysis.

(iii) Based on the problem of the size of the regions, the alternative method is to treat each voxel as a separate node. The only difference from the first method is that this method occurs after the data acquisition. The advantage of this method is that it can construct a very large, high-resolution brain network (more than 10^4 nodes) for each brain. The weaknesses of this method include the following: (1) it could yield noise and thus affect the subsequent brain network analysis; (2) as the resulting brain network is large, it can cause difficulties in brain network analysis, such as looking for modularity.

(iv) The fourth method is to define nodes according to some a priori criteria. For example, some researchers have mapped activation patterns in a specific task and defined activation regions as the nodes of brain networks of interest according to these mappings [60]. Meanwhile, some researchers have applied meta-analysis methods to identify important brain regions as the nodes of brain networks of interest [61]. The advantage of this method is that the determination of the nodes is based on the measurement of brain functions, which can be adjusted according to specific hypotheses about brain networks of interest. The weakness of this method is that the resulting nodes may not be used in different modalities. For example, in fMRI, the nodes can be defined by a specific task and are usually included within the gray matter. However, in dMRI, the resulting nodes may make it difficult to track connections, since most fiber tracking methods are difficult to reconstruct the pathways of the axons within gray matter [62].

(v) The fifth method uses connectivity to define nodes. The essence of this method is to measure the connectivity of each voxel to all other voxels, and then some voxels are clustered together as brain regions with a specific function if these voxels have a similar connectivity [63]. For example, Anwander et al. [64] used an automatic clustering method to identify cortical regions with internally coherent connectivity in DTI and parcellate Broca's area to three subregions, which include BA44, BA45, and the deep frontal operculum. The advantage of this methods is that it can find brain regions with specific functions as the nodes of brain networks of interest. The weakness of this method is that since spatially separating brain regions may have similar connectivity, there is no guarantee that the resulting nodes are composed of a number of voxels that are spatially continuous [65, 66]. To address the weakness of this method, some methods with spatial constraints have been proposed [67, 68].

(vi) The sixth method is to define nodes by combining pieces of multimodal information, such as anatomical homogeneity [69] and synchrony [65]. The advantage of this method is that it can obtain complementary information from the multimodal data so that the location of the nodes is more accurate. The weaknesses of this method include the following: (1) it may have more noise than other methods with single modality data and thus affect the location of the nodes; (2) since it uses multimodal information, the cost of computation is also very large.

Based on the discussion about the above five methods, the definition of nodes is still a very challenging problem in order to obtain accurate results from brain network analysis.

2.2. Edges. The edges of a brain network represent the connectivity between two brain regions. Brain network connectivity can be divided into three types: structural connectivity, functional connectivity, and effective connectivity [39, 51, 70]. Structural connectivity contains two types: (1) the anatomical

connections between neural elements, such as fiber bundles; (2) the interregional covariation of specific morphometric parameters, such as gray matter thickness. Functional connectivity refers to a statistical dependence between neural elements with physiological recordings or neurophysiological signals. The purpose of effective connectivity is to uncover the direct, causal influences that neural elements exert over each other's activity. Since there are relatively few studies of effective connectivity in brain networks, in this article we mainly consider the studies of structural connectivity and functional connectivity.

2.2.1. Edges Based on Structural Connectivity. In brain network study, there are two common types of neuroimaging techniques that are often used to define structural connectivity. The two types of neuroimaging techniques are sMRI and dMRI as shown in Figure 1. How to quantify structural connectivity based on sMRI and dMRI is described as follows:

(i) In sMRI, structural connectivity is indirectly estimated by calculating interregional correlation of specific morphometric parameters, such as gray matter volume or cortical thickness [71–73]. In this method, a measure (such as gray matter volume, cortical thickness, or other similar metrics) of each brain region is extracted, and then correlations between two brain regions are calculated as the edges of the brain network of interest. Morphological measurements of brain regions can be implemented by some open source tools such as FreeSurfer (http://www.freesurfer.net/) and SPM (http://www.fil.ion.ucl.ac.uk/spm/). Pearson correlation and partial correlation are two of the most common methods to compute structure connectivity from sMRI images.

(ii) The most common technique for studying structural connectivity is dMRI [74–77]. In dMRI, the trajectories (connectivity) of axonal fibers can be reconstructed by using tractography, which includes deterministic tractography [78, 79] and probabilistic tractography [80].

(a) The deterministic tractography is simple and effective, and the satisfactory reconstruction can be obtained in some brain disorders. However, the weakness of this tractography is that an initial seed (i.e., a specific voxel) does not change during the reconstruction process. If the seed changes, the reconstruction is likely to produce a deviation. This tractography can be implemented by some common open source tools such as Diffusion Toolkit and Trackvis (http://trackvis.org/), DTIStudio (https://www.dtistudio.org/), and medInria (http://med.inria.fr/).

(b) The advantage of the probabilistic tractography is that the reconstruction results are more stable to noise, and fiber cross problem can be improved to some extent. The weakness of this method is computationally intensive and time consuming. This tractography can be also implemented by some common open source tools such as FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki), MRItrix3 (http://www.mrtrix.org/), and DSIStudio (http://dsi-studio.labsolver.org/).

For more details about the above two tractographies, please see [81, 82]. After this step, structural connectivity can be estimated by several different measures of connectivity strength. The simplest measure is that the number of axonal fibers (FN) connecting two brain regions is used as connectivity strength, which is the weight of the edge of the brain network of interest. Another common measure of connectivity strength is the average fractional anisotropy (FA) value of all voxels over the reconstructed tract between two brain regions.

Based on the above analysis, in both sMRI and dMRI, the resulting edges are undirected and weighted. Thus, the resulting connectivity matrix of each brain is symmetric and generates a weighted undirected network. In some studies, the weighted undirected resulting networks were converted into binary undirected networks by a specific threshold [72, 83, 84].

2.2.2. Edges Based on Functional Connectivity. In brain network study, functional connectivity is often defined by using fMRI, EEG, and MEG as shown in Figure 1. As can be seen from Figure 1, neurophysiological signals (such as time series) can be extracted from fMRI and EEG/MEG. EEG/MEG offer high temporal resolution, which allows brain activity to be sampled on the millisecond ranges that match with the speed of neural signals. fMRI generally offers a higher spatial resolution than EEG/MEG, but its temporal resolution is relatively low.

Functional connectivity reflects the statistical correlation between neurophysiological signals (such as time series) recorded from each brain region. The correlation can be measured by using various methods. The main measurement methods are divided into two categories: linear methods and nonlinear methods. The common linear methods [85, 86] mainly include Pearson correlation, partial correlation, and partial coherence. The common nonlinear methods mainly include synchronization likelihood [21, 34], mutual information [87], and wavelet correlation [88, 89]. For example, the Pearson correlation between brain regional activity time series is calculated as the edges of the brain network of interest, which are weighted and undirected. Thus, the resulting brain networks are also weighted undirected networks. Similarly, such weighted undirected resulting networks can also be converted into binary undirected networks by a specific threshold [33, 85, 87].

3. Brain Network Analysis

In general, the topology of networks can be divided into four types: binary undirected, binary directed, weighted

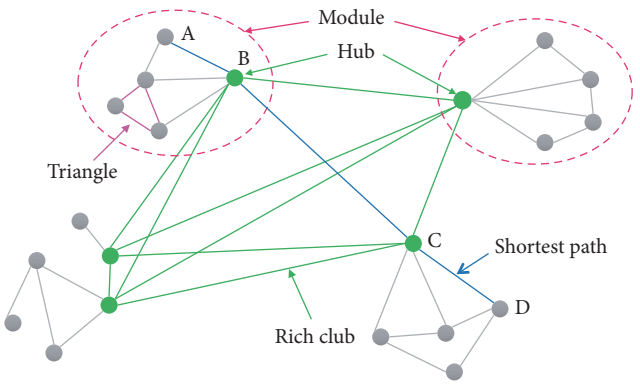


Figure 2: An example of a simple binary undirected network (graph).

undirected, and weighted directed. In this article, since we only focus on structural connectivity and functional connectivity, only two types of brain networks are taken into consideration: weighted undirected and binary undirected.

3.1. Degree, Clustering Coefficient, and Shortest Path Length. Node degree is one of the most elementary and important measures for a brain network and is often denoted as k. The degree of a node is the number of edges connecting the node with all other nodes. In general, the greater the degree of a node is, the more the nodes it connected to are and the more important it may be in the brain network. In a binary undirected network, the degree k_i of a node i is defined as

$$k_i = \sum_{j \neq i} a_{ij}, \tag{1}$$

where $a_{ij} = 1$ if the connection of node i and node j exists; otherwise $a_{ij} = 0$. For example, the degree of node C in Figure 2 is 7. The strength of a network is the average of the degree across all of the nodes in the network. Thus, for a binary undirected brain network, the network strength S can be calculated by

$$S = \frac{1}{N} \sum_{i \in N} k_i. \tag{2}$$

Degree distribution $P(k)$ is also a basic topological characterization and is defined as the fraction of nodes with degree k in the whole brain network in practical application. For example, if there are totally N nodes in a brain network where there are N_k nodes with degree of k,

$$P(k) = \frac{N_k}{N}. \tag{3}$$

For many brain networks, the degree distribution is characterized by a fat tail that indicates the presence of central position nodes. These central position nodes usually play a vital role in the convergence and divergence of information in the brain network [138]. In the field of brain networks, if a node occupies a central position in the overall organization of a brain network, the node can be called hub node [139]. For example, the green nodes (such as nodes B and C) in Figure 2 are considered as hubs in the simple binary undirected network.

A subgraph with 3 nodes and 3 edges is called a triangle as shown in Figure 2 (pink). In a network, the number of triangles t_i around a node i is defined as

$$t_i = \frac{1}{2} \sum_{i \leftrightarrow j \leftrightarrow h} a_{ij} a_{ih} a_{jh}. \tag{4}$$

The local clustering coefficient of a node measures the possibility that any two neighbors of the node are also connected. In this article, for convenience, the local clustering coefficient is described by clustering coefficient. In a binary undirected network, the clustering coefficient $C(i)$ of a node i is equal to the ratio of the number of the actual connected edges between its adjacent nodes to the number of all possible connection edges; that is,

$$C(i) = \frac{2t_i}{k_i (k_i - 1)}. \tag{5}$$

The average clustering coefficient of all nodes in a network is defined as the clustering coefficient of the network:

$$C = \frac{1}{N} \sum_i C(i) = \frac{1}{N} \sum_i \frac{2t_i}{k_i (k_i - 1)}. \tag{6}$$

The clustering coefficient is a measure of functional segregation, which is the ability for specialized processing to occur within densely interconnected groups of brain regions [140].

The shortest path plays an important role in the information transmission of a brain network, and it is a very important measure to describe the internal structure of the brain network. The shortest path can transmit the information more quickly and reduce brain consumption. In a binary undirected network, a path between nodes i and j with the minimum number of edges is called the shortest path between these two nodes and its length l_{ij} is denoted as

$$l_{ij} = \sum_{a_{st} \in l_{i \leftrightarrow j}} a_{st}, \tag{7}$$

where $l_{i \to j}$ is the shortest path between nodes i and j. For example, the shortest path length of nodes A and D in Figure 2 is 3 (i.e., $l_{AD} = 3$). The average shortest path length between node i and other all nodes is denoted as l_i:

$$l_i = \frac{1}{(N-1)} \sum_{i \neq j} l_{ij}. \tag{8}$$

The characteristic path length L of a network is the average shortest path length between all possible pairs of nodes in the network and is defined as

$$L = \frac{1}{N} \sum_i l_i. \tag{9}$$

The characteristic path length is a measure of functional integration, which is the ability to rapidly combine pieces of specialized information from distributed brain regions [140].

3.2. Centrality. The centrality is to measure the importance of nodes in a brain network. The higher the centrality of a node is, the more effective the node is in the information transmission of the brain network. In the brain network analysis, three measures of centrality are often used, including degree centrality, closeness centrality, and betweenness centrality as follows:

(i) Degree centrality is the most common measure of centrality, which uses the degree of a node to describe the importance of the node in the brain network. In brain network analysis, the degree centrality of a brain region measures the direct impact of the brain region on other adjacent brain regions. Thus, in a binary undirected network, the degree centrality $C_d(i)$ of a node i is equivalent to the degree of the node:

$$C_d(i) = k_i = \sum_{j \neq i} a_{ij}. \tag{10}$$

(ii) Closeness centrality [141] reflects the closeness between a node and other nodes in a brain network. Thus, the closer a node is to all other nodes in the brain network, the higher the centrality of the node is. In brain network analysis, the closeness centrality of a brain region measures the indirect impact of the brain region on other brain regions. For a binary undirected network, the closeness centrality $C_c(i)$ of a node i is defined as the inverse of the average shortest path length of the node to all other nodes:

$$C_c(i) = \frac{N-1}{\sum_{j \neq i} l_{ij}}. \tag{11}$$

(iii) Betweenness centrality [142, 143] is a very popular measure, which quantifies the number of times that a node acts as a bridge along the shortest path between two other nodes. In brain network analysis, the betweenness centrality of a brain region measures the impact of the brain region on the flow of information across the brain network. In a binary undirected network, the betweenness centrality $C_b(i)$ of a node i is defined as the proportion of shortest paths between nodes j and h that pass through i:

$$C_b(i) = \frac{2}{(N-1)(N-2)} \sum_{j \neq h \neq i} \frac{n_{hj}(i)}{n_{hj}}, \tag{12}$$

where $n_{hj}(i)$ is the number of shortest paths between h and j that pass through i, n_{hj} is the number of all shortest paths between h and j, and $(N-1)(N-2)/2$ is the number of node pairs that do not include node i.

In brain network analysis, the above three centrality measures are also often used to identify hub brain regions. Firstly, the centrality values of all brain regions are ranked. Then, a specific threshold (e.g., mean + square deviation) is used to determine the hub nodes. That is, the brain regions whose centrality values are larger than the specific threshold are considered as hubs. For example, node B in Figure 2 is a hub of the simple network in terms of degree centrality.

3.3. Efficiency. The efficiency of a network (such as brain network) measures the ability of the network to exchange information. The higher the efficiency of the network is, the stronger the ability of information exchange is. The efficiency of a network mainly considers global efficiency and local efficiency. The global efficiency measures the ability of parallel information exchange across the whole network, while the local efficiency measures the ability of fault tolerance of a network [144].

Both global efficiency and local efficiency are closely related to nodal efficiency. Nodal efficiency measures how well a specific region is integrated within the network via its shortest paths. In a binary undirected network, the efficiency $E_{\text{nodal}}(i)$ of a node i is defined as the normalized sum of the reciprocal of the shortest path lengths from the node to all other nodes of the network:

$$E_{\text{nodal}}(i) = \frac{1}{N-1} \sum_{i \neq j} \frac{1}{l_{ij}}. \tag{13}$$

Obviously, the shorter the shortest path lengths of a node is, the higher the efficiency of the node is. The nodes with high nodal efficiency play an important role in the information integration and distribution [139]. It is worth mentioning that the nodes with high nodal efficiency can also be seen as hubs [130].

The global efficiency of a network [145] is the average nodal efficiencies of all nodes in the network and is defined as

$$E_{\text{glob}} = \frac{1}{N} \sum_{i} E_{\text{nodal}}(i) = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{l_{ij}}. \tag{14}$$

Since the brain is considered as a multiactivity parallel system, it should have a high global efficiency in a brain network [144].

The local efficiency of a node can be regarded as the global efficiency of the subnetwork containing itself and its all direct neighbors. In a binary undirected network, the local efficiency $E_{\text{loc}}(i)$ of a node i can be defined as

$$E_{\text{loc}}(i) = \frac{1}{N_{G_i}\left(N_{G_i}-1\right)} \sum_{j \neq h \in G_i} \frac{1}{l_{jh}}, \tag{15}$$

where G_i is the subgraph that consists of node i and its all direct neighbors.

Similarly, the local efficiency of a network is the average local efficiencies of all nodes in the network and is computed by

$$E_{\text{loc}} = \frac{1}{N} \sum_{i} E_{\text{loc}}(i). \tag{16}$$

It is worth mentioning that the difference between nodal efficiency and local efficiency of a node is that the former measures the ability of information exchange of the node itself, while the latter measures the ability of information exchange of the subnetwork consisting of itself and its all direct neighbors.

3.4. Modularity and Rich Club. A module is a group of nodes with dense internal connections but sparse external connections in a network as shown in Figure 2. The real network often has a number of relatively independent and interrelated modules that have different functions and evolve independently without affecting other modules. At the same time, the modular structure also provides more detailed roles and properties of nodes. For example, some nodes are important in their modules but are not necessarily important for the entire network; these nodes are called provincial hubs; while some other nodes though in their own modules are limited, they are connected to different modules, maintaining the connectivity of the entire network. Therefore these nodes play an important role in information transmission throughout the network and are called connector hubs. A participation index [146] is used to determine whether a node is a provincial hub or a connector hub. The participation index (P_i) of node i is computed by

$$P_i = 1 - \sum_{m=1}^{N_m} \left(\frac{k_{im}}{k_i} \right)^2, \tag{17}$$

where N_m is the number of modules and k_{im} is the number of connections from node i to module m. In general, if P_i is greater than a specific threshold, node i is a connector hub, and otherwise it is a provincial hub.

Biological networks, including human brains, exhibit a high degree of modularity. In complex network analysis, modularity is used to measure the quality of division of a network into modules [147]. Currently, there are different methods to calculate the modularity of a brain network [148]. Here, we introduce two common modularity measures in brain network analysis. Given that a brain network is fully subdivided into several nonoverlapping modules M, these two common modularity measures are computed as follows:

(i)

$$Q = \sum_{s \in M} \left[p_{ss} - \left(\sum_{t \in M} p_{st} \right) \right], \tag{18}$$

where p_{ss} is the proportion of existing links in module s and p_{st} is the proportion of existing links between modules s and t.

(ii)

$$Q = \frac{1}{2N(N-1)} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{N(N-1)} \right) \delta\left(m_i, m_j\right), \tag{19}$$

where m_i and m_j are the modules containing node i and node j, respectively, and if nodes i and j are in the same module, $\delta(m_i, m_j) = 1$; otherwise, $\delta(m_i, m_j) = 0$.

The so-called "rich club" effect in complex networks is that the hubs of a complex network tend to be more densely connected among themselves than nodes of a lower degree [149–151] as shown in Figure 2. In essence, nodes with a large number of edges, which are usually called rich nodes, are much more likely to form closely interconnected subgraphs (also called clubs) than low-degree nodes. The normalized rich club coefficient ($\rho_{\text{rand}}(k)$) is used to quantify the rich club effect of a complex network [149, 150] and is defined as

$$\begin{aligned} \rho_{\text{rand}}(k) &= \frac{\phi(k)}{\phi_{\text{rand}}(k)} \\ \phi(k) &= \frac{2E_{>k}}{N_{>k}\left(N_{>k} - 1\right)}, \end{aligned} \tag{20}$$

where $N_{>k}$ is the number of nodes with degree larger than k, $E_{>k}$ is the number of edges connecting these $N_{>k}$ nodes, $N_{>k}(N_{>k} - 1)/2$ is the maximum possible number of edges among these nodes $N_{>k}$, $\phi(k)$ is the rich club coefficient of a given degree k, and $\phi_{\text{rand}}(k)$ is the rich club coefficient on a maximally randomized network with the same degree distribution of the network under study.

The rich club effect of brain networks plays an important role in global brain information transmission [152], which provides important information on the higher-level topology of brain networks. The rich club serves as an important backbone for a number of coactivation patterns among brain regions [153]. Thus, a rich club in a brain network is also crucial for promoting and integrating various segregated functions [148]. In brain network, a normalized rich club coefficient increasing with the degree k indicates the dominance of a number of highly connected and mutually communicating brain regions, as opposed to a set consisting of many loosely connected and relatively independent brain regions.

3.5. Small-World Network. The concept of small-world networks was first proposed by Watts and Strogatz [154]. The small-world networks have both high clustering characteristics similar to regular networks and shorter shortest path lengths similar to random networks. In other words, The small-world networks combine the respective topological advantages of both regular networks and random networks to ensure the efficiency of information transmission at both local and global levels.

To determine whether a network is a small-world network, the following three criteria are employed:

$$\begin{aligned} \gamma &= \frac{C}{C_{\text{rand}}} \\ \lambda &= \frac{L}{L_{\text{rand}}} \\ \sigma &= \frac{\gamma}{\lambda}, \end{aligned} \tag{21}$$

where C_{rand} and L_{rand} are the average clustering coefficient and characteristic path length of M matched random networks that preserve the same number of nodes, edges, and degree distribution as the real network, γ and λ are the normalized clustering coefficient and normalized characteristic path length of the network, and σ is the small-world index

of the network [155]. A network with small-world property needs to meet two conditions: $\gamma \gg 1$ and $\lambda \approx 1$. Thus, the small-world index $\sigma > 1$.

The brain supports both segregated and integrated information processing. Small-world topology comprises both high clustering (compatible with segregated processing) and short path length (compatible with integrated processing). Therefore, for a given brain network, if the brain network is found to have small-world property, it shows that the brain has better information processing performance, at least without great disruption.

In addition to the above important topological properties, there are many other properties in graph theory to characterize brain networks, such as hierarchy, motif, assortativity, and transitivity. For more details about brain network analysis, please see [140, 148, 156]. The brain network analysis can be also implemented by some common open source tools such as Brain Connectivity Toolbox (http://www.nitrc.org/projects/bct/), GRETNA (https://www.nitrc.org/projects/gretna/), and BrainWave (http://home.kpn.nl/stam7883/brainwave.html).

4. Applications of Brain Network Analysis

Brain network analysis becomes increasingly important when either studying pathophysiology or exploring the network-based biomarkers of brain disorders. The studies of brain network analysis with graph theoretical frameworks have been extensively applied to identify detailed abnormalities of network topologies associated with various brain disorders, including AD, SCZ, PD, and multiple sclerosis. In this section, we summarize the recent progress of structural and functional brain networks with several brain disorders, focusing on the changes in the topological organization of structural and functional brain networks in terms of graph theoretical frameworks.

4.1. Alzheimer's Disease. Alzheimer's disease (AD) is a degenerative brain disease and the most common type of dementia, comprising 60–80% of all dementia cases. In the US, there are more than 5.2 million AD patients in 2014, and it is estimated that 13.8 million Americans are AD patients by 2050 [157]. Thus, early diagnosis and treatment of AD, especially at its prodromal stage such as MCI [158], have become a crucial step to delay or even avoid dementia. The clinical symptoms of AD are impairments of memory, language, and other cognitive functions, which seriously affect the daily life of patients and their families. Existing studies have demonstrated that these impairments are associated with abnormal structural and functional brain networks [83, 95, 103, 159].

4.1.1. Structural Brain Networks in AD. Structural brain networks of human beings can be constructed by using sMRI and dMRI. In this section, we review recent progress in analyzing the networks based on sMRI and dMRI in AD as shown in Table 2.

(i) Using sMRI: He et al. [95] first used cortical thickness measurement to investigate structural brain networks of 92 AD patients and 97 healthy controls (HCs). They found increased clustering coefficient and shortest paths in AD, implying an abnormal small-world property. In addition, they found reduced betweenness centrality in the temporal and parietal heteromodal association cortex regions and increased betweenness centrality in the occipital cortex regions. Yao et al. [83] used gray matter volumes to construct structural brain networks of 91 AD patients, 113 MCI patients, and 98 HCs. Among structural brain networks of three groups, they found the greatest clustering coefficient and the longest absolute path length in AD. Their finding was similar to that by He et al. [95]. In addition, they found the small-world index of the MCI networks was between AD and HC networks. Their finding showed that MCI is a transitional stage between HC and AD. Compared with the HCs, the MCI and AD patients retained hub regions in the frontal lobe but lost hub regions in the temporal lobe. Tijms et al. [96] investigated the topology properties of single-subject gray matter networks in AD and found decreased normalized clustering coefficient and normalized path length. Their finding is contrary to the previous two findings. Moreover, they found decreased small-world index in AD. Pereira et al. [97] constructed structural brain networks of stable MCI (sMCI) patients, late MCI converters (lMCIc), early MCI converters (eMCIc), and AD patients to investigate topology structure across groups. They found that, compared with the HC group, all patient groups exhibited increased path length, reduced transitivity, and increased modularity, and the patient group showed decreased small-world index. In addition, compared with the sMCI group, other three patient groups showed decreased path length and clustering coefficient.

(ii) Using dMRI: Lo et al. [98] used DTI to construct structural brain networks of 25 AD patients and 30 HCs. They found that although the two groups had a small-world property, the AD group showed increased shortest path length compared with the HC group. In addition, they found decreased global efficiency and reduced nodal efficiency in the frontal regions in AD. Bai et al. [99] considered two high risk groups, remitted geriatric depression (RGD) and amnestic MCI (aMCI), and constructed structural brain networks of the two groups using DTI and deterministic tractography. They found reduced network strength, reduced global efficiency, and increased absolute path length for both the RGD and aMCI patients compared with HCs, and there were no significant differences in these global network properties between the two high risk groups. Compared with HCs, they found that the two high risk groups had similar deficits of the regional and connectivity characteristics in the frontal regions. From comparison of RGD and aMCI, they found that the nodal efficiency of networks in the two groups was different in the posterior cingulate cortex

Table 2: Overview of structural brain network studies in AD.

Study	Modality	Subjects	Node definition	Edge definition	Network type	Main findings
He et al., 2008 [95]	sMRI	92 AD 97 HC	54 regions in ANIMAL package	Partial correlation based on cortical thickness	Binary	(1) Increased clustering coefficient and shortest paths in AD. (2) Reduced betweenness centrality in the temporal and parietal regions and increased betweenness centrality in the occipital regions.
Yao et al., 2010 [83]	sMRI	91 AD 113 MCI 98 HC	90 regions in AAL atlas	Pearson correlation based on gray matter volume	Binary	(1) The greatest clustering coefficient and the longest absolute path length in AD. (2) The small-world index of the MCI network was between AD and HC networks. (3) Compared with the HCs, the MCI and AD patients retained hub regions in the frontal lobe but lost hub regions in the temporal lobe.
Tijms et al., 2013 [96]	sMRI	38 AD 38 HC	8683 ± 545 cubes	Intracortical similarity	Binary	(1) Decreased normalized clustering coefficient and normalized path length in AD. (2) Decreased small-world index in AD.
Pereira et al., 2016 [97]	sMRI	282 AD 110 sMCI 71 lMCIc 87 eMCIc 301 HC	82 regions in FreeSurfer	Pearson correlation	Binary	(1) Increased characteristic path length in sMCI, lMCIc, eMCIc, and AD compared with HC. (2) Decreased clustering coefficient in lMCIc, eMCIc, and AD compared with HC. (3) Decreased transitivity and increased modularity in patients compared with HCs. (4) Decreased small-world index in patients compared with HCs. (5) Decreased characteristic path length and clustering coefficient in lMCIc, eMCIc, and AD compared with sMCI.
Lo et al., 2010 [98]	DTI	25 AD 30 HC	78 regions in AAL atlas	FN × FA	Weighted	(1) Increased shortest path length in AD. (2) Decreased global efficiency and reduced nodal efficiency in the frontal regions in AD.
Bai et al., 2012 [99]	DTI	35 RGD 38 aMRI 30 HC	90 regions in AAL atlas	FN	Weighted	(1) Reduced network strength, reduced global efficiency, and increased absolute path length in RGD and aMCI. (2) Similar deficits of the regional and connectivity characteristics in the frontal brain regions in RGD and aMCI. (3) Different nodal efficiency in the posterior cingulate cortex and several prefrontal brain regions between RGD and aMCI.
Daianu et al., 2015 [100]	DTI	42 AD 110 MCI 50 HC	68 regions in FreeSurfer	FN	Binary	(1) AD affected the low degree brain regions, rather than the rich club comprising the high degree brain regions. (2) Global connectivity of AD was disrupted. (3) Detecting network differences of MCI/HC and AD/HC using the normalized rich club coefficient.

Table 2: Continued.

Study	Modality	Subjects	Node definition	Edge definition	Network type	Main findings
Wang et al., 2016 [101]	DTI	26 AD 16 HC	90 regions in AAL atlas	FN	Binary	(1) Higher small-world index in AD. (2) Decreased global efficiency and local efficiency in AD. (3) Increased normalized shortest path length and normalized clustering coefficient in AD.

and several prefrontal regions. Daianu et al. [100] constructed structural brain networks of 42 patients, 110 MCI patients, and 50 HCs and investigated the rich club organization of networks of the three groups. They found that AD affected the low-degree brain regions, rather than the rich club comprising the high degree brain regions; global connectivity of AD was disrupted; the normalized rich club coefficient could be used to detect brain network differences of MCI/HC and AD/HC. To detect abnormal topological organization of structural brain networks of AD patients, Wang et al. [101] used DTI data to construct structural brain networks of 26 AD patients and 16 HCs. They found that although both groups showed small-world property, the AD group exhibited higher small-world index than the HC group. In addition, the AD group displayed decreased global efficiency and local efficiency and increased normalized shortest path length and normalized clustering coefficient.

4.1.2. Functional Brain Networks in AD. Functional brain networks of human beings can be constructed using fMRI and EEG/MEG. In this section, we review recent progress in analyzing the networks based on fMRI and EEG/MEG in AD as shown in Table 3.

(i) Using fMRI: Supekar et al. [88] constructed functional brain networks of 21 AD patients and 18 HCs using task-free fMRI. They found that, in the low frequency band, 0.01–0.05 Hz, the AD group had lower clustering coefficient than HCs, which cause loss of small-world property in AD. Specifically, they found that clustering coefficients in the left and right hippocampus in AD were lower than HC while in the left and right precentral gyrus they were not significantly different. To investigate whether aMCI patients disrupt the topological structure of brain networks, Wang et al. [89] constructed functional brain networks of aMCI patients. Compared with HCs, they found decreased functional connectivity and increased path length in the frequency bands, 0.031–0.063 Hz, in aMCI. Brier et al. [85] used path length, clustering coefficient, and modularity to investigate the topology properties of functional brain networks of AD patients (Clinical Dementia Rating (CDR) = 1). They constructed different functional brain networks for participants with CDR = 0, participants with CDR = 0.5, and participants with CDR = 1. They found decreased clustering coefficient and modularity with increasing CDR, but path length was not significantly different. Golbabaei et al. [86] used the local and global measures to assess functional brain networks of AD patients. They found decreased clustering coefficient and global efficiency and increased characteristic path length in AD compared with HC. In addition, AD patients exhibited decreased node strength, local clustering coefficient, and local efficiency and increased local characteristic path length in olfactory, hippocampus, parahippocampal, amygdala, and superior parietal gyrus.

(ii) Using EEG/MEG: Buldú et al. [34] constructed functional brain networks of MCI patients by using MEG data during a memory task in five frequency bands: α_1 (8–11 Hz), α_2 (11–14 Hz), β_1 (14–25 Hz), β_2 (25–35 Hz), and γ (35–45 Hz). They found that the MCI group exhibited an enhancement of the connection strength, which demonstrated that memory processing of MCI patients needs higher energy. In particular, the MCI group also showed lower normalized clustering coefficient and characteristic path length. de Haan et al. [21] used MEG data to explore functional brain network integrity in AD, focusing on network connectivity, synchronizability, and node centrality. They found a loss of network connectivity and altered synchronizability in most frequency bands and demonstrated a low centrality of the left temporal region in the theta band in AD. To clarify these two problems, how functional connectivity is affected in AD subgroups of disease severity and how network hubs change, Engels et al. [103] used EEG data to investigate functional brain networks of three subgroups of AD patients based on disease severity: mild AD (mi-AD), moderate AD (mo-AD), and severe AD (se-AD). They had three main findings: decreased functional connectivity with increasing AD severity in the alpha band; increased betweenness centrality with increasing AD severity in all regions (except for posterior); decreased hub regions in posterior regions and increased hub regions in most anterior regions with increasing AD severity. To investigate the underlying alteration of the high-level visual (HLV) networks in AD patients, late MCI (LMCI) patients, and early MCI (EMCI) patients, Deng et al. [104] constructed HLV networks of "where" visions across groups. In their study, the

Table 3: Overview of functional brain network studies in AD.

Study	Modality	Subjects	Node definition	Edge definition	Network type	Main findings
Supekar et al., 2008 [88]	fMRI	21 AD 18 HC	90 regions in AAL atlas	Wavelet correlation	Weighted	(1) Compared with HC, the loss of small-world property characterized by a lower clustering coefficient in AD in low frequency band: 0.01–0.05 Hz. (2) Lower clustering coefficients in the left and right hippocampus in AD.
Wang et al., 2013 [89]	fMRI	37 aMCI 47 HC	1024 regions in H-1024 [102]	Wavelet correlation	Weighted	(1) Decreased functional connectivity in the frequency band, 0.031–0.063 Hz, in aMCI. (2) Increased path length in aMCI.
Brier et al., 2014 [85]	fMRI	31 CDR 1 90 CDR 0.5 205 CDR 0	160 regions [61]	Pearson correlation	Binary	(1) Reduced clustering coefficient and modularity with increasing CDR. (2) No significant differences in path length among participants with different CDR.
Golbabaei et al., 2016 [86]	fMRI	21 AD 21 HC	90 regions in AAL atlas	Pearson correlation	Weighted	(1) Decreased clustering coefficient and global efficiency and increased characteristic path length in AD. (2) Decreased node strength, local clustering coefficient, and local efficiency and increased local characteristic path length in olfactory, hippocampus, parahippocampal, amygdala, and superior parietal gyrus in AD.
Buldú et al., 2011 [34]	MEG	19 AD 19 HC	148 sensors	Synchronization likelihood	Weighted	(1) Higher connection strength in MCI. (2) Lower normalized clustering coefficient and characteristic path length in MCI.
de Haan et al., 2012 [21]	MEG	18 AD 18 HC	149 channels	Synchronization likelihood	Weighted	(1) Loss of network connectivity and altered synchronizability in most frequency bands in AD. (2) Low centrality of the left temporal region in the theta band in AD.
Engels et al., 2015 [103]	EEG	117 se-AD 96 mo-AD 105 mi-AD 133 HC	21 channels	Phase lag index	Weighted	(1) Decreased functional connectivity with increasing AD severity in the alpha band. (2) Increased betweenness centrality with increasing AD severity in all regions (except for posterior). (3) Decreased hub regions in posterior regions and increased hub regions in most anterior regions with increasing AD severity.
Deng et al., 2016 [104]	EEG	30 AD 35 LMCI 52 EMCI 44 HC	25 regions	Wavelet correlation	Binary	(1) Increased clustering coefficient and longer characteristic path length in AD compared with HC. (2) No significant difference of clustering coefficient and characteristic path length between EMCI and HC and between LMCI and HC.

AD group showed increased clustering coefficient and longer characteristic path length than the HC group. In addition, compared with the HC group, the LMCI and EMCI groups had no significant difference in terms of clustering coefficient and characteristic path length.

4.2. Schizophrenia. Schizophrenia (https://www.nimh.nih.gov/health/publications/schizophrenia-booklet/index.shtml) (SCZ) is a chronic and severe psychiatric disorder characterized by hallucinations, delusions, and loss of initiative and cognitive dysfunction. Patients with SCZ may seem like losing touch with reality. Families and society are affected by

SCZ, too. Many patients with SCZ have difficulty in doing a job or caring for themselves, so they rely on others for help. Approximately 8 out of 1,000 individuals have SCZ in their lifetime. In principle, exploring the pathological mechanism of SCZ is a key step in the diagnosis and treatment of SCZ. Some existing studies have demonstrated that the pathological mechanism of SCZ is related to abnormal structural and functional brain networks [47, 160–162].

4.2.1. Structural Brain Networks in SCZ. Structural neuroimaging data, such as sMRI and dMRI, have been widely used in the structural brain network study of SCZ. In this section, we review recent progress in analyzing the structural brain networks based on sMRI and dMRI in SCZ as shown in Table 4.

(i) Using sMRI: Bassett et al. [71] used interregional covariation of gray matter volume as structural connectivity to construct structural brain networks of 259 HCs and 203 SCZ patients. In their study, the cortical cortex was divided into multimodal, unimodal, and transmodal. They found that, in the HC group, the three cortical divisions had small-world property, the multimodal network was hierarchy characterized by frontal hubs with low clustering coefficient, and the transmodal network was assortative. In addition, in the SCZ group, abnormal multimodal network organization showed reduced hierarchy, the loss of frontal and the emergence of nonfrontal hubs, and increased connection distance. Zhang et al. [72] hypothesized that the core symptoms of SCZ originate from the inability to integrate information transmission segregated across different brain regions. To demonstrate this hypothesis, they constructed structural brain networks of SCZ patients using the cortical thickness measurement and found increased characteristic path length and clustering coefficient in the SCZ group compared with the HC group. Moreover, they found reduced nodal centrality in several regions of the default network and increased nodal centrality in primary cortex and paralimbic cortex regions in SCZ. To investigate whether developmental abnormalities associated with SCZ occur in the neonatal stage, Shi et al. [105] constructed morphological brain networks of 26 neonates who were at genetic risk for SCZ and found that although the SCZ group exhibited small-world topology, the SCZ group had lower global efficiency, longer connection distance, and fewer number of hub nodes with higher betweenness. Tijms et al. [106] constructed more refined structural brain networks of high risk SCZ (HR-SCZ) by using a $6 \times 6 \times 6\,\mathrm{mm}^3$ cube as a node. They found lower path length in the bilateral inferior frontal gyri, left posterior cingulate region, and superior temporal gyrus and lower clustering coefficient in the right medial superior frontal gyrus, right insula, right fusiform gyrus, left occipital gyrus, and right temporal regions in HR-SCZ compared with HC.

(ii) Using dMRI: Zalesky et al. [107] used corticocortical anatomical connectivity at the scale of axonal fiber bundles to construct structural brain networks of 74 SCZ patients and 32 HCs. They found that although the SCZ group exhibited small-world topology, the SCZ group had lower global efficiency. In addition, they found lower node degree in medial frontal, parietal/occipital, and the left temporal lobe. To test whether connectivity disturbances are associated with familial vulnerability for SCZ, Collin et al. [108] constructed structural brain networks of 40 SCZ patients, 54 unaffected siblings of SCZ patients, and 51 HCs using DTI data. They found reduced connectivity between rich club hubs across groups, which was lowest in the SCZ group, intermediate in the SCZ siblings group, and highest in the HC group. Moreover, in the SCZ group, they found that lower levels of rich club connectivity were associated with longer duration of illness and worse overall functioning. To investigate alterations in hemispheric white matter (WM) topology in SCZ, Sun et al. [109] constructed weighted hemispheric brain anatomical networks of SCZ patients and HCs. They found that although the hemispheric networks showed small-world property, the hemispheric-independent deficit of global integration was significantly different in SCZ. Furthermore, compared with the HC group, the SCZ group had longer characteristic path length, lower global efficiency, and reduced asymmetric nodal efficiency in several frontal regions and the hippocampus. Later, Sun et al. [109] continued to investigate alterations in the topological structure of brain anatomical networks using DTI data in SCZ. They constructed weighted brain anatomical networks of 31 SCZ patients and 28 HCs using deterministic tractography and found similar results to the previous study that the SCZ group had small-world property and that, compared with HCs, the SCZ group had longer characteristic path length and lower global efficiency and was significantly different in the independent deficit of global integration.

4.2.2. Functional Brain Networks in SCZ. fMRI and EEG/MEG have been widely used in the functional brain network study of SCZ. In this section, we review recent progress in analyzing the functional brain networks based on fMRI and EEG/MEG in SCZ as shown in Table 5.

(i) Using fMRI: Lynall et al. [111] measured aspects of both functional connectivity and functional network topology of SCZ patients to test whether SCZ is a disorder of connectivity. They found decreased strength of functional connectivity and increased diversity of functional connections in SCZ patients. Specifically, they found reduced clustering coefficient and small-world index, reduced probability of high degree hubs, and increased robustness in the SCZ group. Furthermore, the SCZ group had reduced degree and clustering coefficient in medial parietal, premotor and

Table 4: Overview of structural brain network studies in SCZ.

Study	Modality	Subjects	Node definition	Edge definition	Network type	Main findings
Bassett et al., 2008 [71]	sMRI	203 SCZ 259 HC	104 regions in Pick atlas	Partial correlation based on gray matter volume	Binary	(1) The multimodal network was hierarchy and the transmodal network was assortative in HC. (2) Abnormal multimodal network organization showed reduced hierarchy, the loss of frontal and the emergence of nonfrontal hubs, and increased connection distance in SCZ.
Zhang et al., 2012 [72]	sMRI	101 SCZ 101 HC	78 regions in AAL atlas	Partial correlation based on cortical thicknesses	Binary	(1) Increased characteristic path length and clustering coefficient in SCZ. (2) Reduced nodal centrality in several regions of the default network in SCZ. (3) Increased nodal centrality in primary cortex and paralimbic cortex regions in SCZ.
Shi et al., 2012 [105]	sMRI	26 SCZ 26 HC	90 regions in AAL atlas	Pearson correlation based on gray matter volume	Binary	(1) Lower global efficiency in SCZ. (2) Longer connection distance in SCZ. (3) Fewer number of hub nodes with higher betweenness in SCZ.
Tijms et al., 2015 [106]	sMRI	144 HR-SCZ 36 HC	$6 \times 6 \times 6\ \mathrm{mm}^3$ cubes	The similarity of grey matter structure	Binary	(1) Lower path length in the bilateral inferior frontal gyri, left posterior cingulate region, and superior temporal gyrus in HR-SCZ. (2) Lower clustering coefficient in the right medial superior frontal gyrus, right insula, right fusiform gyrus, left occipital gyrus, and right temporal regions in HR-SCZ.
Zalesky et al., 2011 [107]	DTI	74 SCZ 32 HC	82 regions in AAL atlas	FN	Binary	(1) Lower global efficiency in SCZ. (2) Lower node degree in medial frontal, parietal/occipital, and the left temporal lobe in SCZ.
Collin et al., 2014 [108]	DTI	40 SCZ 54 SCZ siblings 51 HC	68 regions in FreeSurfer	FN	Weighted	(1) Reduced connectivity between rich club hubs (i.e., lowest in SCZ, intermediate in SCZ siblings, and highest in HC). (2) Lower levels of rich club connectivity related to longer duration of illness and worse overall functioning.
Sun et al., 2015 [109]	DTI	116 SCZ 66 HC	90 regions in AAL atlas	FN	Weighted	(1) Significantly different in the hemispheric-independent deficit of global integration in SCZ. (2) Longer characteristic path length and lower global efficiency in SCZ. (3) Reduced asymmetric nodal efficiency in several frontal regions and the hippocampus in SCZ compared with HC.
Sun et al., 2016 [110]	DTI	31 SCZ 28 HC	90 regions in AAL atlas	FN × FA	Weighted	(1) Significantly different in the independent deficit of global integration in SCZ compared with HC. (2) Longer characteristic path length and lower global efficiency in SCZ.

Table 5: Overview of functional brain network studies in SCZ.

Study	Modality	Subjects	Node definition	Edge definition	Network type	Main findings
Lynall et al., 2010 [111]	fMRI	12 SCZ 15 HC	72 regions in AAL atlas	Wavelet correlation	Binary	(1) Reduced clustering coefficient and small-world index, reduced probability of high degree hubs, and increased robustness in SCZ. (2) Reduced degree and clustering coefficient in medial parietal, premotor and cingulate, and right orbitofrontal cortical regions in SCZ.
Su et al., 2015 [112]	fMRI	49 SCZ 28 HC	90 regions in AAL atlas	Pearson correlation	Weighted	(1) Lower global efficiency in SCZ. (2) The severity of psychopathology, negative symptoms, and depression and anxiety symptoms were related to global efficiency in SCZ.
Hadley et al., 2016 [113]	fMRI	32 SCZ 32 HC	278 regions [114]	Wavelet correlation	Binary	(1) Reduced global efficiency and increased clustering coefficients in SCZ. (2) Aberrant functional integration and segregation in SCZ.
Ganella et al., 2017 [115]	fMRI	42 TR-SCZ 42 HC	116 regions in AAL atlas	Pearson correlation	Binary	(1) Reduced global brain functional connectivity and reduced strength in frontotemporal, frontooccipital, and temporooccipital connections in TR-SCZ. (2) Reduced global efficiency and increased local efficiency in TR-SCZ.
Jhung et al., 2013 [116]	EEG	12 SCZ 13 UHR 13 HC	64 channels	Synchronization likelihood	Binary	(1) Reduced small-world property in the theta band during the working memory task in SCZ compared with HC. (2) The small-world index of the UHR was intermediate value among SCZ, UHR, and HC during the working memory task.
Shim et al., 2014 [35]	EEG	34 SCZ 34 HC	314 dipole sources	Phase locking value	Weighted	(1) Reduced clustering coefficients and increased path lengths in SCZ. (2) The severity of SCZ symptoms was negatively correlated with the clustering coefficient and positively correlated with path length.
Yin et al., 2017 [87]	EEG	14 P-SCZ 14 N-SCZ 14 HC	32 electrodes	Mutual information	Binary	(1) Smaller clustering coefficient, larger average characteristic path length, lower global efficiency, lower local efficiency, and smaller degrees in SCZ. (2) SCZ patients had fewer information interactions than HCs, and P-SCZ had more information interactions than N-SCZ.

cingulate, and right orbitofrontal cortical regions. To clarify the correlation between brain network efficiency and SCZ symptoms, Su et al. [112] constructed functional brain networks of 49 SCZ patients and 28 HCs and found that the SCZ group had lower global efficiency than the HC group. Moreover, they also found that the severity of psychopathology, negative symptoms, depression, and anxiety symptoms were related to global efficiency in SCZ. To test whether the balance between functional integration and segregation of brain networks is impaired in SCZ, Hadley et al. [113] constructed functional brain networks of 32 SCZ patients and 32 HCs. They found reduced global efficiency and increased clustering coefficients in SCZ. Since the global efficiency is a measure of functional integration and the clustering coefficient is a measure of functional segregation, their findings demonstrated aberrant functional integration and segregation in SCZ. To explore disruptions in functional connectivity and altered efficiency of functional brain networks in "treatment-resistant" SCZ (TR-SCZ), Ganella et al. [115] constructed functional brain networks of 42 TR-SCZ patients and 42 HCs. They found reduced global brain functional connectivity

and reduced strength in frontotemporal, frontooccipital, and temporooccipital connections in TR-SCZ. In addition, they found reduced global efficiency and increased local efficiency in TR-SCZ.

(ii) Using EEG/MEG: Jhung et al. [116] constructed functional brain networks of 13 individuals at ultrahigh risk (UHR) for psychosis, 12 SCZ patients, and 13 HCs to investigate the small-world functional networks across groups. They found that, compared with HCs, SCZ patients had reduced small-world property in the theta band during a working memory task. Furthermore, they found that the small-world index of the UHR during the working memory task showed intermediate value between those of HC and SCZ. Subsequently, Shim et al. [35] investigated small-world functional networks during auditory oddball tasks and their relations with the severity of symptoms in SCZ. They found reduced clustering coefficients and increased path lengths in SCZ. This finding showed disrupted small-world functional network in SCZ. In addition, the severity of SCZ symptoms was negatively correlated with the clustering coefficient and positively correlated with path length. To test whether positive SCZ (P-SCZ) had more information interaction and negative SCZ (N-SCZ) had less information interaction between brain regions compared with HC, Yin et al. [87] used EEG data to constructed functional brain networks of P-SCZ patients, N-SCZ patients, and HCs. They found smaller clustering coefficient, larger average characteristic path length, lower global efficiency, lower local efficiency, and smaller degrees in SCZ and concluded that SCZ patients had fewer information interactions than HCs, and P-SCZ had more information interactions than N-SCZ.

4.3. Parkinson's Disease. Parkinson's disease (PD) is the second most common progressive neurodegenerative disorder, trailing only AD [163, 164]. Advances in neuroimaging techniques and electrophysiological techniques are rapidly expanding the complexity of neurophysiologic understanding of PD. These techniques help to better understand the neurophysiologic mechanisms of PD and its treatments.

4.3.1. Structural Brain Networks in PD. The study of structural brain networks provides another perspective for PD. In this section, we review recent progress in analyzing the structural brain networks based on sMRI and dMRI in PD as shown in Table 6.

(i) Using sMRI: to investigate the topological organization of PD patients, Zhang et al. [117] constructed morphological brain networks of PD patients. In their study, PD patients showed increased global efficiency and local efficiency, increased nodal local efficiency in several regions, decreased local nodal efficiency in several regions, and increased global nodal efficiency in several regions. Pereira et al. [118] used global measures and regional measures to investigate the topology structure of structural brain networks of 123 PD patients and 56 HCs. In their study, to test whether MCI is associated with disruption in structural brain networks, 123 PD patients were classified into 33 PD patients with MCI (PD-MCI) and 90 PD patients with HC (PD-HC) using the Movement Disorders Society Task Force criteria. They found that, compared with the HC group, the PD-MCI group had reduced connectivity strength between cortical and subcortical regions. In addition, they found that, compared with PD-HC patients and HCs, PD-MCI patients had longer characteristic path length and reduced global efficiency and lower regional efficiency in frontal and parietal regions. Specifically, both PD-MCI and PD-HC had a reorganization of the highly connected regions in the brain networks. To investigate the topological differences of male PD (PD-M) patients and female PD (PD-F) patients, Yadav et al. [119] constructed structural brain network across groups by using cortical thickness. In their study, compared with PD-F patients, PD-M patients showed lower connectivity strength and clustering coefficients and longer path length. In addition, compared with PD-F patients, PD-M patients exhibited lower nodal betweenness in left caudal middle frontal, left rostral middle frontal, and right parahippocampal regions. Moreover, hubs of the PD-M group were right fusiform and right isthmus cingulate region and left inferior temporal and left rostral anterior cingulate, while hubs of the PD-F group were right parahippocampal, right superior temporal, and left rostral middle frontal regions.

(ii) Using dMRI: to reveal topological changes in structural brain networks in PD patients, Li et al. [120] used DTI data to construct structural brain networks of 35 PD patients and 26 HCs by using deterministic tractography. They found that, compared with HCs, PD patients had lower connectivity strength in the feeder and local connections. Furthermore, they found that, in the two modules, the limbic/paralimbic/subcortical module and the cognitive control/attention module, the PD group had decreased connections compared with the HC group. In addition, they found increased shortest path length and decreased global efficiency in PD. To assess whether structural topological brain network changes are detectable in PD patients, Nigro et al. [121] used centrality, segregation, and integration measures to assess structural brain networks of PD patients. They found decreased network strength, global efficiency, and global clustering coefficient in PD patients compared with HCs. In addition, both groups had 18 hub regions, of which 14 are the same while the other 4 are different.

4.3.2. Functional Brain Networks in PD. The study of functional brain networks provides another perspective for PD. In this section, we review recent progress in analyzing the

Table 6: Overview of structural brain network studies in PD.

Study	Modality	Subjects	Node definition	Edge definition	Network type	Main findings
Zhang et al., 2015 [117]	sMRI	16 PD 20 HC	264 regions in [65]	Intracortical similarity	Binary	(1) Increased global efficiency and local efficiency in PD. (2) Increased nodal local efficiency in the right inferior frontal gyrus (orbital part) and precentral gyrus, left insula and post cingulated cortex, and cerebellum in PD. (3) Decreased local nodal efficiency in the right Heschl's gyrus and precuneus gyrus, and bilateral medial superior frontal gyrus in PD. (4) Increased global nodal efficiency in the right inferior occipital cortex, inferior frontal gyrus (orbital part), precentral gyrus, and Heschl's gyrus in PD.
Pereira et al., 2015 [118]	sMRI	33 PD-MCI 90 PD-HC 56 HC	162 regions in FreeSurfer	Pearson correlation	Binary	(1) Reduced connectivity strength between cortical and subcortical regions in PD-MCI compared to HC. (2) Larger characteristic path length and reduced global efficiency and lower regional efficiency in frontal and parietal regions in the PD-MCI group compared with other two groups. (3) A reorganization of the highly connected regions in both PD-MCI and PD-HC.
Yadav et al., 2016 [119]	sMRI	43 PD-M 21 PD-F 46 HC	68 regions in FreeSurfer	Pearson correlation	Binary	(1) Lower connectivity strength and clustering coefficients and higher path length in PD-M compared with PD-F. (2) Lower nodal betweenness in left caudal middle frontal, left rostral middle frontal, and right parahippocampal regions in PD-M compared with PD-F. (3) Hubs were right fusiform and right isthmus cingulate region and left inferior temporal and left rostral anterior cingulate in PD-M. (4) Hubs were right parahippocampal, right superior temporal, and left rostral middle frontal regions in PD-F.
Li et al., 2017 [120]	DTI	35 PD 26 HC	90 regions in AAL atlas	FA	Weighted	(1) Lower connectivity strength in the feeder and local connections in PD. (2) Decreased connections in the two modules: the limbic/paralimbic/subcortical module and the cognitive control/attention module in PD. (3) Increased shortest path length and decreased global efficiency in PD.
Nigro et al., 2016 [121]	DTI	21 PD 30 HC	90 regions in AAL atlas	FA × FN	Weighted	(1) Decreased network strength, global efficiency, and global clustering coefficient in PD. (2) 14 same hub regions and 4 different hub regions between PD and HC.

functional brain networks based on fMRI and EEG/MEG in PD as shown in Table 7.

(i) Using fMRI: to investigate the efficiency of functional brain networks of PD patients, Skidmore et al. [122] constructed functional brain networks of 14 PD patients and 15 HCs. They found decreased global efficiency and decreased nodal efficiency in the left supplementary motor cortex, contiguous precentral regions, the calcarine cortices, secondary visual regions, and the certain regions within the cerebellum in PD patients. Later, to investigate altered

Table 7: Overview of functional brain network studies in PD.

Study	Modality	Subjects	Node definition	Edge definition	Network type	Main findings
Skidmore et al., 2011 [122]	fMRI	14 PD 15 HC	116 regions in AAL atlas	Wavelet correlation	Weighted	(1) Decreased global efficiency in PD. (2) Decreased nodal efficiency the left supplementary motor cortex, contiguous precentral regions, the calcarine cortices, secondary visual regions, and the certain regions within the cerebellum in PD.
Göttlich et al., 2013 [123]	fMRI	37 PD 20 HC	343 regions	Zero-lag Pearson correlation	Binary	(1) Lower global efficiency in PD in PD. (2) Increased connectivity within the sensorimotor network and decreased interaction of the visual network with other brain modules in PD. (3) Lower connectivity between the cuneus and the ventral caudate, medial orbitofrontal cortex, and the temporal lobe. (4) Decreased degree in the occipital lobe and increased degree in the superior parietal cortex, posterior cingulate gyrus, supramarginal gyrus, and supplementary motor area.
Luo et al., 2015 [124]	fMRI	47 PD 47 HC	200 regions in [67]	Pearson correlation	Binary	(1) Lower clustering coefficient and local efficiency in PD. (2) Reduced node centralities and connectivity strength in temporal-occipital and sensorimotor regions in PD.
Koshimori et al., 2016 [125]	fMRI	42 PD 23 HC	120 regions in [61]	Pearson correlation	Binary	(1) Higher nodal degree in the right and left dorsolateral prefrontal cortex in PD. (2) Reduced local efficiency the right mid-insula in PD. (3) Reduced nodal betweenness centrality in the right presupplementary motor area in PD.
Olde Dubbelink et al., 2014 [126]	MEG	70 PD 21 HC	78 regions in AAL atlas	Phase lag index	Weighted	(1) Lower local clustering coefficient with preserved path length in the delta frequency band in PD. (2) Decreased local clustering coefficient in multiple frequency bands with decreased path length in the alpha2 frequency band in PD.
Utianski et al., 2016 [127]	EEG	18 PD-D 57 PD-HC 57 HC	8 epochs	Phase lag index	Weighted	(1) Higher connectivity strength in the theta band in PD-HC compared with HC. (2) higher gamma, lambda, and modularity in PD-HC compared with HC. (3) lower functional connectivity in the alpha1 band in PD-D compared with PD-HC. (4) Lower gamma and lambda in the alpha1 band and higher modularity in both alpha bands in PD-D compared with PD-HC.

brain functional connectivity of PD patients, Göttlich et al. [123] constructed functional brain networks of 37 PD patients and 20 HCs. They found lower global efficiency in PD. Moreover, by analyzing brain network modules, they found out only increased connectivity within the sensorimotor network and decreased interaction of the visual network with other brain modules but also lower connectivity

between the cuneus and the ventral caudate, medial orbitofrontal cortex, and the temporal lobe in PD. In addition, they found decreased degree in the occipital lobe and increased degree in the superior parietal cortex, posterior cingulate gyrus, supramarginal gyrus, and supplementary motor area in PD. Subsequently, Luo et al. [124] continued to investigate the topological organization of functional brain networks of PD patients. They found lower clustering coefficient and local efficiency in PD patients compared with HCs. Moreover, they found reduced node centrality and connectivity strength in temporal-occipital and sensorimotor regions in PD patients. To investigate functional changes in cognitive and sensorimotor networks in PD patients, Koshimori et al. [125] used nodal degree, local efficiency, and betweenness centrality measures to assess functional brain networks of PD patients. They found that, compared with HCs, PD patients showed higher nodal degree in the right and left dorsolateral prefrontal cortex, reduced local efficiency the right mid-insula, and reduced nodal betweenness centrality in the right presupplementary motor area.

(ii) Using EEG/MEG: to explore the spatial organization of alterations in functional connectivity between brain regions, Olde Dubbelink et al. [126] used clustering coefficient and shortest path length to measure functional brain networks of 70 PD patients and 21 HCs. They found lower local clustering coefficient with preserved path length in the delta frequency band. Moreover, they found decreased local clustering coefficient in multiple frequency bands with decreased path length in the alpha2 frequency band. To determine the differences between PD patients who are healthy control (PD-HC) and HCs and between PD-HC and PD dementia (PD-D), Utianski et al. [127] used EEG data to construct functional brain networks across groups. In their study, compared with HCs, PD-HC patients showed higher connectivity strength in the theta band but no differences in other frequency bands. Moreover, PD-HC patients exhibited higher gamma, lambda, and modularity than HCs. Compared with the PD-HC group, the PD-D group showed lower functional connectivity, gamma, and lambda in the alpha1 band and higher modularity in both alpha bands.

4.4. Multiple Sclerosis. Multiple sclerosis (MS) is an inflammatory and degenerative disease of the central nervous system (CNS). It is characterized by multiple lesions mainly affecting the WM, accompanying structural and functional disconnection between various regions in the CNS, resulting in kinds of signs and symptoms.

4.4.1. Structural Brain Networks in MS. The study of structural brain networks provides another perspective for MS. In this section, we review recent progress in analyzing the structural brain networks based on sMRI and dMRI in MS as shown in Table 8.

(i) Using sMRI: to investigate the correlation between the WM lesion load and the topological efficiency of structural brain networks in MS patients, He et al. [128] divided all MS patients into six subgroups based on corresponding total WM lesion loads (TWMLL) and constructed structural brain networks across groups. They found decreased integrated absolute local and global efficiency and decreased integrated relative local efficiency with increasing TWMLL in MS. Tewarie et al. [129] constructed structural brain networks based on cortical thicknesses to investigate the topological differences of MS patients and HCs. They found higher normalized clustering coefficient and higher normalized shortest path length in MS.

(ii) Using dMRI: to investigate the alterations in the topological organization of the WM structural networks, Shu et al. [130] used DTI and deterministic tractography to constructed the WM structural networks of 39 MS patients and 39 HCs. They found that the MS patients and the HCs showed efficient small-world property in their WM structural networks. Moreover, compared with HCs, MS patients had decreased global efficiency and decreased local efficiency in the sensorimotor, visual, default-mode, and language regions. Later, Shu et al. [132] continued to investigate the topological alterations of structural networks in 41 clinically isolated syndrome (CIS) patients, 32 MS patients, and 35 HCs. They found that, compared with HCs, both CIS and MS patients showed decreased network strength, global and local efficiency, and clustering coefficient and increased shortest path length. Moreover, compared with the HCs, the MS patients exhibited increased gamma and sigma, and, compared with the CIS patients, the MS patients exhibited reduced network strength and global and local efficiency and increased shortest path length, gamma, and sigma. To explore the underlying brain mechanisms of major depression MS (MD-MS) patients and nondepressed MS (ND-MS) patients, Nigro et al. [131] used DTI data to construct structural brain networks across groups. In their study, both MS patient groups showed small-world property. In addition, MS patients exhibited increased path length compared with HCs, and MD-MS patients showed increased local path length in the right hippocampus and right amygdala compared with ND-MS patients and HCs. To investigate changes in structural connectivity in MS, Llufriu et al. [133] used FA values as connectivity strength between brain regions to construct structural brain networks of 72 MS patients and 38 HCs. In their study, compared with HCs, MS patients showed decreased transitivity and global efficiency and increased path length. Moreover, MS patients displayed decreased nodal strength in 26 of 84 gray matter regions and increased betweenness centrality in right pallidum and left insula.

Table 8: Overview of structural brain network studies in MS.

Study	Modality	Subjects	Node definition	Edge definition	Network type	Main findings
He et al., 2009 [128]	sMRI	102 MS 42 HC	54 regions	Pearson correlation based on cortical thicknesses	Binary	(1) Decreased integrated absolute local and global efficiency with increasing TWMLL in MS. (2) Decreased integrated relative local efficiency with increasing TWMLL in MS.
Tewarie et al., 2014 [129]	sMRI	102 MS 42 HC	78 regions in AAL atlas	Pearson correlation based on cortical thicknesses	Weighted	(1) Higher normalized clustering coefficient in MS. (2) Higher normalized shortest path length in MS.
Shu et al., 2011 [130]	DTI	39 MS 39 HC	90 regions in AAL atlas	FN	Weighted	(1) Decreased global efficiency in MS. (2) Decreased local efficiency in the sensorimotor, visual, default-mode, and language areas in MS.
Nigro et al., 2015 [131]	DTI	20 MD-MS 22 ND-MS 16 HC	90 regions in AAL atlas	FN	Weighted	(1) Increased path length in MS patients compared with HCs. (2) Increased local path length in the right hippocampus and right amygdala in MD-MS compared with ND-MS and HC.
Shu et al., 2016 [132]	DTI	41 CIS 32 MS 35 HC	90 regions in AAL atlas	FN	Weighted	(1) Decreased network strength, global and local efficiency, and clustering coefficient and increased shortest path length in both CIS and MS. (2) Increased gamma and sigma in MS compared with HC. (3) Reduced network strength and global and local efficiency and increased shortest path length, gamma, and sigma in MS compared with CIS.
Llufriu et al., 2017 [133]	DTI	72 MS 38 HC	84 regions in FreeSurfer	FA	Weighted	(1) Decreased transitivity and global efficiency and increased path length in MS. (2) Increased betweenness centrality in right pallidum and left insula in MS. (3) Decreased nodal strength in 26 of the 84 brain regions in MS.

4.4.2. Functional Brain Networks in MS. The study of functional brain networks provides another perspective for MS. In this section, we review recent progress in analyzing the functional brain networks based on fMRI and EEG/MEG in MS as shown in Table 9.

(i) Using fMRI: to investigate the modularity of MS patients, Gamboa et al. [134] constructed functional brain networks of MS patients and found increased modularity in MS patients compared with HCs. To explore the topological organization of functional brain network connectivity, Rocca et al. [135] constructed functional brain networks of 246 MS patients and 55 HCs. They found that, compared with HCs, MS patients lost hubs in the superior frontal gyrus and precuneus and anterior cingulum in the left hemisphere and showed new hubs in the left temporal pole and cerebellum, located at different hemisphere for basal ganglia hubs. Furthermore, MS patients exhibited decreased nodal degree in the bilateral caudate nucleus and right cerebellum. Shu et al. [132] constructed functional networks in CIS patients, MS patients, and HCs to investigate the topological alterations across groups. They found that, compared with the HCs, the MS patients showed decreased local efficiency and clustering coefficient. Moreover, they found that the CIS group had no significant differences with the other two groups in any global metrics. Later, Liu et al. [136] continued to investigate the topological organization of CIS patients and MS patients. They found that CIS patients showed intermediate global efficiency between MS patients and HCs, and global efficiency of MS patients was the lowest. In addition, MS patients exhibited lower local efficiency than HCs.

(ii) Using EEG/MEG: to investigate functional connectivity changes in MS, Schoonheim et al. [137] used MEG data to construct functional brain networks of MS patients. In their study, compared with HCs, MS

Table 9: Overview of functional brain network studies in MS.

Study	Modality	Subjects	Node definition	Edge definition	Network type	Main findings
Gamboa et al., 2014 [134]	fMRI	16 MS 20 HC	116 regions in AAL atlas	Pearson correlation	Binary	(1) Increased modularity in MS.
Rocca et al., 2016 [135]	fMRI	246 MS 55 HC	116 regions in AAL atlas	Pearson correlation	Binary	(1) Lost hubs in superior frontal gyrus, precuneus, and anterior cingulum in the left hemisphere in MS. (2) New hubs in the left temporal pole and cerebellum in MS. (3) Different hemisphere for basal ganglia hubs in MS. (4) Decreased nodal degree in the bilateral caudate nucleus and right cerebellum in MS.
Shu et al., 2016 [132]	fMRI	41 CIS 32 MS 35 HC	90 regions in AAL atlas	Pearson correlation	Weighted	(1) Decreased local efficiency and clustering coefficient in MS compared with HC. (2) No significant differences with the other two groups in any global metrics in the CIS group.
Liu et al., 2017 [136]	fMRI	35 CIS 37 MS 36 HC	90 regions in AAL atlas	Pearson correlation	Weighted	(1) CIS showed intermediate global efficiency between MS and HC. (2) Lower global efficiency and local efficiency in MS compared with HC.
Schoonheim et al., 2013 [137]	MEG	34 MS 28 HC	137 channels	Synchronization likelihood	Weighted	(1) Increased connectivity strength in theta, lower alpha, and beta bands in MS. (2) Decreased connectivity strength in the upper alpha band in MS. (3) Increased path length and clustering coefficient in the lower alpha band in MS.
Tewarie et al., 2014 [129]	MEG	102 MS 42 HC	78 regions in AAL atlas	Phase lag index	Weighted	(1) Higher normalized path length in the theta band in MS. (2) Lower normalized clustering coefficient in the alpha2 band in MS.

patients showed increased functional connectivity strength in theta, lower alpha, and beta bands and decreased functional connectivity strength in the upper alpha band. Furthermore, MS patients exhibited increased path length and clustering coefficient in the lower alpha band compared with HCs. Later, Tewarie et al. [129] constructed functional brain networks of MS patients and HCs to investigate the topological differences across groups. They found higher normalized path length in the theta band and lower normalized clustering coefficient in the alpha2 band in MS.

In addition to the above four brain disorders, brain network analysis has also been applied to other brain disorders, such as attention deficit/hyperactivity (ADHD) [165–167], epilepsy [168–170], and autism [171–173].

5. Conclusions and Outlook

In summary, the development of noninvasive neuroimaging and electrophysiological techniques (such as sMRI, dMRI, fMRI, and EEG/MEG) has enabled us to construct human brain structural and functional connectivity networks, while the complex network analysis has revealed a number of important topological properties hidden in the human brain structural and functional networks, such as small-world property, modularity, hubs, and rich club. The study of complex brain networks will not only promote the construction of the human brain connectome but also deepen our understanding of the important issues such as the information processing mode of the brain and the working mechanism of various cognitive functions. Moreover, to explore the brain network topology caused by brain disorder abnormal changes, the methods with brain network analysis have been applied to different brain disorder studies, such as the above-mentioned four brain disorders in Section 4. Brain network analysis not only provides a new perspective for revealing the pathophysiologic mechanisms of brain disorders at the system level, but also establishes some brain network neuroimaging and electrophysiological markers to describe different brain disorders. For example, for the above-mentioned four brain disorders in Section 4, most findings indicate that, compared with the HC group, the disorder group exhibited decreased small-world index and decreased global efficiency at the global level, while, at the local level, the

disorder group showed the loss of hub nodes and decreased local efficiency compared with the HC group. Therefore, brain network analysis can provide important auxiliary guidance for early diagnosis and treatment of brain disorders.

As can be seen from Section 4, although many findings are obtained for a specific brain disorder after using brain network analysis, parts of findings obtained by different studies are not consistent or even opposite. For example, Tijms et al. [96] found that, compared with the HC group, the AD group exhibited decreased small-world index, while Wang et al. [101] found that the AD group showed higher small-world index compared with the HC group; Lynall et al. [111] found that, compared with the HC group, the SCZ group exhibited decreased clustering coefficient, while Hadley et al. [113] found that the SCZ group showed increased clustering coefficient compared with the HC group. One of the main causes of the above similar opposite results is that the experimental data is too small. However, although the results of the different studies are opposite, a single property cannot be used to determine whether there is a difference between the disorder group and the HC group in the field of brain network analysis. In general, in order to determine the difference between the disorder group and the HC group, multiple properties must be used to complete it. Hence, in order to obtain more accurate brain network analysis results, there are many problems and challenges to be solved urgently as follows:

(i) How to construct a brain network that conforms to the working mechanism of the brain is a primary problem in brain network analysis. In Section 2, we introduce two basic elements of the brain network, nodes and edges, and present a variety of their common definitions, such as node definitions based on different brain atlas and edge definitions based on different correlations between two nodes. The various existing node definitions in constructing brain networks can only reflect one aspect of the brain regions themselves, such as cortical thickness and time series. Similarly, a specific edge definition can also only reflect one aspect of the connectivity between the brain regions, such as the number of fibers and Pearson correlation. Thus, how to evaluate the impact of different nodes and edges on the brain network and determine the most reasonable definition of nodes and edges is a key problem, as well as an important challenge in constructing the brain network.

(ii) Studies have shown that the brain structural and functional network topology of most brain disorders have abnormal changes, but there is still no unified conclusion on the trends and amplitudes of the brain network topology properties of various brain disorders. For example, Shu et al. [132] used DTI and fMRI to investigate the brain network topology properties of MS, and the results are inconsistent as shown in Table 5. Thus, it is urgent to solve the problem of how to integrate multimodal data to analyze and understand the pathophysiologic mechanisms of brain disorders, and establish reliable and effective neuroimaging and electrophysiological diagnostic markers in brain network research.

(iii) The structure and function of a brain are inseparable; the structure is the basis of function and the function is the representation of the structure. It has been shown that the structure and function of the human brains are closely related [174, 175]. Thus, it is challenging to combine the structure network and function network of a brain for evaluating the similarity and specificity of brain structure-function network comprehensively and understanding the effect of structural network organization on brain function formation and brain function shaping on brain structure.

(iv) The functional activity of a brain is a dynamic process, and most existing functional networks only describe the topological properties of brain function activity in a certain period of time. Thus, how to construct a dynamic brain function network to find the regularity of the brain function topology properties with time changes in a smaller time scale is one of the directions of future brain network research. It can further explore the mechanisms of brain real-time functional activities.

(v) Because of their simplicity, many researchers focused on the undirected brain networks. However, the information transmission of each activity of a brain is directional, and the undirected network analysis is unable to obtain the results with the direction of information flow. As the undirected network analysis cannot reveal the direction of information transmission hidden in the brain structural and functional networks, it cannot really reflect the real brain activities. In order to reflect the real brain activities and for more in-depth understanding of the regularity of the brain structural organization patterns and functional activities, it is necessary to construct the directional structural and functional brain networks to understand the brain activities and to further reveal how to transmit information in the brain activities. How to construct effective directed brain networks and how to carry out effective directed brain network analysis are two important problems and challenges for researchers.

(vi) As can be seen from several tables (Tables 2–9), the sample sizes of most existing brain network researches are too small, most of which are not more than 100. Therefore, the results of brain network analysis may be incomplete, or even wrong, which to some extent restricts the development of brain network analysis. Thus, in order to make the results of brain network analysis closer to reality, the sample size of the experimental data should be urgently expanded.

(vii) Most brain network studies are based on a single layer network. More recently, the multilayer network studies [176–178] have also been proposed. Since

multilayer networks can integrate different anatomical/functional types of links or different frequency bands, the multilayer network studies are likely to become one of the most promising future research directions.

Study of brain networks is a part of brain science, which incorporates a wide range of disciplines such as neuroscience and graph theory. It is reasonable to expect that brain network analysis would bring more remarkable achievements in the near future.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to express their gratitude for the support from the National Natural Science Foundation of China under Grant nos. 61232001, 61420106009, and 61622213.

References

[1] F. A. C. Azevedo, L. R. B. Carvalho, L. T. Grinberg et al., "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," *Journal of Comparative Neurology*, vol. 513, no. 5, pp. 532–541, 2009.

[2] S. Herculano-Houzel, "The human brain in numbers: a linearly scaled-up primate brain," *Frontiers in Human Neuroscience*, vol. 3, article 31, 2009.

[3] M. P. van den Heuvel, R. S. Kahn, J. Goñi, and O. Sporns, "High-cost, high-capacity backbone for global brain communication," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 109, no. 28, pp. 11372–11377, 2012.

[4] O. Sporns, "The human connectome: Origins and challenges," *NeuroImage*, vol. 80, pp. 53–61, 2013.

[5] M. G. Mattar and D. S. Bassett, *Brain network architecture: Implications for human learning*, preprint arXiv:1609.01790.

[6] J. Lehrer, "Neuroscience: Making connections," *Nature*, vol. 457, no. 7229, pp. 524–527, 2009.

[7] O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgetag, "Organization, development and function of complex brain networks," *Trends in Cognitive Sciences*, vol. 8, no. 9, pp. 418–425, 2004.

[8] J. D. Power, D. A. Fair, B. L. Schlaggar, and S. E. Petersen, " The development of Human Functional Brain Networks," *Neuron*, vol. 67, no. 5, pp. 735–748, 2010.

[9] O. Sporns, "The human connectome: a complex network," *Annals of the New York Academy of Sciences*, vol. 1224, no. 1, pp. 109–125, 2011.

[10] T. E. J. Behrens and O. Sporns, "Human connectomics," *Current Opinion in Neurobiology*, vol. 22, no. 1, pp. 144–153, 2012.

[11] D. C. van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, and K. Ugurbil, "The WU-Minn human connectome project: an overview," *NeuroImage*, vol. 80, pp. 62–79, 2013.

[12] D. S. Bassett and O. Sporns, "Network neuroscience," *Nature Neuroscience*, vol. 20, no. 3, pp. 353–364, 2017.

[13] M.-M. Poo, J.-L. Du, N. Y. Ip, Z.-Q. Xiong, B. Xu, and T. Tan, "China brain project: basic neuroscience, brain diseases, and brain-inspired computing," *Neuron*, vol. 92, no. 3, pp. 591–596, 2016.

[14] O. Sporns, "From simple graphs to the connectome: networks in neuroimaging," *NeuroImage*, vol. 62, no. 2, pp. 881–886, 2012.

[15] Y. D. Reijmer, A. Leemans, K. Caeyenberghs, S. M. Heringa, H. L. Koek, and G. J. Biessels, "Disruption of cerebral networks and cognitive impairment in Alzheimer disease," *Neurology*, vol. 80, no. 15, pp. 1370–1377, 2013.

[16] J. Liu, M. Li, J. Wang, F. Wu, T. Liu, and Y. Pan, "A survey of MRI-based brain tumor segmentation methods," *Tsinghua Science and Technology*, vol. 19, no. 6, pp. 578–595, 2014.

[17] M. E. Shenton, M. Kubicki, and N. Makris, "Understanding alterations in brain connectivity in attention-deficit/hyperactivity disorder using imaging connectomics," *Biological Psychiatry*, vol. 76, no. 8, pp. 601-602, 2014.

[18] J. Liu, M. Li, W. Lan, F. Wu, Y. Pan, and J. Wang, "Classification of Alzheimer's disease using whole brain hierarchical network," *IEEE Transactions on Computational Biology and Bioinformatics*, no. 99, 2016.

[19] C. Garcia-Ramos, J. J. Lin, T. S. Kellermann, L. Bonilha, V. Prabhakaran, and B. P. Hermann, "Graph theory and cognition: A complementary avenue for examining neuropsychological status in epilepsy," *Epilepsy & Behavior*, vol. 64, pp. 329–335, 2016.

[20] J. Liu, J. Wang, Z. Tang, B. Hu, F. Wu, and Y. Pan, "Improving Alzheimeres disease classification by combining multiple measures," *IEEE Transactions on Computational Biology and Bioinformatics*, 2017.

[21] W. de Haan, W. M. van der, H. Flier, P. F. Van Mieghem, P. Scheltens, and C. J. Stam, "Disruption of functional brain networks in alzheimers disease: what can we learn from graph spectral analysis of resting-state magnetoencephalography?" *Brain connectivity*, vol. 2, no. 2, pp. 45–55, 2012.

[22] J. Liu, M. Li, Y. Pan, F. Wu, X. Chen, and J. Wang, "Classification of Schizophrenia based on individual hierarchical brain networks constructed from structural MRI images," *IEEE Transactions on NanoBioscience*, 2017.

[23] A. Fornito, J. Yoon, A. Zalesky, E. T. Bullmore, and C. S. Carter, "General and specific functional connectivity disturbances in first-episode schizophrenia during cognitive control performance," *Biological Psychiatry*, vol. 70, no. 1, pp. 64–72, 2011.

[24] J. Liu, J. Wang, B. Hu, F. Wu, and Y. Pan, "Alzheimer's disease classification based on individual hierarchical networks constructed with 3-D," *IEEE Transactions on NanoBioscience*, vol. 16, no. 6, pp. 1–10, 2017.

[25] P. Skudlarski, K. Jagannathan, K. Anderson et al., "Brain connectivity is not only lower but different in schizophrenia: a combined anatomical and functional approach," *Biological Psychiatry*, vol. 68, no. 1, pp. 61–69, 2010.

[26] J. Liu, Y. Pan, M. Li et al., "Applications of deep learning to mri images: a survey," *Big Data Mining and Analytics*, vol. 1, no. 1, 2017.

[27] E. M. Haacke, R. W. Brown, M. R. Thompson et al., *Magnetic Resonance Imaging: Physical Principles and Sequence Design*, vol. 82, Wiley-Liss, New York, NY, USA, 1999.

[28] S. Mori, B. J. Crain, V. P. Chacko, and P. C. M. Van Zijl, "Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging," *Annals of Neurology*, vol. 45, no. 2, pp. 265–269, 1999.

[29] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional Magnetic Resonance Imaging*, vol. 1, Sinauer Associates, Sunderland, Mass, USA, 2004.

[30] N. K. Logothetis, "What we can do and what we cannot do with fMRI," *Nature*, vol. 453, no. 7197, pp. 869–878, 2008.

[31] E. Niedermeyer and F. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, Lippincott Williams & Wilkins, 2005.

[32] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain," *Reviews of Modern Physics*, vol. 65, no. 2, pp. 413–497, 1993.

[33] W. de Haan, Y. A. L. Pijnenburg, R. L. M. Strijers et al., "Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory," *BMC Neuroscience*, vol. 10, no. 101, 2009.

[34] J. M. Buldú, R. Bajo, F. Maestú et al., "Reorganization of functional networks in mild cognitive impairment," *PLoS ONE*, vol. 6, no. 5, Article ID e19584, 2011.

[35] M. Shim, D.-W. Kim, S.-H. Lee, and C.-H. Im, "Disruptions in small-world cortical functional connectivity network during an auditory oddball paradigm task in patients with schizophrenia," *Schizophrenia Research*, vol. 156, no. 2-3, pp. 197–203, 2014.

[36] E. van Diessen, T. Numan, E. van Dellen et al., "Opportunities and methodological challenges in EEG and MEG resting state functional brain network research," *Clinical Neurophysiology*, vol. 126, no. 8, pp. 1468–1481, 2015.

[37] J. Wang, X. Zuo, and Y. He, "Graph-based network analysis of resting-state functional MRI," *Frontiers in Systems Neuroscience*, vol. 4, article 16, 2010.

[38] E. T. Bullmore and D. S. Bassett, "Brain graphs: graphical models of the human brain connectome," *Annual Review of Clinical Psychology*, vol. 7, pp. 113–140, 2011.

[39] A. Fornito, A. Zalesky, and M. Breakspear, "Graph analysis of the human connectome: Promise, progress, and pitfalls," *NeuroImage*, vol. 80, pp. 426–444, 2013.

[40] J. A. Maldjian, E. M. Davenport, and C. T. Whitlow, "Graph theoretical analysis of resting-state MEG data: Identifying interhemispheric connectivity and the default mode," *NeuroImage*, vol. 96, pp. 88–94, 2014.

[41] M. Li, R. Zheng, H. Zhang, J. Wang, and Y. Pan, "Effective identification of essential proteins based on priori knowledge, network topology and gene expressions," *Methods*, vol. 67, no. 3, pp. 325–333, 2014.

[42] X. Peng, J. Wang, W. Peng, F. Wu, and Y. Pan, "Protein–protein interactions: detection, reliability assessment and applications," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 798–819, 2016.

[43] X. Peng, J. Wang, J. Huan, and F.-X. Wu, "Double-layer clustering method to predict protein complexes based on power-law distribution and protein sublocalization," *Journal of Theoretical Biology*, vol. 395, pp. 186–193, 2016.

[44] X. Peng, X. Yan, and J. Wang, "Framework to identify protein complexes based on similarity preclustering," *Tsinghua Science and Technology*, vol. 22, no. 1, pp. 42–51, 2017.

[45] O. Sporns, G. Tononi, and R. Kötter, "The human connectome: a structural description of the human brain," *PLoS Computational Biology*, vol. 1, no. 4, article e42, 2005.

[46] M. P. van den Heuvel, C. J. Stam, M. Boersma, and H. E. Hulshoff Pol, "Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain," *NeuroImage*, vol. 43, no. 3, pp. 528–539, 2008.

[47] A. Fornito, A. Zalesky, C. Pantelis, and E. T. Bullmore, "Schizophrenia, neuroimaging and connectomics," *NeuroImage*, vol. 62, no. 4, pp. 2296–2314, 2012.

[48] M. Xia and Y. He, "Functional connectomics from a "big data" perspective," *NeuroImage*, 2016.

[49] P. Hagmann, L. Cammoun, X. Gigandet et al., "MR connectomics: Principles and challenges," *Journal of Neuroscience Methods*, vol. 194, no. 1, pp. 34–45, 2010.

[50] E. C. W. van Straaten and C. J. Stam, "Structure out of chaos: Functional brain network analysis with EEG, MEG, and functional MRI," *European Neuropsychopharmacology*, vol. 23, no. 1, pp. 7–18, 2013.

[51] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[52] Q. K. Telesford, J. H. Burdette, and P. J. Laurienti, "An exploration of graph metric reproducibility in complex brain networks," *Frontiers in Neuroscience*, no. 7, article 67, 2013.

[53] F. Vecchio, F. Miraglia, D. Quaranta et al., "Cortical connectivity and memory performance in cognitive decline: A study via graph theory from EEG data," *Neuroscience*, vol. 316, pp. 143–150, 2016.

[54] L. D. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: a survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007.

[55] A. Zalesky, L. Cocchi, A. Fornito, M. M. Murray, and E. Bullmore, "Connectivity differences in brain networks," *NeuroImage*, vol. 60, no. 2, pp. 1055–1062, 2012.

[56] D. Mears and H. B. Pollard, "Network science and the human brain: Using graph theory to understand the brain and one of its hubs, the amygdala, in health and disease," *Journal of Neuroscience Research*, vol. 94, no. 6, pp. 590–605, 2016.

[57] C. T. Butts, "Revisiting the foundations of network analysis," *American Association for the Advancement of Science: Science*, vol. 325, no. 5939, pp. 414–416, 2009.

[58] K. Brodmann, *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*, Barth, 1909.

[59] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou et al., "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.

[60] D. B. Dwyer, B. J. Harrison, M. Yücel et al., "Large-scale brain network dynamics supporting adolescent cognitive control," *The Journal of Neuroscience*, vol. 34, no. 42, pp. 14096–14107, 2014.

[61] N. U. F. Dosenbach, B. Nardos, A. L. Cohen et al., "Prediction of individual brain maturity using fMRI," *Science*, vol. 329, no. 5997, pp. 1358–1361, 2010.

[62] C. Reveley, A. K. Seth, C. Pierpaoli et al., "Superficial white matter fiber systems impede detection of long-range cortical connections in diffusion MR tractography," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 112, no. 21, pp. E2820–E2828, 2015.

[63] R. E. Passingham, K. E. Stephan, and R. Kötter, "The anatomical basis of functional localization in the cortex," *Nature Reviews Neuroscience*, vol. 3, no. 8, pp. 606–616, 2002.

[64] A. Anwander, M. Tittgemeyer, D. Y. Von Cramon, A. D. Friederici, and T. R. Knösche, "Connectivity-based parcellation of Broca's area," *Cerebral Cortex*, vol. 17, no. 4, pp. 816–825, 2007.

[65] J. D. Power, A. L. Cohen, S. M. Nelson et al., "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, pp. 665–678, 2011.

[66] B. T. T. Yeo, F. M. Krienen, J. Sepulcre et al., "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *Journal of Neurophysiology*, vol. 106, no. 3, pp. 1125–1165, 2011.

[67] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Human Brain Mapping*, vol. 33, no. 8, pp. 1914–1928, 2012.

[68] D. Moreno-Dominguez, A. Anwander, and T. R. Knösche, "A hierarchical method for whole-brain connectivity-based parcellation," *Human Brain Mapping*, vol. 35, no. 10, pp. 5000–5025, 2014.

[69] M. F. Glasser, T. S. Coalson, E. C. Robinson et al., "A multimodal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, 2016.

[70] K. J. Friston, "Functional and effective connectivity: a review," *Brain Connectivity*, vol. 1, no. 1, pp. 13–36, 2011.

[71] D. S. Bassett, E. Bullmore, B. A. Verchinski, V. S. Mattay, D. R. Weinberger, and A. Meyer-Lindenberg, "Hierarchical organization of human cortical networks in health and Schizophrenia," *The Journal of Neuroscience*, vol. 28, no. 37, pp. 9239–9248, 2008.

[72] Y. Zhang, L. Lin, C.-P. Lin et al., "Abnormal topological organization of structural brain networks in schizophrenia," *Schizophrenia Research*, vol. 141, no. 2-3, pp. 109–118, 2012.

[73] A. Alexander-Bloch, J. N. Giedd, and E. Bullmore, "Imaging structural co-variance between human brain regions," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 322–336, 2013.

[74] D. S. Tuch, T. G. Reese, M. R. Wiegell, and V. J. Wedeen, "Diffusion MRI of complex neural architecture," *Neuron*, vol. 40, no. 5, pp. 885–895, 2003.

[75] D. S. Tuch, "Q-ball imaging," *Magnetic Resonance in Medicine*, vol. 52, no. 6, pp. 1358–1372, 2004.

[76] V. J. Wedeen, P. Hagmann, W.-Y. I. Tseng, T. G. Reese, and R. M. Weisskoff, "Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging," *Magnetic Resonance in Medicine*, vol. 54, no. 6, pp. 1377–1386, 2005.

[77] S. Mori and J. Zhang, "Principles of diffusion tensor imaging and its applications to basic neuroscience research," *Neuron*, vol. 51, no. 5, pp. 527–539, 2006.

[78] T. E. Conturo, N. F. Lori, T. S. Cull et al., "Tracking neuronal fiber pathways in the living human brain," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 96, no. 18, pp. 10422–10427, 1999.

[79] P. J. Basser, S. Pajevic, C. Pierpaoli, J. Duda, and A. Aldroubi, "In vivo fiber tractography using DT-MRI data," *Magnetic Resonance in Medicine*, vol. 44, no. 4, pp. 625–632, 2000.

[80] T. E. J. Behrens, H. J. Berg, S. Jbabdi, M. F. S. Rushworth, and M. W. Woolrich, "Probabilistic diffusion tractography with multiple fibre orientations: what can we gain?" *NeuroImage*, vol. 34, no. 1, pp. 144–155, 2007.

[81] J.-D. Tournier, S. Mori, and A. Leemans, "Diffusion tensor imaging and beyond," *Magnetic Resonance in Medicine*, vol. 65, no. 6, pp. 1532–1556, 2011.

[82] S. Jbabdi and H. Johansen-Berg, "Tractography: where do we go from here?" *Brain Connectivity*, vol. 1, no. 3, pp. 169–183, 2011.

[83] Z. Yao, Y. Zhang, L. Lin et al., "Abnormal cortical networks in mild cognitive impairment and Alzheimer's disease," *PLoS Computational Biology*, vol. 6, no. 11, Article ID e1001006, 2010.

[84] N. Shu, Y. Liang, H. Li et al., "Disrupted topological organization in white matter structural networks in amnestic mild cognitive impairment: relationship to subtype," *Radiology*, vol. 265, no. 2, pp. 518–527, 2012.

[85] M. R. Brier, J. B. Thomas, A. M. Fagan et al., "Functional connectivity and graph theory in preclinical Alzheimer's disease," *Neurobiology of Aging*, vol. 35, no. 4, pp. 757–768, 2014.

[86] S. Golbabaei, A. Dadashi, and H. Soltanian-Zadeh, "Measures of the brain functional network that correlate with Alzheimer's neuropsychological test scores: An fMRI and graph analysis study," in *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2016*, pp. 5554–5557, August 2016.

[87] Z. Yin, J. Li, Y. Zhang, A. Ren, K. M. Von Meneen, and L. Huang, "Functional brain network analysis of schizophrenic patients with positive and negative syndrome based on mutual information of EEG time series," *Biomedical Signal Processing and Control*, vol. 31, pp. 331–338, 2017.

[88] K. Supekar, V. Menon, D. Rubin, M. Musen, and M. D. Greicius, "Network analysis of intrinsic functional brain connectivity in Alzheimer's disease," *PLoS Computational Biology*, vol. 4, no. 6, Article ID e1000100, 2008.

[89] J. Wang, X. Zuo, Z. Dai et al., "Disrupted functional brain connectome in individuals at risk for Alzheimer's disease," *Biological Psychiatry*, vol. 73, no. 5, pp. 472–481, 2013.

[90] B. Fischl, A. van der Kouwe, C. Destrieux et al., "Automatically parcellating the human cerebral cortex," *Cerebral Cortex*, vol. 14, no. 1, pp. 11–22, 2004.

[91] C. Destrieux, B. Fischl, A. Dale, and E. Halgren, "Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature," *NeuroImage*, vol. 53, no. 1, pp. 1–15, 2010.

[92] R. S. Desikan, F. Ségonne, B. Fischl et al., "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, 2006.

[93] D. W. Shattuck, M. Mirza, V. Adisetiyo et al., "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, 2008.

[94] L. Fan, H. Li, J. Zhuo et al., "The human brainnetome atlas: a new brain atlas based on connectional architecture," *Cerebral Cortex*, vol. 26, no. 8, pp. 3508–3526, 2016.

[95] Y. He, Z. Chen, and A. Evans, "Structural insights into aberrant topological patterns of large-scale cortical networks in Alzheimer's disease," *The Journal of Neuroscience*, vol. 28, no. 18, pp. 4756–4766, 2008.

[96] B. M. Tijms, C. Möller, H. Vrenken et al., "Single-subject grey matter graphs in Alzheimer's disease," *PLoS ONE*, vol. 8, no. 3, Article ID e58921, 2013.

[97] J. B. Pereira, M. Mijalkov, E. Kakaei et al., "Disrupted network topology in patients with stable and progressive mild cognitive impairment and Alzheimer's disease," *Cerebral Cortex*, vol. 26, no. 8, pp. 3476–3493, 2016.

[98] C.-Y. Lo, P.-N. Wang, K.-H. Chou, J. Wang, Y. He, and C.-P. Lin, "Diffusion tensor tractography reveals abnormal topological organization in structural cortical networks in Alzheimer's disease," *The Journal of Neuroscience*, vol. 30, no. 50, pp. 16876–16885, 2010.

[99] F. Bai, N. Shu, Y. Yuan et al., "Topologically convergent and divergent structural connectivity patterns between patients with remitted geriatric depression and amnestic mild cognitive impairment," *The Journal of Neuroscience*, vol. 32, no. 12, pp. 4307–4318, 2012.

[100] M. Daianu, N. Jahanshad, T. M. Nir et al., "Rich club analysis in the Alzheimer's disease connectome reveals a relatively

undisturbed structural core network," *Human Brain Mapping*, vol. 36, no. 8, pp. 3087–3103, 2015.

[101] T. Wang, F. Shi, Y. Jin et al., "Multilevel deficiency of white matter connectivity networks in Alzheimer's disease: a diffusion MRI study with DTI and HARDI models," *Neural Plasticity*, vol. 2016, Article ID 2947136, 14 pages, 2016.

[102] A. Zalesky, A. Fornito, I. H. Harding et al., "Whole-brain anatomical networks: does the choice of nodes matter?" *NeuroImage*, vol. 50, no. 3, pp. 970–983, 2010.

[103] M. M. A. Engels, C. J. Stam, W. M. van der Flier, P. Scheltens, H. de Waal, and E. C. W. van Straaten, "Declining functional connectivity and changing hub locations in Alzheimer's disease: An EEG study," *BMC Neurology*, vol. 15, no. 1, article no. 145, 2015.

[104] Y. Deng, L. Shi, Y. Lei, and D. Wang, "Altered topological organization of high-level visual networks in Alzheimer's disease and mild cognitive impairment patients," *Neuroscience Letters*, vol. 630, pp. 147–153, 2016.

[105] F. Shi, P.-T. Yap, W. Gao, W. Lin, J. H. Gilmore, and D. Shen, "Altered structural connectivity in neonates at genetic risk for schizophrenia: a combined study using morphological and white matter networks," *NeuroImage*, vol. 62, no. 3, pp. 1622–1633, 2012.

[106] B. M. Tijms, E. Sprooten, D. Job et al., "Grey matter networks in people at increased familial risk for schizophrenia," *Schizophrenia Research*, vol. 168, no. 1-2, article no. 6513, pp. 1–8, 2015.

[107] A. Zalesky, A. Fornito, M. L. Seal et al., "Disrupted axonal fiber connectivity in schizophrenia," *Biological Psychiatry*, vol. 69, no. 1, pp. 80–89, 2011.

[108] G. Collin, R. S. Kahn, M. A. De Reus, W. Cahn, and M. P. Van Den Heuvel, "Impaired rich club connectivity in unaffected siblings of schizophrenia patients," *Schizophrenia Bulletin*, vol. 40, no. 2, pp. 438–448, 2014.

[109] Y. Sun, Y. Chen, S. L. Collinson, A. Bezerianos, and K. Sim, "Reduced hemispheric asymmetry of brain anatomical networks is linked to schizophrenia: A connectomics study," *Cerebral Cortex*, vol. 1, article 14, 2015.

[110] Y. Sun, Y. Chen, R. Lee, A. Bezerianos, S. L. Collinson, and K. Sim, "Disruption of brain anatomical networks in schizophrenia: A longitudinal, diffusion tensor imaging based study," *Schizophrenia Research*, vol. 171, no. 1-3, pp. 149–157, 2016.

[111] M.-E. Lynall, D. S. Bassett, R. Kerwin et al., "Functional connectivity and brain networks in schizophrenia," *The Journal of Neuroscience*, vol. 30, no. 28, pp. 9477–9487, 2010.

[112] T.-W. Su, T.-W. Hsu, Y.-C. Lin, and C.-P. Lin, "Schizophrenia symptoms and brain network efficiency: A resting-state fMRI study," *Psychiatry Research: Neuroimaging*, vol. 234, no. 2, pp. 208–218, 2015.

[113] J. A. Hadley, N. V. Kraguljac, D. M. White, L. Ver Hoef, J. Tabora, and A. C. Lahti, "Change in brain network topology as a function of treatment response in schizophrenia: a longitudinal resting-state fMRI study using graph theory," *Schizophrenia*, vol. 2, no. 1, article 16014, 2016.

[114] X. Shen, F. Tokoglu, X. Papademetris, and R. T. Constable, "Groupwise whole-brain parcellation from resting-state fMRI data for network node identification," *NeuroImage*, vol. 82, pp. 403–415, 2013.

[115] E. P. Ganella, C. F. Bartholomeusz, C. Seguin et al., "Functional brain networks in treatment-resistant schizophrenia," *Schizophrenia Research*, vol. 184, pp. 73–81, 2017.

[116] K. Jhung, S.-H. Cho, J.-H. Jang et al., "Small-world networks in individuals at ultra-high risk for psychosis and first-episode schizophrenia during a working memory task," *Neuroscience Letters*, vol. 535, no. 1, pp. 35–39, 2013.

[117] D. Zhang, J. Wang, X. Liu, J. Chen, and B. Liu, "Aberrant brain network efficiency in Parkinson's disease patients with tremor: A multi-modality study," *Frontiers in Aging Neuroscience*, vol. 7, article no. 169, 2015.

[118] J. B. Pereira, D. Aarsland, C. E. Ginestet et al., "Aberrant cerebral network topology and mild cognitive impairment in early Parkinson's disease," *Human Brain Mapping*, vol. 36, no. 8, pp. 2980–2995, 2015.

[119] S. K. Yadav, N. Kathiresan, S. Mohan et al., "Gender-based analysis of cortical thickness and structural connectivity in Parkinson's disease," *Journal of Neurology*, vol. 263, no. 11, pp. 2308–2318, 2016.

[120] C. Li, B. Huang, R. Zhang et al., "Impaired topological architecture of brain structural networks in idiopathic Parkinson's disease: a DTI study," *Brain Imaging and Behavior*, vol. 11, no. 1, pp. 113–128, 2017.

[121] S. Nigro, R. Riccelli, L. Passamonti et al., "Characterizing structural neural networks in de novo Parkinson disease patients using diffusion tensor imaging," *Human Brain Mapping*, vol. 37, no. 12, pp. 4500–4510, 2016.

[122] F. Skidmore, D. Korenkevych, Y. Liu, G. He, E. Bullmore, and P. M. Pardalos, "Connectivity brain networks based on wavelet correlation analysis in Parkinson fMRI data," *Neuroscience Letters*, vol. 499, no. 1, pp. 47–51, 2011.

[123] M. Göttlich, T. F. Münte, M. Heldmann, M. Kasten, J. Hagenah, and U. M. Krämer, "Altered resting state brain networks in Parkinson's disease," *PLoS ONE*, vol. 8, no. 10, Article ID e77336, 2013.

[124] C. Y. Luo, X. Y. Guo, W. Song et al., "Functional connectome assessed using graph theory in drug-naive Parkinson's disease," *Journal of Neurology*, vol. 262, no. 6, pp. 1557–1567, 2015.

[125] Y. Koshimori, S.-S. Cho, M. Criaud et al., "Disrupted nodal and hub organization account for brain network abnormalities in Parkinson's disease," *Frontiers in Aging Neuroscience*, vol. 8, article no. 259, 2016.

[126] K. T. E. Olde Dubbelink, A. Hillebrand, D. Stoffers et al., "Disrupted brain network topology in Parkinson's disease: a longitudinal magnetoencephalography study," *Brain*, vol. 137, no. 1, pp. 197–207, 2014.

[127] R. L. Utianski, J. N. Caviness, E. C. W. van Straaten et al., "Graph theory network function in parkinson's disease assessed with electroencephalography," *Clinical Neurophysiology*, vol. 127, no. 5, pp. 2228–2236, 2016.

[128] Y. He, A. Dagher, Z. Chen et al., "Impaired small-world efficiency in structural cortical networks in multiple sclerosis associated with white matter lesion load," *Brain*, vol. 132, no. 12, pp. 3366–3379, 2009.

[129] P. Tewarie, M. D. Steenwijk, B. M. Tijms et al., "Disruption of structural and functional networks in long-standing multiple sclerosis," *Human Brain Mapping*, vol. 35, no. 12, pp. 5946–5961, 2014.

[130] N. Shu, Y. Liu, K. Li et al., "Diffusion tensor tractography reveals disrupted topological efficiency in white matter structural networks in multiple sclerosis," *Cerebral Cortex*, vol. 21, no. 11, pp. 2565–2577, 2011.

[131] S. Nigro, L. Passamonti, R. Riccelli et al., "Structural 'connectomic' alterations in the limbic system of multiple sclerosis

patients with major depression," *Multiple Sclerosis Journal*, vol. 21, no. 8, pp. 1003–1012, 2015.

[132] N. Shu, Y. Duan, M. Xia et al., "Disrupted topological organization of structural and functional brain connectomes in clinically isolated syndrome and multiple sclerosis," *Scientific Reports*, vol. 6, Article ID 29383, 2016.

[133] S. Llufriu, E. Martinez-Heras, E. Solana et al., "Structural networks involved in attention and executive functions in multiple sclerosis," *NeuroImage: Clinical*, vol. 13, pp. 288–296, 2017.

[134] O. L. Gamboa, E. Tagliazucchi, F. von Wegner et al., "Working memory performance of early MS patients correlates inversely with modularity increases in resting state functional connectivity networks," *NeuroImage*, vol. 94, pp. 385–395, 2014.

[135] M. A. Rocca, P. Valsasina, A. Meani, A. Falini, G. Comi, and M. Filippi, "Impaired functional integration in multiple sclerosis: a graph theory study," *Brain Structure and Function*, vol. 221, no. 1, pp. 115–131, 2016.

[136] Y. Liu, H. Wang, Y. Duan et al., "Functional brain network alterations in clinically isolated syndrome and multiple sclerosis: A graph-based connectome study," *Radiology*, vol. 282, no. 2, pp. 534–541, 2017.

[137] M. M. Schoonheim, J. J. G. Geurts, D. Landi et al., "Functional connectivity changes in multiple sclerosis patients: A graph analytical study of MEG resting state data," *Human Brain Mapping*, vol. 34, no. 1, pp. 52–61, 2013.

[138] N. A. Crossley, A. Mechelli, J. Scott et al., "The hubs of the human connectome are generally implicated in the anatomy of brain disorders," *Brain*, vol. 137, no. 8, pp. 2382–2395, 2015.

[139] M. P. van den Heuvel and O. Sporns, "Network hubs in the human brain," *Trends in Cognitive Sciences*, vol. 17, no. 12, pp. 683–696, 2013.

[140] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[141] M. A. Beauchamp, "An improved index of centrality," *Behavioural Science*, vol. 10, pp. 161–163, 1965.

[142] J. M. Anthonisse, "The rush in a directed graph," *Stichting Mathematisch Centrum. Mathematische Besliskunde*, no. BN 9/71, pp. 1–10, 1971.

[143] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[144] S. Achard and E. Bullmore, "Efficiency and cost of economical brain functional networks," *PLoS Computational Biology*, vol. 3, no. 2, p. e17, 2007.

[145] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical Review Letters*, vol. 87, no. 19, Article ID 198701, 2001.

[146] O. Sporns, C. J. Honey, and R. Kötter, "Identification and classification of hubs in brain networks," *PLoS ONE*, vol. 2, no. 10, Article ID e1049, 2007.

[147] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 2, Article ID 026113, pp. 1–26113, 2004.

[148] A. Fornito, A. Zalesky, and E. Bullmore, *Fundamentals of Brain Network Analysis*, Academic Press, 2016.

[149] S. Zhou and R. J. Mondragón, "The rich-club phenomenon in the internet topology," *IEEE Communications Letters*, vol. 8, no. 3, pp. 180–182, 2004.

[150] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, "Detecting rich-club ordering in complex networks," *Nature Physics*, vol. 2, no. 2, pp. 110–115, 2006.

[151] M. Kocher, E. Gleichgerrcht, T. Nesland et al., "Individual variability in the anatomical distribution of nodes participating in rich club structural networks," *Frontiers in Neural Circuits*, vol. 9, article no. 16, 2015.

[152] M. P. van den Heuvel and O. Sporns, "Rich-club organization of the human connectome," *The Journal of Neuroscience*, vol. 31, no. 44, pp. 15775–15786, 2011.

[153] M. Senden, G. Deco, M. A. De Reus, R. Goebel, and M. P. Van Den Heuvel, "Rich club organization supports a diverse set of functional network configurations," *NeuroImage*, vol. 96, pp. 174–182, 2014.

[154] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[155] M. D. Humphries, K. Gurney, and T. J. Prescott, "The brainstem reticular formation is a small-world, not scale-free, network," *Proceedings of the Royal Society of London B: Biological*, vol. 273, no. 1585, pp. 503–511, 2006.

[156] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: structure and dynamics," *Physics Reports*, vol. 424, no. 4-5, pp. 175–308, 2006.

[157] A. Association, "2014 Alzheimers Disease Facts and Figures," *Alzheimers & Dementia*, vol. 10, no. 2, pp. e47–e92, 2014.

[158] R. C. Petersen, "Mild cognitive impairment as a diagnostic entity," *Journal of Internal Medicine*, vol. 256, no. 3, pp. 183–194, 2004.

[159] B. M. Tijms, A. M. Wink, W. de Haan et al., "Alzheimer's disease: connecting findings from graph theoretical studies of brain networks," *Neurobiology of Aging*, vol. 34, no. 8, pp. 2023–2036, 2013.

[160] M. P. Van Den Heuvel and A. Fornito, "Brain networks in schizophrenia," *Neuropsychology Review*, vol. 24, no. 1, pp. 32–48, 2014.

[161] A. L. Wheeler and A. N. Voineskos, "A review of structural neuroimaging in schizophrenia: From connectivity to connectomics," *Frontiers in Human Neuroscience*, vol. 8, article no. 653, 2014.

[162] R. Zhang, Q. Wei, Z. Kang et al., "Disrupted brain anatomical connectivity in medication-naïve patients with first-episode schizophrenia," *Brain structure & function*, vol. 220, no. 2, pp. 1145–1159, 2015.

[163] A. W. Willis, "Parkinson disease in the elderly adult.," *Missouri Medicine*, vol. 110, no. 5, pp. 406–410, 2013.

[164] C. P. Weingarten, M. H. Sundman, P. Hickey, and N.-K. Chen, "Neuroimaging of Parkinson's disease: Expanding views," *Neuroscience & Biobehavioral Reviews*, vol. 59, pp. 16–52, 2015.

[165] Q. Cao, N. Shu, L. An et al., "Probabilistic diffusion tractography and graph theory analysis reveal abnormal white matter structural connectivity networks in drug-naive boys with attention deficit/hyperactivity disorder," *The Journal of Neuroscience*, vol. 33, no. 26, pp. 10676–10687, 2013.

[166] G. Alba, E. Pereda, S. Mañas et al., "The variability of EEG functional connectivity of young ADHD subjects in different resting states," *Clinical Neurophysiology*, vol. 127, no. 2, pp. 1321–1330, 2016.

[167] R. Beare, C. Adamson, M. A. Bellgrove et al., "Altered structural connectivity in ADHD: a network based analysis," *Brain Imaging and Behavior*, vol. 11, no. 3, pp. 846–858, 2017.

[168] E. Van Diessen, W. J. E. M. Zweiphenning, F. E. Jansen, C. J. Stam, K. P. J. Braun, and W. M. Otte, "Brain network organization in focal epilepsy: A systematic review and meta-analysis," *PLoS ONE*, vol. 9, no. 12, Article ID e114606, 2014.

[169] B. C. Bernhardt, L. Bonilha, and D. W. Gross, "Network analysis for a network disorder: The emerging role of graph theory in the study of epilepsy," *Epilepsy & Behavior*, vol. 50, pp. 162–170, 2015.

[170] J. Wirsich, A. Perry, B. Ridley et al., "Whole-brain analytic measures of network communication reveal increased structure-function correlation in right temporal lobe epilepsy," *NeuroImage: Clinical*, vol. 11, pp. 707–718, 2016.

[171] J. D. Rudie, J. A. Brown, D. Beck-Pancer et al., "Altered functional and structural brain network organization in autism," *NeuroImage: Clinical*, vol. 2, no. 1, pp. 79–94, 2013.

[172] L. M. Hernandez, J. D. Rudie, S. A. Green, S. Bookheimer, and M. Dapretto, "Neural signatures of autism spectrum disorders: Insights into brain network dynamics," *Neuropsychopharmacology*, vol. 40, no. 1, pp. 171–189, 2015.

[173] J. D. Lewis, A. C. Evans, J. R. Pruett et al., "The Emergence of Network Inefficiencies in Infants With Autism Spectrum Disorder," *Biological Psychiatry*, vol. 82, no. 3, pp. 176–185, 2017.

[174] H.-J. Park and K. Friston, "Structural and functional brain networks: from connections to cognition," *Science*, vol. 342, no. 6158, Article ID 1238411, 2013.

[175] A. Messé, D. Rudrauf, H. Benali, and G. Marrelec, "Relating Structure and Function in the Human Brain: Relative Contributions of Anatomy, Stationary Dynamics, and Non-stationarities," *PLoS Computational Biology*, vol. 10, no. 3, Article ID e1003530, 2014.

[176] M. De Domenico, S. Sasai, and A. Arenas, "Mapping multiplex hubs in human functional brain networks," *Frontiers in Neuroscience*, vol. 10, article no. 326, 2016.

[177] F. Battiston, V. Nicosia, M. Chavez, and V. Latora, "Multilayer motif analysis of brain networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 4, article 047404, 2017.

[178] M. De Domenico, "Multilayer modeling and analysis of human brain networks," *Giga Science*, vol. 6, no. 5, pp. 1–8, 2017.

Hindawi
Complexity
Volume 2017, Article ID 8917258, 14 pages
https://doi.org/10.1155/2017/8917258

Research Article

Building Up a Robust Risk Mathematical Platform to Predict Colorectal Cancer

Le Zhang,[1,2,3] Chunqiu Zheng,[2] Tian Li,[2] Lei Xing,[4] Han Zeng,[2] Tingting Li,[3] Huan Yang,[5] Jia Cao,[5] Badong Chen,[4] and Ziyuan Zhou[6]

[1]*College of Computer Science, Sichuan University, Chengdu 610065, China*
[2]*College of Computer and Information Science, Southwest University, Chongqing 400715, China*
[3]*College of Mathematics and Statistics, Southwest University, Chongqing 400715, China*
[4]*School of Electronic and Information Engineering, Xi'an Jiaotong University, 28 Xianning West Road, Beilin District, Xi'an 710049, China*
[5]*Toxicology Institute, College of Preventive Medicine, Third Military Medical University, 30 Gaotanyan Street, Shapingba District, Chongqing 400038, China*
[6]*Department of Environment Health, College of Preventive Medicine, Third Military Medical University, 30 Gaotanyan Street, Shapingba District, Chongqing 400038, China*

Correspondence should be addressed to Badong Chen; chenbd@mail.xjtu.edu.cn and Ziyuan Zhou; ziyuanzhou@tmmu.edu.cn

Received 30 April 2017; Revised 10 July 2017; Accepted 17 August 2017; Published 16 October 2017

Academic Editor: Fang-Xiang Wu

Colorectal cancer (CRC), as a result of a multistep process and under multiple factors, is one of the most common life-threatening cancers worldwide. To identify the "high risk" populations is critical for early diagnosis and improvement of overall survival rate. Of the complicated genetic and environmental factors, which group is mostly concerning colorectal carcinogenesis remains contentious. For this reason, this study collects relatively complete information of genetic variations and environmental exposure for both CRC patients and cancer-free controls; a multimethod ensemble model for CRC-risk prediction is developed by employing such big data to train and test the model. Our results demonstrate that (1) the explored genetic and environmental biomarkers are validated to connect to the CRC by biological function- or population-based evidences, (2) the model can efficiently predict the risk of CRC after parameter optimization by the big CRC-related data, and (3) our innovated heterogeneous ensemble learning model (HELM) and generalized kernel recursive maximum correntropy (GKRMC) algorithm have high prediction power. Finally, we discuss why the HELM and GKRMC can outperform the classical regression algorithms and related subjects for future study.

1. Introduction

During past decades, new strategies are developed to decrease the incidence and to improve the prognosis of colorectal cancer (CRC), from popularizing regular screening in individuals older than 50 years for prevention to taking some new technologies like laparoscopic surgery, neoadjuvant chemotherapies, and bio-targeted therapy into consideration for more precise and individualized treatment. However, CRC is still one of the important contributors to cancer worldwide [1–7]. CRC ranks 4 in cancer incidences and accounts for approximately 8–10% cancer-related death [8], and the 5-year survival rate (40–50%) is still not as satisfied as expected. CRC is now recognized as a result of multistep process under very complicated gene-environment interactions; either genetic variation and environmental factors or dietary pattern and unfavorable lifestyle may jointly play the important roles in colorectal neoplasia [9–12]. Accordingly, to efficiently identify CRC-risk factors is the first step for prevention and early diagnosis which is critical for decreasing CRC morbidity and mortality [13, 14]. Based on this hypothesis, a consortium that includes institutions from South Korea, Japan, and China cooperatively performs a multicenter case-control study (KOJACH study) during 2000–2004 to explore the CRC-risk factors in East Asia populations [15–18]. In this cooperative study, information of family history, life styles,

food, nutrition intakes, and single nucleotide polymorphisms (SNPs) of each participant is collected for both CRC cases and cancer-free controls. Then this study plans to develop such a CRC predictive model that can not only investigate which potential risk factors have the significant impact on the occurrence of CRC regarding the collected data but also efficiently and reliably predict the risk of CRC before being diagnosed as early as possible.

There are some mathematical models already developed and used to process different type of data for CRC occurrence prediction. For low dimensional data, Wu et al. [19] and Huang et al. [20] propose the logistic regression and the greedy Bayesian model. To process high dimensional dichotomous data, Hahn and his colleagues [21–23] propose to use multifactor dimensionality reduction (MDR) method for mapping them into the low dimensional space and Li et al. [24] propose a novel forward U test to estimate the possibility of the risk of CRC. In addition, Andrew et al. [25], Meredith et al. [26], and Rutledge et al. [27] employ the linear regression models to predict the occurrence of CRC. However, these previous models cannot simultaneously process our big high dimensional CRC data with both continuous and discrete data type to obtain enough high predictive accuracy.

For this reason, to avoid the shortcomings of the previous research when they are used for such complicated data collected in the KOJACH study as mentioned above, we propose a robust CRC cancer predictive model based on our latest study [28] with the following three innovations. Firstly, we use a common standard to collect clinical CRC data with information of genetic variations and environmental exposure [29], since the quickly collected high dimensional data not only have the large volume including 369 CRC patients and 929 cancer-free controls, but also have 305 data types. Secondly, the biological classification, dimensionality reduction, and regression analysis stages are integrated into the CRC predictive model to make it robust and reliable. Thirdly, both heterogeneous ensemble learning model (HELM) and a generalized kernel recursive maximum correntropy (GKRMC) algorithm are developed to increase the predictive accuracy of the model.

The research results indicate that (1) both genetic and environmental related factors play the significant role in the occurrence of CRC; (2) CRC risk can be accurately and efficiently identified with this model by using these explored biomarkers as the classifiers; and (3) our innovated HELM and GKRMC have higher predictive power than the classical regression algorithms.

Finally, we analyze the outperformance reasons for both HELM and GKRMC algorithm and discuss the future study for the CRC predictive model.

2. Materials and Methods

The data used in this study is from the hospital-based case-control study of colorectal cancer in Chongqing, China, by the Department of Toxicology at the Third Military Medical University [18]. The clinical case data is comprised of 369 pathologically diagnosed colorectal cancer patients. The control data consists of 929 cancer-free patients with frequency matched by age, gender, and birthplace. All controls are selected from the orthopedics and general surgery department of the same hospitals and those who have cancer history or any cancer-related diseases are excluded. All recruitments sign a written informed consent.

Food intake is evaluated by our previously developed Semi-Quantitative Food Frequency Questionnaire [30]. The SNP information of full-length genes plus 2,000 bp in the upper stream of each candidate gene is obtained from the HapMap [31]. After setting the minor allele frequency at 0.01 [32], the Haploview software [33] is used to screen the tag SNPs and only one SNP is selected in each of linkage disequilibrium blocks. As a result, there is a total of 46 tag SNPs from the 127 reported SNPs of the three key alcohol-metabolism genes (ADH1B, ALDH2, and CYP2E1) [34–36]. DNA is extracted from 2.5 mL whole blood according to the manufacturer's instructions of Promega DNA Purification Wizard kit. The DNA purification and Polymerase Chain Reactions (PCR) are done by Eppendorf 5333 Mastercycler. Genotyping of the selected TagSNPs is done by ABI 3130xl Gene Analyzer. This study protocol is approved by the Third Military Medical University Ethics Committee.

The items in the dataset include general information (such as gender and age), polymorphism distribution of genes related to ethanol metabolism (the distribution of homozygotes and heterozygotes of gene loci), and demographic characteristics, food, and lifestyle habits (smoking and alcohol consumption). To avoid any bias, a standard questionnaire is generated in which each survey item has a specific definition. The examination is carried out as a face-to-face query. Several survey items, such as the amount of alcohol and cigarettes consumed, are quantitatively estimated. Using age 60 as the demarcation point, the surveyed patients are divided into the elderly group and the young/middle-aged group. Alcohol consumption is divided into healthy drinking (including people who do not drink and people who drink no more than 15 g per day) and nonhealthy drinking (including people who drink more than 15 g per day). Based on smoking habits, the participants are divided into nonsmokers and smokers (including those who had quit smoking).

This study employs these data to build the predictive CRC model with biological classification, dimensionality reduction, and regression analysis stages, which will be illustrated in detail in the next section.

2.1. Biological Classification. The biological classification is carried out from the perspective of medical science to divide the original dataset into four subclasses, which are as follows: (1) polymorphism distribution of genes related to ethanol metabolism: the data of the SNPs are listed in Supplementary S1 in Supplementary Material available online at https://doi.org/10.1155/2017/8917258; (2) demographic characteristics information: the data of the demographic characteristics are listed in Supplementary S2; (3) lifestyle habits: the data of the lifestyles are listed in Supplementary S3; (4) food: the data of the foods are listed in Supplementary S4.

2.2. Dimensionality Reduction for the Original Data. This study employs three broadly used dimensionality reduction

methods, namely, principal component analysis, entropy of information, and relief method to obtain the mutually explored biomarkers for each subclass.

(1) Sparse Principal Component Analysis (SPCA) Method. Principal component analysis (PCA) [37–39] is a dimensionality reduction technique to ease complexity in multivariate data analyses by replacing the original variables with a small group of principal components. SPCA uses the Lasso [40] to produce modified principal components with sparse loadings. PCs are the uncorrelated linear combinations of original variables ranked by their variances in the descending order:

$$\begin{aligned} &\mathrm{PC}_i = l_{1i}X_1 + l_{2i}X_2 + \cdots + l_{mi}X_m \\ &\max \quad \left(\mathrm{var}\left(\mathrm{PC}_i\right)\right) \\ &\text{s.t.} \quad \sum_{j=1}^{m} l_{ji}^2 = 1, \\ &\qquad \sum_{j=1}^{m} l_{ji} \cdot l_{jk} = 0, \\ &\qquad\qquad\qquad\qquad 0 \le k < i, \end{aligned} \tag{1}$$

where $X_1, X_2, \ldots, X_m$ are the original variables and $l_{1i}, l_{2i}, \ldots, l_{mi}$ are the coefficients of principal components PC_i corresponding to the original variables estimated by the R-system packages.

(2) Entropy Method. Entropy measures the uncertainty associated with a random variable [41–43] as

$$H(X) = -E\left[\log_p(X)\right] = -\sum_{x\in\chi} p(x)\log p(x), \tag{2}$$

where $p(x) = P(X = x), x \in \chi$, is the probability mass function of the random variable X and χ is a finite set (e.g., $\{1, 2, \ldots, n\}$) or an enumerable infinite set (e.g., $\{1, 2, \ldots\}$). High entropy $H(X)$ indicates high uncertainty about the random variable X.

(3) Relief Method. Relief algorithm [44] is applied to classification of two kinds of data. Relief is a kind of feature weighting algorithm, which gives different weights according to the relevance of features and categories. Also, the relevance of features and categories in relief algorithms is based on the ability of features to distinguish between close samples. Relief algorithm process is as follows:

$$w_i = w_i + \left|x^{(i)} - \mathrm{NM}^{(i)}(x)\right| + \left|x^{(i)} - \mathrm{NH}^{(i)}(x)\right|, \quad \text{for } i = 1 : T. \tag{3}$$

The key idea of relief is to iteratively estimate feature weights according to their ability to discriminate between neighboring patterns. In each of the iterations, a pattern x is randomly selected and then two nearest neighbors of x are found, one from the same class (termed the nearest hit or NH) and the other from a different class (termed the nearest miss or NM). w_i represents the weight of the ith feature.

2.3. Regression Analysis. After biological classification and data dimensional reduction stages, we used the logistic regression (LR), support vector machine (SVM), heterogeneous ensemble learning model (HELM), kernel recursive lease squares (KRLS) [45], and our innovated generalized kernel recursive maximum correntropy (GKRMC) algorithm to build up the predictive regression model.

(1) Logistic Regression. The logistic regression (LR) [46, 47] (see (4)) can be considered as a type of semilinear regression (Huang et al., 2006), which assumes that dependent variable has 0 and 1 states.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{4}$$

where $x_1, x_2, \ldots, x_k$ are covariates and $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown coefficients for the covariates and p is the probability of the dependent variable equaling a "success" or "case."

(2) Support Vector Machine. Support vector machine (SVM) [48] is a machine learning method proposed by Vapnik in the early 1990s and successively extended by other researchers. The general form of the equation of the separating line is given as

$$f(x) = (W \cdot X) + b, \tag{5}$$

where $(W \cdot X)$ represents the inner product of the vector W and the X vector. If the linear discriminator function is normalized so that all samples meet $|f(x) \ge 1|$, then the margin between the classification face $(W \cdot X) + b = 1$ and $(W \cdot X) + b = -1$ is $2\|W\|$ (namely, the classification interval).

Minimizing the distance $2/\|W\|$, it is equivalent to maximizing $1/2||W||^2$, and then we can get the optimal classification face. Thus, the problem of seeking the optimal classification face is transformed into the following optimization problem:

$$\min \quad \frac{1}{2}w'w + c\sum_{i=1:N}\xi_i. \tag{6}$$

(3) Heterogeneous Ensemble Learning Model (HELM). Ensemble learning [49] employs multiple learners to solve a problem. The generalization ability of an ensemble is usually significantly better than that of a single learner [50]. The adaboost algorithm [51] is a type of ensemble learning. Based on previous studies, most of the ensemble learning algorithms are the integration of several of the same (homomorphic ensemble) or different (anomaly ensemble) weak classifiers. Here we propose such a HELM algorithm based on the adaboost algorithm that integrates the advantages of both homomorphic and anomaly ensemble. HELM algorithm process is illustrated in Figure 1.

Input. Sample set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x_n is the examples and $y_n \in \{0, 1\}$ is the label; weak classifier $\mathscr{L} \in \{\mathscr{L}_1 = \mathrm{svm}, \mathscr{L}_2 = \text{logistic regression}, \mathscr{L}_3 = \mathrm{KRLS}\}$. T is the iteration number.

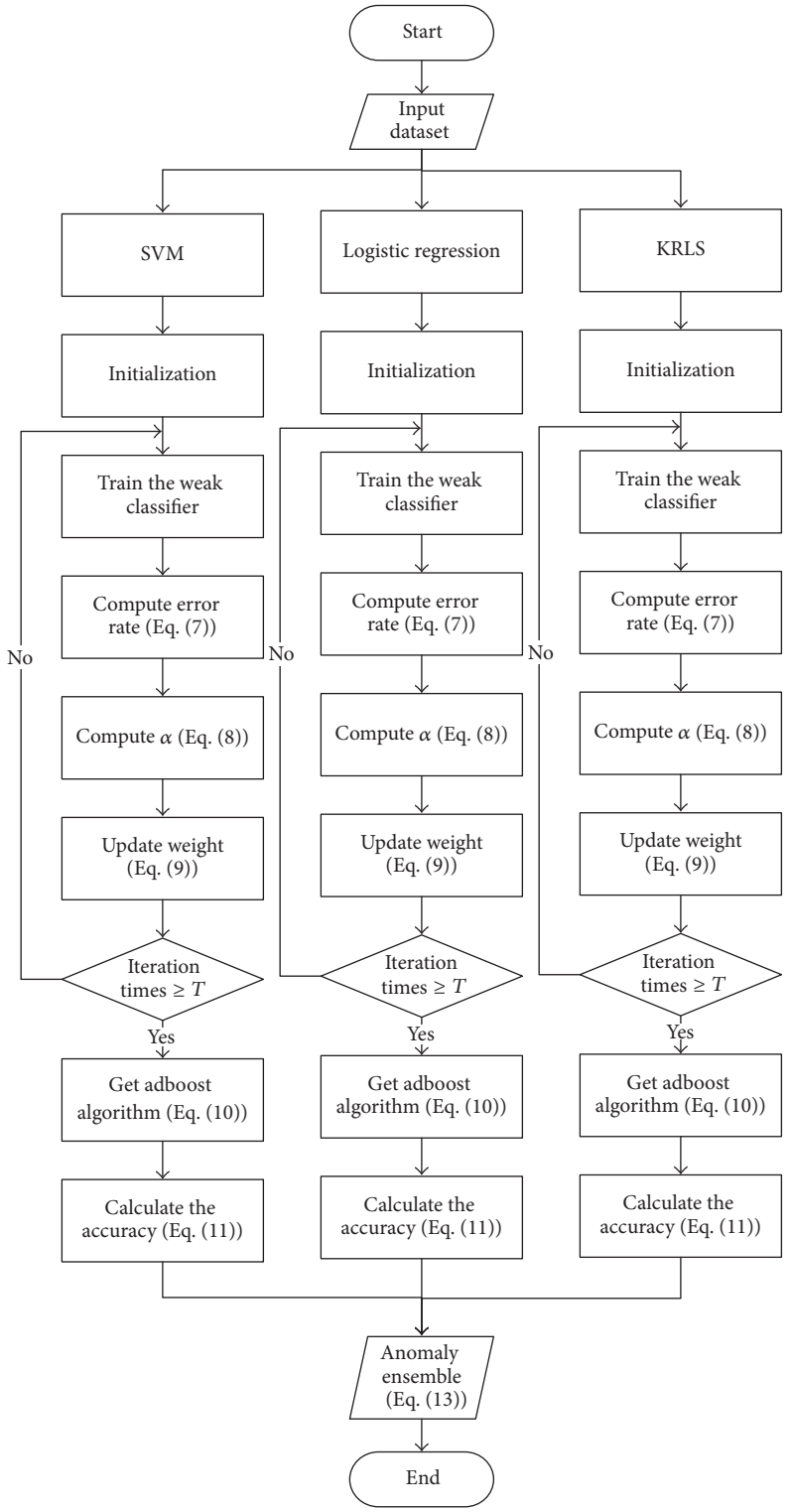


Figure 1: Workflow of HELM algorithm.

Process

(1) For $m = 1, \ldots, \mathscr{L}$,

(2) initialize the weight distribution $D_1(i) = 1/n$ (n is the number of examples; i is the index of the example),

(3) for $t = 1, \ldots, T$

(4) based on the sample distribution D_t and $\mathscr{L}_m$, we train the weak classifier h_t,

(5) compute the error (ε_t) for h_t

$$\varepsilon_t = \frac{\text{number of incorrectly classified example}}{\text{total number of examples}}, \quad (7)$$

(6) compute the weight (α_t) for h_t

$$\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}, \quad (8)$$

(7) update the weight for each sample

$$D_{t+1}(i) = \frac{D_t(i)}{\text{sum}(D)} \begin{cases} \exp(-\alpha_t), & \text{if } h_t(x_i) = y_i, \\ \exp(\alpha_t), & \text{if } h_t(x_i) \neq y_i, \end{cases} \quad (9)$$

(8) end,

(9) obtain the ensemble learning classifier H_m by adboost algorithm [49, 50]

$$H_m(x) = \text{sign}(f(x)) = \text{sign}\sum_{t=1}^{T}\alpha_t h_t(x), \quad (10)$$

(10) calculate the accuracy of H_m

$$P_{H_m} = \frac{\text{number of correctly classified example}}{\text{total number of examples}}, \quad (11)$$

(11) end,

(12) assign a weight w_{H_m} to each H_m

$$w_{H_m} = \frac{P_{H_m}}{P_{H_1} + P_{H_2} + P_{H_3}}. \quad (12)$$

Output. Anomaly ensemble:

$$\text{HELM}(x) = \text{sign}\sum_{m=1}^{3} w_{H_m} H_m(x). \quad (13)$$

(4) Generalized Kernel Recursive Maximum Correntropy (GKRMC) Algorithm. It is well known that linear regression models can quickly estimate the occurrence rate of CRC. Nonetheless, using nonlinear model should sacrifice the computing cost to obtain the high predictive accuracy. Regarding the nature of our collected data, this study developed a nonlinear regression algorithm, GKRMC (Pseudocode 1), which can significantly increase the predictive accuracy with a reasonable computing cost. GKRMC is based on the kernel recursive least squares (KRLS) algorithm [45, 52–55] and the novel concept of the generalized correntropy [56]. Equation (14) gives the corresponding weighted and regularized cost function.

$$J = \max_{\Omega}\sum_{i=1}^{j}\beta^{i-j}G_{\alpha,\beta}\left(d_j - \Omega^T\varphi_j\right) - \frac{1}{2}\beta^i\gamma_2\left|\left|\Omega\right|\right|^2, \quad (14)$$

where $G_{\alpha,\beta}(\varepsilon) = (\alpha/2\beta\Gamma(1/\alpha))\exp(-|\varepsilon/\beta|^\alpha) = \gamma_{\alpha.\beta}\exp(-\lambda|\varepsilon|^\alpha)$, $\Gamma(\cdot)$ is the gamma function, $\alpha > 0$ is the shape parameter, β is the forgetting factor and it is set to 1, φ_i stands for $\varphi(u_i)$, with φ being the nonlinear mapping induced by a Mercer kernel, γ_2 is the regularization factor, i, j denote the numerical order of the samples, and $\gamma_{\alpha,\beta} = \alpha/(2\beta\Gamma(1/\alpha))$ is the normalization constant. Setting its gradient with respect to Ω equal to zero, one can obtain the solution as

$$\Omega_i = \left(\Phi_i B_i \Phi_i^T + \gamma_2\beta^i{\sigma_1}^\alpha\mathbf{I}\right)^{-1}\Phi_i B_i d_i, \quad (15)$$

where $\Phi_i = [\varphi_1, \varphi_2, \ldots, \varphi_i]$, $\sigma_1 = \beta^{\alpha/2}$ and $\mathbf{I}$ is an identity matrix.

$$B_i = \text{diag}\begin{bmatrix} \beta^{i-1}\left(d_1 - \Omega^T\varphi_1\right)^{\alpha-2} \times \left(\frac{\alpha^2}{4\sigma_1\Gamma(1/\alpha)}\right) \times \exp\left(-\left|\frac{d_1 - \Omega^T\varphi_1}{\sigma_1}\right|^\alpha\right) \\ \beta^{i-2}\left(d_2 - \Omega^T\varphi_2\right)^{\alpha-2} \times \left(\frac{\alpha^2}{4\sigma_1\Gamma(1/\alpha)}\right) \times \exp\left(-\left|\frac{d_2 - \Omega^T\varphi_2}{\sigma_1}\right|^\alpha\right) \\ \vdots \\ \left(d_1 - \Omega^T\varphi_1\right)^{\alpha-2} \times \left(\frac{\alpha^2}{4\sigma_1\Gamma(1/\alpha)}\right) \times \exp\left(-\left|\frac{d_i - \Omega^T\varphi_i}{\sigma_1}\right|^\alpha\right) \end{bmatrix}. \quad (16)$$

Generalized Kernel Recursive Maximum Correntropy
Initialization:
$\mathbf{Q}_1 = (\gamma_2 \beta {\sigma_1}^{\alpha} + G_{\sigma_2}(\mathbf{u}_1 - \mathbf{u}_1))^{-1}$
$\mathbf{a}_1 = \mathbf{Q}_1 d_1$
Computation:
Iterate for $i > 1$:
$\mathbf{h}_i = [G_{\sigma_2}(\mathbf{u}_i - \mathbf{u}_1), \ldots, G_{\sigma_2}(\mathbf{u}_i - \mathbf{u}_{i-1})]^T$
$y_i = \mathbf{h}_i^T \mathbf{a}_{i-1}$
$e_i = d_i - y_i$
$\mathbf{z}_i = \mathbf{Q}_{i-1}\mathbf{h}_i$
$\theta_i = (\exp(-{e_i}^{\alpha}/2{\sigma_1}^2))^{-1}$
$r_i = \gamma_2 \beta^i {\sigma_1}^{\alpha}\theta_i + G_{\sigma_2}(\mathbf{u}_i - \mathbf{u}_i) - \mathbf{z}_i^T \mathbf{h}_i$
$\mathbf{Q}_i = r_i^{-1} \begin{bmatrix} \mathbf{Q}_{i-1} r_i + \mathbf{z}_i \mathbf{z}_i^T & -\mathbf{z}_i \\ -\mathbf{z}_i^T & 1 \end{bmatrix}$
$\mathbf{a}_i = \begin{bmatrix} \mathbf{a}_{i-1} - \mathbf{z}_i r_i^{-1} e_i \\ r_i^{-1} e_i \end{bmatrix}$

Pseudocode 1: Pseudocode of GKRMC.

Using the matrix inversion lemma [54], we have

$$\left(\Phi_i B_i \Phi_i^T + \gamma_2 \beta^i {\sigma_1}^{\alpha} \mathbf{I}\right)^{-1} \Phi_i B_i = \Phi_i \left(\Phi_i^T \Phi_i + \gamma_2 \beta^i {\sigma_1}^{\alpha} B_i^{-1}\right)^{-1}. \tag{17}$$

Substituting (17) into (15) yields

$$\Omega_i = \Phi_i \left(\Phi_i^T \Phi_i + \gamma_2 \beta^i {\sigma_1}^{\alpha} B_i^{-1}\right)^{-1} d_i. \tag{18}$$

The weight vector can be expressed explicitly as a linear combination of the transformed data; that is, $\Omega_i = \Phi_i a_i$, where the coefficients vector $a_i = (\Phi_i^T \Phi_i + \gamma_2 \beta^i \sigma_1^{\alpha} B_i^{-1})^{-1} d_i$ can be computed using the kernel trick. Denote $Q_i = (\Phi_i^T \Phi_i + \gamma_2 \beta^i \sigma_1^{\alpha} B_i^{-1})^{-1}$; we have

$$Q_i = \begin{bmatrix} \Phi_{i-1}^T \Phi_{i-1} + \gamma_2 \beta^i \sigma_1^{\alpha} B_{i-1}^{-1} & \Phi_{i-1}^T \varphi_i \\ \varphi_i^T \Phi_{i-1} & \varphi_i^T \varphi_i + \gamma_2 \beta^i \sigma_1^{\alpha} \theta_i \end{bmatrix}^{-1}, \tag{19}$$

where $\theta_i = (d_i - \Omega^T \varphi_i)^{\alpha-2} \times (\alpha^2 / 2\sigma_1 \Gamma(1/\alpha)) \times \exp(-|(d_i - \Omega^T \varphi_i)/\sigma_1|^{\alpha})$. It is easy to observe that

$$Q_i^{-1} = \begin{bmatrix} Q_{i-1}^{-1} & h_i \\ h_i^T & \varphi_i^T \varphi_i + \gamma_2 \beta^i \sigma_1^{\alpha} \theta_i \end{bmatrix}, \tag{20}$$

where $h_i = \Phi_{i-1}^T \varphi_i$. Using the block matrix inversion identity, we can derive

$$Q_i^{-1} = \begin{bmatrix} Q_{i-1} r_i + z_i z_i^T & -z_i \\ -z_i^T & 1 \end{bmatrix}, \tag{21}$$

where $z_i = Q_{i-1} h_i$ and

$$r_i = \gamma_2 \beta^i \sigma_1^{\alpha} \theta_i + \varphi_i^T \varphi_i - z_i^T h_i. \tag{22}$$

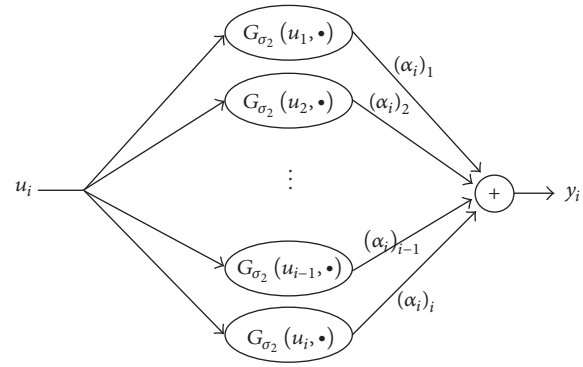


Figure 2: Network topology of GKRMC at ith iteration.

So,

$$a_i = Q_i d_i = r_i^{-1} \begin{bmatrix} Q_{i-1} r_i + z_i z_i^T & -z_i \\ -z_i^T & 1 \end{bmatrix} \begin{bmatrix} d_{i-1} \\ d_i \end{bmatrix} = \begin{bmatrix} a_{i-1} - z_i r_i^{-1} e_i \\ r_i^{-1} e_i \end{bmatrix}, \tag{23}$$
$$e_i = d_i - \Omega^T \varphi_i.$$

Then we obtain the GKRMC algorithm, in which the coefficients update follows (23) and r_i is computed by (22). This study uses $G_{\sigma_2}(\cdot)$ to denote the Gaussian kernel for RKHS [57], with σ_2 being the kernel size. The GKRMC produces a RBF [58] type network, which is a linear combination of the kernel functions (Figure 2). a_i denotes the coefficient vector of the network at iteration i and $(a_i)_j$ denotes the jth scalar in a_i.

3. Results

3.1. The Results of the Biological Classification. In past decades, a number of candidate factors implicated in CRC risk are proposed by epidemiology studies, which can be divided into two groups in total, genetic factors and non-genetic factors. The genetic factors' group consists of many SNPs, and the nongenetic factors' group is comprised of several kinds of environment factors. According to the biological characteristics and the manner that human beings are exposed to environmental factors in whole lifetime, the raw big CRC-related genetic and environmental data can be classified into four biological categories: SNPs, demographic characteristics, lifestyles, and foods as in Table 1.

3.2. Results of Original Data Dimensionality Reduction. To process the dataset of SNPs, demographic characteristics, lifestyle and food, SPCA, and entropy and relief methods are employed, respectively.

Table 2 shows the principal components for the SNPs, demographic characteristics, and lifestyle and food by SPCA method, respectively. The result of the SPCA is listed in Supplementary S5.

Table 1: Results of biological classification.

Categories	Illustration
SNPs	Polymorphism distribution of genes
Demographic characteristics	Including factors like age, sex, body weight, income levels, and educations, which represents the individually biological or social-psychological features
Lifestyles	Behavioral factors, such as smoking and alcohol drinking
Foods	The amount of food intake

Table 2: The results by SPCA method.

SNPs	rs10046, rs10505477, rs1152579, rs1229984, rs1255998, rs1256030, rs1256049, rs1271572, rs12953717, rs1329149, rs16941669, rs17033, rs1801132, rs2075633, rs2077647, rs3798758, rs3820033, rs4767939, rs4767944, rs4939827, rs676387, rs6905370, rs6983267, rs7296651, rs7837688, rs827421, rs886205, rs928554, rs9322354, rs9340799
Demographic characteristics	Cholesterol, blood triglyceride, psychological trauma, depression, age, exercise, BMI, physical activity, activity, marriage status, emotion status
Lifestyles	Smoking, drinking, coffee consumption, drinking and smoking in the same time point, tea consumption
Foods	Grains, melons, bean products, roots, vegetables, fruits, eggs and milk, mushrooms, oil, seasoning, meat, seafood, pickles

We consider that the features with high weight will result in the colorectal cancer when the relief algorithm is applied to extract key features from the dataset. The result of relief algorithm is shown in Figure 3. In the upper part of Figures 3(a), 3(b), 3(c), and 3(d), the horizontal axis shows the feature numerical number and the vertical axis shows the feature weight. In the lower part of Figures 3(a), 3(b), 3(c), and 3(d), the horizontal axis shows the feature weight and the vertical axis shows the feature value, while the bars in Figure 3 represent the numbers of the features according to the feature weight.

Table 3 shows the results of dimensionality reduction by entropy method for the SNPs, demographic characteristics, and lifestyle and food, respectively. The entropy $H(X)$ in (2) is for data dimensionality reduction.

Regarding the results of Figure 3, Table 4 shows the common factors for the SNPs, demographic characteristics, and lifestyle and food by relief method, respectively.

Figure 4 shows the interaction results for the three dimensionality reduction methods. Figure 4(a) indicates that rs1256030 is the mutually explored biomarker by SPCA, entropy, and relief; rs10046, rs1152579, rs676387, rs6905370, rs928554, and rs6983267 are the mutually explored biomarkers by SPCA and entropy and rs4939827, rs4767944, rs1801132, rs4767939, rs10505477, rs3798758, and rs2075633 are the mutually explored biomarker by SPCA and relief.

Table 3: The results by entropy method.

SNPs	rs6983267, rs1256030, rs10046, rs928554, rs1152579, rs690537, rs676387
Demographic characteristics	Age, BMI, blood triglyceride, depression, mental stress, psychological trauma
Lifestyles	Drinking and smoking in the same time point, drinking
Foods	Vegetables, nuts, mushrooms, seasoning, pickles, grains

Table 4: The results by relief method.

SNPs	rs10505477, rs1256030, rs1801132, rs2071454, rs2075633, rs2228480, rs2249695, rs2486758, rs3798758, rs4767939, rs4767944, rs4939827
Demographic characteristics	Age, BMI, physical activity, activity, family number, emotion status, temperament, mental stress, psychological trauma, depression, cholesterol
Lifestyles	Drinking, tea consumption, drinking and smoking in the same time point
Foods	Nuts, vegetables, meat, eggs and milk, seafood

Figure 4(b) indicates that age, depression, blood triglyceride, and BMI are the mutually explored biomarkers by SPCA, entropy, and relief; blood triglyceride is the mutually explored biomarker by SPCA and entropy; cholesterol, activity, emotion status, and physical activity are the mutually explored biomarkers by SPCA and relief and mental stress is the mutually explored biomarkers by entropy and relief.

Figure 4(c) indicates that drinking and drinking and smoking in same time point are the mutually explored biomarkers by SPCA, entropy, and relief; tea consumption is the mutually explored biomarker by SPCA and relief.

Figure 4(d) indicates that vegetables are the mutually explored biomarkers by SPCA, entropy, and relief; mushrooms, seasoning, pickles, and grains are the mutually explored biomarkers by SPCA and entropy; eggs and milk, meat, and seafood are the mutually explored biomarkers by SPCA and relief and nuts is the mutually explored biomarker by entropy and relief.

We have 36 features mutually explored by every two of the SPCA, entropy, and relief methods.

By U test [59], Table 5 shows that 13 out of 36 features have small p value.

Table 6 shows that 13 features with small p value are important biomarkers.

3.3. Results of Regression. According to the dimensionality reduction analysis, there are 13 biomarkers selected as the classifier for these four biological datasets. Next, we employ LR, SVM, KRLS, HELM, and GKRMC algorithm to build up the predictive cancer model based on these selected classifiers.

Table 7 presents four measures (accuracy, sensitivity, specificity, and precision) to assess how good or how "accurate" the classifier is.

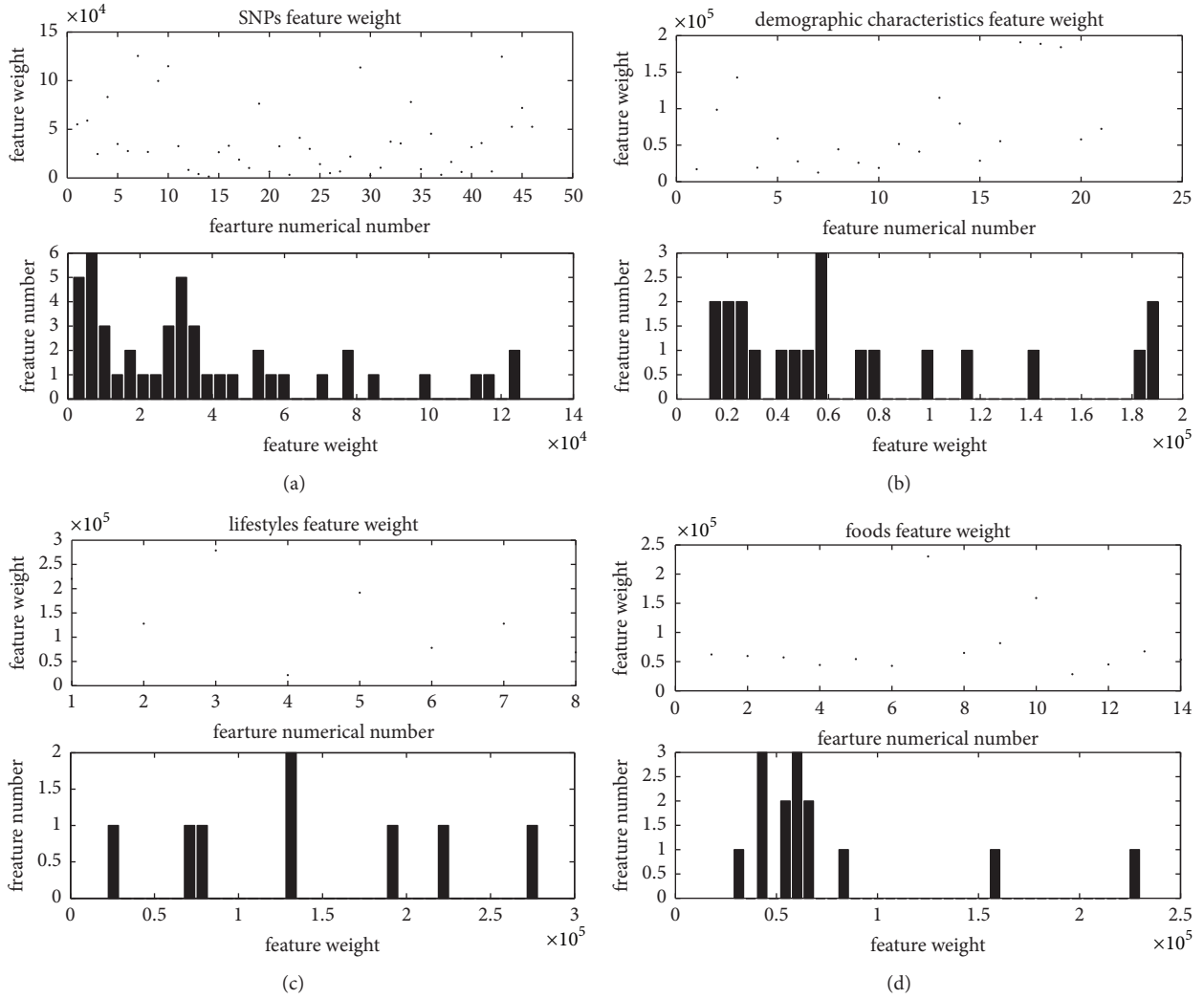


Figure 3: Feature selection by relief algorithm: (a) SNPs feature (note: the feature numerical number in the upper figure is regarding Supplementary S1 from columns B(1) to AU(46)), (b) demographic characteristics feature (note: the Feature numerical number in the upper figure is regarding Supplementary S2 from columns A(1) to U(21)), (c) lifestyle feature (note: the feature numerical number in the upper figure is regarding Supplementary S3 from columns B(1) to I(8)), and (d) food feature (note: the feature numerical number in the upper figure is regarding Supplementary S4 from columns B(1) to O(14)).

There are 1298 cases-control samples, 369 of which are case and 929 of which are control. Cross validation [60] method randomly chooses 75% of samples (973 samples) as the training dataset and the rest (325 samples) are used for testing dataset. Since cross validation introduces the random effect, we have to repeat the experiment 10 times. Figure 5 shows that GKRMC always has the greatest sensitive, precision, and accuracy values as well as greater specificity value compared to KRLS. Moreover, Table 8 lists the average value and standard deviation of the classification measurement for each algorithm.

4. Discussion and Conclusion

For CRC tumorigenesis, both genetic and environmental factors, as well as their interaction, playing important role in CRC risk is already the common view of most previously studies [61], but to figure out how to predict the occurrence of CRC by using the risk factors is still a challenge today. In the present study, we use big data of 1298 samples from a CRC case-control study in which relatively complete information of genetic and demographic characteristics and life style and food intake is simultaneously collected; furthermore, we expect to develop such a CRC-risk predictive model that not only can explore which risk factors included in the collected big dataset have significant impact on the occurrence of CRC, but also can accurately predict the occurrence of CRC as early as possible.

Such big datasets are classified into four different categories in the biological classification stage. And 13 of all explored potential biomarkers consisting of 4 SNPs, 6 demographic characteristics, 1 lifestyle factor, and 2 foods are screened out in data dimensionality reduction stage.

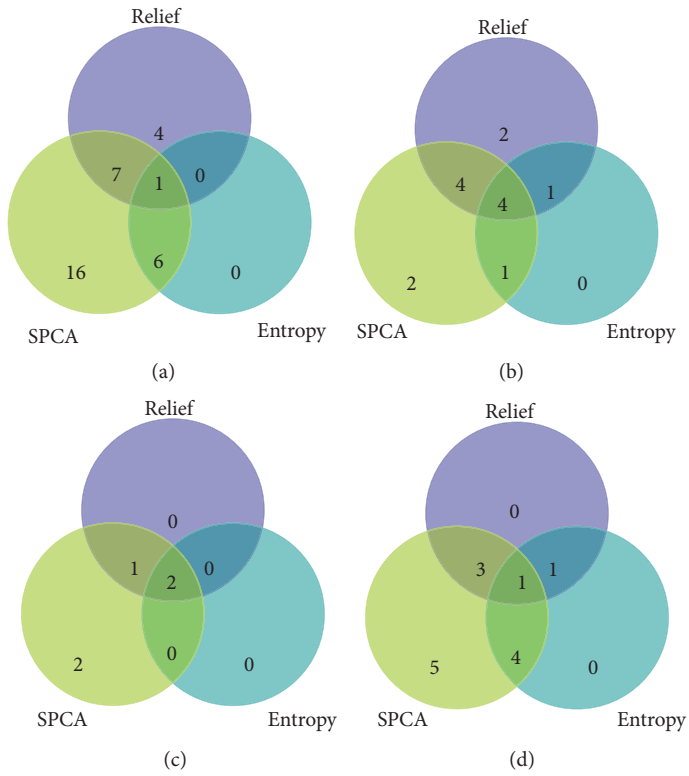


Figure 4: Venn plots of (a) SNPs, (b) demographic characteristics, (c) lifestyle, and (d) food.

Table 5: p value of 13 important biomarkers.

Biomarkers	p value
rs10046	0.0172
rs1256030	0.0004
rs6766387	0.0015
rs6983267	0.0000
age	0.0152
BMI	0.0019
Physical activity	0.0030
Emotion status	0.0247
Mental stress	0.0213
Cholesterol	0.0000
Drinking and smoking in the same time point	0.0000
Vegetables	0.0000
Seafood	0.0023

Table 6: Mutually explored biomarkers.

SNPS	rs10046, rs1256030, rs676387, 6983267
Demographic characteristics	Age, BMI, physical activity, emotion status, mental stress, cholesterol
Lifestyle	Drinking and smoking in the same time point
Foods	Vegetables, seafood

Unlike pure mathematical formulae, the biological rationality of such model depends on whether the selected biomarkers can be biologically explained as validated etiology of colorectal cancer supported by either population-based association study or biological function-based mechanisms experimental study. And then, these explored biomarkers can be used as the classifiers for the predictive model to access the risk of colorectal cancer in the regression analysis stage.

In fact, results from substantial epidemiology studies focusing on CRC risk/protective factors provide evidences for the associations between each category and risk of CRC. For the genetic variations, at least 2 (rs10046, rs6983267) of the 4 currently selected SNPs listed in Table 5 were reported to have significant association with CRC risk in either genome-wide association studies or candidate gene based study [59, 62]. Particularly, SNP rs6983267 is one of the most significant variations associated with increasing CRC risk in Caucasians, Asians, and Africans [63]. Regarding the other two selected SNPs (rs1256030, rs676387) located, respectively, in estrogen receptor beta gene (ESR2) and 17 β-hydroxysteroid dehydrogenases gene (HSD17B1) (both are estrogen metabolism pathway genes), though there is no direct evidence supporting their association with CRC, they both are found significantly associated with cancers such as liver and ovarian cancers [64, 65]. Moreover, considerable evidence from epidemiological and metabolic studies support that the estrogen metabolism pathway genes undoubtedly play an important role in CRC and other cancers [66], which implies the potential that the two SNPs may affect the susceptibility of CRC.

For demographic factors, almost all the 6 selected factors have been reported to be the unfavorable factors for CRC risk in a bunch of previous studies [67, 68].

Table 7: The definition of the classification measurement.

Measure	Formula	Illustration
Sensitivity	$\frac{TP}{P}$	TP: the number of true positives P: the number of positives
Specificity	$\frac{TN}{N}$	TN: the number of true negatives N: the number of negatives
Precision	$\frac{TP}{TP + FP}$	TP: the number of true positives FP: the number of false positives
Accuracy	$\frac{TP + TN}{P + N}$	TP: the number of true positives TN: the number of true negatives P: the number of positives N: the number of negatives

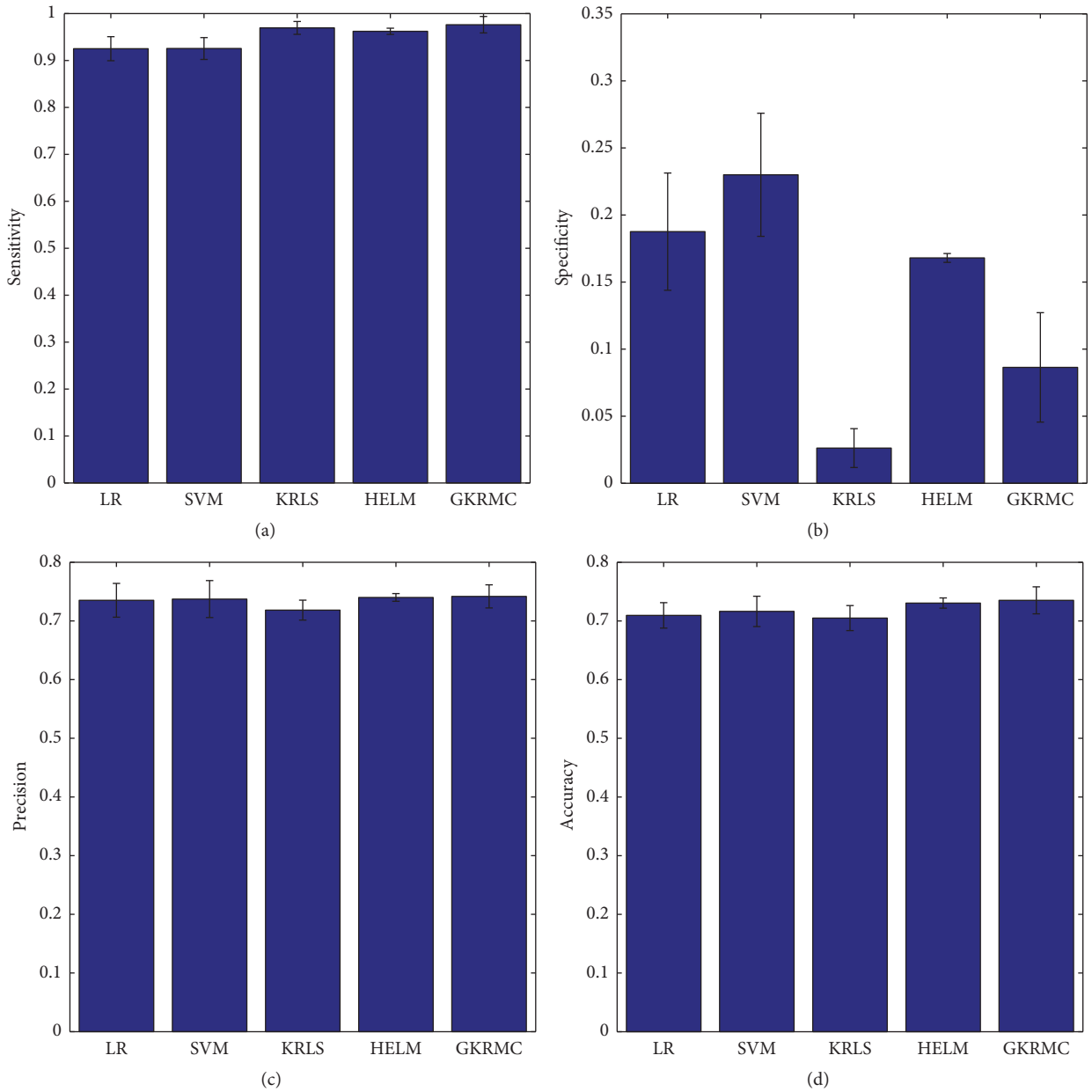


Figure 5: Predictive performance for the LR, SVM, KRLS, HELM, and GKRMC.

Table 8: The mutually explored biomarkers.

	LR	SVM	KRLS	HELM	GKRMC
Sensitivity	0.9251 ± 0.0256	0.9255 ± 0.0233	0.9694 ± 0.0137	0.9621 ± 0.0066	0.9762 ± 0.0175
Specificity	0.1876 ± 0.0437	0.2300 ± 0.0459	0.0262 ± 0.0145	0.1680 ± 0.0033	0.0864 ± 0.0408
Precision	0.7351 ± 0.0288	0.7372 ± 0.0315	0.7184 ± 0.0170	0.7400 ± 0.0066	0.7418 ± 0.0197
Accuracy	0.7095 ± 0.0217	0.7163 ± 0.0258	0.7049 ± 0.0213	0.7305 ± 0.0087	0.7351 ± 0.0230

For lifestyles, alcohol drinking and smoking are proved as two significant risk factors of CRC [18, 68]. Alcohol drinking, in a dose-response manner, evidently contributes to the increase of CRC risk. Meanwhile, obvious positive associations between CRC risk and cigarette smoking are observed in most measures [69].

For food, extensive epidemiologic and experimental studies confirm their important roles in the development of CRC. For example, higher consumption of vegetables and seafood is always associated with relatively lower CRC risk due to their relatively high content of antioxidant nutrients such as dietary fiber, vitamins, and long-chain unsaturated fatty acids [70–73]. On the contrary, the excessive consumption of smoked/salted/processed meat is linked to higher risk of colorectal neoplasia [73].

In general, it is demonstrated that the 13 currently explored biomarkers can be used as the classifiers in the regression analysis stage, which is supported by these manually reviewed experimental evidences [59, 63, 67, 69–71].

Although LR and SVM may perform very well for linear systems, their performance will get worse when applied to the nonlinear and non-Gaussian situations [74], which is rather common in real world applications. Therefore, we suggest using nonlinear regression algorithm to process our dataset, which is comprised of continuous and discrete data with multivariate data type. However, using classical nonlinear algorithm such as KRLS will suffer from outliers.

To overcome the shortcoming of both linear and conventional nonlinear regression algorithms, this study proposes an ensemble learning model (HELM) and a generalized kernel recursive maximum correntropy (GKRMC) algorithm to increase the predictive power of the model. Next, we analyze the reason why HELM and GKRMC can outperform LR, SVM, and KRLS algorithms.

HELM is an ensemble learning algorithm, which integrates linear and nonlinear classifiers to classify the data points. Based on the previous study [75], the diversity of weak classifiers is one of the evaluation criteria for ensemble algorithm. HELM includes both linear (SVM and logistic regression) and nonlinear (KRLS) classifiers and its superior performance has been shown in Figure 5.

The cost function of GKRMC (see (14)) is so robust that is not sensitive to large outliers as KRLS, since an exponentially weighted mechanism $G_{\alpha,\beta}(d_i - \Omega^T \varphi_i)$ of (14) can assign greater weight to the samples with smaller error but not to the samples with greater error. Since the big dataset usually consists of outliers [29, 76], GKRMC can achieve the higher predictive accuracy with the less standard deviation (Table 8) than KRLS. As mentioned before, the predictive power of GKRMC should be better than LR, SVM, and KRLS due to the nature of nonlinear regression (Figure 5).

In conclusion, this study proposes a robust CRC-risk predictive model to analyze the big data with information of genetic variations and environmental exposure for the CRC patients and cancer-free controls. The research results indicate that both genetic and environmental related factors explored by our model play the significant roles in the occurrence of CRC and the innovated HELM and GKRMC can increase the predictive power of the model.

However, this novel predictive model is the first step in predicting the risk of CRC tumor growth. Except for the environment factors and SNPs involved in the current model, if other factors such as pathway-pathway and pathway-environment interactions are included, there will be a higher chance to find a set of variations which may be integrative biomarkers, as proved in other researches [77, 78]. A limitation of our study is that there is only a finite number of tag SNPs located in a relatively small number of genes, which results in the nonuse of employing pathway interaction into model construction. Also, how to improve the GKRMC's specificity is an important topic for future study, which will further improve the whole system's performance. While extensions will be necessary to account in greater detail for the complexity of the CRC involved, we believe that if properly combined with more experimental data such as RNA sequence analysis and recent modeling techniques[79–86], advanced in silico platforms such as this one will evolve into powerful integrative research platforms that improve our understanding of CRC tumorigenesis.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the General Program from National Natural Science Foundation of China (nos. 81273156, 30771841, 61372138, and 61372152), Chongqing Excellent Youth Award and the Chinese Recruitment Program of Global Youth Experts, and the Fundamental Research Funding of the Chinese Central Universities (nos. XDJK2014B012 and XDJK2016A00).

References

[1] M. A. Arafa, M. I. Waly, J. Sahar, A. K. Ahmed, and S. Sunny, "Dietary and lifestyle characteristics of colorectal cancer in

Jordan: a case-control study," *Asian Pacific Journal of Cancer Prevention*, vol. 12, no. 8, pp. 1931–1936, 2011.

[2] M. M. Center, A. Jemal, and E. Ward, "International trends in colorectal cancer incidence rates," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 18, pp. 1688–1694, 2009.

[3] M. Li and J. Gu, "Changing patterns of colorectal cancer in China over a period of 20 years," *World Journal of Gastroenterology*, vol. 11, no. 30, pp. 4685–4688, 2005.

[4] S. Liu, R. Zheng, M. Zhang, S. Zhang, X. Sun, and W. Chen, "Incidence and mortality of colorectal cancer in China, 2011," *Journal of Thoracic Disease*, vol. 5, pp. 330–336, 2014.

[5] D. M. Parkin, C. A. Stiller, and J. Nectoux, "International variations in the incidence of childhood bone tumours," *International Journal of Cancer*, vol. 53, no. 3, pp. 371–376, 1993.

[6] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, and J. Lortet-Tieulent, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[7] Y. Zhao, X. Deng, Z. Wang, Q. Wang, and Y. Liu, "Genetic polymorphisms of DNA repair genes XRCC1 and XRCC3 and risk of colorectal cancer in chinese population," *Asian Pacific Journal of Cancer Prevention*, vol. 13, no. 2, pp. 665–669, 2012.

[8] J. He and W. Chen, "Chinese cancer registry annual report 2012," Press of Military Medical Sciences, Beijing, China, pp. 68–71, 2012.

[9] S. D. Markowitz and M. M. Bertagnolli, "Molecular basis of colorectal cancer," *The New England Journal of Medicine*, vol. 361, no. 25, pp. 2404–2460, 2009.

[10] M. Van Engeland, S. Derks, K. M. Smits, G. A. Meijer, and J. G. Herman, "Colorectal cancer epigenetics: Complex simplicity," *Journal of Clinical Oncology*, vol. 29, no. 10, pp. 1382–1391, 2011.

[11] A. A. Ghazarian, N. I. Simonds, K. Bennett et al., "A review of NCI's extramural grant portfolio: Identifying opportunities for future research in genes and environment in cancer," *Cancer Epidemiology Biomarkers and Prevention*, vol. 22, no. 4, pp. 501–507, 2013.

[12] E. J. Kuipers, W. M. Grady, D. Lieberman et al., "Colorectal cancer," *Nature Reviews Disease Primers*, vol. 1, pp. 1–25, 2015.

[13] H. Brenner, M. Kloor, and C. P. Pox, "Colorectal cancer," *The Lancet*, vol. 383, no. 9927, pp. 1490–1502, 2014.

[14] V. T. DeVita Jr. and S. A. Rosenberg, "Two hundred years of cancer research," *New England Journal of Medicine*, vol. 366, no. 23, pp. 2207–2214, 2012.

[15] J. Kim, Y. A. Cho, D.-H. Kim et al., "Dietary intake of folate and alcohol, MTHFR C677T polymorphism, and colorectal cancer risk in Korea," *American Journal of Clinical Nutrition*, vol. 95, no. 2, pp. 405–412, 2012.

[16] K. Matsuo, T. Suzuki, H. Ito et al., "Association between an 8q24 locus and the risk of colorectal cancer in Japanese," *BMC Cancer*, vol. 9, article no. 379, 2009.

[17] K. Wakai, K. Hirose, K. Matsuo et al., "Dietary risk factors for colon and rectal cancers: a comparative case-control study," *Journal of Epidemiology*, vol. 16, no. 3, pp. 125–135, 2006.

[18] H. Yang, Y. Zhou, Z. Zhou et al., "A Novel Polymorphism rs1329149 of CYP2E1 and a Known Polymorphism rs671 of ALDH2 of Alcohol Metabolizing Enzymes Are Associated with Colorectal Cancer in a Southwestern Chinese Population," *Cancer Epidemiology Biomarkers and Prevention*, vol. 18, no. 9, pp. 2522–2527, 2009.

[19] Y.-Z. Wu, H. Yang, L. Zhang et al., "Application of crossover analysis-logistic regression in the assessment of gene-environmental interactions for colorectal cancer," *Asian Pacific Journal of Cancer Prevention*, vol. 13, no. 5, pp. 2031–2037, 2012.

[20] H. Huang, P. Chanda, A. Alonso, J. S. Bader, and D. E. Arking, "Gene-based tests of association," *PLoS genetics*, vol. 7, no. 7, Article ID e1002177, 2011.

[21] L. W. Hahn, M. D. Ritchie, and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," *Bioinformatics*, vol. 19, no. 3, pp. 376–382, 2003.

[22] J. H. Moore, J. C. Gilbert, C.-T. Tsai et al., "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *Journal of Theoretical Biology*, vol. 241, no. 2, pp. 252–261, 2006.

[23] M. D. Ritchie, L. W. Hahn, N. Roodi et al., "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, 2001.

[24] M. Li, C. Ye, W. Fu, R. C. Elston, and Q. Lu, "Detecting genetic interactions for quantitative traits with *U*-statistics," *Genetic Epidemiology*, vol. 35, no. 6, pp. 457–468, 2011.

[25] A. S. Andrew, H. H. Nelson, K. T. Kelsey et al., "Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility," *Carcinogenesis*, vol. 27, no. 5, pp. 1030–1037, 2006.

[26] W. Meredith, R. Rutledge, S. M. Fakhry, S. Emery, and S. Kromhout-Schiro, "The conundrum of the Glasgow Coma Scale in intubated patients: a linear regression prediction of the Glasgow verbal score from the Glasgow eye and motor scores," *Journal of Trauma*, vol. 44, no. 5, pp. 839–845, 1998.

[27] R. Rutledge, C. W. Lentz, S. Fakhry, and J. Hunt, "Appropriate use of the glasgow coma scale in intubated patients: a linear regression prediction of the glasgow verbal score from the glasgow eye and motor scores," *The Journal of Trauma*, vol. 41, no. 3, pp. 514–522, 1996.

[28] C. Zheng, L. Xing, T. Li, H. Yang, J. Cao et al., "Developing a robust colorectal cancer (CRC) risk predictive model with the big genetic and environment related CRC data," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1885–1893, IEEE, Shenzhen, China, 2016.

[29] C. A. Mattmann, "Computing: A vision for data science," *Nature*, vol. 493, no. 7433, pp. 473–475, 2013.

[30] Z. Y. Zhou, B. Q. Takezaki TMo, H. M. Sun, W. C. Wang, L. P. Sun, and S. X. Liu, "Development of a semi-quantitative food frequency questionnaire to determine variation in nutrient intakes between urban and rural areas of Chongqing, China," *Asia Pacific Journal of Clinical Nutrition*, vol. 13, pp. 273–283, 2004.

[31] G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein, "The International HapMap Project Web site," *Genome Research*, vol. 15, no. 11, pp. 1592-1593, 2005.

[32] O. De la Cruz and P. Raska, "Population structure at different minor allele frequency levels," *BMC Proceedings*, vol. 8, pp. 1–5, 2014.

[33] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics*, vol. 21, no. 2, pp. 263–265, 2005.

[34] C. J. Mulligan, R. W. Robin, M. V. Osier et al., "Allelic variation at alcohol metabolism genes (ADH1B, ADH1C, ALDH2) and alcohol dependence in an American Indian population," *Human Genetics*, vol. 113, no. 4, pp. 325–336, 2003.

[35] M. Crous-Bou, G. Rennert, D. Cuadras et al., "Polymorphisms in alcohol metabolism genes ADH1B and ALDH2, alcohol consumption and colorectal cancer," *PLoS ONE*, vol. 8, no. 11, Article ID e80158, 2013.

[36] T. S. Kang, S. W. Woo, H. J. Park, Y. Lee, and J. Roh, "Comparison of genetic polymorphisms of CYP2E1, ADH2, and ALDH2 genes involved in alcohol metabolism in Koreans and four other ethnic groups," *Journal of Clinical Pharmacy and Therapeutics*, vol. 34, no. 2, pp. 225–230, 2009.

[37] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[38] I. T. Jolliffe, *Principal Component Analysis*, Springer, Berlin, Germany, 2nd edition, 2010.

[39] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987.

[40] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational Graphical Statistics*, p. 2007, 2012.

[41] T. M. Cover and J. A. Thomas, *Thomas JA. Elements of Information Theory*, Cognitive Science - A Multidisciplinary Journal, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition, 2005.

[42] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[43] Z.-H. Zou, Y. Yun, and J.-N. Sun, "Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment," *Journal of Environmental Sciences* , vol. 18, no. 5, pp. 1020–1023, 2006.

[44] Y. Sun, "Iterative RELIEF for feature weighting: algorithms, theories, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1035–1051, 2007.

[45] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, 2004.

[46] K. Koh, S. J. Kim, and S. P. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, no. 8, pp. 1519–1555, 2007.

[47] J. Pearce and S. Ferrier, "Evaluating the predictive performance of habitat models developed using logistic regression," *Ecological Modelling*, vol. 133, no. 3, pp. 225–245, 2000.

[48] A. Al-Anazi and I. D. Gates, "A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs," *Engineering Geology*, vol. 114, no. 3-4, pp. 267–277, 2010.

[49] T. G. Dietterich, "Machine-learning research: four current directions," *AI Magazine*, vol. 18, no. 4, pp. 97–136, 2000.

[50] X. Wu, V. Kumar, J. R. Quinlan et al., "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[51] Y. Freund and R. E. Schapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, vol. 14, Springer, Berlin, Germany, 1995.

[52] W. Liu, P. P. Pokharel, and J. C. Principe, "The kernel least-mean-square algorithm," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 543–554, 2008.

[53] B. Chen, S. Zhao, P. Zhu, and J. C. Príncipe, "Quantized kernel least mean square algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 22–32, 2012.

[54] B. Chen, S. Zhao, P. Zhu, and J. C. Principe, "Quantized kernel recursive least squares algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1484–1491, 2013.

[55] W. Liu, J. C. Principe, and S. Haykin, "Kernel Adaptive Filtering: A Comprehensive Introduction," in *Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, 2010.

[56] B. Chen, L. Xing, H. Zhao, N. Zheng, and J. C. Príncipe, "Generalized correntropy for robust adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3376–3387, 2016.

[57] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Space in Probability and Statistics*, Kluwer Academic, Boston, Mass, USA, 2004.

[58] M. J. L. Orr, "Introduction to Radial Basis Function Networks," *Journal of Vitamin Research*, vol. 4, pp. 2797–2800, 1967.

[59] N. H. Ramzi, J. K. Chahil, S. H. Lye et al., "Role of genetic & environment risk factors in the aetiology of colorectal cancer in Malaysia," *Indian Journal of Medical Research*, vol. 139, pp. 873–882, 2014.

[60] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1137-43, International Joint Conference on Artificial Intelligence, Stanford, Calif, USA, 2001.

[61] H. Mahdi, B. A. Fisher, H. Källberg et al., "Specific interaction between genotype, smoking and autoimmunity to citrullinated α-enolase in the etiology of rheumatoid arthritis," *Nature Genetics*, vol. 41, no. 12, pp. 1319–1324, 2009.

[62] R. Cui, Y. Okada, S. G. Jang et al., "Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population," *Gut*, vol. 60, no. 6, pp. 799–805, 2011.

[63] C. A. Haiman, L. le Marchand, J. Yamamato et al., "A common genetic risk factor for colorectal and prostate cancer," *Nature Genetics*, vol. 39, no. 8, pp. 954–956, 2007.

[64] G. Lurie, L. R. Wilkens, P. J. Thompson et al., "Genetic polymorphisms in the estrogen receptor beta (ESR2) gene and the risk of epithelial ovarian carcinoma," *Cancer Causes and Control*, vol. 20, no. 1, pp. 47–55, 2009.

[65] L. S. Zhang, F. Yuan, X. Guan et al., "Association of genetic polymorphisms in HSD17B1, HSD17B2 and SHBG genes with hepatocellular carcinoma risk," *Pathology and Oncology Research*, vol. 20, no. 3, pp. 661–666, 2014.

[66] A. Barzi, A. M. Lenz, M. J. Labonte, and H.-J. Lenz, "Molecular pathways: Estrogen pathway in colorectal cancer," *Clinical Cancer Research*, vol. 19, no. 21, pp. 5842–5848, 2013.

[67] T. E. Robsahm, B. Aagnes, A. Hjartåker, H. Langseth, F. I. Bray, and I. K. Larsen, "Body mass index, physical activity, and colorectal cancer by anatomical subsites: A systematic review and meta-analysis of cohort studies," *European Journal of Cancer Prevention*, vol. 22, no. 6, pp. 492–505, 2013.

[68] Z.-Y. Zhou, H. Yang, J. Cao, K. Tajima, K. Matsuo, and W.-C. Wang, *Dietary Risks: Folate, Alcohol and Gene Polymorphisms*, INTECH Open Access, 2012.

[69] H. Raskov, H. C. Pommergaard, J. Burcharth, and J. Rosenberg, "Colorectal carcinogenesis—update and perspectives," *World Journal of Gastroenterology*, vol. 20, no. 48, pp. 18151–18164, 2014.

[70] M. Song, W. S. Garrett, and A. T. Chan, "Nutrients, foods, and colorectal cancer prevention," *Gastroenterology*, vol. 148, no. 6, pp. 1244–1260, 2015.

[71] Q. Ben, J. Zhong, J. Liu et al., "Association between consumption of fruits and vegetables and risk of colorectal adenoma a prisma-compliant meta-Analysis of observational studies," *Medicine (United States)*, vol. 94, no. 42, p. e1599, 2015.

[72] S. S. Young, "Re: Low-fat dietary pattern and cancer incidence in the Women's Health Initiative Dietary Modification Randomized Controlled Trial," *Journal of the National Cancer Institute*, vol. 100, no. 4, p. 284, 2008.

[73] E. D. Kantor, J. W. Lampe, U. Peters, T. L. Vaughan, and E. White, "Long-chain omega-3 polyunsaturated fatty acid intake and risk of colorectal cancer," *Nutrition and Cancer*, vol. 66, no. 4, pp. 716–727, 2014.

[74] W. Linde, "Stable non-gaussian random processes: stochastic models with infinite variance," *Bulletin of the London Mathematical Society*, vol. 28, no. 430, 1994.

[75] E. C. Ensembles, E. S. O. Hyperspectrale, S. Yu, and S. Cao, "Feature Selection and Classifier Ensembles: A Study on Hyperspectral Remote Sensing Data," 2003.

[76] L. V. Subramaniam, "Big Data and Veracity Challenges," 2014, http://wwwisicalacin/~acmsc/TMW2014/LVSpdf.

[77] K.-Q. Liu, Z.-P. Liu, J.-K. Hao, L. Chen, and X.-M. Zhao, "Identifying dysregulated pathways in cancers from pathway interaction networks," *BMC Bioinformatics*, vol. 13, no. 1, article 126, 2012.

[78] R. Visakh and K. A. Abdul Nazeer, "Identifying epigenetically dysregulated pathways from pathway–pathway interaction networks," *Computers in Biology and Medicine*, vol. 76, pp. 160–167, 2016.

[79] B. Jiang, W. Dai, A. Khaliq, M. Carey, X. Zhou, and L. Zhang, "Novel 3D GPU based numerical parallel diffusion algorithms in cylindrical coordinates for health care simulation," *Mathematics and Computers in Simulation*, vol. 109, pp. 1–19, 2015.

[80] B. Jiang, A. Struthers, Z. Sun et al., "Employing graphics processing unit technology, alternating direction implicit method and domain decomposition to speed up the numerical diffusion solver for the biomedical engineering research," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 27, no. 11, pp. 1829–1849, 2011.

[81] H. Peng, T. Peng, J. Wen et al., "Characterization of p38 MAPK isoforms for drug resistance study using systems biology approach," *Bioinformatics*, vol. 30, no. 13, pp. 1899–1907, 2014.

[82] Y. Xia, C. Yang, N. Hu et al., "Exploring the key genes and signaling transduction pathways related to the survival time of glioblastoma multiforme patients by a novel survival analysis model," *BMC Genomics*, vol. 18, article no. 950, 2017.

[83] L. Zhang, B. Jiang, Y. Wu et al., "Developing a multiscale, multi-resolution agent-based brain tumor model by graphics processing units," *Theoretical Biology and Medical Modelling*, vol. 8, no. 1, article no. 46, 2011.

[84] L. Zhang, M. Qiao, H. Gao et al., "Investigation of mechanism of bone regeneration in a porous biodegradable calcium phosphate (CaP) scaffold by a combination of a multi-scale agent-based model and experimental optimization/validation," *Nanoscale*, vol. 8, no. 31, pp. 14877–14887, 2016.

[85] L. Zhang, Y. Xue, B. Jiang et al., "Multiscale agent-based modelling of ovarian cancer progression under the stimulation of the STAT 3 pathway," *International Journal of Data Mining and Bioinformatics*, vol. 9, no. 3, pp. 235–253, 2014.

[86] L. Zhang and S. Zhang, "Using game theory to investigate the epigenetic control mechanisms of embryo development: Comment on: 'Epigenetic game theory: How to compute the epigenetic control of maternal-to-zygotic transition' by Qian Wang et al," *Physics of Life Reviews*, vol. 20, pp. 140–142, 2017.

Hindawi
Complexity
Volume 2017, Article ID 2498957, 9 pages
https://doi.org/10.1155/2017/2498957

Research Article

miRNA-Disease Association Prediction with Collaborative Matrix Factorization

Zhen Shen,[1] You-Hua Zhang,[2] Kyungsook Han,[3] Asoke K. Nandi,[4] Barry Honig,[5] and De-Shuang Huang[1]

[1] *Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*
[2] *School of Information and Computer, Anhui Agricultural University, Changjiang West Road 130, Hefei, Anhui, China*
[3] *Department of Computer Science and Engineering, Inha University, Incheon, Republic of Korea*
[4] *Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, UK*
[5] *Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Avenue, Room 815, New York, NY 10032, USA*

Correspondence should be addressed to De-Shuang Huang; dshuang@tongji.edu.cn

Received 31 March 2017; Accepted 2 May 2017; Published 28 September 2017

Academic Editor: Fang-Xiang Wu

As one of the factors in the noncoding RNA family, microRNAs (miRNAs) are involved in the development and progression of various complex diseases. Experimental identification of miRNA-disease association is expensive and time-consuming. Therefore, it is necessary to design efficient algorithms to identify novel miRNA-disease association. In this paper, we developed the computational method of Collaborative Matrix Factorization for miRNA-Disease Association prediction (CMFMDA) to identify potential miRNA-disease associations by integrating miRNA functional similarity, disease semantic similarity, and experimentally verified miRNA-disease associations. Experiments verified that CMFMDA achieves intended purpose and application values with its short consuming-time and high prediction accuracy. In addition, we used CMFMDA on Esophageal Neoplasms and Kidney Neoplasms to reveal their potential related miRNAs. As a result, 84% and 82% of top 50 predicted miRNA-disease pairs for these two diseases were confirmed by experiment. Not only this, but also CMFMDA could be applied to new diseases and new miRNAs without any known associations, which overcome the defects of many previous computational methods.

1. Introduction

MicroRNAs (miRNAs) are a class of short noncoding RNAs (19~25 nt), which normally regulate gene expression and protein production by targeting messenger RNAs (mRNAs) at the posttranscriptional level [1–9]. Since the first two miRNA lin-4 and let-7 were found in 1993 and 2000 [10, 11], thousands of miRNAs have been detected in eukaryotic organisms ranging from nematodes to humans. The latest venison of miRBase contains 26845 entries and more than 2000 miRNAs have been detected in human [12–14]. With the development of bioinformatics and the progress of miRNA-related projects, researches are gradually focused on the function of miRNAs. Existing studies have shown that miRNAs are involved in many important biological processes [15, 16], like cell differentiation [17], proliferation [18], signal transduction [19], viral infection [20], and so on. Therefore, it is easy to find that miRNAs have close relationship with various human complex diseases [12, 21–26]. For example, researchers found that mir-433 is upregulated in gastric carcinoma by regulating the expression of GRB2, which is a known tumour-associated protein [27]. Mir-126 can not only function as an inhibitor to suppress the growth of colorectal cancer cells by its overexpression, but also can help to differentiate between malignant and normal colorectal tissue [28]. Besides, the change of mir-17~92 miRNA cluster expression has close relationship with kidney cyst growth in polycystic kidney disease [29]. Considering the close relationship between miRNA and disease, we should try all means to excavate all latent associations between miRNA

and disease and to facilitate the diagnose, prevention, and treatment human complex disease [30–33]. However, using experimental methods to identify miRNA-disease association is expensive and time-consuming. As the miRNA-related theories are becoming more and more common, such as the prediction model about miRNA and disease, the function of miRNA in biological processes, and signaling pathways, new therapies are urgently needed for the treatment of complex disease; it is necessary to develop powerful computational methods to reveal potential miRNA-disease associations [12, 15, 20, 34–40].

Previous studies had shown that functionally similar miRNAs always appear in similar diseases; therefore many computational models were proposed to identify novel miRNA-disease associations [13, 41–46]. For example, Jiang et al. [31] analyzed and improved disease-gene prediction model, introduced the principle of hypergeometric distribution and how to use it, and discussed its application in prediction model and its actual effect. In order to realize the prediction function of the improved model, they used different types of dataset including miRNA functional similarity data, disease phenotype similarity data, and the known human disease-miRNA association data. Therefore, the prediction accuracy of this method is greatly impacted by miRNA neighbor information and miRNA-target interaction prediction. Chen et al. [47] reported a new method HGIMDA to identify novel miRNA-disease association by using heterogeneous graph inference. This algorithm can get better prediction accuracy by integrating known miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity for diseases and miRNAs. In addition, HGIMDA could be applied for new diseases and new miRNAs which do not have any known association. Li et al. [48] proposed the computational model Matrix completion for MiRNA-disease association prediction (MCMDA) to predict miRNA-disease associations. This model only uses known miRNA-disease associations and achieved better prediction performance. The limitation of MCMDA is that it could not be applied for new diseases and new miRNAs which do not have any known association. You et al. [49] developed model Path-Based MiRNA-Disease Association Prediction (PBMDA) to predict miRNA-disease associations by integrating known human miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity for miRNAs and diseases. Depth-first search algorithm was used in this model to identify novel miRNA-disease associations. Benefiting from effective algorithm and reliable biological datasets, PBMDA has better prediction performance. Furthermore, Xu et al. [50] introduced an approach to identify disease-related miRNAs by the miRNA target-dysregulated network (MTDN). Furthermore, in order to distinguish and identify disease-related miRNAs from candidate, a SVM classifier based on radial basis function and the lib SVM package had been proposed. Researches have shown that miRNAs can functionally interact with environmental factors (EFs) to affect and determine human complex disease. Chen [51] proposed model miREFRWR to predict the association between disease and miRNA-EF interactions. Random walks theory was applied on miRNA similarity network and EF similarity network. In addition, drug chemical structure similarity, miRNA function similarity, and networked-based similarity were also used in miREFRWR. Based on these biological datasets and efficient calculation method, miREFRWR could be an effective tool in computational biology. What is more, Chen et al. [52] also proposed a computational model RKNNMDA to predict the potential associations between miRNA and disease. Four biological datasets, experimentally verified human miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity for miRNAs and diseases were integrated into RKNNMDA. It can be found that the prediction accuracy of RKNNMDA is excellent. Moreover, RKNNMDA could be applied for new diseases which do not have any known related miRNA information.

Generally speaking, current prediction model on miRNA-disease association is still demonstrating some shortcomings. For example, unreliable datasets have a great influence on the accuracy of prediction model, such as miRNA-target interactions and disease-genes associations. In addition, for miRNAs and diseases which do not have any known associations, we cannot use some of the existing models to predict its relevant information. In other words, we need to design and develop a new effective computational model. According to the assumption that functionally similar miRNAs always appear in similar diseases, we introduce the model of Collaborative Matrix Factorization for MiRNA-Disease Association prediction (CMFMDA) to reveal novel miRNA-disease association by integrating experimentally validated miRNA-disease associations, miRNA functional similarity information, and disease semantic similarity information. For CMFMDA, we can obtain its test results with three different ways: 5-fold CV, Local LOOCV, and global LOOCV. The AUCs of these three methods are 0.8697, 0.8318, and 0.8841, respectively, which suggest that CMFMDA is a reliable and efficient prediction model. And then, we use two cases: Esophageal Neoplasms and Kidney Neoplasms, to evaluate the performance of CMFMDA. In both of these two important diseases, 42 and 41 out of top 50 predicted miRNA-disease associations were confirmed by recent experimental literatures, respectively. In addition, experiments show that CMFMDA can be applied for diseases and miRNAs without any known association.

2. Materials and Methods

2.1. Human miRNA-Disease Associations. We obtained information about the associations between miRNA and disease from HMDD, including 5430 experimentally confirmed human miRNA-diseases associations about 383 diseases and 495 miRNAs. Adjacency matrix A is proposed to describe the association between miRNA and disease. If miRNA $m(i)$ is associated with disease $d(j)$, the entity $A(m(i), d(j))$ is 1, otherwise 0. Furthermore, we declared two variables nm and nd to represent the number of miRNAs and diseases investigated in this paper, respectively.

2.2. MiRNA Functional Similarity. Base on the assumption that miRNAs with similarity functions are regarded to be involved in similar diseases, Wang et al. [42] present a method to calculate the miRNA functional similarity score. We downloaded miRNA functional similarity scores from http://www.cuilab.cn/files/images/cuilab/misim.zip and constructed matrix SM to represent the miRNA function similarity network, where the entity $SM(m(i), m(j))$ represents the functional similarity score between miRNA $m(i)$ and $m(j)$.

2.3. Disease Semantic Similarity. In this paper, disease can be described as a Directed Acyclic Graph (DAG) and $\mathrm{DAG}(D) = (D, T(D), E(D))$ was used to describe disease D, where $T(D)$ is the node set including all ancestor nodes of D and D itself and $E(D)$ is the corresponding links set including the direct edges from parent nodes to child nodes. The semantic value of disease D in $\mathrm{DAG}(D)$ is defined as follows:

$$DV1(D) = \sum_{d \in T(D)} D1_D(d),$$

$$D1_D(d) = \begin{cases} 1 & \text{if } d = D \\ \max\left\{\Delta * D1_D\left(d'\right) \mid d' \in \text{ children of } d\right\} & \text{if } d \neq D, \end{cases} \tag{1}$$

where Δ is the semantic contribution factor. For disease D, the contribution of itself to the semantic value of disease D is 1. However, with the growth of the distance between D and other disease, the contribution will fall. Therefore, disease terms in the same layer would have the same contribution to the semantic value of disease D.

If there is much in common between two diseases in DAG, their semantic similarity will become larger. Therefore, the semantic similarity between diseases $d(i)$ and $d(j)$ can be defined as follows:

$$SD\left(d(i), d(j)\right) = \frac{\sum_{t \in T(i) \cap T(j)} \left(D1_i(t) + D1_j(t)\right)}{DV1(i) + DV1(j)}, \tag{2}$$

where SD is the disease semantic similarity matrix.

2.4. CMFMDA. In this study, we developed the computational model of Collaborative Matrix Factorization for MiRNA-Disease Association prediction (CMFMDA) to predict novel miRNA-disease associations [53]. The flow of CMFMDA is shown in Figure 1.

In the first step in Figure 1, we will get the final miRNA similarity matrix SM and diseases similarity matrix SD by integrating miRNA functional similarity network, disease semantic similarity network, and experimentally verified miRNA-disease associations.

Then, we use WKNKN [54] to estimate the association probability for these unknown cases based on their known neighbors.

Thirdly, Collaborative Matrix Factorization was used to obtain the final prediction F. This step contains three parts:

(1) For the input matrix Y, this step adopts singular value decomposition to get the initial value of A and B.

$$[U, S, V] = SVD(Y, k),$$
$$A = US_k^{1/2}, \tag{3}$$
$$B = VS_k^{1/2}.$$

(2) We use L to represent the objection function and use a_i and b_j to represent the ith and jth row vectors of A and B. Two alternative update rules (one for updating matrix A and one for updating matrix B) were derived by setting $\partial L/\partial A = 0$ and $\partial L/\partial B = 0$. According to alternating least squares, these two update rules are run alternatingly until convergence.

$$\begin{aligned} \min_{A,B} \quad & \left\|Y - AB^T\right\|_F^2 + \lambda_l\left(\|A\|_F^2 + \|B\|_F^2\right) \\ & + \lambda_m\left\|SM - AA^T\right\|_F^2 + \lambda_d\left\|SD - BB^T\right\|_F^2 \\ & A = \left(YB + \lambda_d S_d A\right)\left(B^T B + \lambda_l I_k + \lambda_d A^T A\right)^{-1} \\ & B \\ & = \left(Y^T A + \lambda_m S_m B\right)\left(A^T A + \lambda_l I_k + \lambda_m B^T B\right)^{-1}. \end{aligned} \tag{4}$$

Finally, the predicted matrix for miRNA-disease associations is then obtained by multiplying A and B.

3. Results

3.1. Performance Evaluation. Based on the known miRNA-disease associations obtained from HMDD database [55], the predictive performance of CMFMDA is evaluated through two ways: Local and global LOOCV. Not only that, three computational models: WBSMDA [4], RLSMDA [12], and NCPMDA [56], were introduced to compare the prediction performance with CMFMDA. To obtain relevant miRNA information for the chosen disease d, all association related to disease d was left out, and the rest of the associations serve as a training set to get prediction association by CMFMDA. For cross-validation, the difference between local LOOCV and global LOOCV is that all diseases would be investigated simultaneously or not. Furthermore, Receiver-Operating Characteristics (ROC) were used to express the difference between true positive rate (TPR, sensitivity) and false positive rate (FPR, 1 – specificity) at different thresholds. In this case, sensitivity indicates that the percentage of the test miRNA-disease association which obtained ranks higher than the given threshold. Meanwhile, specificity indicates the percentage of miRNA-disease associations below the threshold. What is more, Area under the ROC curve (AUC) could be calculated to demonstrate the prediction performance of CMFMDA. AUC = 1 showed that the model has perfect prediction ability; AUC = 0.5 indicates random prediction ability.

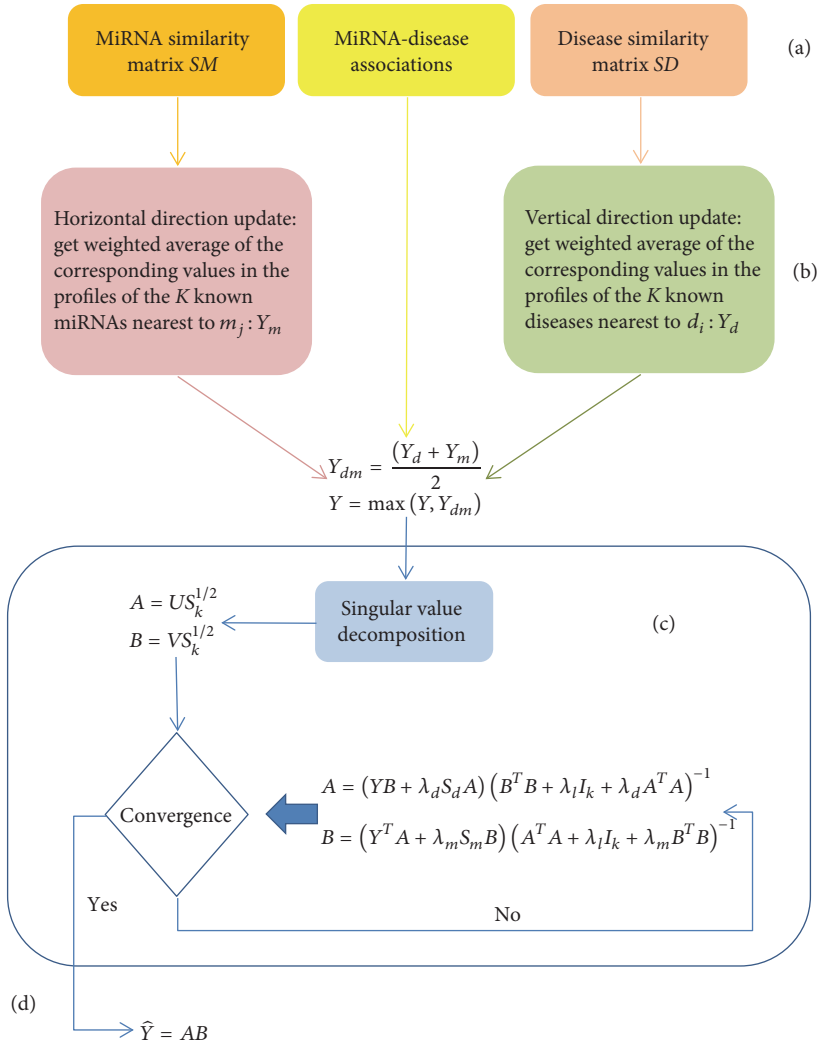


Figure 1: Flowchart of potential miRNA-disease associations prediction based on CMFMDA.

To illustrate the performance of CMFMDA, we compare it with the existed computational model: NCPMDA, RLSMDA, and WBSMDA. The comparison result has been shown in Figure 2. As a result, these four models obtained AUCs of 0.8841, 0.8630, 0.8501, and 0.7799 in the global LOOCV, respectively. For local LOOCV, these four models obtained AUCs of 0.8318, 0.8198, 0.8068, and 0.7213, respectively. In general, CMFMDA has not only high prediction performance, but also better ability to identify novel miRNA-disease association.

3.2. Case Studies. All diseases in this paper have been investigated by CMFMDA to predict some novel miRNAs which have association with the disease. Here, two case studies, Esophageal Neoplasms and Kidney Neoplasms, were proposed to demonstrate the prediction performance of CMFMDA. In addition, we use two important miRNA-disease association databases to validate the prediction results: miR2Disease [57] and dbDEMC [58]. A final note about validation datasets is that only the associations which were absent from the HMDD database would be used. In other words, validation datasets have no correlation with the datasets which have been used for prediction.

Esophageal Neoplasms is a serious disease in digestive system, which leads to high death rate [59–61]. Early diagnosis and treatment is essential for improving patient's survival [62, 63]. Here, we use CMFMDA to identify potential miRNAs associated with Esophageal Neoplasms. As a result, 9 out of the top 10 and 42 out of the top 50 predicted related miRNAs were experimentally confirmed to be associated with Esophageal Neoplasms (See Table 1). For example, mir-133b can inhibit the cell growth and invasion of esophageal squamous cell carcinoma (ESCC) by targeting Fascin homolog 1 [59]. The expression level of mir-335 is an independent prognostic factor in ESCC, which might be a potential valuable biomarker for ESCC [64].

As a common urologic malignancy, the morbidity and mortality of Kidney Neoplasm have been shown to rise gradually [65–68]. Renal cell carcinoma (RCC) can be divided into several different types of cancer [69–71], including chromophobe RCC (CHRCC), collecting duct carcinoma (CDC),

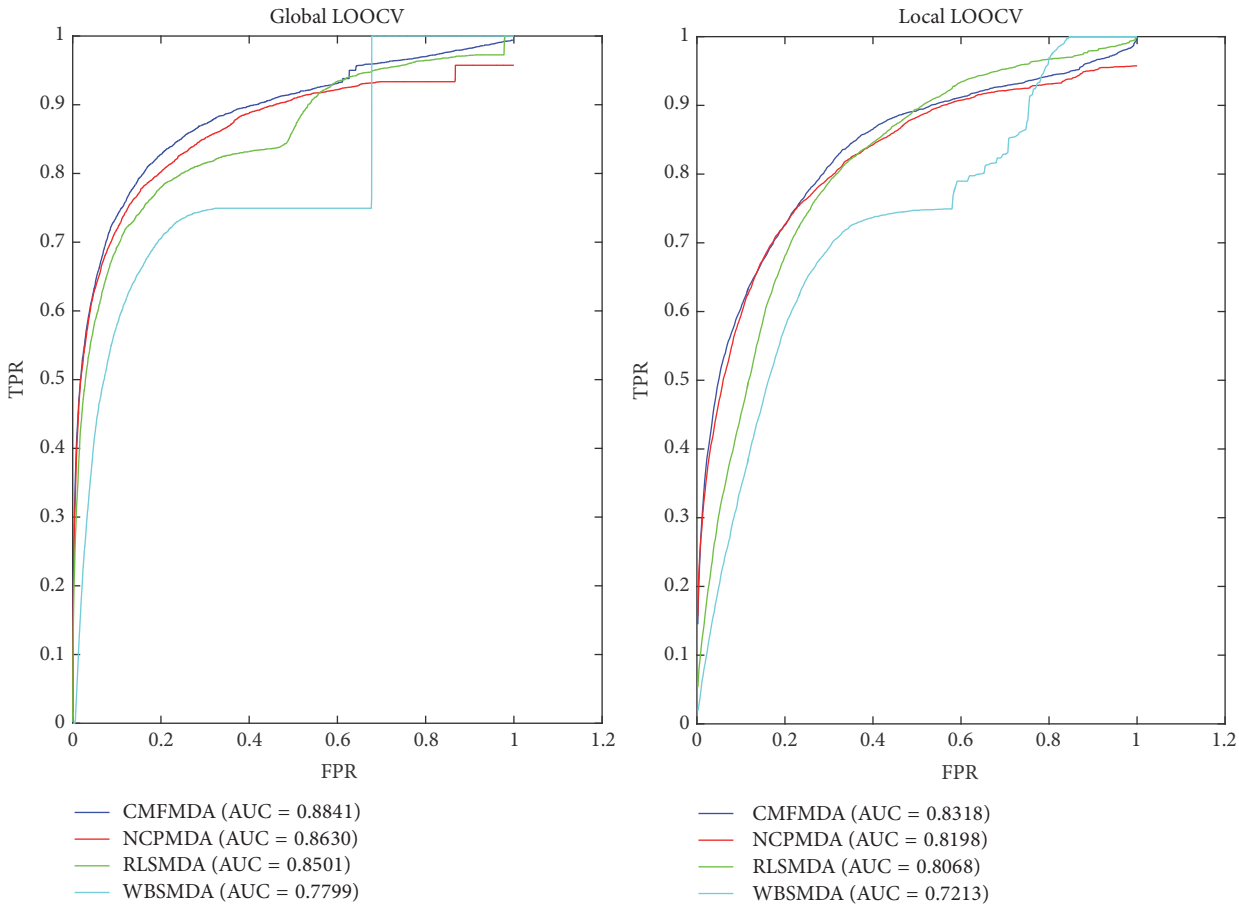


Figure 2: Performance comparisons between CMFMDA and three state-of-the-art disease-miRNA association prediction models (NCPMDA, RLSMDA, and WBSMDA) in terms of ROC curve and AUC based on local and global LOOCV, respectively. As a result, CMFMDA achieved AUCs of 0.8841 and 0.8318 in the global and local LOOCV, significantly outperforming all the previous classical models.

clear cell RCC (CCRCC), and papillary RCC (PRCC). Previous studies have shown that miRNAs play a significant part in Kidney Neoplasm [72–74]. In this paper, CMFMDA was employed to identify potential miRNAs associated with Kidney Neoplasms. As a result, 9 out of the top-10 candidates and 41 out of the top-50 candidates of Kidney Neoplasm related miRNAs were confirmed by dbDEMC and miR2Ddisease (See Table 2). For example, the serum level of mi-210 may be used as a novel noninvasive biomarker for the detection of CCRCC [75]. Experiment results demonstrate that mir-9 expression is correlative not only with the development of CCRCC, but also with the development of metastatic recurrence [76].

The results of cross-validation and independent case studies show that CMFMDA can satisfy the needs to identify potential miRNA-disease associations. Furthermore, all diseases in HMDD have been investigated by CMFMDA to predict potential miRNAs (See Supplementary Table 1 in Supplementary Material available online at https://doi.org/10.1155/2017/2498957). We hope that potential disease-miRNA association predicted by CMFMDA could be confirmed by further biological experiments.

4. Discussion

According to the assumption that functionally similar miRNAs are often associated with similar diseases, we proposed the computational model of Collaborative Matrix Factorization for MiRNA-Disease Association prediction (CMFMDA) to identify potential miRNA-disease associations by integrating miRNA functional similarity, disease semantic similarity, and experimentally verified miRNA-disease associations. We compare CMFMDA with the existing computational model: NCPMDA, RLSMDA, and WBSMDA, and concluded that CMFMDA has better prediction performance from these four models' obtained AUCs in the global LOOCV or local LOOCV, respectively. There are some reasons for the reliable performance of CMFMDA. Firstly, several types of experimentally confirmed biological datasets are used in CMFMDA, including known miRNA-disease associations, miRNA functional similarity network, and disease semantic similarity network, which help improve the prediction performance and reduce variance. Then, CMFMDA can work not only for known miRNA-disease association, but also for diseases and miRNAs without any known association. Finally,

Table 1: We implemented CMFMDA to predict potential Esophageal Neoplasms-related miRNAs. As a result, 9 out of the top 10 and 42 out of the top 50 predicted Esophageal Neoplasms related miRNAs were confirmed based on miR2Disease and dbDEMC (1st column: top 1–25; 2nd column: top 26–50).

miRNA	Evidence	miRNA	Evidence
hsa-mir-142	dbDEMC	hsa-mir-30c	dbDEMC
hsa-mir-1	dbDEMC	hsa-mir-212	Unconfirmed
hsa-mir-16	dbDEMC	hsa-mir-424	dbDEMC
hsa-mir-127	dbDEMC	hsa-mir-429	dbDEMC
hsa-mir-497	dbDEMC	hsa-mir-498	dbDEMC
hsa-mir-200b	dbDEMC	hsa-mir-340	Unconfirmed
hsa-mir-376c	Unconfirmed	hsa-mir-222	dbDEMC
hsa-mir-148b	dbDEMC	hsa-mir-146b	dbDEMC
hsa-mir-335	dbDEMC	hsa-mir-195	dbDEMC
hsa-mir-93	dbDEMC	hsa-mir-10b	dbDEMC
hsa-mir-125b	dbDEMC	hsa-mir-218	Unconfirmed
hsa-mir-18a	dbDEMC	hsa-mir-181a	dbDEMC
hsa-mir-17	dbDEMC	hsa-mir-137	dbDEMC
hsa-mir-30a	dbDEMC	hsa-let-7e	dbDEMC
hsa-mir-133b	dbDEMC	hsa-mir-181b	dbDEMC
hsa-mir-135a	dbDEMC	hsa-mir-106a	dbDEMC
hsa-mir-107	dbDEMC, miR2Disease	hsa-mir-19b	dbDEMC
hsa-mir-224	dbDEMC	hsa-mir-95	dbDEMC
hsa-mir-199b	dbDEMC	hsa-mir-122	Unconfirmed
hsa-mir-221	dbDEMC	hsa-mir-152	dbDEMC
hsa-mir-18b	dbDEMC	hsa-mir-370	dbDEMC
hsa-mir-191	dbDEMC	hsa-mir-30d	dbDEMC
hsa-let-7g	dbDEMC	hsa-mir-15b	dbDEMC
hsa-let-7f	Unconfirmed	hsa-mir-629	Unconfirmed
hsa-mir-494	dbDEMC	hsa-mir-204	Unconfirmed

as a global prediction model, CMFMDA could be used to predict all disease-related miRNA at the same time.

Although CMFMDA has better prediction performance, the limitation still exists in it and needs to be improved in the future. Firstly, CMFMDA may cause bias to miRNAs with more known associated diseases. Secondly, the known miRNA-disease associations with experimental evidences are still insufficient. The prediction performance of CMFMDA will be improved by integrating more reliable biological information [77–86]. Finally, how to more reasonably extract and integrate information from biological datasets should be investigated in the future.

5. Conclusions

Research has shown that the abnormal expression of miRNA plays a crucial role in the occurrence and development of human complex diseases. The in-depth study and analysis of diseases-related miRNA could help find new biomarker and therapies and then improve the survival rate of patients.

Table 2: We implemented CMFMDA to prioritize candidate miRNAs for Kidney Neoplasms based on known associations in the HMDD database. As a result, 9 out of the top 10 and 41 out of the top 50 predicted Kidney Neoplasms related miRNAs were confirmed based on miR2Disease and dbDEMC (1st column: top 1–25; 2nd column: top 26–50).

miRNA	Evidence	miRNA	Evidence
hsa-mir-429	dbDEMC	hsa-mir-34c	dbDEMC
hsa-mir-200b	dbDEMC, miR2Disease	hsa-mir-10b	dbDEMC
hsa-mir-200a	dbDEMC	hsa-mir-9	dbDEMC
hsa-mir-210	dbDEMC, miR2Disease	hsa-mir-199b	dbDEMC
hsa-mir-203	dbDEMC	hsa-mir-99b	dbDEMC
hsa-mir-218	dbDEMC	hsa-mir-126	dbDEMC, miR2Disease
hsa-mir-127	dbDEMC	hsa-mir-224	dbDEMC
hsa-mir-155	dbDEMC	hsa-mir-195	dbDEMC
hsa-mir-205	dbDEMC, miR2Disease	hsa-mir-145	dbDEMC
hsa-mir-196a	dbDEMC	hsa-mir-7	dbDEMC, miR2Disease
hsa-mir-143	dbDEMC	hsa-mir-142	Unconfirmed
hsa-mir-135a	Unconfirmed	hsa-mir-139	dbDEMC
hsa-mir-92a	Unconfirmed	hsa-mir-16	dbDEMC
hsa-mir-19a	dbDEMC	hsa-mir-183	dbDEMC
hsa-mir-29c	dbDEMC, miR2Disease	hsa-mir-296	Unconfirmed
hsa-mir-146a	dbDEMC	hsa-mir-367	Unconfirmed
hsa-mir-99a	dbDEMC	hsa-mir-15b	dbDEMC
hsa-mir-101	dbDEMC, miR2Disease	hsa-mir-204	dbDEMC
hsa-mir-199a	dbDEMC, miR2Disease	hsa-mir-196b	dbDEMC
hsa-mir-27a	dbDEMC, miR2Disease	hsa-mir-339	dbDEMC
hsa-mir-100	dbDEMC	hsa-mir-26a	dbDEMC, miR2Disease
hsa-mir-452	dbDEMC	hsa-mir-302c	Unconfirmed
hsa-mir-34b	dbDEMC	hsa-mir-302b	Unconfirmed
hsa-mir-93	dbDEMC	hsa-mir-107	dbDEMC
hsa-mir-34a	dbDEMC	hsa-mir-133b	Unconfirmed

Therefore, it is necessary to develop more effective computational models to identify potential miRNA-disease associations. In this paper, we presented a computational model CMFMDA to identify novel miRNA-disease associations. Except for disease semantic similarity and miRNA functional similarity, CMFMDA also uses known miRNA-disease associations to predict miRNA-disease associations. LOOCV was chosen to evaluate the predict performance of CMFMDA. The results of LOOCV and case studies show that CMFMDA has better prediction performance than other models. In other words, as an effective tool, CMFMDA can be used not only to predict potential miRNA-disease associations, but

also to identify new biomarker that gave new direction for diagnosis and treatment of human complex disease.

Disclosure

De-Shuang Huang is the corresponding author of this paper. Professor A. K. Nandi is a Distinguished Visiting Professor at Tongji University, Shanghai, China.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Acknowledgments

This work was supported by the grants of the National Science Foundation of China, nos. 61520106006, 61732012, 31571364, 61532008, U1611265, 61672382, 61772370, 61402334, 61472173, and 61472282, and China Postdoctoral Science Foundation [Grant nos. 2015M580352, 2017M611619, and 2016M601646].

References

[1] V. Ambros, "microRNAs: tiny regulators with great potential," *Cell*, vol. 107, no. 7, pp. 823–826, 2001.

[2] V. Ambros, "The functions of animal microRNAs," *Nature*, vol. 431, no. 7006, pp. 350–355, 2004.

[3] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[4] X. Chen, C. C. Yan, X. Zhang et al., "WBSMDA: within and between score for MiRNA-disease association prediction," *Scientific Reports*, vol. 6, Article ID 21106, 2016.

[5] R. Ibrahim, N. A. Yousri, M. A. Ismail, and N. M. El-Makky, "MiRNA and gene expression based cancer classification using self-learning and co-training approaches," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013*, pp. 495–498, Shanghai, China, December 2013.

[6] Y. Katayama, M. Maeda, K. Miyaguchi et al., "Identification of pathogenesis-related microRNAs in hepatocellular carcinoma by expression profiling," *Oncology Letters*, vol. 4, no. 4, pp. 817–823, 2012.

[7] G. Meister and T. Tuschl, "Mechanisms of gene silencing by double-stranded RNA," *Nature*, vol. 431, no. 7006, pp. 343–349, 2004.

[8] D. S. Huang and H.-J. Yu, "Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 2, pp. 457–467, 2013.

[9] S.-P. Deng and D. S. Huang, "SFAPS: an R package for structure/function analysis of protein sequences based on informational spectrum method," *Methods*, vol. 69, no. 3, pp. 207–212, 2014.

[10] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene *lin*-4 encodes small RNAs with antisense complementarity to *lin*-14," *Cell*, vol. 75, no. 5, pp. 843–854, 1993.

[11] B. J. Reinhart, F. J. Slack, M. Basson et al., "The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*," *Nature*, vol. 403, no. 6772, pp. 901–906, 2000.

[12] X. Chen and G.-Y. Yan, "Semi-supervised learning for potential human microRNA-disease associations inference," *Scientific Reports*, vol. 4, article 5501, 2014.

[13] S. Bandyopadhyay, R. Mitra, U. Maulik, and M. Q. Zhang, "Development of the human cancer microRNA network," *Silence*, vol. 1, article 6, 2010.

[14] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic Acids Research*, vol. 42, pp. D68–D73, 2013.

[15] X. Chen, C. Clarence Yan, X. Zhang et al., "RBMMMDA: predicting multiple types of disease-microRNA associations," *Scientific Reports*, vol. 5, Article ID 13877, 2015.

[16] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[17] E. A. Miska, "How microRNAs control cell division, differentiation and death," *Current Opinion in Genetics and Development*, vol. 15, no. 5, pp. 563–568, 2005.

[18] A. M. Cheng, M. W. Byrom, J. Shelton, and L. P. Ford, "Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis," *Nucleic Acids Research*, vol. 33, no. 4, pp. 1290–1297, 2005.

[19] Q. Cui, Z. Yu, E. O. Purisima, and E. Wang, "Principles of microRNA regulation of a human cellular signaling network," *Molecular Systems Biology*, vol. 2, article 46, 2006.

[20] X. Chen, M. X. Liu, and G. Y. Yan, "RWRMDA: predicting novel human microRNA-disease associations," *Molecular BioSystems*, vol. 8, no. 10, pp. 2792–2798, 2012.

[21] J. Li, Y. Liu, X. Xin et al., "Evidence for positive selection on a number of microRNA regulatory interactions during recent human evolution," *PLoS Genetics*, vol. 8, no. 3, Article ID e1002578, 2012.

[22] K. Chen and N. Rajewsky, "Natural selection on human microRNA binding sites inferred from SNP data," *Nature Genetics*, vol. 38, no. 12, pp. 1452–1456, 2006.

[23] M. A. Saunders, H. Liang, and W.-H. Li, "Human polymorphism at microRNAs and microRNA target sites," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 9, pp. 3300–3305, 2007.

[24] P. Sethupathy and F. S. Collins, "MicroRNA target site polymorphisms and human disease," *Trends in Genetics*, vol. 24, no. 10, pp. 489–497, 2008.

[25] S.-P. Deng, L. Zhu, and D. S. Huang, "Predicting hub genes associated with cervical cancer through gene co-expression networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 1, pp. 27–35, 2016.

[26] S.-P. Deng, L. Zhu, and D. S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," *BMC Genomics*, vol. 16, no. 3, supplement 3, article S4, 2015.

[27] H. Luo, H. Zhang, Z. Zhang et al., "Down-regulated miR-9 and miR-433 in human gastric carcinoma," *Journal of Experimental and Clinical Cancer Research*, vol. 28, article 82, 2009.

[28] X.-M. Li, A.-M. Wang, J. Zhang, and H. Yi, "Down-regulation of miR-126 expression in colorectal cancer and its clinical significance," *Medical Oncology*, vol. 28, no. 4, pp. 1054–1057, 2011.

[29] V. Patel, D. Williams, S. Hajarnis et al., "MiR-17~92 miRNA cluster promotes kidney cyst growth in polycystic kidney

disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 26, pp. 10765–10770, 2013.

[30] G. A. Calin and C. M. Croce, "MicroRNA signatures in human cancers," *Nature Reviews Cancer*, vol. 6, no. 11, pp. 857–866, 2006.

[31] Q. Jiang, Y. Hao, G. Wang et al., "Prioritization of disease microRNAs through a human phenome-microRNAome network," *BMC Systems Biology*, vol. 4, supplement 1, article S2, 2010.

[32] Q. Jiang, G. Wang, S. Jin, Y. Li, and Y. Wang, "Predicting human microRNA-disease associations based on support vector machine," *International Journal of Data Mining and Bioinformatics*, vol. 8, no. 3, pp. 282–293, 2013.

[33] C.-H. Zheng, D. S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 599–607, 2009.

[34] X. Chen, "Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA," *Scientific Reports*, vol. 5, Article ID 13186, 2015.

[35] X. Chen, C. C. Yan, C. Luo, W. Ji, Y. Zhang, and Q. Dai, "Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity," *Scientific Reports*, vol. 5, Article ID 11338, 2015.

[36] X. Chen and G.-Y. Yan, "Novel human lncRNA-disease association inference based on lncRNA expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617–2624, 2013.

[37] X. Chen, M.-X. Liu, Q.-H. Cui, and G.-Y. Yan, "Prediction of disease-related interactions between microRNAs and environmental factors based on a semi-supervised classifier," *PLoS ONE*, vol. 7, no. 8, Article ID e43425, 2012.

[38] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Scientific Reports*, vol. 5, Article ID 16840, 2015.

[39] C.-H. Zheng, L. Zhang, V. T.-Y. Ng, S. C.-K. Shiu, and D. S. Huang, "Molecular pattern discovery based on penalized matrix decomposition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1592–1603, 2011.

[40] L. Zhu, S.-P. Deng, and D. S. Huang, "A two-stage geometric method for pruning unreliable links in protein-protein networks," *IEEE Transactions on Nanobioscience*, vol. 14, no. 5, pp. 528–534, 2015.

[41] M. Lu, Q. Zhang, M. Deng et al., "An analysis of human microRNA and disease associations," *PLoS ONE*, vol. 3, no. 10, Article ID e3420, 2008.

[42] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[43] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[44] C. Pasquier and J. Gardès, "Prediction of miRNA-disease associations with a vector space model," *Scientific Reports*, vol. 6, Article ID 27036, 2016.

[45] T. D. Le, J. Zhang, L. Liu, and J. Li, "Computational methods for identifying miRNA sponge interactions," *Briefings in Bioinformatics*, p. bbw042, 2016.

[46] D. S. Huang and C. H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15, pp. 1855–1862, 2006.

[47] X. Chen, C. C. Yan, X. Zhang, Z.-H. You, Y.-A. Huang, and G.-Y. Yan, "HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction," *Oncotarget*, vol. 7, no. 40, pp. 65257–65269, 2016.

[48] J.-Q. Li, Z.-H. Rong, X. Chen, G.-Y. Yan, and Z.-H. You, "MCMDA: matrix completion for MiRNA-disease association prediction," *Oncotarget*, vol. 8, pp. 21187–21199, 2017.

[49] Z.-H. You, Z.-A. Huang, Z. Zhu et al., "PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction," *PLOS Computational Biology*, vol. 13, no. 3, Article ID e1005455, 2017.

[50] J. Xu, C.-X. Li, J.-Y. Lv et al., "Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer," *Molecular Cancer Therapeutics*, vol. 10, no. 10, pp. 1857–1866, 2011.

[51] X. Chen, "miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method," *Molecular BioSystems*, vol. 12, no. 2, pp. 624–633, 2016.

[52] X. Chen, Q.-F. Wu, and G.-Y. Yan, "RKNNMDA: ranking-based KNN for MiRNA-disease association prediction," *RNA Biology*, pp. 1–11, 2017.

[53] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1025–1033, Chicago, Ill, USA, August 2013.

[54] A. Ezzat, P. Zhao, M. Wu, X. Li, and C. Kwoh, "Drug-target interaction prediction with graph regularized matrix factorization," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 3, pp. 646–656, 2016.

[55] Y. Li, C. Qiu, J. Tu et al., "HMDD v2.0: a database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, pp. D1070–D1074, 2014.

[56] C. Gu, B. Liao, X. Li, and K. Li, "Network consistency projection for human miRNA-disease associations inference," *Scientific Reports*, vol. 6, Article ID 36054, 2016.

[57] Q. Jiang, Y. Wang, Y. Hao et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D98–D104, 2009.

[58] Z. Yang, F. Ren, C. Liu et al., "dbDEMC: a database of differentially expressed miRNAs in human cancers," *BMC Genomics*, vol. 11, supplement 4, article S5, 2010.

[59] M. Kano, N. Seki, N. Kikkawa et al., "MiR-145, miR-133a and miR-133b: tumor-suppressive miRNAs target FSCN1 in esophageal squamous cell carcinoma," *International Journal of Cancer*, vol. 127, no. 12, pp. 2804–2814, 2010.

[60] P. C. Enzinger and R. J. Mayer, "Esophageal cancer," *The New England Journal of Medicine*, vol. 349, no. 23, pp. 2241–2252, 2003.

[61] B. He, B. Yin, B. Wang, Z. Xia, C. Chen, and J. Tang, "microRNAs in esophageal cancer (Review)," *Molecular Medicine Reports*, vol. 6, no. 3, pp. 459–465, 2012.

[62] Z. Xie, G. Chen, X. Zhang et al., "Salivary microRNAs as promising biomarkers for detection of esophageal cancer," *PLoS ONE*, vol. 8, no. 4, Article ID e57502, 2013.

[63] J. Wan, W. Wu, Y. Che, N. Kang, and R. Zhang, "Insights into the potential use of microRNAs as a novel class of biomarkers in esophageal cancer," *Diseases of the Esophagus*, vol. 29, no. 5, pp. 412–420, 2016.

[64] B.-J. Zhang, H.-Y. Gong, F. Zheng, D.-J. Liu, and H.-X. Liu, "Up-regulation of miR-335 predicts a favorable prognosis in esophageal squamous cell carcinoma," *International Journal of Clinical and Experimental Pathology*, vol. 7, no. 9, pp. 6213–6218, 2014.

[65] A. Jemal, R. Siegel, E. Ward et al., "Cancer statistics, 2006," *CA: A Cancer Journal for Clinicians*, vol. 56, no. 2, pp. 106–130, 2006.

[66] N. M. A. White, T. T. Bao, J. Grigull et al., "MiRNA profiling for clear cell renal cell carcinoma: biomarker discovery and identification of potential controls and consequences of miRNA dysregulation," *The Journal of Urology*, vol. 186, no. 3, pp. 1077–1083, 2011.

[67] N. M. A. White, H. W. Z. Khella, J. Grigull et al., "MiRNA profiling in metastatic renal cell carcinoma reveals a tumour-suppressor effect for miR-215," *British Journal of Cancer*, vol. 105, no. 11, pp. 1741–1749, 2011.

[68] R. Siegel, D. Naishadham, and A. Jemal, "Cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 62, no. 1, pp. 10–29, 2012.

[69] W. M. Linehan, "Genetic basis of kidney cancer: role of genomics for the development of disease-based therapeutics," *Genome Research*, vol. 22, no. 11, pp. 2089–2100, 2012.

[70] W. M. Linehan, M. M. Walther, and B. Zbar, "The genetic basis of cancer of the kidney," *The Journal of Urology*, vol. 170, no. 6, part 1, pp. 2163–2172, 2003.

[71] W. M. Linehan and B. Zbar, "Focus on kidney cancer," *Cancer Cell*, vol. 6, no. 3, pp. 223–228, 2004.

[72] U. Senanayake, S. Das, P. Vesely et al., "miR-192, miR-194, miR-215, miR-200c and miR-141 are downregulated and their common target ACVR2B is strongly expressed in renal childhood neoplasms," *Carcinogenesis*, vol. 33, no. 5, pp. 1014–1021, 2012.

[73] H. Hidaka, N. Seki, H. Yoshino et al., "Tumor suppressive microRNA-1285 regulates novel molecular targets: aberrant expression and functional significance in renal cell carcinoma," *Oncotarget*, vol. 3, pp. 44–57, 2012.

[74] M. Redova, A. Poprach, J. Nekvindova et al., "Circulating miR-378 and miR-451 in serum are potential biomarkers for renal cell carcinoma," *Journal of Translational Medicine*, vol. 10, article 55, 2012.

[75] A. Zhao, G. Li, M. Péoc'h, C. Genin, and M. Gigante, "Serum miR-210 as a novel biomarker for molecular diagnosis of clear cell renal cell carcinoma," *Experimental and Molecular Pathology*, vol. 94, no. 1, pp. 115–120, 2013.

[76] M. A. T. Hildebrandt, J. Gu, J. Lin et al., "Hsa-miR-9 methylation status is associated with cancer development and metastatic recurrence in patients with clear cell renal cell carcinoma," *Oncogene*, vol. 29, no. 42, pp. 5724–5728, 2010.

[77] Y. Liu, D. A. Tennant, Z. Zhu, J. K. Heath, X. Yao, and S. He, "DiME: A scalable disease module identification algorithm with application to glioma progression," *PLoS ONE*, vol. 9, no. 2, Article ID e86693, 2014.

[78] L. Wong, Z.-H. You, Z. Ming, J. Li, X. Chen, and Y.-A. Huang, "Detection of interactions between proteins through rotation forest and local phase quantization descriptors," *International Journal of Molecular Sciences*, vol. 17, no. 1, p. 21, 2015.

[79] Z. H. You, L. Zhu, C. H. Zheng, H. Yu, S. Deng, and Z. Ji, "Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set," *BMC Bioinformatics*, vol. 15, supplement 15, p. S9, 2014.

[80] L. Zhu, W.-L. Guo, S.-P. Deng, and D. S. Huang, "ChIP-PIT: enhancing the analysis of ChIP-Seq data using convex-relaxed pair-wise interaction tensor decomposition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 1, pp. 55–63, 2016.

[81] D. S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Current Protein and Peptide Science*, vol. 15, no. 6, pp. 553–560, 2014.

[82] L. Zhu, Z. You, D. S. Huang, and B. Wang, "*t*-LSE: a novel robust geometric approach for modeling protein-protein interaction networks," *PLoS ONE*, vol. 8, no. 4, Article ID e58368, 2013.

[83] D. S. Huang and J.-X. Du, "A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks," *IEEE Transactions on Neural Networks*, vol. 19, no. 12, pp. 2099–2115, 2008.

[84] D. S. Huang, *Systematic Theory of Neural Networks for Pattern Recognition*, Publishing House of Electronic Industry of China, Beijing, China, 1996.

[85] D. S. Huang and W. Jiang, "A general CPL-AdS methodology for fixing dynamic parameters in dual environments," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 5, pp. 1489–1500, 2012.

[86] D. S. Huang, "Radial basis probabilistic neural networks: model and application," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 13, no. 7, pp. 1083–1101, 1999.

Hindawi
Complexity
Volume 2017, Article ID 3017632, 9 pages
https://doi.org/10.1155/2017/3017632

Research Article

Exploring the Limitations of Peripheral Blood Transcriptional Biomarkers in Predicting Influenza Vaccine Responsiveness

Luca Marchetti,[1] Emilio Siena,[2] Mario Lauria,[1,3] Denise Maffione,[2] Nicola Pacchiani,[2] Corrado Priami,[1,3,4] and Duccio Medini[2]

[1]*The Microsoft Research-University of Trento Centre for Computational and Systems Biology (COSBI), 38068 Rovereto, Italy*
[2]*GSK Vaccines, 53100 Siena, Italy*
[3]*Department of Mathematics, University of Trento, Povo, 38123 Trento, Italy*
[4]*Department of Computer Science, Stanford University, Stanford, CA, USA*

Correspondence should be addressed to Luca Marchetti; marchetti@cosbi.eu

Received 26 April 2017; Revised 2 August 2017; Accepted 10 August 2017; Published 28 September 2017

Academic Editor: Min Li

Systems biology has been recently applied to vaccinology to better understand immunological responses to the influenza vaccine. Particular attention has been paid to the identification of early signatures capable of predicting vaccine immunogenicity. Building from previous studies, we employed a recently established algorithm for signature-based clustering of expression profiles, SCUDO, to provide new insights into why blood-derived transcriptome biomarkers often fail to predict the seroresponse to the influenza virus vaccination. Specifically, preexisting immunity against one or more vaccine antigens, which was found to negatively affect the seroresponse, was identified as a confounding factor able to decouple early transcriptome from later antibody responses, resulting in the degradation of a biomarker predictive power. Finally, the broadly accepted definition of seroresponse to influenza virus vaccine, represented by the maximum response across the vaccine-targeted strains, was compared to a composite measure integrating the responses against all strains. This analysis revealed that composite measures provide a more accurate assessment of the seroresponse to multicomponent influenza vaccines.

1. Introduction

Vaccines represent one of the most effective interventions to control infectious diseases. Despite the many successes [1, 2], however, we are still missing an effective vaccine against current global pandemics such as HIV, malaria, and tuberculosis.

Most vaccines have been developed empirically, against pathogens characterized by limited antigenic variation and that can be neutralized by antibodies alone. Protection against those organisms displaying a fast rate of antigenic variation, such as the HIV virus, or complex host-pathogen interaction biology, such as *Plasmodium falciparum*, requires vaccines that activate multiple arms of the immune response and that do not rely only on the production of neutralizing serum antibodies [3]. Such vaccines will only be accomplished through a careful design involving new-generation antigens, designed to induce optimal and broadly protective immune responses [4], and adjuvants, which are able to guide the type of adaptive response to produce the most effective forms of immunity for a specific pathogen [5].

Paramount to this novel approach will be the exploitation of new high-throughput technologies combined with the use of advanced algorithms to extract meaningful information from the large sets of generated data. Conventional immunological methods, such as ELISA and flow cytometry, can only assess a limited number of components of the immune system at a given time and, as such, they are not suited for addressing the complexity of the human immune system. Novel technologies, in contrast, allow easily quantifying the abundance of cells, RNA, proteins, and other metabolites across different tissues with high throughput. The availability of high dimensional data, coupled with computational modeling, holds the potential to provide new insights into

Table 1: Transcriptional data sets used in this study.

Set	Accession	Reference	Samples (subjects)	Platform	Description
Set #1	GSE29614	Nakaya et al., Nat Immunol 2011	27 (9)	Affymetrix U133 Plus 2.0	Time course of young adults vaccinated with influenza TIV during 2007/08 flu season
Set #2	GSE29617	Nakaya et al., Nat Immunol 2011	80 (28)	Affymetrix U133 Plus 2.0	Time course of young adults vaccinated with influenza TIV during 2008/09 flu season
Set #3	GSE74817	Nakaya et al., Immunity 2015	621 (212)	Affymetrix U133 Plus PM	Time course of adults vaccinated with influenza TIV during the years 2009–2012
Set #4	GSE48018	Bucasas et al., J Infect Dis 2011	431 (116)	Illumina HumanHT-12 V3.0	Time course of adults vaccinated with influenza TIV

the mechanisms underlying the immunological response to vaccination [6].

In addition to being foundational to a rational vaccine design, an increased understanding of the mechanisms involved in the response to vaccination will also guide the optimization of immunogenicity and reactogenicity profiles of existing vaccines. Also, deciphering the early events following vaccination will provide molecular signatures predictive of vaccine efficacy or safety which, in turn, will enable a prospective identification of subjects with a suboptimal response profile and speed up future clinical trials [7].

Recently, several studies [8–11] have analyzed the transcriptome profile of peripheral blood, or blood-derived cells, to study immunity to the trivalent influenza vaccine (TIV) in humans. These studies decoded the global pattern of transcriptional response to TIV vaccination and independently reported that the extent of upregulation of a set of interferon-inducible genes, one to three days after vaccination, and of genes involved in plasmablasts differentiation and activity, seven days after vaccination, correlates with the magnitude of serum functional antibody titers measured after one month. Among those, Nakaya and colleagues [10, 11] extended this further, by applying a machine learning algorithm to identify sets of genes that could predict subjects' seroresponse across independent TIV vaccination trials.

Despite their extensive use, most machine learning approaches only provide information in the form of disease- or treatment-specific gene signatures, along with their associated predictive power. This limited set of information hardly provides any supporting evidence in the investigation of the causes of an incorrect prediction. This is especially true for vaccine response prediction, due to the complexity and multicomponent nature of the immune system and the genetic heterogeneity across patients.

With this study we sought to expand our understanding of the early events following vaccination by looking for transcriptional signatures, derived from peripheral blood mononuclear cells (PBMCs), whose combined response is associated with the magnitude of functional antibody responses measured four weeks after vaccination. Using two sets of published profiles covering two vaccination campaigns (2007/2008 and 2008/2009, resp., for set #1 and set #2, see Table 1), we identified a biomarker consisting of 207 transcripts whose early (3–7 days after vaccination) response was capable of prospectively discriminating low- from high-responder subjects. Following best practices in biomarker identifications, we used one cohort (2008/2009 campaign) for training and a different one (2007/2008 campaign) for validation. The good performance of the identified biomarker was further confirmed using a third set of profiles from a separate study including vaccination campaigns spanning a total of three years (2009/2010/2011, set #3). Despite the consistently high prediction accuracy achieved by the biomarker across cohorts, we noticed that not all the subjects were correctly classified. We investigated the possible reasons for the observed misclassifications and identified preexisting immunity against one or more of the vaccine antigens to be a confounding factor. While this finding confirms earlier reports on preexisting immunity negatively affecting the seroresponse to vaccination, we additionally identified a difference in the response between subjects with and without preexisting immunity. Specifically, we found preexisting immunity to affect the 24 h response and to have little to no effect on whole blood transcriptomes profiled 3 or 14 days after vaccination. Therefore, preexisting immunity can decouple early transcriptome from later antibody responses, resulting in the degradation of a biomarker predictive power. This finding was confirmed using a fourth set of profiles (set #4).

We finally related our predictions to a composite HAI response measure, which integrates the responses to all strains represented in the vaccine, and showed how this can be a more informative measure for influenza vaccine responsiveness compared to the widely accepted maximum-across strains one.

2. Results and Discussion

We analyzed a total of four transcriptome datasets, whose essential features are reported in Table 1.

2.1. A Biomarker of 207 Transcripts, Primarily Involved in Cell Proliferation and Cytokine Signaling, Predicts Seroresponse to Seasonal Influenza Vaccination. We employed an optimized version of SCUDO [12–15], an algorithm for signature-based clustering of expression profiles that relies on a completely new way of addressing the classification problem. The algorithm is based on the concept of subject-specific signatures, rather than disease-specific signatures, to divide samples. Briefly, the method first seeks to summarize the characteristics of each sample by means of a subject-specific, rank-based signature and then it performs a systematic, all-to-all signature comparison to segregate samples into different

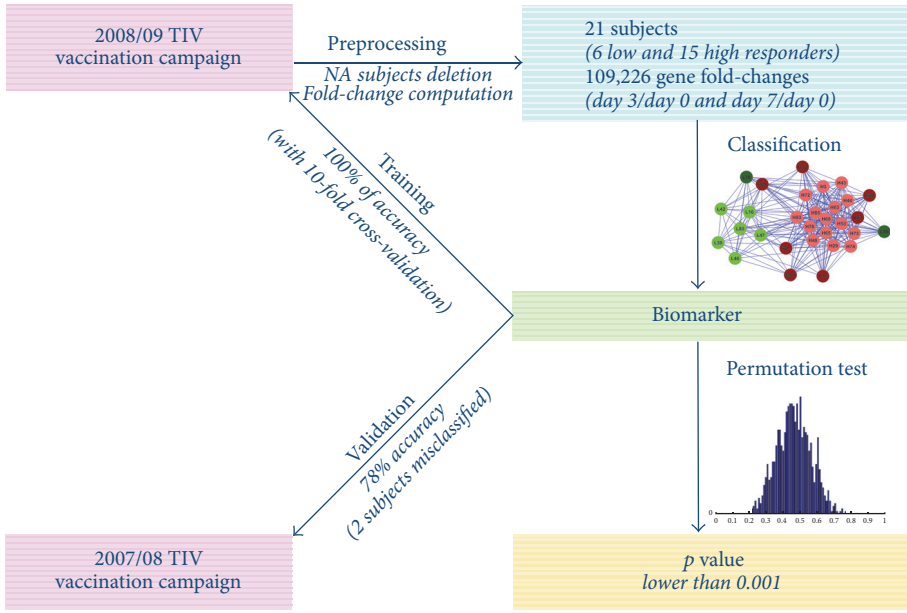


Figure 1: *Graphical representation of the analysis pipeline implemented to compute biomarkers (see Materials and Methods).* The analysis started with the dataset related to the 2008/09 TIV vaccination campaign (set #2). After a preliminary preprocessing of the dataset for removing subjects with intermediate answer to the vaccine ("NA subjects," maximum $HAI_{Day\ 28/Day\ 0}$ across strains equal to 4) and for computing gene fold changes (day 3/day 0 and day 7/day 0), we applied the classification algorithm SCUDO to obtain the biomarker. The classification process on the training dataset exhibited 100% accuracy with a 10-fold cross-validation scheme. Moreover, its statistical significance was assessed by a permutation test providing a p value < 0.001. Finally, the identified biomarker was validated using an independent dataset (2007/08 TIV vaccination campaign, set #1). The classification on the validation dataset exhibited 78% accuracy.

groups. The result of the comparison can be represented in the form of a graph, which allows for a qualitative, as opposed to solely quantitative, interpretation of the classification performance. Therefore, the algorithm combines the benefits of rank-based classification, dimensionality reduction, and the power of network analysis. The computation of subject-specific signatures also allows the characterization of specific differences between subjects of the same class by highlighting intraclass variability that may cause incorrect predictions. Subject-specific signatures can then be merged together to obtain a unique disease-specific biomarker.

After stratifying vaccines into high responders (maximum hemagglutinin inhibition (HAI) titer fold increase > 4) and low responders (maximum HAI fold increase < 4) according to [10], we applied SCUDO for predicting subjects' membership to the correct category, based on transcriptome response profiles derived from PBMCs (see Materials and Methods and Figure 1). SCUDO produced a biomarker of 207 transcripts whose regulation 3 or 7 days after immunization was able to predict seroconversion with 100% and 78% accuracy within the training (2008-2009 trial, set #2) and an independent validation (2007-2008 trial, set #1) sets, respectively (see Supplementary File S1 in Supplementary Material available online at https://doi.org/10.1155/2017/3017632). The length of the biomarker produced by the algorithm was optimized to maximize both the classification accuracy and the amount of information about the mechanisms responsible for seroconversion. In a previous work [12] we showed that the length of the signatures is not critical, because often the same level of classification accuracy can be obtained with a range of signature sizes. For this reason, we also sought to identify the minimal set of genes that could predict vaccination outcome without a significant loss in the prediction accuracy compared to the identified biomarker. The algorithm identified a set of 12 transcripts, which showed prediction accuracy comparable to that of the longer biomarker (see Supplementary File S2). Interestingly, all 12 selected transcripts were also included in the 207 transcripts set. Similarly to the 207 transcripts biomarker, this minimal biomarker relied on the fold change of some transcripts at day 3 and of some others at day 7 after immunization. We verified that transcript regulation at day 3 or day 7 alone is not enough for an accurate subject classification (not shown).

One hundred and sixty-seven (167) of the 207 transcripts could be mapped into functionally annotated genes (see Supplementary File S1). Functional analysis revealed that the biomarker is enriched for genes involved in cell proliferation (IFIT3, PML, PTPN6, PTPRU, TBRG1, WNK2, and ZEB1) and in cytokine signaling (DDX58, IFIT3, PIK3R1, PIK3R5, PML, PTPN6, and SKP1), supporting previous evidence of cell- and interferon-mediated immune responses to be linked to humoral responses to influenza vaccination. Of note, also the CAMK4 kinase, which was experimentally shown to be involved in the regulation of antibody response [10], was included in the biomarker.

2.2. Preexisting Immunity Interferes with Response Outcome Predictability. Subject classification was based on a similarity

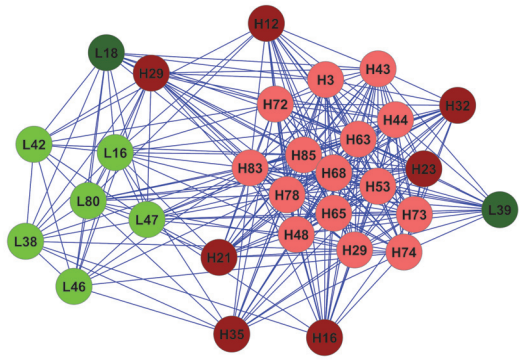


Figure 2: *Prediction of seroresponse to TIV vaccination.* The graph represents sample classification by means of the 207 transcripts biomarker. Nodes represent subjects and the length of connecting edges is proportional to the distance between subject signatures. Node colors indicate subjects' class (green for low responders and red for high responders); darker colors represent samples included in the validation set (2007-2008 TIV campaign, set #1), while lighter colors represent samples in the training set (2008-2009 TIV campaign, set #2). Node labels indicate the sample class (H for high responders and L for low responders) and sample identifier.

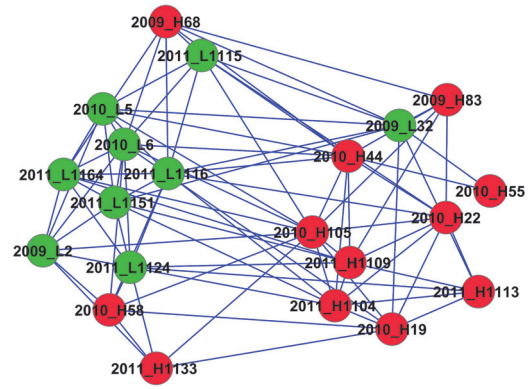


Figure 3: *Additional validation of the 207 transcripts biomarker.* The graph represents samples classification by means of the 207 transcripts biomarker of subjects included in the testing group of [11], which is related to the TIV vaccination campaigns of 2009, 2010, and 2011 (set #3). Nodes represent subjects and the length of connecting edges is proportional to the distance between subjects' signatures. Node colors indicate subjects' class (green for low responders and red for high responders). Node labels indicate the year of the vaccination campaign, sample class (H for high responders and L for low responders), and sample identifier. Four out of 21 subjects were misclassified (subjects 2009_H68, 2010_H58, 2011_H1133, and 2009_L32).

map that could be represented in the form of a graph in which nodes correspond to subjects and edge length encodes the level of similarity between subject-specific signatures (short edge = high similarity, long edge = low similarity, and no edge = negligible similarity). This provided a graphical representation of the underlying sample structure (Figure 2) showing how the biomarker successfully stratified the tested subjects into high responders and low responders. One of the two misclassified subjects, subject H29, showed some similarity with members from both clusters, reflecting a suboptimal signature performance. The other misclassified subject, subject L39, was closely associated with the high responders group despite its modest HAI response. This same subject was consistently misclassified also by the 12 transcripts biomarker.

As a further validation of the identified biomarker, we tested it on an additional set of transcriptional profiles derived from three independent vaccination campaigns. Specifically, we used the testing cohort described in Nakaya et al. [11], which is related to the TIV vaccination campaigns of 2009, 2010, and 2011 (set #3). The testing cohort includes 21 subjects out of a total of 212, selected because only for them the classification in high/low responders was available. Despite the fact that none of the subjects from these vaccination campaigns were used for training and that the data was collected as part of a different study, we obtained a classification accuracy of 81% (4 out of 21 subjects were misclassified; see Figure 3), which is in close agreement with what was reported in the original study [11].

Motivated by the consistently good performance of the biomarker across cohorts, we conducted an investigation on the possible reasons explaining why some subjects escape correct classification. Focusing on subject L39 as a representative case, we identified a correlation with an unusually high level of preexisting antibodies against the strain H3N2 (HAI titer = 2560), which possibly interfered with the subject's seroresponse. This particular evidence was in line with the several independent studies that reported that the existence of preexisting immunity (defined as the presence of preimmune antigen-specific antibodies) is associated with a reduced vaccine responsiveness [16–18]. Subject L39's proximity to the high responders cluster (Figure 2) indicates that these experienced a similar transcriptional response to the vaccine stimulus, suggesting that the seroresponse inhibition played by preexisting immunity depends on mechanisms that are not captured by the day 3 or day 7 PBMCs transcriptomes.

In order to substantiate this conclusion, which was derived from a single subject, we run a further investigation aimed at characterizing the influence of preimmune antigen-specific antibodies on the transcriptional response to influenza vaccination. Whole Blood (WB) transcriptomes, assessed in a pool of 119 healthy individuals vaccinated with a 2008-2009 trivalent influenza vaccine [8], were used for this analysis (set #4). From the initial pool, subjects were selected and stratified based on their baseline antigen-specific antibody level. Fifteen (15) subjects were found to have negligible preexisting immunity to the vaccine (maximum HAI titer across the three antigens ≤ 16) while 20 showed signs of preexposure to one or more of the vaccine antigens (maximum HAI titer across the three antigens ≥ 512). Transcriptome responses (fold changes from baseline measured 1, 3, and 14 days after vaccination) were then compared among the two groups. Due to inconsistencies (use of WB rather than PBMCs and lack of day 7 transcriptome profiling) between this and the previously introduced studies, it was not possible

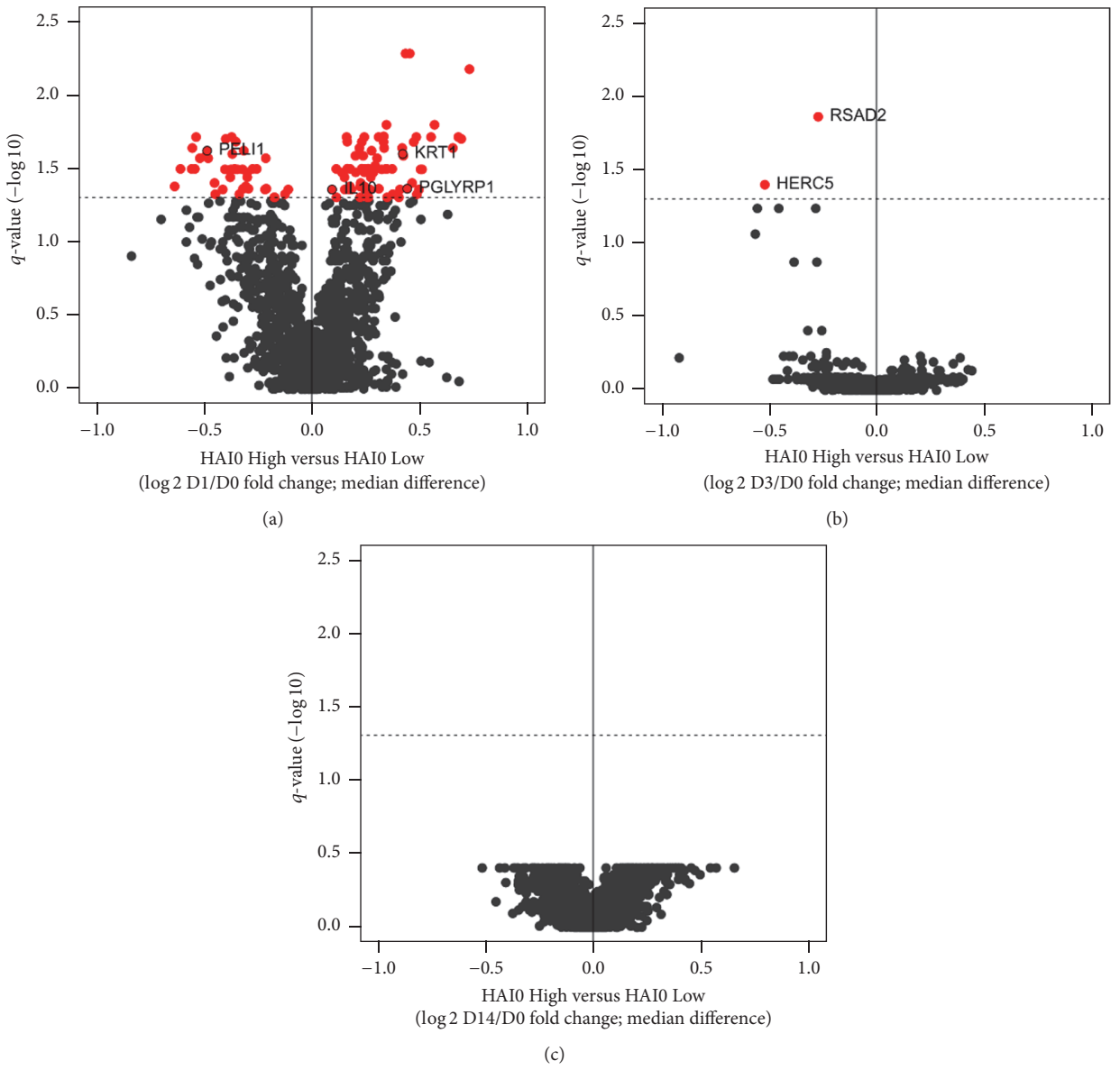


Figure 4: *Differentially responsive genes in subjects with low and high preexisting immunity to the influenza vaccine, set #4.* Volcano plots showing the differences in gene responsiveness (fold change from baseline) between subjects with high (High) and low (Low) baseline antibody titers ($HAI_{Day\ 0}$). *Y*-axes represent the *q*-values computed by adjusting Wilcoxon test *p* values using the Benjamini-Hochberg procedure. *X*-axes represent the difference in the median fold change from baseline response between subjects with high and low baseline immunity. Results are reported for day 1 (a), day 3 (b), and day 14 (c) time points.

to apply SCUDO to this dataset for confirmatory purposes. We note that our use of WB gene expression profiles is not ideal to confirm a hypothesis originally formulated while observing PBMC profiles. However previous studies have shown that there is a substantial overlap between the transcriptional responses for these two cell lines for a number of conditions similar to ours (see, e.g., Bondar et al. [19], for the case of inflammatory response). Thus the general evolution of transcriptional profiles in WB following vaccination is highly suggestive of those in PBMC and provides strong support if not conclusive evidence for our hypothesis of negligible influence of preexisting immunity on transcriptional response at days 3–7 after vaccination.

Differential gene expression analysis identified 118 transcripts whose response 1 day following vaccination differed among subjects with high and low preimmune status (Figure 4(a)). Among these, 77 were found to be more responsive in subjects with high baseline immunity, while 41 were less responsive (see Supplementary File S3). Functional analysis of these genes did not result in any enriched gene ontology (data not shown). Nonetheless, multiple genes participating in the inflammatory response (IL10, KRT1, PELI1,

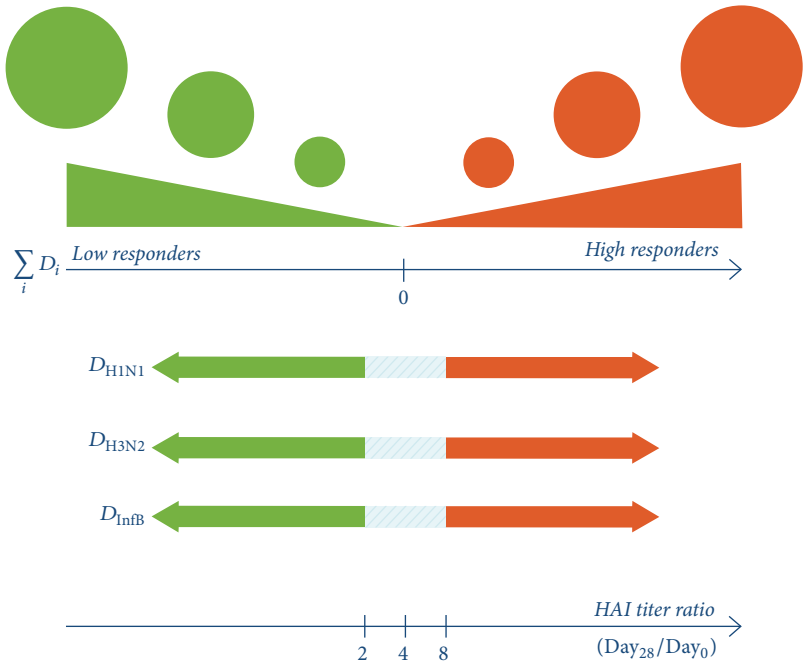


Figure 5: *Rationale for the computation of the titer response score.* The score integrates the responses obtained with the three strains represented in the TIV (H1N1, H3N2, and influenza B) by summing the logarithmic deviations of the $HAI_{Day\ 28/Day\ 0}$ ratios from the threshold value of the class to which the subject belongs (2 for low responders, 8 for high responders). The circles of different radius on the top of the figure represent how the TRS was encoded as node size in Figure 6.

and PLGYRP1) were present. Interestingly, IL10, KRT1, and PGLYRP1, all negative regulators of inflammatory response, were more responsive in subjects with preexisting immunity, while PELI1, a TRIF-dependent Toll-like receptor agonist, was less responsive in this group (Figure 4(a)). This specific response pattern suggests that preexisting immunity may act as negative feedback on innate immune activation when triggered by a recurring antigen. Differently, later time points (≥3 days after vaccination) did not show such a strong effect (Figures 4(b) and 4(c)). Only two interferon-inducible genes, HERC5 and RSAD2, were found to be more responsive in individuals with no preexisting immunity 3 days after vaccination.

Overall, these findings are in agreement with our initial hypothesis that while preexisting immunity can inhibit seroresponse to influenza vaccination, it does not affect peripheral blood-derived transcriptome profiles assessed 3 or more days after vaccination. This directly translates in the inability of transcriptional biomarkers, based on information collected from peripheral bloodstream 3 or more days after vaccination, to correctly account for the preimmune status of a vaccines.

It must be noted, however, that most of the subjects (28 out of 30) employed in this specific analysis were seroresponders (maximum $HAI_{Day\ 28/Day\ 0}$ fold increase across strains > 4), making the proper assessment of transcriptional differences between responder and nonresponder subjects in case of presence or absence of preexisting immunity not possible.

2.3. Maximum HAI Response across Strains Is an Oversimplified Measure for Assessing Influenza Vaccine Responsiveness. The outcome of a classification strongly depends on the definition of the response categories. For this reason we devised a titer response score (TRS; see Materials and Methods, Figure 5 and Supplementary File S4) that integrated the responses obtained from all the three strains represented in the TIV and investigated how this is related to the maximum response across strains. Figure 6 represents TRS responses, encoded as node radius, mapped on the classification graph of Figure 2. According to the TRS, subjects showing more extreme responses will be assigned higher scores and, therefore, will appear as bigger nodes in the graph. Interestingly, subjects with high TRS were placed in peripheral regions of the graph, with smaller nodes populating more central regions. This is indicative of the fact that subjects with high TRS were better discriminated (greater distance between high and low responder groups) by the biomarker. Coherently with these observations, both misclassified subjects (H29 and L39) had a small TRS, indicating that their membership in their predicted class, the wrong one, was not strong. Subject H29, which was classified as high responder based on the response relative to the H3N2 strain ($HAI_{Day\ 28/Day\ 0} = 32$), did not respond as strongly against the other two strains (see Supplementary File S4). Similarly, while most of the low responders did not respond to any of the strains ($HAI_{Day\ 28} = HAI_{Day\ 0}$), subject L39 showed a 2-fold increase in the HAI titer against the H1N1 strain (see Supplementary File S4).

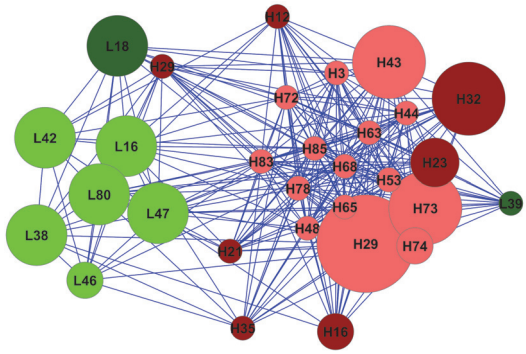


Figure 6: *Prediction of seroresponse to TIV vaccination.* The graph represents sample classification by means of the 207 transcripts biomarker as in Figure 2, but here the node size encodes the titer response score (bigger nodes correspond to higher indices, as displayed in Figure 5). Nodes represent subjects and the length of connecting edges is proportional to the distance between subject signatures. Node colors indicate subjects class (green for low responders and red for high responders); darker colors represent samples included in the validation set (2007-2008 TIV campaign, set #1), while lighter colors represent samples in the training set (2008-2009 TIV campaign, set #2). Node labels indicate the sample class (H for high responders and L for low responders) and the sample identifier.

Overall, this evidence indicates that while using the maximum HAI fold change across strains may be accurate enough for the discrimination between high responders and low responders, composite measures integrating information from all strains represented in the vaccine provide a more informative measure.

3. Conclusions

Overall, we have identified a set of 207 transcripts, derived from PBMCs, whose early (3–7 days) regulation after vaccination was predictive of the magnitude of later (1 month) serum antibody response. Overall, this biomarker performed similarly to the one described in the original work published by Nakaya et al. [10], by providing a 100% within-dataset classification accuracy and a cross-datasets accuracy of 81% (see Figure 3). Moreover, the applied classification method provided a graphical representation of the classification result, which enabled for a qualitative, as opposed to solely quantitative, interpretation of the results obtained. This provided the opportunity to investigate on the possible reasons behind the observed misclassifications and allowed identifying preexisting immunity, against one or more of the vaccine antigens represented in the vaccine, to be a possible confounding factor. Specifically, preexisting immunity, which was found to negatively affect the seroresponse, was shown to have little to no effect on whole blood transcriptome profiles assessed 3, or more, days after vaccination. This implies a decoupling of early transcriptome from later antibody responses, highlighting a major limitation in transcriptome-based biomarkers.

Differently from 3 and later days responses, 24-hour whole blood transcriptomes revealed substantial differences in gene responsiveness between subjects with preexisting immunity and subjects without it. Part of these differences included a decreased responsiveness of the transcriptional inflammatory response in subjects with preexisting immunity, providing a first clue of a possible mechanistic link between preimmune HAI titers and reduced serological response to influenza vaccination.

The present study represents an example of how innovative computational tools can improve our data mining capabilities and helps to reveal latent factors that can impact the response to vaccination. We envision that extending this analysis pipeline to other studies will help in identifying additional confounding factors and produce more accurate predictions.

4. Materials and Methods

Definition and testing of the biomarkers were based on data published by Nakaya et al. [10, 11]. Gene expression derived from peripheral blood mononuclear cells and HAI response data were downloaded from the Gene Expression Omnibus repository (GEO, https://www.ncbi.nlm.nih.gov/geo) using the two accession identifiers GSE29614 and GSE29617 for the 2007-2008 and 2008-2009 vaccination trials, respectively. Gene expression data were imported using the ArrayExpress Bioconductor package and processed using the RMA normalization procedure. Gene-level expression data were derived by computing the geometric mean of multiple probes mapping to the same gene, where applicable. The dataset related to the additional validation presented in Figure 3 was downloaded from the GEO repository using the accession identifier GSE74817. Analysis of differential gene expression between subjects with high and low baseline HAI titers was performed on the dataset published in [8] and downloaded from GEO using the accession identifier GSE48018. In that study, 119 healthy individuals vaccinated with a 2008-2009 trivalent influenza vaccine were profiled for both whole blood transcriptome and HAI responses. Fifteen (15) subjects had a preimmune HAI titer ≤ 16 (maximum across strains) and were considered to be naive to the vaccine. In order to generate a contrast group comparable in size, 20 subjects with a HAI titer ≥ 512 (maximum across strains) were selected as high-preimmune individuals. After removing genes that did not show substantial variation across the entire dataset (log 2 interquartile range ≤ 0.5), fold changes from baseline in gene expression were computed for both groups 1, 3, and 14 days after vaccination. Gene responsiveness was then compared between the two groups through the Mann-Whitney-Wilcoxon test and resulting p values were corrected for multiple testing using the Benjamini-Hochberg procedure. Transcripts with a q-value ≤ 0.05 were assumed to be affected by preexisting immunity.

According to [10], seroresponder and nonseroresponder subjects were identified based on whether they achieved a maximum $\mathrm{HAI}_{\mathrm{Day\ 28/Day\ 0}}$ fold increase across strains > 4 or <4, respectively. Afterwards, day 3 and day 7 fold changes from baseline transcript abundances were employed for the prediction of subjects' class membership using an enhanced version of the previously described classification

algorithm SCUDO [12, 13, 15]. The present version of SCUDO has been extended to increase its discriminatory power and equipped with a parameter optimizer that allows the automatic selection of the algorithm parameters according to user-defined criteria as we did to compute the transcriptional biomarker presented in [20, 21].

SCUDO is a rank-based classification method and this makes the computation normalization-free. Therefore, we did not apply any data preprocessing with batch effect removal algorithms (e.g., quantile normalization) before running SCUDO [12]. This feature of the classification algorithm played a crucial role in the additional validation presented in Figure 3, in which samples from separate datasets could be tested without taking batch effects into account.

The classification accuracy of the analysis was evaluated using a 10-fold cross-validation scheme over the training dataset (2008-2009 trial) and by validating over another independent dataset (2007-2008 trial). Statistical significance of the identified biomarkers was assessed by comparing the classification accuracy of the method with the accuracy of an empirical null distribution obtained by 10000 random permutations of the transcript labels (permutation test p value < 0.001). An overview of the analysis pipeline is provided in Figure 1. The robustness of the identified 207 transcripts biomarker was further assessed by applying it to independent TIV vaccination campaigns (2009, 2010, and 2011 campaigns included in GSE74817).

The titer response score (TRS; Figures 5 and 6 and Supplementary File S4) was generated by integrating the responses obtained with all the three strains represented in the TIV vaccine. Differently from a previously proposed integrated antibody response measure [8], the TRS of each subject was computed as the sum of the logarithmic deviations of the three considered $\mathrm{HAI}_{\mathrm{Day\ 28/Day\ 0}}$ ratios from the threshold value of the class to which the subject belongs (2 for low responders, 8 for high responders). An additional penalty was added for subjects with high (>160) baseline HAI titers. Additional information can be found in the Supplementary File S4.

Conflicts of Interest

Emilio Siena, Nicola Pacchiani, and Duccio Medini are employees of GSK Vaccines s.r.l. The other authors do not have any conflicts of interest.

Authors' Contributions

Luca Marchetti and Emilio Siena contributed equally to this work.

References

[1] J. G. Breman and I. Arita, "The certification of smallpox eradication and implications for guinea worm, poliomyelitis, and other diseases: Confirming and maintaining a negative," *Vaccine*, vol. 29, no. 4, pp. D41–D48, 2011.

[2] J. Smith, R. Leke, A. Adams, and R. H. Tangermann, "Certification of polio eradication: Process and lessons learned," *Bulletin of the World Health Organization*, vol. 82, no. 1, pp. 24–30, 2004.

[3] R. Rappuoli, "Bridging the knowledge gaps in vaccine design," *Nature Biotechnology*, vol. 25, no. 12, pp. 1361–1366, 2007.

[4] P. R. Dormitzer, J. B. Ulmer, and R. Rappuoli, "Structure-based antigen design: a strategy for next generation vaccines," *Trends in Biotechnology*, vol. 26, no. 12, pp. 659–667, 2008.

[5] D. T. O'Hagan and C. B. Fox, "New generation adjuvants - From empiricism to rational design," *Vaccine*, vol. 33, no. 2, pp. B14–B20, 2015.

[6] B. Pulendran, S. Li, and H. I. Nakaya, "Systems vaccinology," *Immunity*, vol. 33, no. 4, pp. 516–529, 2010.

[7] R. Rappuoli and A. Aderem, "A 2020 vision for vaccines against HIV, tuberculosis and malaria," *Nature*, vol. 473, no. 7348, pp. 463–469, 2011.

[8] K. L. Bucasas, L. M. Franco, C. A. Shaw et al., "Early patterns of gene expression correlate with the humoral immune response to influenza vaccination in humans," *Journal of Infectious Diseases*, vol. 203, no. 7, pp. 921–929, 2011.

[9] G. Obermoser, S. Presnell, K. Domico et al., "Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines," *Immunity*, vol. 38, no. 4, pp. 831–844, 2013.

[10] H. I. Nakaya, J. Wrammert, E. K. Lee et al., "Systems biology of vaccination for seasonal influenza in humans," *Nature Immunology*, vol. 12, no. 8, pp. 786–795, 2011.

[11] H. I. Nakaya, T. Hagan, S. S. Duraisingham et al., "Systems Analysis of Immunity to Influenza Vaccination across Multiple Years and in Diverse Populations Reveals Shared Molecular Signatures," *Immunity*, vol. 43, no. 6, pp. 1186–1198, 2015.

[12] M. Lauria, "Rank-based transcriptional signatures: A novel approach to diagnostic biomarker definition and analysis," *Systems Biomedicine*, vol. 1, pp. 35–46, 2013.

[13] M. Lauria, "Rank-based miRNA signatures for early cancer detection," *BioMed Research International*, vol. 2014, Article ID 192646, 2014.

[14] A. L. Tarca, M. Lauria, M. Unger et al., "Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge," *Bioinformatics*, vol. 29, no. 22, pp. 2892–2899, 2013.

[15] M. Lauria, P. Moyseos, and C. Priami, "SCUDO: A tool for signature-based clustering of expression profiles," *Nucleic Acids Research*, vol. 43, no. 1, pp. W188–W192, 2015.

[16] W. E. P. Beyer, A. M. Palache, M. J. W. Sprenger et al., "Effects of repeated annual influenza vaccination on vaccine sero-response in young and elderly adults," *Vaccine*, vol. 14, no. 14, pp. 1331–1339, 1996.

[17] S. Sasaki, X.-S. He, T. H. Holmes et al., "Influence of prior influenza vaccination on antibody and B-cell responses," *PLoS ONE*, vol. 3, no. 8, Article ID e2975, 2008.

[18] S. Ng, D. K. M. Ip, V. J. Fang et al., "The Effect of Age and Recent Influenza Vaccination History on the Immunogenicity and Efficacy of 2009-10 Seasonal Trivalent Inactivated Influenza Vaccination in Children," *PLoS ONE*, vol. 8, no. 3, Article ID e59077, 2013.

[19] G. Bondar, M. Cadeiras, N. Wisniewski et al., "Comparison of whole blood and peripheral blood mononuclear cell gene expression for evaluation of the perioperative inflammatory response in patients with advanced heart failure," *PLoS ONE*, vol. 9, no. 12, Article ID e115097, 2014.

[20] L. Caberlotto, L. Marchetti, M. Lauria, M. Scotti, and S. Parolo, "Integration of transcriptomic and genomic data suggests candidate mechanisms for APOE4-mediated pathogenic action in

Alzheimer's disease," *Scientific Reports*, vol. 6, Article ID 32583, 2016.

[21] A. Matone, E. Derlindati, L. Marchetti et al., "Identification of an early transcriptomic signature of insulin resistance and related diseases in lymphomonocytes of healthy subjects," *PLOS ONE*, vol. 12, no. 8, p. e0182559, 2017.

Hindawi
Complexity
Volume 2017, Article ID 5024867, 10 pages
https://doi.org/10.1155/2017/5024867

Research Article

FAACOSE: A Fast Adaptive Ant Colony Optimization Algorithm for Detecting SNP Epistasis

Lin Yuan,[1] Chang-An Yuan,[2] and De-Shuang Huang[1]

[1]*Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Caoan Road 4800, Shanghai 201804, China*
[2]*Science Computing and Intelligent Information Processing of Guang Xi Higher Education Key Laboratory, Guangxi Teachers Education University, Nanning, Guangxi 530001, China*

Correspondence should be addressed to De-Shuang Huang; dshuang@tongji.edu.cn

Received 31 March 2017; Accepted 24 July 2017; Published 7 September 2017

Academic Editor: Jianxin Wang

The epistasis is prevalent in the SNP interactions. Some of the existing methods are focused on constructing models for two SNPs. Other methods only find the SNPs in consideration of one-objective function. In this paper, we present a unified fast framework integrating adaptive ant colony optimization algorithm with multiobjective functions for detecting SNP epistasis in GWAS datasets. We compared our method with other existing methods using synthetic datasets and applied the proposed method to Late-Onset Alzheimer's Disease dataset. Our experimental results show that the proposed method outperforms other methods in epistasis detection, and the result of real dataset contributes to the research of mechanism underlying the disease.

1. Introduction

Accompanied by the rapid development of genomics and gene chip technology, Genome-Wide Association Studies (GWAS) predicted massive genetic variations related to complex traits [1, 2]. Although this method has achieved great success. It can only explain a small part of the mechanism under the complex diseases known as "missing heritability" [3]. That is to say, marginal genetic effects of GWAS identified single nucleotide polymorphisms (SNPs) account for small part of pathogenic causes. For single-locus SNPs related disease [4], GWAS can identify SNPs that are responsible for disease trait. However, complex diseases are often due to the small and complex effects of large SNPs, such as type 2 diabetes [5], prostate cancer, and rheumatoid arthritis (RA) [6]. More and more studies have shown that epistasis exists in SNPs interaction. Many SNPs will interact with each other in the process of affecting the disease traits [7]. Some SNPs will affect the disease and dominate the effect of others. The relationship of one SNP repressing the effect of another SNP is known as epistasis. In many complex human diseases, the effect of epistasis among complex human diseases is unclear. The proposed methods for SNP related disease may have poor performance due to failure to identify epistasis.

During the past decade, a lot of approaches have been proposed to detect epistasis. Some methods focus on the interaction between two certain SNPs. Zhang et al. [8] proposed a Bayesian partition method for epistatic eQTL modules. Kang et al. [9] proposed four different models to measure epistasis effect between two loci and suggest a statistical strategy to infer the hierarchical relationships. Recently, Lin et al. [10] reported forty-five SNP-SNP interaction models by considering the inheritance modes and model structures. Though these methods have been successful in studying epistasis between two SNPs. The GWAS data is high dimension data which contains hundreds of thousands or even million SNPs; at the same time, GWAS data only contains dozens or hundreds of individual sample data, for example, the small number sample data and the high dimension features; it needs vast amounts of time to identify the interaction between each pair of SNPs [11–13]. The computational burden is out of bounds.

More and more machine learning methods are applied to research epistasis. Many methods were proposed to model

epistasis effect from the perspective of the overall data. Moore et al. [14] applied regression method to identify the relationship between gene expression and epistasis effect. Michael et al. [15] applied Bayesian networks to identify the epistasis effect network from the original SNPs data. Although these methods solved some problems, they still did not show significant effects with the large scale Genome-Wide Association Study datasets owing to the same "high-dimensional small sample size problem." With the rapid development of multiobjective optimization method and machine learning discipline, ant colony optimization (ACO) algorithm was applied to epistasis research. Wang et al. [16] proposed AntEpiSeeker; AntEpiSeeker combines heuristic search with the ant colony optimization to identify SNPs which dominate other SNPs. Experimental results on real rheumatoid arthritis dataset show that AntEpiSeeker is better than other methods. The drawback of this method is that other methods show different performance on different disease models. Zhang and Liu [17] developed the Bayesian inference method which identifies the epistatic interactions in case-control studies. However, the BEAM method needs a lot of time in GWAS dataset. In this paper we extend SNP epistasis study to a fast adaptive ant colony optimization algorithm for detecting SNP epistasis. We search SNP epistasis with two-objective functions and fast adaptive ant colony optimization.

The experiments on several simulated datasets show the good performance of our method. We also compare our method with the benchmark methods, including BEAM, generic ACO, and AntEpiSeeker. Experimental results show that our method has better performance in GWAS datasets containing epistasis effect among SNPs.

2. Methods

2.1. Ant Colony Optimization. In the research of artificial intelligence and large scale problem solving, the ant colony optimization (ACO) algorithm is inspired by the ants food search behaviour in nature. Assume that the food search paths constitute a graph; the ant colony optimization algorithm can reduce time of search paths through graphs [18]. This algorithm with other ant colony optimization algorithms is kind of swarm intelligence methods, and it is member of metaheuristic optimizations. Marco Dorigo proposed the ant colony optimization algorithm in 1992 in his Ph.D. thesis. In the GWAS datasets, the datasets often contain tens of hundreds to millions of SNPs. It is not feasible to identify the relationship of every pair of SNPs within an acceptable time. ACO algorithm was used here to reduce the complexity of exhaustive search. In kingdom of insects, in the process of finding food, ants look like they are walking randomly, and in the back and forth path of searching for food, the ants will leave pheromones on the path. If the path is found by other ants, other ants tend to follow the path but not walk randomly; going further, if they find food through this path, they will also leave pheromones; the pheromone value on this path is enhanced. Subject to other factors in nature, pheromone value starts to evaporate and the path's attractive strength starts to decrease. The longer the path is, the more the time the ants are looking for food. As a comparison, the time the ants take to walk through the short path is greatly shortened, and pheromone values will be larger on shorter paths than longer paths. Pheromone evaporation results in dynamic changes in the path. Path dynamic changes can avoid the convergence of solutions to a locally optimal solution. If there is no pheromone values evaporation, the food search path selected by first ants would tend to be the only path or the most attractive path. This phenomenon will lead to limitation of the solution space. The mechanism of pheromone evaporation in ant colony is unclear, but pheromone evaporation is a very important application in artificial intelligence systems. Though the ant colony optimization algorithm has achieved great success in application [19–21].

The travelling salesperson problem (TSP) is a problem with some cities and physical distances between each pair of cities. The question is what is the shortest possible path where travelling salesperson visits each city once and finally returns to the origin city? Suppose there are n cities; there are $(n-1)!/2$ solutions to the problem. The feasible solutions will increase exponentially when the number of city increases, making the computation impractical. Obviously, it is an NP-hard problem of combinatorial optimizations.

Suppose that m ants are randomly placed in n cities, the kth ant in the ith city; the probability if ant chooses the next city j is

$$p_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha}(t)\,\eta_{ij}^{\beta}(t)}{\sum_{o \in \text{candidate}_k} \tau_{io}^{\alpha}(t)\,\eta_{io}^{\beta}(t)}, & j \in \text{candidate}_k, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

$$\eta_{ij}(t) = \frac{1}{d_{ij}},$$

where $\tau_{ij}(t)$ indicates the surplus information on path ij in moment t. $\eta_{ij}(t)$ indicates the heuristic function. d_{ij} indicates the physical distance between city i and city j. tabu_k indicates the cities set which indicates ant k has visited. candidate_k indicates the set of cities which ant k can visit next.

Over time, after n moments, the ants complete a cycle; the information of each path should be adjusted according to

$$\begin{aligned} \tau_{ij}(t+n) &= (1-\rho)\,\tau_{ij}(t) + \Delta\tau_{ij}, \\ \Delta\tau_{ij} &= \sum_{k=1}^{m} \Delta\tau_{ij}^{k}, \end{aligned} \tag{2}$$

where $\Delta\tau_{ij}$ indicates information increment of path ij after this cycle.

$$\Delta\tau_{ij} = \begin{cases} \dfrac{Q}{l_k}, & ij \in L_k \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where L_k indicates ant k's paths in this cycle. l_k indicates the path length of ant k in this cycle. The parameters needed to be determined are $\alpha, \beta, \rho, m, Q$; the number of ants is less than or equal to city number; Q is a large suitable

number. ACO is always used in large scale data problems. However, slowness is still a bottleneck in the application of the ant colony algorithm for large scale search optimization problems. Pheromone update strategy is one of the keys to determine the convergence rate.

In the process of applying ant colony optimization to specific problems, the search space should be as large as possible. At the same time, ACO should consider time efficiency. ACO should balance the optimal solutions and solve speed. On the basis of previous studies [22–24]. We only consider pheromone evaporation factor ρ and pheromone importance factor α. In (2), ρ is used to balance the effects of old pheromone value and current pheromone value. When ρ is too small, the residual pheromone value is too much and leads to local minimum solution. We adopt adaptive ρ, when the algorithm does not improve the current optimal solution within n iterations.

$$\rho(t+n) = \begin{cases} g\rho(t), & \rho(t) \le \rho_{\max} \\ \rho_{\max}, & \text{otherwise,} \end{cases} \tag{4}$$

where $\rho_{\max}$ equals 0.85 in practice. g equals 1.02 as tune parameter. When the pheromone value reaches the critical value, the pheromone importance factor begins to play a role. With the increase of pheromone importance factor α, the algorithm will jump out of the local optimal solution and has ability to search for global optimal solution.

$$\alpha(t+n) = \begin{cases} g_1\alpha(t), & \alpha(t) \le \alpha_{\max} \\ \alpha_{\max}, & \text{otherwise,} \end{cases} \tag{5}$$

where g_1 is a constant larger than one and $\alpha_{\max}$ is less than or equal to five. In the process of calculation, first, we follow the standard ant colony optimization algorithm for N iterations. N is predefined number. If the current optimal solution is not improved after N iterations, update the parameters according to formulas (4) and (5). Then update all pheromone value according to (2).

Given pheromone values and transfer rules, we can use the ant colony optimization algorithm to find a group of SNPs which affect the disease. Assume there are P SNPs in the global Genome-Wide Association Studies dataset, we can construct a p-dimensional symmetric matrix M to store every ant's pheromone value. The element m_{ij} of matrix M denotes the interaction which is related to disease between ith SNP and jth SNP. At the beginning of our method, every element of matrix M is assigned to a constant value m_0; equivalent value shows the epistasis in every pair of SNPs and there is equal possibility relationship between the SNPs and disease.

At the final pheromone iteration, the ACO algorithm will obtain the optimal solutions through forward selection strategy. The advantage of ACO algorithm in this paper is that the result contains nondominated solutions which have the potentially equivalent possibility and potentially highest related strength with disease and omit dominated solutions.

The disadvantages of traditional ant colony optimization algorithm are long search time and tendency to fall into the local optimal solution. The drawback of this working mode is that the current pheromone evaporation factor and pheromone importance factor are predefined. As an improved strategy, we extended the "dynamic adaptive strategy" to ant colony optimization. The advantage of this strategy is the fast convergence rate and searching for global optimization solution. Compared with traditional ACO, the new strategy can provide more accurate result.

2.2. Two-Objective Function Optimization. The results of ant colony optimization need to be evaluated. We combine two-objective methods to assess the final epistasis results. In general, one of two-objective functions combines Akaike Information Criterion (AIC) score and logistic regression function to measure relationship between phenotypic trait and genotype data; Akaike Information Criterion indicates the effectiveness and complexity of the model [25, 26]. In our method, on the basis of the standard logistic regression, following the North et al. [27] strategy, we use ADDINT logistic regression model to search the relationship between disease and SNP nodes. The second objective function uses frequency measurement based on mutual information theory to model the relationship between genotype data and phenotypic trait from the perspective of information theory. The second objective function used to represent the selected SNP subsets can explain how much information is about the disease trait. Our proposed method obtains information from data rather than a lot of priori information. The above two-objective functions are designed from the different perspective to measure the quality of the search results, and the simulation data experiment results show that our two-objective functions have a better performance than other methods on simulated and real biological datasets.

In order to avoid the bad impact of high dimension small size sample problem, the identification of disease-associated SNPs is known as a heuristic optimization problem. In our proposed method, proposed method yields optimal solutions which is nondominated solutions; the proposed two-objective functions method actually is kind of multiobjective optimization; the proposed method uses ant colony optimization to search for optimal solution [28].

Our proposed fast adaptive ACO framework contains two stages. In the first stage, we use modified ACO optimization algorithm with two-objective functions to search for nondominated SNP subset. After generating the nondominated SNP subset, we apply Fisher exact test [29, 30] to the dataset containing nondominated SNP generated in the algorithm first stage. The Fisher exact test will be used to identify the relationship between disease and SNPs.

2.2.1. AIC Score. The Akaike Information Criterion (AIC) is used to measure quality of dataset statistical models. AIC is from information theory, and it estimates loss of information when a statistical model is used to express the data generation process. The mechanism of Akaike Information Criterion is that it deals with the trade-off between the goodness of fit of the model and the complexity of the model. Based on the nature of the AIC, we construct AIC model from the perspective of GWAS dataset. The goal of our method is to measure the relationship between the genotype data of

genome and phenotype disease trait. Logistic regression is widely used to quantitatively analyze the correlation between dependent variable and independent variable. Based on above methods, we construct AIC score model containing logistic regression and gradient penalty function. Logistic regression can compute the maximized log-likelihood of the model; k is used to express the number of free parameters. AIC score deals with the trade-off between the fitness effect of the model and the complexity of the model. We follow Jing and Shen [28] strategy:

$$\text{AICscore} = 2k - 2\log lik, \tag{6}$$

where k denotes the number of free parameters.

2.2.2. Explanation Score. In GWAS research, the relationship between two loci and disease, in SNP research, each locus has three values, 0, 1, and 2; 0 means major allele homozygous, 1 means heterozygote, and 2 means minor allele homozygous [31]. For two loci, there are nine cases of their combination; the disease related SNP locus often changes when the disease occurs. In the case of double locus combination, x_i means the number of ith combinations of two SNP loci, Y means case or control state, y_1 means state case, and y_2 means state control. The potential interrelationships of two discrete random variables X and Y are defined as $H(X;Y)$; the relationship between locus combination and disease is measured based on the information of locus frequency. $H(X;Y)$ is described as below:

$$H(X;Y) = \sum_{i=1}^{I} \left(\left| x_{iy_1} - x_{iy_2} \right| \right), \tag{7}$$

where I means the total number of locus combinations. To avoid unbalanced sample, the size affects score. For example, if data size of case is larger than control, we extract the same size of control data from case samples randomly. To avoid the impact of randomness, we extract sample several times and average the results. The large value H means the potential association probability between disease and SNPs is large. Equation can also be applied to more than two locus combinations. We name this score explain score.

2.3. Pareto Optimality for SNP Epistasis Detection. Pareto optimality defines such a situation. Pareto optimality is proposed to solve the following questions where it is impossible to make all objective function values of multiobjective optimization optimal values [32, 33]. Pareto optimality is first applied to the area of income distribution and economy. Now Pareto optimality has been extended to engineering and multiobjective optimization research. On the basis of previous proposed methods, the modified ant colony optimization algorithm with first objective function and second objective function, the first objective function is AIC score with logistic regression and related parameters; the second objective function is explain score. For the first objective functions, the lower score of the objective function indicates the high potential relationship between disease phenotype trait and SNPs [34]. For the second objective functions, the higher score of the objective function indicates the high potential relationship between the disease phenotype trait and SNPs. The target of fast two-stage ant colony optimization algorithm is to find the epistasis effect among SNPs and extract real SNP subset with respect to the above proposed methods.

In the real GWAS datasets, an identified SNP subset may perform the best compared with other method solutions in terms of one-objective function, but SNP subset may perform poorly in terms of another objective function. Thus, the target is how to select better SNP subset with respect to both objective functions. In practical application, rare subset performs better than other solutions while satisfying both conditions. Thus, for a framework with two-objective functions, it is hard and even impossible to calculate the global optimal solution. On the basis of previous studies [28, 34, 35], we adopt Pareto optimality to find the practical optimal solution. We first compare the two solutions, in terms of GWAS SNP subset, a solution named S_1, and another solution named S_2; comparing S_1 and S_2 only have two consequences; one result is one solution dominates the other; another result is S_1 does not dominate S_2; in turn, the solution S_2 does not dominate S_1. Based on the mind of Pareto optimality, we consider S_1 dominates S_2 if they satisfy the following two conditions. The first condition is the value of $f_e(S_1)$ is not higher than $f_e(S_2)$ for those two-objective functions. The second condition is the objective function $f_e(S_1)$ is lower than $f_e(S_2)$ for at least one-objective function. The function f_e denotes the objective function: modified AIC score objective function and explain score objective function. The e equal to one denotes the first objective function; the e equal to two denotes the second objective function. If solutions S_1 and S_2 satisfy the above two conditions, we say solution S_1 is a nondominated solution; in turn, we say solution S_2 is a dominated solution. Based on above Pareto optimality approach and two-objective functions, all solutions can be divided into two kinds; one is nondominated set and another is dominated set. Finally, nondominated sets contain many solutions and all the solutions from our proposed method with respect to two-objective functions; now our goal is to find a nondominated set which is the best under certain conditions.

Next, we will use the judgment rule mentioned earlier to sort the solutions of nondominated sets to find the optimal nondominated set. Specifically, in the first case, $f_1(S_2)$ is larger than $f_1(S_1)$; at the same time, $f_e(S_2)$ is larger than $f_2(S_1)$. In the second case, $f_1(S_2)$ equals $f_1(S_1)$; at the same time, $f_2(S_2)$ is larger than $f_2(S_1)$. In the third case, $f_1(S_2)$ is larger than $f_1(S_1)$; at the same time, $f_2(S_2)$ equals $f_2(S_1)$.

2.4. Fisher Exact Test for Experimental Results. Fisher exact test is used in contingency tables to get a statistical significance [36–38]. Although in practice it is used in small size sample, it is can also be used in all sample sizes. Ronald Fisher first proposed this method and Fisher exact test is one kind of exact tests.

In terms of our GWAS datasets research article, on the basis of unified framework which contains fast adaptive ant colony optimization (ACO) algorithm, Akaike Information Criterion (AIC) score, explain score, and Pareto optimality, we can obtain the final result which is a nondominated SNP set; in this section, we will use Fisher exact test to exhaustively

search for the epistasis effect. Fisher exact test is based on hypergeometric distribution; the P value in the Fisher exact test is accurate for all individual samples. Fisher exact test is used on the basis of contingency table. The null hypothesis is that the identified SNP subset and disease are not associated. The alternative hypothesis is that SNP subset affects the expression of the disease when the Fisher exact test's P value is significant, when P value is less than predetermined value such as 0.05 or smaller value. Our proposed method will identify significance SNP subsets.

2.5. Power Test. In previous section, we introduce each part of our proposed fast adaptive ant colony optimization algorithm for detecting SNP epistasis. Our proposed unified framework contains fast adaptive ant colony optimization algorithm, Akaike Information Criterion (AIC) score, explain score, Pareto optimality, and modified Fisher exact test. In this section, we introduce how to verify the significance of the results. We construct 100 datasets according to the same parameters. Then we use the traditional power test to measure the effect of methods. The power test is defined as follows:

$$\text{Power} = \frac{|SD|}{100}, \tag{8}$$

where $|SD|$ denotes the number of disease related datasets which were correctly selected from 100 datasets. Only using the single test criterion may not clearly show the quality of results. We use precision recall standard to measure true positive rate and false positive rate. Precision recall criteria have been widely used in classification model evaluation model [39, 40]. In pattern recognition and information retrieval with binary classification, precision, also called positive predictive value, is the fraction of retrieved instances that are relevant; while recall, also known as sensitivity, is the fraction of relevant instances that are retrieved [26]. Both precision and recall are therefore based on an understanding and measure of relevance. We use precision recall criteria to determine whether the classification results are good or bad. The precision recall criteria can avoid the imbalance problem of precision recall numbers. In our research, the number of precision and recall always differs greatly. In terms of the SNP epistasis research, precision is also known as positive predictive value, equivalent to the true disease related SNP subsets; recall is also known as sensitivity or negative, equivalent to the true disease unrelated SNP subsets. If we use only one judgment criterion, thus false positive rate, single indicator cannot make the real result clear. We use false positive rate and true positive rate to measure the real result. This is why we use precision and recall. We also use F_1 score (also F score or F measure) to measure the precision recall test accuracy. The precision and recall will be introduced next with confusion matrix (Figure 1).

$$\begin{aligned} \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ F_1 &= \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \end{aligned} \tag{9}$$

		Predicted class: Associated	Predicted class: Nonassociated
True class	Associated	True positive (TP)	False negative (FN)
True class	Nonassociated	False positive (FP)	True negative (TN)

Figure 1: Precision recall explanation matrix.

The precision, also known as specificity, denotes true positive number ratio in the result through the number of true positives divided by the sum of true positive number and false positive number; precision is often used to report false positive rate of an algorithm's false positive rate. The recall, also known as sensitivity, denotes true positive ration in the sum of true positives and false negative. In terms of SNPs selection problem, the larger the recall value is, the larger the number of real true disease-related SNP combinations can be found. Simultaneously, the larger the precision value, the larger the number of real true disease-related SNP combinations account for a high proportion of the identified SNP combinations. The criterion F measure is the harmonic mean of precision and recall, which is a synthesized measure combining both precision and recall [41].

3. Simulation Experiments

3.1. Compared with One-Objective Function. In this section, we use simulation data to compare our proposed method with other existing methods. In order to avoid data favor caused by the model, we adopt BEAM package to generate simulation datasets [17]. Data was simulated following three genetic models: (1) additive model, (2) epistatic interactions with multiplicative effects, and (3) epistatic interactions with threshold effects. In order to introduce our experiments, the additive model is referred to as ADDME. The model about epistatic interactions with multiplicative effects is referred to as EIME. The epistatic interactions with threshold effects are referred to as EITEME. In the next section, we will use the short name to indicate the corresponding data model.

Because our method is two-objective-based SNP epistasis search method, first, we compared our proposed method with existing single objective-based exhaustive SNP epistasis search method to demonstrate the effectiveness of two-objective function SNP epistasis subset search method. Second, we compare our proposed method with recently proposed method BEAM [17], generic ACO algorithm, and AntEpiSeeker [16]. In the one-objective function SNP epistasis search method, the objective function is used to score every SNP combinations; in general, the score for every SNP combination is not the same. Based on the nature of the method, low score indicates the association between SNP combination and disease is relatively small; high score indicates the association between SNP combination and disease is relatively large. Then the one-objective function ranks all SNP combinations based on the scores. However, the two-objective-based SNP epistasis search method is to find a set of nondominated results, and every nondominated SNP

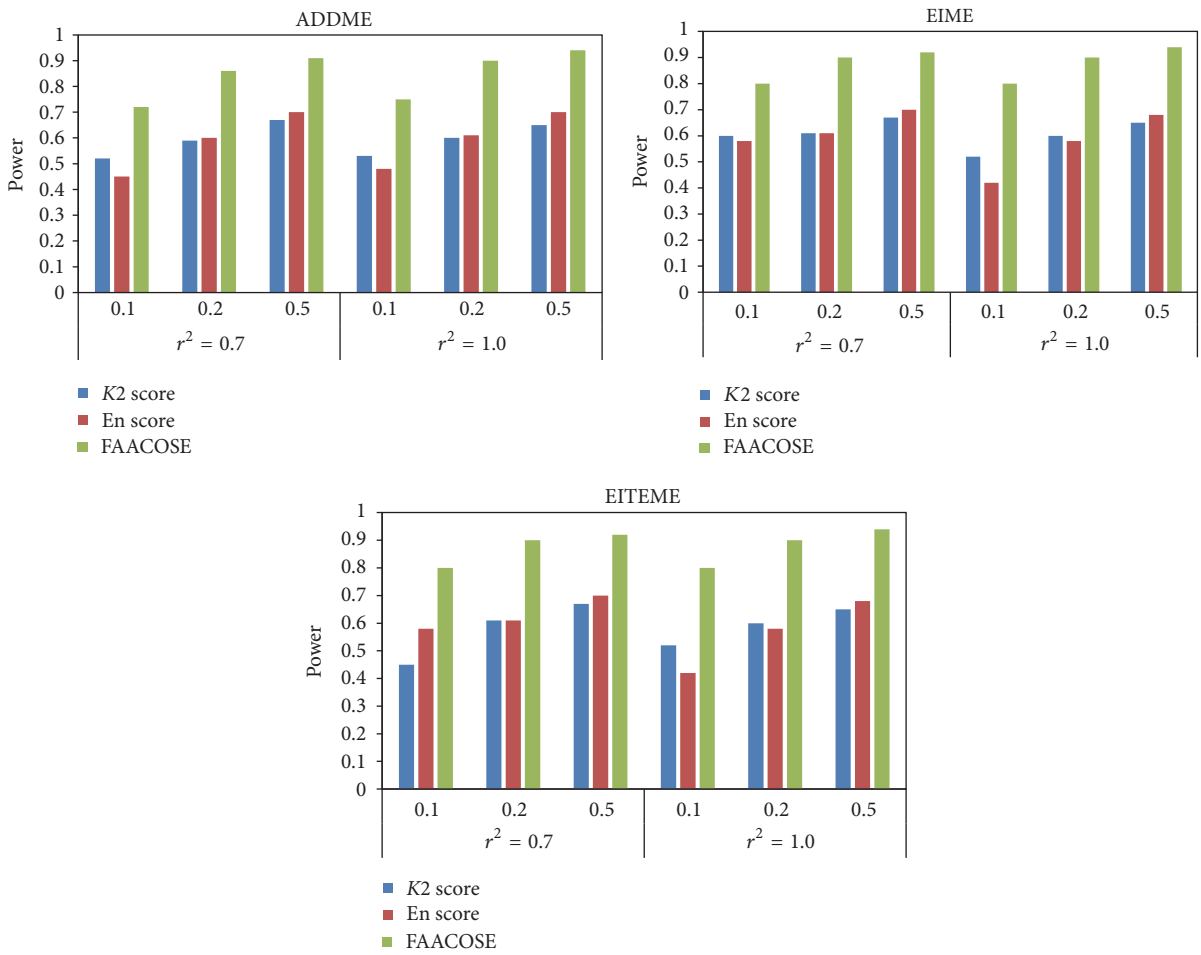


Figure 2: Power test comparisons between one-objective and two-objective methods on three different model with MAF value 0.1, 0.2, and 0.5.

epistasis results' score is the same. To ensure fairness, for the one-objective function, we collect the same number as two-objective-based SNP epistasis search method from the top of one-objective-based SNP rank. The comparing results show that the two-objective-based SNP epistasis search method is better than one-objective-based SNP epistasis search method in three simulation data models. In terms of two single objective-based SNP epistasis search methods, the results of one-objective-based SNP epistasis search methods are similar with the other one-objective-based SNP epistasis search methods. The simulation data experiment results show the effectiveness of two-objective-based SNP epistasis search method, and the poor experimental results show the insufficiency of one-objective functions. The experiment results are shown in Figure 2. The abscissa of Figure 2 is minor allele frequency (MAF) which is assigned 0.1, 0.2, and 0.5. We generate the simulate dataset and study the parameter setting following many previous studies [17, 42–44]. For each simulate dataset of parameter combination, we generated 100 datasets which contain 2,000 experimental samples (1,000 case samples and 1,000 control samples) and 1000 SNPs were simulated. We evaluate the algorithm performance through calculating the ratio of real number identified following the significance level 0.01 which is adjusted after Bonferroni correction. The parameter λ was set to 0.3 for ADDME and 0.2 for EIME and EITEME. The parameter range of linkage disequilibrium between SNPs is r^2 from 0.7 to 1.

3.2. Compared with Benchmark Methods. After comparing with single objective function. We compare our proposed method with existing method. The performance of our proposed method was evaluated by comparison with benchmark methods [45]. In many previous studies, the authors have already discussed the parameter settings problem. In this section, we set the parameters according to the existing strategy. We evaluated performance of FAACOSE by comparing with two recent methods, BEAM, generic ACO algorithm, and the AntEpiSeeker; we use BEAM package and previous parameter strategy to generate simulate dataset. Be aware of the fact that the generic ACO algorithm could not select larger size SNP set. We use simulated dataset introduced in Section 3.1. We evaluate the algorithm performance through calculating the ratio of real number identified following the significance level 0.01 which is adjusted after Bonferroni correction. We generate simulate datasets following three genetic models: ADDME, EIME, and EITEME. Other parameters

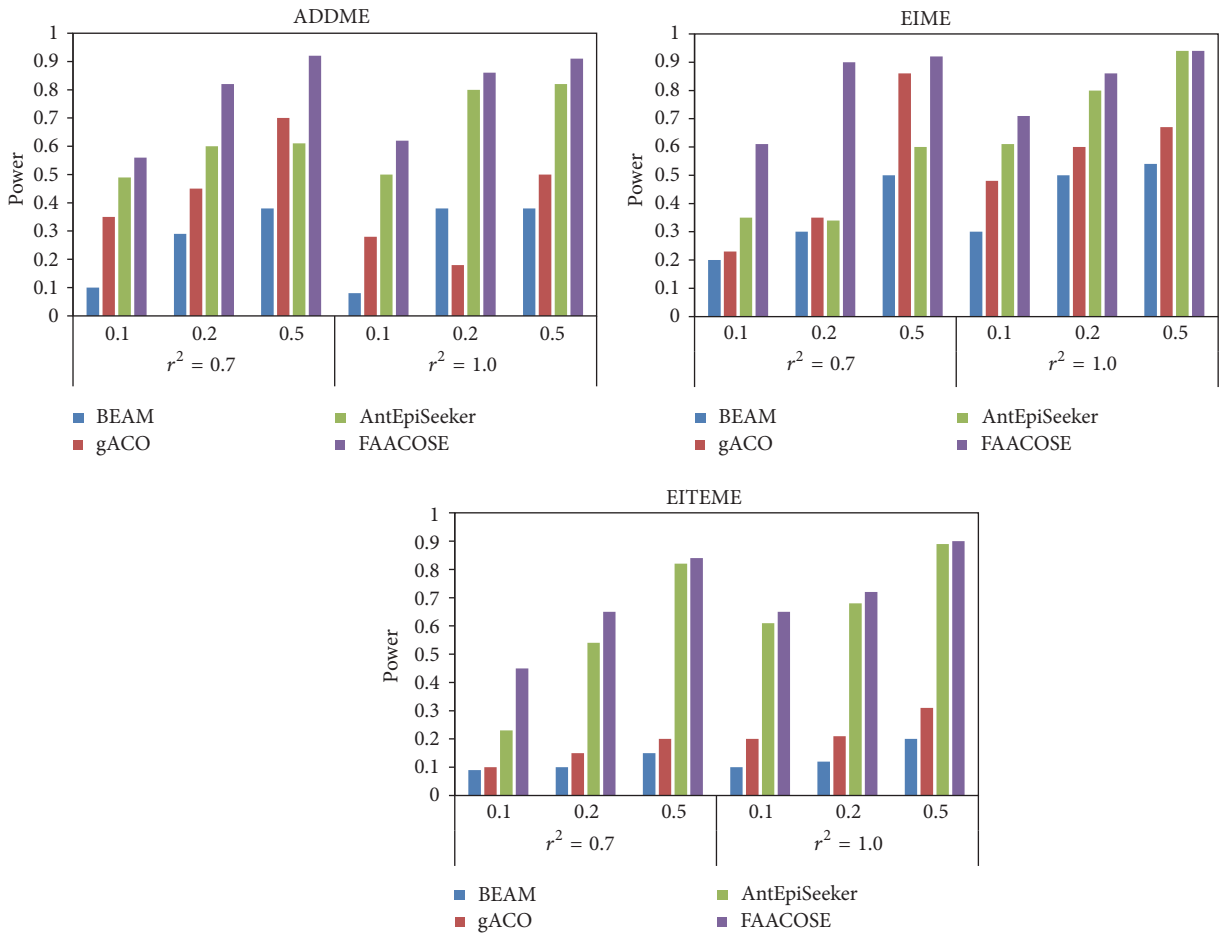


Figure 3: Power comparisons between existing methods and FAACOSE on three models.

for data simulation were the effective size λ, a measure of marginal effects as defined by Marchini et al. [42], linkage disequilibrium between SNPs measured by r^2, and minor allele frequencies (MAFs). λ was set to 0.3 for ADDME and 0.2 for EIME and EITEME. For r^2, two values (0.7 and 1.0) were used for each model. For MAFs, three values (0.1, 0.2, and 0.5) were considered. The parameters for BEAM were set as default. The parameter settings for AntEpiSeeker were large dataset size = 6, small dataset size = 3, count large = 150, count small = 300, epistasis model = 2, ant count = 1000, $\alpha = 1$, $\rho = 0.05$, and $\tau 0 = 100$ (also available in the software package documentation of AntEpiSeeker). The parameters of the generic ACO algorithm were set as ant count = 1000, $\alpha = 1$, $\rho = 0.05$, $\tau_0 = 100$, count (number of iterations) = 900, and epistasis model = 2. The comparison of detection power for BEAM, genetic ACO algorithm, and the AntEpiSeeker is presented in Figure 3. The results show that FAACOSE outperforms BEAM and the generic ACO in all parameter settings and is superior to AntEpiSeeker in most parameter settings.

In this section, we compare our proposed method with benchmark methods. First, we use power test to detect how many real SNP subsets can be found with our proposed method. Second, we use precision, recall, and F_1 score to evaluate the results. Precision denotes how many right SNP subsets in the total final identified SNP subsets. Recall denotes the number of right SNP subsets that are identified. F_1 score is an indicator used in statistics to measure the accuracy of two classification models. It takes into account the precision and recall of the classification model simultaneously. F_1 score can be seen as a weighted average of precision and recall, its maximum is 1, and minimum is 0.l. We show the results of FAACOSE with other methods on $r^2 = 0.7$ and MAF = 0.2 in Table 1.

The F_1 score of FAACOSE is better than other methods. We run the same experiment on datasets with different parameter combination. In all eighteen datasets FAACOSE has the highest F_1 score in fifteen of them. In real GWAS dataset experiment, the sample size of real dataset is huge. The efficiency of the method is also to be considered. The experimental results indicate that our proposed method is more effective method in real GWAS dataset. AntEpiSeeker is the most efficient algorithm among three methods. In different data samples, we compare run time of AntEpiSeeker and FAACOSE. And averaging the results, FAACOSE is faster 30% than AntEpiSeeker.

Table 1: F_1 score comparison between FAACOSE and other methods.

Model	Method	Recall	Precision	F_1 score
ADDME	BEAM	0.29	0.15	0.20
	gACO	0.45	0.36	0.40
	AntEpiSeeker	0.6	0.55	0.57
	FAACOSE	0.82	0.74	0.78
EIME	BEAM	0.3	0.45	0.36
	gACO	0.35	0.32	0.33
	AntEpiSeeker	0.34	0.56	0.42
	FAACOSE	0.9	0.82	0.86
EITEME	BEAM	0.1	0.14	0.12
	gACO	0.15	0.20	0.17
	AntEpiSeeker	0.54	0.46	0.50
	FAACOSE	0.65	0.62	0.63

4. Application to Real SNP Dataset

Late-Onset Alzheimer's Disease (LOAD) is the most frequent form of Alzheimer's disease, which is frequently identified in people older than 65 years; the LOAD or AD is a kind of chronic neurodegenerative diseases which is frequently not obvious in the onset of the disease and slowly changes dementia over time. It is the cause of 60% to 70% of cases of dementia. The most common early symptom is difficulty in remembering recent events (short-term memory loss). As the disease advances, symptoms can include problems with language, disorientation (including easily getting lost), mood swings, loss of motivation, not managing self-care, and behavioural issues. LOAD is a multifactor genetic disease; its etiology and pathogenesis have not yet been fully understood. The apolipoprotein (APOE) gene is a definite risk factor for LOAD. The APOE gene has three forms. The $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$; the effect of $\varepsilon 2$ is positive; $\varepsilon 2$ can effectively prevent the occurrence of the disease. There has been research report that genetic variant $\varepsilon 4$ has induced effect on disease. Between 40 and 80% of people with AD possess at least one APOE $\varepsilon 4$ allele [46]. Previous studies have reported some significant SNPs in the field of Genome-Wide Association Studies [47]. Reference [47] reported that 10 SNPs in the area of GAB2 gene have an epistasis effect with APOE e4 in relation to Late-Onset Alzheimer's Disease. We applied our proposed method to the LOAD GWAS dataset from website https://www.tgen.org/ [47]. After data preprocessing, the real biological dataset contains 1368 samples [48, 49]. Of these, 836 samples were identified case studies; the remaining 532 samples were normal sample [50, 51]. Each sample of real biological dataset contains 309,316 SNPs with genotype information, APOE status, and LOAD status [52]. For the next calculation, we code the APOE gene state with a binary variable; the value 1 represents the $\varepsilon 4$ variant and in turn the value 0 represents the other three variants [53]. An SNP locus was coded as a quaternary variable considering the missing state. The high potential LOAD disease related SNP is shown in Table 2.

Table 2: The number of selected SNPs of FAACOSE in LOAD dataset.

SNP rs#			
rs7756992	rs611154	rs191840	rs7294919
rs1887922	rs304900	rs1999764	rs1385600
rs2373115	rs7101429	rs609812	rs613375
rs1007837	rs2510038	rs4945261	rs10793294
rs520227	rs191740	rs7924284	rs829465
rs602106	rs7174511	rs606889	rs602192

5. Discussions

In this paper, we proposed a novel ant colony optimization based fast search method for the discovery of epistasis interactions in large scale real GWAS dataset. FAACOSE was evaluated through comparison with existing three approaches on both simulated and real datasets. FAACOSE, which adopts a fast adaptive optimization procedure, is a modified algorithm derived from the generic ACO. And with two-objective function, to demonstrate the advantages of fast adaptive ant colony optimization algorithm, we also compared the performance of the FAACOSE with that of the generic ACO.

In future studies, we intend to find more powerful modeling approaches, ant colony optimization algorithm with faster convergence, objective functions which can better measure data structure of GWAS dataset, more efficient optimal SNP subset search, and identification strategies that can be combined and flexibly embedded into our SNP epistasis search framework to find more accurate SNP subset. With the rapid development of bioinformatics, more and more biological information related to disease is identified. More and more studies will consider prior knowledge. An important future research direction is that we will try to apply expert prior knowledge to GWAS dataset with our proposed method, that is, the fast adaptive ant colony optimization algorithm for detecting SNP epistasis. Expert prior knowledge can improve the power and efficiency of epistasis detection.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is partly supported by National Natural Science Foundation of China (Grant nos. 61520106006, 31571364, 61732012, 61532008, U1611265, 61672382, 61402334, 61472280, 61472173, 61572447, 61672203, 61472282, and 61373098) and China Postdoctoral Science Foundation (Grant nos. 2014M561513, 2015M580352, 2017M611619, and 2016M601646) Guangxi Bagui Scholars Program Special Fund.

References

[1] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, no. 2, pp. 95–108, 2005.

[2] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genetics*, vol. 5, no. 6, Article ID e1000529, 2009.

[3] T. A. Manolio, F. S. Collins, N. J. Cox et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[4] B. S. Shastry, "SNP alleles in human disease and evolution," *Journal of Human Genetics*, vol. 47, no. 11, pp. 561–566, 2002.

[5] B. Stubbs, D. Vancampfort, M. De Hert, and A. J. Mitchell, "The prevalence and predictors of type two diabetes mellitus in people with schizophrenia: a systematic review and comparative meta-analysis," *Acta Psychiatrica Scandinavica*, vol. 132, no. 2, pp. 144–157, 2015.

[6] K. P. Liao, "Cardiovascular disease in patients with rheumatoid arthritis," *Trends in Cardiovascular Medicine*, vol. 27, no. 2, pp. 136–140, 2017.

[7] Y. Mao, N. R. London, L. Ma, D. Dvorkin, and Y. Da, "Detection of SNP epistasis effects of quantitative traits using an extended Kempthorne model," *Physiological Genomics*, vol. 28, no. 1, pp. 46–52, 2006.

[8] W. Zhang, J. Zhu, E. E. Schadt, and J. S. Liu, "A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules," *PLoS Computational Biology*, vol. 6, no. 1, Article ID e1000642, 2010.

[9] M. Kang, C. Zhang, H.-W. Chun, C. Ding, C. Liu, and J. Gao, "EQTL epistasis: Detecting epistatic effects and inferring hierarchical relationships of genes in biological pathways," *Bioinformatics*, vol. 31, no. 5, pp. 656–664, 2015.

[10] H. Lin, D. Chen, P. Huang et al., "SNP interaction pattern identifier (SIPI): an intensive search for SNP–SNP interaction patterns," *Bioinformatics*, 2016.

[11] R. L. Prentice and L. Qi, "Aspects of the design and analysis of high-dimensional SNP studies for disease risk estimation," *Biostatistics*, vol. 7, no. 3, pp. 339–354, 2006.

[12] S.-P. Deng, L. Zhu, and D.-S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," *BMC Genomics*, vol. 16, no. 3, article no. S4, 2015.

[13] S.-P. Deng and D.-S. Huang, "SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method," *Methods*, vol. 69, no. 3, pp. 207–212, 2014.

[14] J. H. Moore, J. M. Lamb, N. J. Brown, and D. E. Vaughan, "A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 Levels," *Clinical Genetics*, vol. 62, no. 1, pp. 74–79, 2002.

[15] B. M. Michael, R. E. Neapolitan, X. Jiang, and V. Shyam, "Learning genetic epistasis using Bayesian network scoring criteria," *BMC Bioinformatics*, vol. 12, no. 1, 89 pages, 2011.

[16] Y. Wang, X. Liu, K. Robbins, and R. Rekaya, "AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm," *BMC Research Notes*, vol. 3, article 117, 2010.

[17] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics*, vol. 39, no. 9, pp. 1167–1173, 2007.

[18] M. Dorigo, M. Birattari, and C. Blum, "Ant colony optimization and swarm intelligence," *SpringerVerlag*, vol. 5217, no. 8, pp. 767–771, 2004.

[19] T. Stützle, M. López-Ibáñez, P. Pellegrini et al., "Parameter adaptation in ant colony optimization," *Autonomous Search*, vol. 9783642214349, pp. 191–215, 2012.

[20] C. Blum and M. Sampels, "An ant colony optimization algorithm for shop scheduling problems," *Journal of Mathematical Modelling and Algorithms*, vol. 3, no. 3, pp. 285–308, 2004.

[21] R. Musa, J.-P. Arnaout, and H. Jung, "Ant colony optimization algorithm to solve for the transportation problem of cross-docking network," *Computers and Industrial Engineering*, vol. 59, no. 1, pp. 85–92, 2010.

[22] G. N. Varela and M. C. Sinclair, "Ant colony optimisation for virtual-wavelength-path routing and wavelength allocation," in *Proceedings of the 1999 Congress on Evolutionary Computation (CEC '99)*, pp. 1809–1816, Washington, DC, USA, July 1999.

[23] K. M. Sim and W. H. Sun, "Ant colony optimization for routing and load-balancing: survey and new directions," *Systems Man & Cybernetics Part A Systems Humans IEEE Transactions on*, vol. 33, no. 5, pp. 560–572, 2003.

[24] S.-H. Ngo, X. Jiang, and S. Horiguchi, "Adaptive routing and wavelength assignment using ant-based algorithm," in *Proceedings of the 2004 12th IEEE International Conference on Networks, ICON 2004 - Unity in Diversity*, pp. 482–486, November 2004.

[25] S. I. Vrieze, "Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)," *Psychological Methods*, vol. 17, no. 2, pp. 228–243, 2012.

[26] D.-S. Huang and J.-X. Du, "A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks," *IEEE Transactions on Neural Networks*, vol. 19, no. 12, pp. 2099–2115, 2008.

[27] B. V. North, D. Curtis, and P. C. Sham, "Application of logistic regression to case-control association studies involving two causative loci," *Human Heredity*, vol. 59, no. 2, pp. 79–87, 2005.

[28] P.-J. Jing and H.-B. Shen, "MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies," *Bioinformatics*, vol. 31, no. 5, pp. 634–641, 2015.

[29] N. Ryman, "CHIFISH: A computer program testing for genetic heterogeneity at multiple loci using chi-square and Fisher's exact test," *Molecular Ecology Notes*, vol. 6, no. 1, pp. 285–287, 2006.

[30] C. R. Mehta and N. R. Patel, "A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 427–434, 1983.

[31] B. Sobrino, M. Brión, and A. Carracedo, "SNPs in forensic genetics: A review on SNP typing methodologies," *Forensic Science International*, vol. 154, no. 2-3, pp. 181–194, 2005.

[32] O. Shoval, H. Sheftel, G. Shinar et al., "Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space," *Science*, vol. 336, no. 6085, pp. 1157–1160, 2012.

[33] D.-S. Huang and W. Jiang, "A general CPL-AdS methodology for fixing dynamic parameters in dual environments," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 5, pp. 1489–1500, 2012.

[34] L. Zhu, W.-L. Guo, S.-P. Deng, and D.-S. Huang, "ChIP-PIT: enhancing the analysis of chip-seq data using convex-relaxed pair-wise interaction tensor decomposition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 1, pp. 55–63, 2016.

[35] C. Angione, G. Carapezza, J. Costanza, P. Lio, and G. Nicosia, "Pareto optimality in organelle energy metabolism analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 1032–1044, 2013.

[36] R. A. Fisher, "On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P," *Journal of the Royal Statistical Society*, vol. 85, no. 1, p. 87, 1922.

[37] A. Agresti, "A survey of exact inference for contingency tables," *Statistical Science*, vol. 7, no. 1, pp. 131–153, 1992.

[38] B. Wenzheng, C. Yuehui, and W. Dong, "Prediction of protein structure classes with flexible neural tree," *Bio-Medical Materials and Engineering*, vol. 24, no. 6, pp. 3797–3806, 2014.

[39] L. Zhu, Z.-H. You, D.-S. Huang, and B. Wang, "*t*-LSE: a novel robust geometric approach for modeling protein-protein interaction networks," *PLoS ONE*, vol. 8, no. 4, Article ID e58368, 2013.

[40] C.-H. Zheng, L. Zhang, V. T.-Y. Ng, C. K. Shiu, and D.-S. Huang, "Molecular pattern discovery based on penalized matrix decomposition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1592–1603, 2011.

[41] D.-S. Huang and H.-J. Yu, "Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 2, pp. 457–467, 2013.

[42] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature Genetics*, vol. 37, no. 4, pp. 413–417, 2005.

[43] R. Jiang, W. Tang, X. Wu, and W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinformatics*, vol. 10, no. 1, article S65, 2009.

[44] J. Kruppa, A. Ziegler, and I. R. König, "Risk estimation and risk prediction using machine-learning methods," *Human Genetics*, vol. 131, no. 10, pp. 1639–1654, 2012.

[45] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15, pp. 1855–1862, 2006.

[46] R. W. Mahley, K. H. Weisgraber, and Y. Huang, "Apolipoprotein E4: a causative factor and therapeutic target in neuropathology, including Alzheimer's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 15, pp. 5644–5651, 2006.

[47] E. M. Reiman, J. A. Webster, A. J. Myers et al., "GAB2 alleles modify Alzheimer's Risk in APOE $\varepsilon 4$ carriers," *Neuron*, vol. 54, no. 5, pp. 713–720, 2007.

[48] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 599–607, 2009.

[49] S.-P. Deng, L. Zhu, and D.-S. Huang, "Predicting hub genes associated with cervical cancer through gene co-expression networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 1, pp. 27–35, 2016.

[50] L. Zhu, S.-P. Deng, and D.-S. Huang, "A two-stage geometric method for pruning unreliable links in protein-protein networks," *IEEE Transactions on Nanobioscience*, vol. 14, no. 5, pp. 528–534, 2015.

[51] D.-S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Current Protein and Peptide Science*, vol. 15, no. 6, pp. 553–560, 2014.

[52] D.-S. Huang, *Systematic Theory of Neural Networks for Pat-tern Recognition*, Publishing House of Electronic Industry of China, May 1996.

[53] D.-S. Huang, "Radial basis probabilistic neural networks: model and application," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 13, no. 7, pp. 1083–1101, 1999.

Hindawi
Complexity
Volume 2017, Article ID 4120506, 11 pages
https://doi.org/10.1155/2017/4120506

Research Article

Predicting Protein Complexes in Weighted Dynamic PPI Networks Based on ICSC

Jie Zhao,[1] Xiujuan Lei,[1] and Fang-Xiang Wu[2,3]

[1]*School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710119, China*
[2]*School of Mathematical Sciences, Nankai University, Tianjin 300071, China*
[3]*Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A9*

Correspondence should be addressed to Xiujuan Lei; xjlei@snnu.edu.cn and Fang-Xiang Wu; faw341@mail.usask.ca

Received 31 March 2017; Accepted 2 July 2017; Published 28 August 2017

Academic Editor: Juan A. Almendral

Protein complexes play a critical role in understanding the biological processes and the functions of cellular mechanisms. Most existing protein complex detection algorithms cannot reflect dynamics of protein complexes. In this paper, a novel algorithm named Improved Cuckoo Search Clustering (ICSC) algorithm is proposed to detect protein complexes in weighted dynamic protein-protein interaction (PPI) networks. First, we constructed weighted dynamic PPI networks and detected protein complex cores in each dynamic subnetwork. Then, ICSC algorithm was used to cluster the protein attachments to the cores. The experimental results on both DIP dataset and Krogan dataset demonstrated that ICSC algorithm is more effective in identifying protein complexes than other competing methods.

1. Introduction

Proteins are indispensable to cellular life. Biological functions of cells are carried out by protein complexes rather than single proteins [1]. Detecting these protein complexes can help to predict protein functions and explain biological processes, which has great significance in biology, pathology, and proteomics [2]. Therefore, the study of protein complexes has become one of most important subjects. Many of experimental methods combined with computational strategies have been proposed to predict and identify protein complexes, such as affinity purification and mass spectrometry [3–5]. However, they are costly and have difficulty in capturing the protein complexes instantaneous and dynamic changes [6].

The high throughput techniques have generated a large amount of protein-protein interaction (PPI) data, gene expression data, and protein structure data, which enable scholars to find protein complexes based on the topological properties of PPI networks and structural information of proteins [7]. Bader and Hogue proposed MCODE [8] method to detect protein complexes based on the proteins' connectivity and density in PPI networks. Liu et al. [9] presented a method called CMC to identify protein complexes based on maximal cliques. Protein complexes integrate multiple gene products to perform cellular functions and may have overlapping. Nepusz et al. [10] developed a clustering algorithm ClusterONE to detect overlapping protein complexes. Gavin et al. [11] suggested that there are two types of proteins in complexes: core components and attachments [11]. According to the core-attachment structure of protein complexes, Leung et al. [12] designed CORE algorithm which calculated the p value to detect cores. Wu et al. [13] proposed COACH algorithm to detect dense subgraphs as core components. The biological processes are dynamic and PPIs are changing over time [14]. Therefore, it is necessary to shift the study of protein complexes from static PPI networks to the dynamic characteristics of PPI networks [15]. Wang et al. constructed dynamic PPI network based on time series gene expression data to detect protein complexes [16]. Zhang et al. proposed CSO [17] algorithm by constructing ontology attributed PPI networks based on GO annotation information. Some classical clustering algorithms such as Markov clustering (MCL) [18] and fuzzy clustering [19, 20] were also developed to detect protein complexes.

However, with the birth of the biological simulation technology, bioinspired algorithms provided a new perspective for solving protein complex detection problem [21]. In 2016, Lei et al. proposed F-MCL [22] clustering model based on Markov clustering and firefly algorithm which automatically adjusted the parameters by introducing the firefly algorithm. At the same year, Lei et al. proposed FOCA [6] clustering model which was based on the fruit flies' foraging behavior and protein complexes' core-attachment structure. The previous studies proved that the protein complex detection methods based on the bioinspired algorithms had shown a relatively better performance.

Cuckoo Search (CS) algorithm is a new intelligence optimization algorithm which has been successfully applied to the global optimization problem, clustering, and other fields [21]. In this study, according to the core-attachment structure of protein complexes and CS mechanism, a new clustering method named Improved Cuckoo Search Clustering (ICSC) algorithm was proposed to detect protein complexes in weighted dynamic PPI networks, in which the corresponding relationships between CS algorithm and clustering procedure of PPI data are established.

2. Methods

2.1. Constructing Weighted Dynamic PPI Network. The static PPI networks data produced by high throughput experiments generally contain a high rate of false positive and false negative interactions [9], which makes it inaccurate to predict protein complexes and impossible to reflect the real dynamic changes of PPIs in a cell. To address this problem, some scholars used the computational methods to evaluate the interactions [23]. On the other hand, the protein dynamic information such as gene expression data, subcellular localization data, and transcription regulation data were integrated to reveal the dynamics of PPIs [24–26]. Tang et al. [27] constructed time course PPI network (TC-PIN) by using gene expression data over three successive metabolic cycles. The expression values of genes were compared with a single-threshold to determine whether a gene was expressed. Some essential genes were filtered out by the single-threshold for their low expression levels. Wang et al. [28] developed a three-sigma method to define an active threshold for each gene and then constructed dynamic PPI network (DPIN) by using active proteins based on the static PPI network in combination with gene expression data. Many previous studies have revealed that the three-sigma principle had better prediction performance. In this study, we use three-sigma principle to construct the DPIN. The gene expression data includes three successive metabolic cycles; each cycle has 12 timestamps, so the DIPN includes 12 subnetworks.

A protein p is considered to be active in a dynamic PPI subnetwork only if its gene expression value is greater than or equal to the active threshold Active_Th(p) [28]:

$$\text{Active\_Th}(p) = \mu(p) + 3\sigma(p)(1 - F(p)), \quad (1)$$

where $\mu(p)$ is the algorithmic mean of gene expression values of protein p over timestamps 1 to n and $\sigma(p)$ is the standard deviation of its gene expression values. $F(p)$ is defined as follows:

$$F(p) = \frac{1}{1 + \sigma^2(p)}. \quad (2)$$

A static PPI network is usually described as an undirected graph $G(V, E)$ which consists of a set of nodes V and a set of edges E, the nodes in V represent the proteins and the edges in $E = \{e(v_i, v_j)\}$ represent the connections between pairs of proteins v_i and v_j. $G_t(V_t, E_t)$ is denoted as the dynamic PPI subnetwork at timestamp t ($t = 1, 2, \ldots, n$). Protein v_i interacts with protein v_j in a dynamic PPI subnetwork G_t only if they are active in the same timestamp t and connect with each other in the static PPI network.

As shown in Figure 1, three-sigma principle was applied to calculate the active threshold Active_Th(p) for each protein and to determine the active timestamps. After that, 12 dynamic subnetworks were constructed.

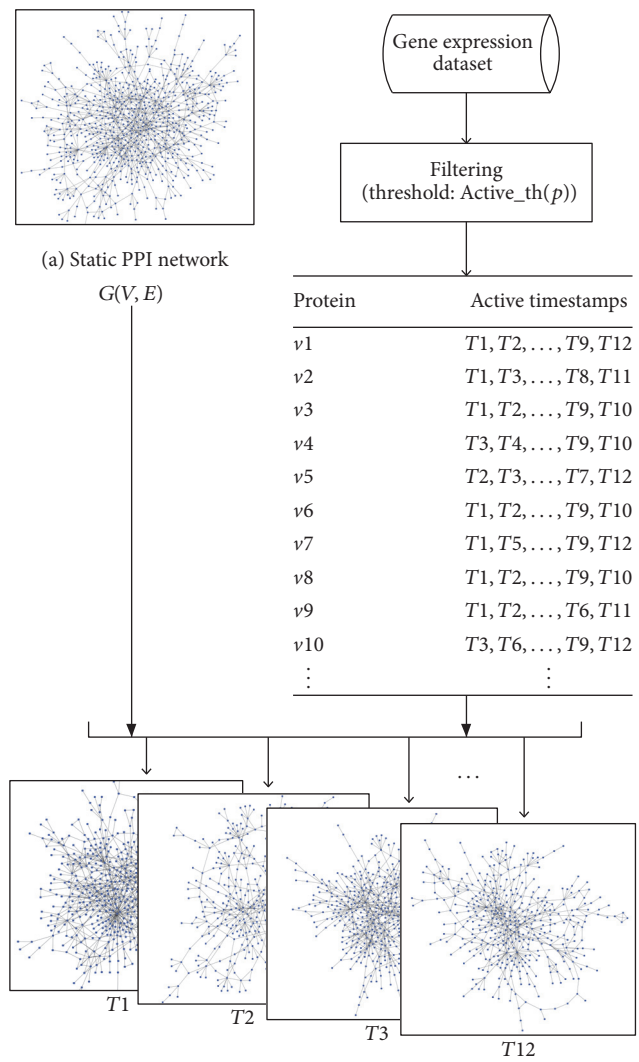


Figure 1: DPIN construction. (a) The static PPI network. (b) The subnetworks contained in the DPIN.

Clustering coefficient has been used as an effective tool to analyze the topology of PPI networks [29]. Radicchi et al. proposed the edge clustering coefficient (ECC) [30]. In PPI network, the ECC of an edge connecting proteins v_i and v_j can be expressed as follows:

$$\mathrm{ECC}_{ij} = \frac{Z_{ij}}{\min\left(\left|N_i\right| - 1, \left|N_j\right| - 1\right)}, \tag{3}$$

where Z_{ij} is the number of triangles built on edge (v_i, v_j); $|N_i|$ and $|N_j|$ are the degrees of protein v_i and v_j, respectively. Edge clustering coefficient is a local variable which characterizes the closeness of two proteins v_i and v_j.

The Pearson correlation coefficient (PCC) was calculated to evaluate how strong two interacting proteins are coexpressed [31]. The PCC value of a pair of genes $x = \{x_1, x_2, \ldots, x_n\}$ and $y = \{y_1, y_2, \ldots, y_n\}$, which encode the corresponding paired proteins v_i and v_j interacting in the PPI network, is defined as

$$\mathrm{PCC}(x, y) = \frac{\sum_{k=1}^{n} (x_k - \mu(x))(y_k - \mu(y))}{\sqrt{\sum_{k=1}^{n} (x_k - \mu(x))^2} \sqrt{\sum_{k=1}^{n} (y_k - \mu(y))^2}}, \tag{4}$$

where $\mu(x)$ and $\mu(y)$ are the mean gene expression value of proteins v_i and v_j, respectively. The value of PCC ranges from −1 to 1; if $\mathrm{PCC}(x, y)$ is a positive value, there is a positive correlation between proteins v_i and v_j.

The protein complex is a group of proteins which show high coexpression patterns and share high degree of functional similarity, so we integrate GO-slims data from the point of view of protein functions. If two interacted proteins v_i and v_j have some common GO terms, their functions are more similar. Let GSM_{ij} denote this correlation which can be computed as follows:

$$\mathrm{GSM}_{ij} = \frac{\left|\mathrm{GSM}_i \cap \mathrm{GSM}_j\right|^2}{\left|\mathrm{GSM}_i\right| \times \left|\mathrm{GSM}_j\right|}, \tag{5}$$

where $|\mathrm{GSM}_i|$ and $|\mathrm{GSM}_j|$ represent the number of GO terms for proteins v_i and v_j, respectively. In the dynamic PPI subnetwork G_t, the weight between proteins v_i and v_j is defined as follows:

$$W_{ij} = \frac{\mathrm{PCC}_{ij} + \mathrm{ECC}_{ij} + \mathrm{GSM}_{ij}}{3}. \tag{6}$$

Up to now, the weighted dynamic PPI network was constructed.

2.2. Cuckoo Search Algorithm. CS algorithm was a novel bioinspired metaheuristic optimization algorithm proposed in 2009 [32], which was based on the obligatory brood parasitic behaviors of some cuckoo species in combination with the Lévy flight behaviors.

During the breeding period, some certain species of cuckoos lay their eggs in host nests. The cuckoos usually look for host birds which have similar incubation period and brood period. Moreover, their eggs are similar to each other in many aspects of color, shape, size, and cicatricle. The cuckoo flight strategy demonstrates the typical characteristics of Lévy flights. Lévy flights comprise sequences of randomly orientated straight-line movements. Actually, the strategies of frequently occurring but relatively short straight-line movements, as well as randomly alternating with more occasionally occurring longer movements, can maximize the efficiency of resource search [33].

Specifically, for a cuckoo i when generating new solutions $x(t + 1)$, a Lévy flight is performed by using the following equation:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus \text{Lévy}(\beta), \quad (i = 1, 2, \ldots, n), \tag{7}$$

where $\alpha > 0$ is the step size which should be related to the scales of the problem of interests. In most cases, we can use $\alpha = 1$; $\oplus$ means the Hadamard product operator. The Lévy flight is a type of random walk which has a power law step length distribution with a heavy tail and the value of β between 1 and 3.

2.3. The ICSC Algorithm. Our ICSC is developed to detect protein complexes in weighted dynamic PPI network through the use of improved CS algorithm. It has been widely accepted that protein complexes are organized in the core-attachment structure.

The core is a small subgraph in a PPI network with high density. As shown in Figure 2(a), four highly connected subgraphs constitute cores, denoted by *core1*, *core2*, *core3*, and *core4* (red round proteins in the dashed circle). Several peripheral connection protein nodes are attachments (blue square proteins) in this PPI network. The blue square proteins and black diamond proteins are all noncore proteins.

In ICSC algorithm, each cuckoo was viewed as a noncore protein (marked with black round in Figure 2(b)), and the nest was viewed as the core proteins (marked with black circles in Figure 2(b)), while the cuckoo population is denoted as a group of clustering results. The noncore proteins become attachments if a cuckoo finds an appropriate nest to lay eggs. Figure 2 illustrates the corresponding relationships between ICSC algorithm and the clustering procedure of a PPI network. Algorithm 1 indicates the function of the proposed algorithm ICSC. The ICSC method operates in three phases. In the first step, some dense subgraphs were selected as initial nests. Then the cuckoos are generated based on these nests. Last the improved Cuckoo Search strategy was applied to generate protein complexes. The complexes in different dynamic subnetworks may have a high level of similarity, so a refinement procedure is applied in order to filter out redundancies and generate the final set of protein complexes.

"Initial nest" subfunction (Algorithm 1) tries to generate initial nests. The initial nests can be seen as the core proteins for each protein complex. The weight of dynamic PPI subnetwork $G_t(V_t, E_t)$ has considered the PCC, ECC, and GSM, so the weight threshold wth can be used to find some protein pairs which have highly functional similarity

Input. The weighted PPI sub-network: $G_t(V_t, E_t)$, $t = 1, 2, \ldots, 12$;
Output. The detected protein complexes: *Complex*
Begin
(1) **for** each G_t **do**
(2) **Initialization**: (1) maximum iterations: maxiter; cuckoo populations' size: *np*;
(3) (2) weight threshold: wth;
(4) (3) Initial nest *nest*: **for** each $e(v_i, v_j) \in E_t$ **do**
(5) **if** $w_{ij} \geq$ (mean(w)/wth) **then** insert (v_i, v_j) into nest **end if**
(6) **end for**
(7) Merge operation;
(8) (4) Initial solutions *Nest*: Nest(:, :, i) = nest, $i = 1, 2, \ldots, np$;
(9) **while** iter ≤ maxiter **do**
(10) **for** $i = 1$ to *np*
(11) Generation cuckoos $Cuckoo_i$: each $v \in V_t$, **if** $v \notin$ Nest(:, :, i) **then** insert v into Cuckoo_i **end if**
(12) **for** each $\text{cuckoo}_j \in \text{Cuckoo}_i$ **do**
(13) **for** each $\text{nest}_k \in$ Nest(:, :, i) **do**
(14) Calculate closeness($\text{cuckoo}_j, \text{nest}_k$);
(15) **if** closeness($\text{cuckoo}_j, \text{nest}_k$) > 0 **then**
(16) Roulette wheel selection $cuckoo_j$, set nest_t = union($\text{nest}_k, \text{cuckoo}_j$);
(17) Calculate objective function $F(\text{nest}_t)$;
(18) **if** $F(\text{nest}_t) > F(\text{nest}_k)$ **then**
(19) insert cuckoo_j into nest_k;
(20) **end if**
(21) **end if**
(22) **end for**
(23) **end for**
(24) Calculate the objective function F(Nest(:, :, i))
(25) **end for**
(26) Find the largest objective function $F\text{max} = \max(F(\text{Nest}(:, :, i)))$, $i = 1, 2, \ldots, np$;
(27) Find the best solution *Nestbest*, F(Nestbest) = Fmax;
(28) **end while**
(29) Complex_t = Nestbest;
(30) **end for**
(31) Complex = $(\text{Complex}_1, \text{Complex}_2, \ldots, \text{Complex}_{12})$
(32) Refinement procedure;
End

Algorithm 1: ICSC algorithm.

and high coexpression. For $e(v_i, v_j) \in E_t$, if the weight w_{ij} is larger than mean(w)/wth, the node pair (v_i, v_j) is denoted as one initial nest, where mean(w) is the average weight of G_t. Protein complex cores often correspond to the small, dense, and reliable subgraphs in PPI networks, but the node pairs may have overlaps with each other. So the node clustering coefficient (NCC) was used to filter out the overlapping nests, which is defined as follows:

$$\text{NCC}(v) = \frac{2n_v}{k_v(k_v - 1)}, \tag{8}$$

where k_v is the degree of node v, n_v is the number of links connecting the k_v neighbors of node v to each other. Because the PPI network has a large number of nodes and edges, many nodes may have the same value of node clustering coefficient. In this study, the weighted node clustering coefficient (WNCC) was defined to distinguish the importance of nodes in the dynamic PPI network. For two initial nests (v_i, v_j) and (v_i, v_k), if $\text{WNCC}(v_i) \geq \text{WNCC}(v_j)$ and $\text{WNCC}(v_i) \geq \text{WNCC}(v_k)$, they are merged into (v_i, v_j, v_k). The WNCC of node v is defined as

$$\text{WNCC}(v) = \frac{\sum \text{We}}{k_v(k_v - 1)}, \quad e \in n_v, \tag{9}$$

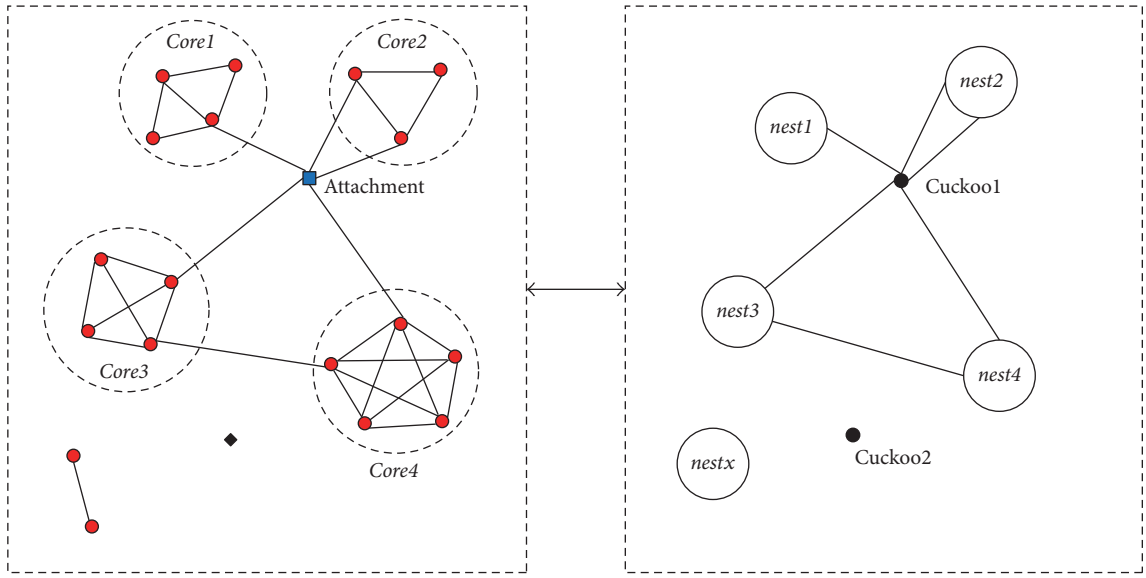
where We is the weight of edge $e \in n_v$; k_v and n_v have the same meanings as in NCC.

After nest detection in the previous steps, the nests are fixed. It is time to find cuckoos around the nests. In $G_t(V_t, E_t)$, if protein $v_i \in V_t$ is not in any nests, it is denoted as a cuckoo.

As a "cuckoo" in $G_t(V_t, E_t)$, there are many "nests" around "cuckoo"; the similarities between "cuckoo" and "nest" is measured based on the closeness between cuckoo_i and nest_j, defined as follows:

$$\text{closeness}\left(\text{cuckoo}_i, \text{nest}_j\right) = \frac{\left|N_{\text{cuckoo}_i} \cap \text{nest}_j\right|}{\left|\text{nest}_j\right|}, \tag{10}$$

where N_{cuckoo_i} is the set of all cuckoo_i's neighbors, $|N_{\text{cuckoo}_i} \cap \text{nest}_j|$ is the number of vertices in nest_j connected with



(a) Core-attachment structure (b) ICSC algorithm

Figure 2: The corresponding relationships between ICSC algorithm and the clustering procedure of a PPI network.

cuckoo_i, and $|\text{nest}_j|$ is the number of vertices in nest_j. In order to keep the diversity of population, the roulette wheel selection was used. For a cuckoo_i, if $\text{closeness}(\text{cuckoo}_i, \text{nest}_j) > 0$, the nest_j is selected to construct the roulette wheel.

The objective function F is defined as follows:

$$F\left(C^1, C^2, \ldots, C^k\right) = \sum_{i=1}^{k} \frac{C_{\text{in}}^i}{C_{\text{in}}^i + C_{\text{out}}^i},$$
$$C_{\text{in}}^i = \frac{2 \times |E|}{|V| \times (|V| - 1)}, \tag{11}$$
$$C_{\text{out}}^i = \frac{W_{ki}}{|V|},$$

where $(C^1, C^2, \ldots, C^k)$ is a clustering result determined by a nest; C^i represents a cluster. $|E|$ is the number of edges in the cluster C^i; $|V|$ is the number of nodes in the cluster C^i. W_{ki} is the number of edges with one node in C^i and another node outside C^i. Finally, the same or highly overlapping protein complexes are filtered out.

2.4. Time Complexity Analysis of ICSC Algorithm. The time complexity is used to estimate the efficiency of the ICSC algorithm. The maximal iterations *maxiter* is for the external loop; each iteration produces np solutions. In order to generate solutions, there are three main operations, generating the cuckoo, calculating the closeness, and calculating the objective function. Let nv be the number of proteins in G_t and ne be the number of interactions in G_t. The time complexity of generating the cuckoos is $O(\text{nv})$. The time complexity of calculating closeness is $O(\text{nc} * \text{nn})$, where nc is the number of cuckoos; nn is the number of nests. The time complexity of calculating the objective function is $O((\text{nv} - \text{nc})^2)$. In summary, the time complexity of ICSC algorithm is $O(\text{maxiter} * \text{np} * (\text{nv} + \text{nc} * \text{nn} + (\text{nv} - \text{nc})^2))$, which is equivalent to $O(\text{maxiter} * \text{np} * \text{nv}^2)$.

3. Experiments and Results

The proposed ICSC algorithm was implemented in Matlab R2015b and executed on a quad-core processor 3.30 GHz PC with 8 G RAM.

3.1. Experimental Dataset. In this study, four PPI datasets DIP [34] (version of 20160114), Krogan et al. [35], MIPS [36], and Gavin et al. [11] were employed to evaluate our algorithm. All the data used were *Saccharomyces cerevisiae* which have false positive and false negative interactions in the datasets. In this study, self-interactions and repetitive interactions are removed for data preprocessing. After preprocessing, the DIP dataset consists of 5028 proteins and 22302 interactions, the Krogan dataset consists of 2674 proteins and 7075 interactions, the MIPS dataset consists of 4546 proteins and 12319 interactions, and the Gavin dataset consists of 1430 proteins and 6531 interactions.

Gene expression data was retrieved from GEO (Gene Expression Omnibus, GSE3431) [37]. After preprocessing, the dataset contains 7074 genes in 3 cell life cycles, each cycle having 12 time points. The GSE3431 dataset contains 4876 proteins in the DIP dataset (coverage rate: 4876/5028 = 96.98%), 2644 proteins in the Krogan dataset (the coverage rate: 2644/2674 = 98.88%), 4446 proteins in the MIPS dataset (the coverage rate: 4446/4546 = 97.80%), and 1418 proteins in the Gavin dataset (the coverage rate: 1418/1430 = 99.16%).

The GO database is currently one of most comprehensive ontology databases in bioinformatics. GO-slims data are

Table 1: The number of proteins and interactions in SPIN and DPIN on four datasets.

Datasets		SPIN	DPIN timestamp t											
			1	2	3	4	5	6	7	8	9	10	11	12
DIP	Protein	5028	860	1029	863	671	645	598	530	1000	1194	638	690	489
	Interactions	22302	1103	1608	1337	839	835	752	627	1861	2447	950	1026	569
Krogan	Protein	2674	336	379	320	256	206	189	202	580	626	304	330	250
	Interactions	7075	334	464	331	234	210	184	213	1025	1081	314	373	258
MIPS	Protein	4546	737	897	781	583	570	531	470	839	1014	523	616	402
	Interactions	12319	1097	1443	1183	754	684	642	504	1238	1637	878	1207	700
Gavin	Protein	1430	177	228	215	135	112	102	96	379	419	174	190	146
	Interactions	6531	242	334	317	150	135	118	135	1019	1043	230	264	184

cut-down version of the GO ontologies [17], which is available at http://www.yeastgenome.org/download-data/curation. GO-slim data provide GO terms to explain gene product feature in biological process (BP), molecular function (MF), and cellular component (CC). we used GO-slims to annotate PPI data.

The standard protein complex CYC2008 [38] is used to evaluate our clustering results, which includes 408 protein complexes and covers 1492 proteins.

In this study, three-sigma principle is used to construct the dynamic PPI networks based on four static PPI networks (SPIN) DIP, Krogan, MIPS, and Gavin in combination with GSE3431 gene expression dataset. There are 12 timestamps per cycle in GSE3431, so each dynamic PPI network contains 12 subnetworks, as shown in Table 1. These 12 subnetworks have different sizes.

3.2. Evaluation Metrics. Three commonly used metrics *sensitivity (SN)*, *specificity (SP), and F-measure* [8, 25, 39] are used to measure the efficiency of the proposed ICSC algorithm and evaluate the performance of the clustering results:

$$\begin{aligned} \mathrm{SN} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{SP} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\ F\text{-measure} &= \frac{2\times \mathrm{SN}\times \mathrm{SP}}{\mathrm{SN}+\mathrm{SP}}, \end{aligned} \tag{12}$$

where TP is the number of predicted protein complexes which are matched with 408 standard protein complexes, FP is the number of predicted protein complexes which are not matched with anyone of 408 standard protein complexes, and FN is the number of standard protein complexes which are not matched with predicted protein complexes [8, 25]. The overlapping score OS is used to evaluate the matching degree between predicted protein complexes and standard protein complexes:

$$\mathrm{OS}\left(\mathrm{pc}, \mathrm{sc}\right) = \frac{\left|V_{\mathrm{pc}} \cap V_{\mathrm{sc}}\right|^2}{\left|V_{\mathrm{pc}}\right| \times \left|V_{\mathrm{sc}}\right|}, \tag{13}$$

where V_{pc} and V_{sc} denote the node sets of predicted protein complex pc and standard protein complex sc, respectively. The threshold of OS is set for 0.2 [8, 40]; that is, if OS(pc, sc) is greater than 0.2, the predicted protein complex pc is considered to match standard protein complex sc. OS(pc, sc) = 1 shows that the predicted protein complex pc is perfectly matched with the standard protein complex sc. The p value [41], which illustrates the probability that a protein complex is enriched by a given functional group, was used to evaluate the biological significance of the predicted protein complexes in this study:

$$p\text{-value} = 1 - \sum_{i=0}^{k-1} \frac{\binom{F}{i}\binom{N-F}{C-i}}{\binom{N}{C}}, \tag{14}$$

where N, C, and F are the sizes of the whole PPI network, a protein complex, and a functional group in the network, respectively, and k is the number of proteins in the functional group in the protein complex [41]. For a protein complex, the smaller the p value is, the higher the biological significance is. The protein complex is considered to be insignificant if p value is greater than 0.01.

3.3. Parameter Analysis. The proposed algorithm ICSC has three parameters, the maximum iterations *maxiter*, the cuckoo populations' size np, and the weight threshold wth. The maximum number of iterations *maxiter* measures the convergence performance of the algorithm, and the populations' size np can guarantee the diversity of the population. The convergence curve of ICSC algorithm on the first subnetwork of the dynamic PPI network was shown in Figure 3. The horizontal axis is the number of iterations, and the vertical axis is the objective function value. Figure 3 illustrates that the ICSC algorithm converges with 30 iterations. The populations' size np is from 5 to 30; the objective function reaches its maximum value at np = 15. In this study, we set maxiter = 100, np = 15.

In ICSC method, cuckoo_i chooses the most suitable nest_j to form a protein complex; the quality of nest_j directly determines the accuracy of protein complexes, and the value of weight threshold wth directly affects the quality of the nest. If the value of wth is too small, a small amount of protein pairs is selected in a nest; the clustering results are not accurate. On the contrary, if the value of wth is too large, lots of meaningless protein complexes are predicted. Therefore, it is critical to select the appropriate value of wth. Matching Rate

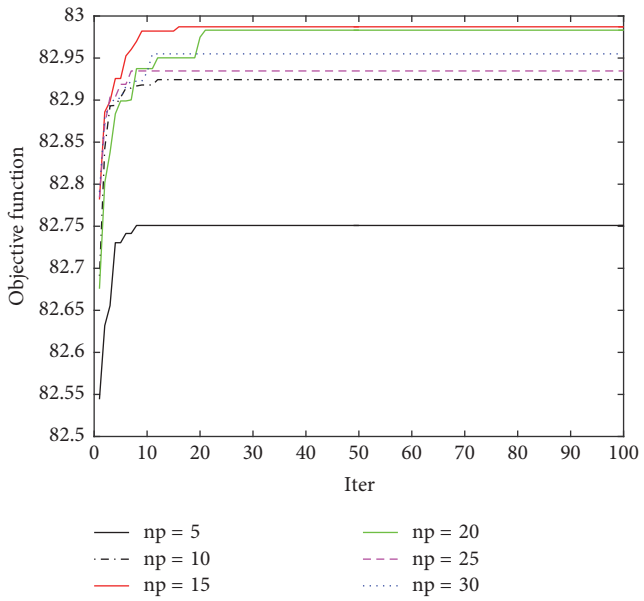


Figure 3: Convergence curve of ICSC algorithm on number 1 subnetwork of DIP dataset.

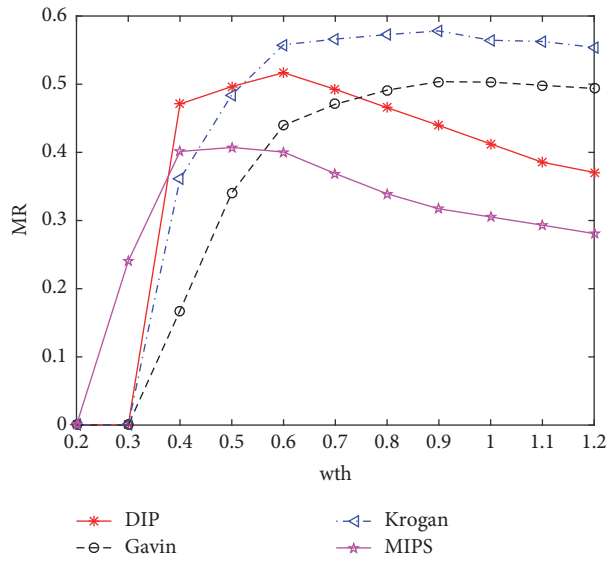


Figure 4: Influence of parameters wth on four DPIN.

(MR) is defined to verify the influence of different values of wth. Nest is the set of initial nests of the dynamic PPI network; SC is the set of standard protein complexes CYC2008, and MR(Nest, SC) is defined as follows:

$$\mathrm{MR}\,(\mathrm{Nest}, \mathrm{SC}) = \frac{(\mathrm{NI}/\,|\mathrm{Nest}| + \mathrm{SI}/\,|\mathrm{SC}|)}{2}, \tag{15}$$

where NI is the number of nests which are included in the standard protein complexes, |Nest| denotes the number of nests in Nest, SI is the number of standard protein complexes which are included in Nest, and |SC| denotes the number of protein complexes in SC. The experiments on four dynamic PPI networks with wth from 0.2 to 1.2 were carried out to verify the influence of parameters wth. The results were showed in Figure 4. From Figure 4, in Krogan and Gavin datasets, the MR tends to be stable while wth is greater than or equal to 0.8. In DIP datasets the MR reaches its maximum value at wth = 0.6 and then gradually declines, and the downward trend is from 0.6 to 0.8. The MR curve in MIPS dataset is similar to DIP. Therefore, the value of wth is set as 0.8 in this study.

3.4. Clustering Results. The performance of ICSC is compared with six other previously proposed methods: MCODE, MCL, CORE, CSO, ClusterONE, and COACH. All the six methods were run on the dynamic PPI networks constructed by three-sigma principle based on DIP, Krogan, MIPS, and Gavin datasets. The clustering results are shown in Table 2, where PC is the total number of predicted protein complexes, MPC is the count of predicted protein complexes which were matched, and MSC is the number of matched standard protein complexes. Perfect is the count of predicted protein complexes and standard complexes are perfectly matched; that is, OS(pc, sc) = 1. AS represents the average size of the predicted protein complexes. The comparison results are also showed in Table 2, from which it is clear that ICSC performs better than other six methods in terms of *sensitivity (SN)* and MPC. The *F*-measure of ICSC is the highest on DIP, Krogan, and MIPS while on the Gavin the *F*-measure of ICSC it was a bit less than that of ClusterONE. The Perfect values of ICSC on DIP and MIPS are 64 and 50, respectively, and are far superior to other algorithms.

In Table 2, the *perfect* value of ICSC on DIP is 64. The degree distribution of perfectly matched protein complexes is calculated in Table 3. The *degree* refers to the number of protein nodes contained in the protein complex. There are 408 protein complexes in the standard protein complexes CYC2008; 172 complexes contain 2 protein nodes accounting for 42.16%. However, the MCODE, CSO, and COACH cannot predict this part of protein complexes. The degree of 149 protein complexes greater than or equal to 4 accounted for 36.52% of all standard protein complexes, only a small part of which can be predicted by MCL, CORE, and ClusterONE. It is clear that ICSC algorithm achieved the best performance in these two aspects.

In order to clearly show the clustering results, we visualize the 265th standard protein complex of CYC2008 "nuclear exosome complex" in Figure 5. As shown in Figure 5(a), there are 12 proteins in this standard protein complex. The clustering results of other five methods MCODE (b), MCL (c), CORE (d), ClusterONE (e), and ICSC (f) are all from Krogan dataset. The blue nodes are proteins that are correctly predicted, the red nodes are proteins that are not identified, and the green nodes are the proteins that are wrongly identified. MCODE method only successfully predicted six proteins. Although MCL successfully predicted all 12 proteins in the protein complex, MCL also produced 3 incorrect proteins. The accuracy of CORE is the lowest; only 2 proteins are successfully predicted. Our method ICSC accurately

Table 2: The performance comparison of several typical algorithms on four datasets.

Dataset	Algorithms	SN	SP	*F*-measure	PC	MPC	MSC	Perfect	AS
DIP	MCODE	0.2318	0.6182	0.3372	165	102	70	6	6.7212
	MCL	0.7031	0.2505	0.3694	1541	386	245	14	4.4361
	CORE	0.7381	0.2769	0.4027	1517	420	259	39	2.443
	CSO	0.4403	0.6257	0.5169	342	214	136	11	4.652
	ClusterONE	0.6093	0.3385	0.4352	972	329	197	15	3.5422
	COACH	0.5009	0.5591	0.5284	474	265	144	13	4.9789
	ICSC	*0.8385*	*0.4186*	*0.5585*	*1997*	*836*	*247*	*64*	*3.5613*
Krogan	MCODE	0.2749	0.7937	0.4084	160	127	73	10	5.125
	MCL	0.566	0.4559	0.5051	658	300	178	40	3.9544
	CORE	0.5417	0.4121	0.4681	677	279	172	39	2.6041
	CSO	0.3284	0.8254	0.4699	189	156	89	10	5.2646
	ClusterONE	0.5232	0.4632	0.4914	585	271	161	28	3.935
	COACH	0.3566	0.81	0.4952	221	179	85	11	5.3575
	ICSC	*0.6314*	*0.5966*	*0.6135*	*761*	*454*	*143*	*23*	*3.3338*
MIPS	MCODE	0.1714	0.5333	0.2595	135	72	60	4	5.437
	MCL	0.5451	0.2017	0.2945	1259	254	196	17	4.7434
	CORE	0.6235	0.249	0.3558	1217	303	225	29	2.5859
	CSO	0.2835	0.5163	0.366	246	127	87	6	4.5528
	ClusterONE	0.4483	0.2796	0.3444	744	208	152	17	3.1317
	COACH	0.3145	0.3662	0.3384	396	145	92	5	6.5253
	ICSC	*0.7181*	*0.3028*	*0.4260*	*1691*	*512*	*207*	*50*	*3.7534*
Gavin	MCODE	0.2612	0.7548	0.3881	155	117	77	6	5.3484
	MCL	0.4411	0.6417	0.5228	321	206	147	25	5.0312
	CORE	0.4336	0.5735	0.4938	347	199	148	26	2.8184
	CSO	0.3109	0.773	0.4434	185	143	91	6	5.9405
	ClusterONE	0.4797	0.6413	0.5488	368	236	152	19	5.2826
	COACH	0.3477	0.6966	0.4585	234	163	94	5	6.312
	ICSC	*0.5033*	*0.5704*	*0.5347*	*540*	*308*	*104*	*10*	*3.6093*

Table 3: The degree distribution of predicted protein complexes perfectly matched (OS = 1) on DIP datasets.

Algorithm	Prefect	Degree (=2)	Degree (≥4)
MCODE	6	0	3
MCL	14	10	1
CORE	39	32	1
CSO	11	0	5
ClusterONE	15	11	2
COACH	13	0	6
ICSC	64	**47**	**6**
CYC2008	408	172 (42.16%)	149 (36.52%)

predicted 9 proteins and achieved the best performance in identifying protein complexes.

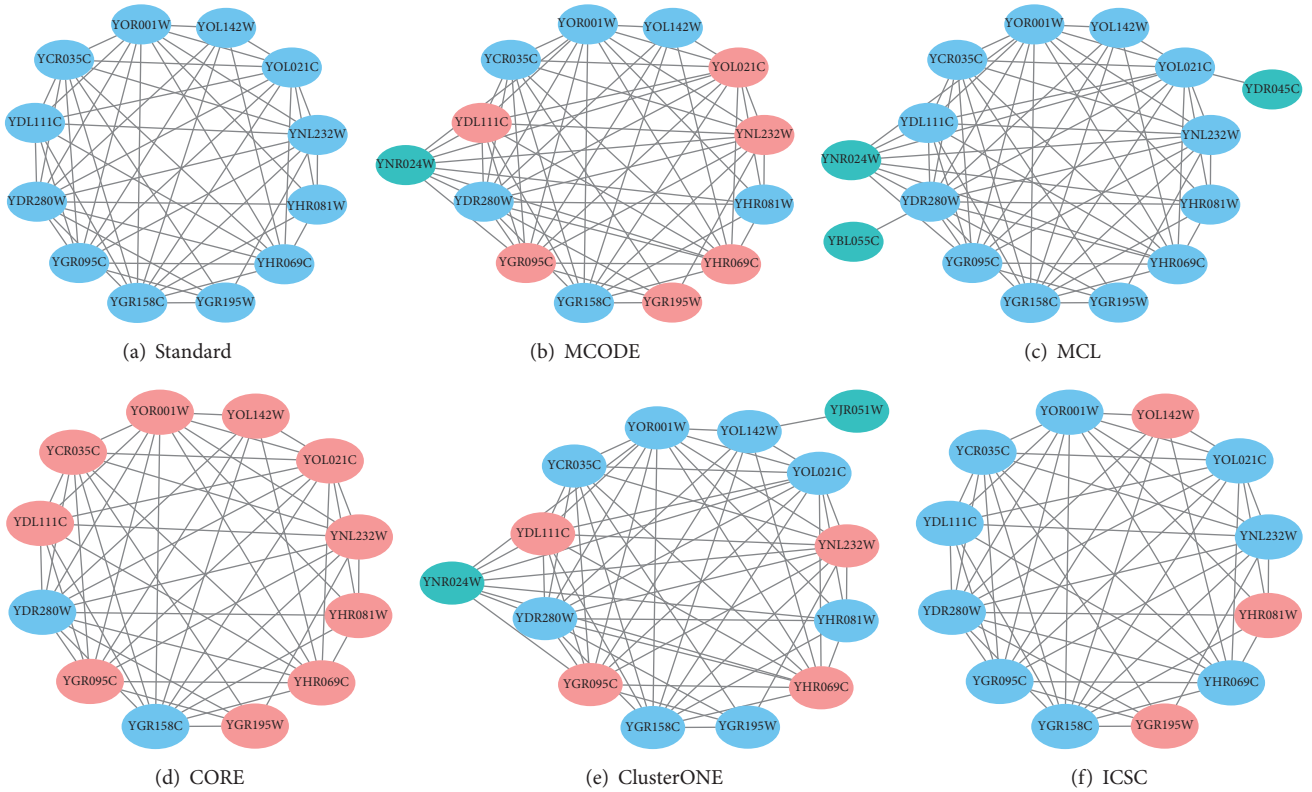
To evaluate the biological significance and functional enrichment of protein complexes identified by ICSI, we randomly selected five predicted protein complexes and calculated the p value of on biological process ontologies based on Krogan datasets by using GO: termFinder (http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl). The results are showed in Table 4. The proteins in bold have well matched standard protein complexes. From Table 4, it is obvious that four protein complexes have larger OS values and lower p values, which illustrates that the ICSC algorithm is effective, and these protein complexes are reliable and biologically meaningful.

4. Conclusion

Protein complexes are involved in multiple biological processes, and thus detection of protein complexes is essential to understanding cellular mechanisms. There are many methods to identify protein complexes but cannot reflect dynamics of protein complexes. In this study, we have presented a novel protein complex identification method ISCS according to the core-attachment structure of protein complexes. First, a weighted dynamic PPI network is constructed, which integrates the gene expression data and GO terms information. Then, we find functional cores and cluster protein

Table 4: Function enrichment analysis of five predicted protein complexes detected form Krogan dataset.

Number	Predicted protein complex	p value	Gene Ontology	OS
(1)	**YOR001W YOL021C YNL232W YHR069C YGR158C YGR095C YDR280W YDL111C YCR035C**	4.25e − 24	GO:0071042 Cluster frequency: 100%	0.75
(2)	**YNL136W** YML041C **YJL081C YHR090C YGR002C** YDR334W	5.98e − 07	GO:0006325 Cluster frequency: 100%	0.23
(3)	**YNL317W YLR277C YKL059C YKL018W YJR093C YGR156W YDR301W YDR195W** YBL010C	2.45e − 16	GO:0031124 Cluster frequency: 88.9%	0.41
(4)	**YPR034W YMR091C YMR033W YLR357W YLR321C YLR033W YKR008W YFR037C YDR303C YCR052W YCR020W-B**	1.35e − 28	GO:0031498 Cluster frequency: 100.0%	0.64
(5)	**YPR110C YOR340C YOR210W YNL248C YJR063W YJL148W YDR156W YBR154C**	4.81e − 16	GO:0098781 Cluster frequency: 100.0%	0.57



(a) Standard (b) MCODE (c) MCL

(d) CORE (e) ClusterONE (f) ICSC

Figure 5: Visualization of the 265th standard protein complex "nuclear exosome complex."

attachments based on the CS algorithm. Compared with the other competing clustering methods, ICSC can effectively identify the protein complexes and has higher precision and accuracy.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (61672334, 61502290, and 61401263) and the Industrial Research Project of Science and Technology in Shaanxi Province (2015GY016).

References

[1] E. A. Winzeler, D. D. Shoemaker, A. Astromoff et al., "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis," *Science*, vol. 285, no. 5429, pp. 901–906, 1999.

[2] A. Lakizadeh, S. Jalili, and S.-A. Marashi, "PCD-GED: Protein complex detection considering PPI dynamics based on time series gene expression data," *Journal of Theoretical Biology*, vol. 378, pp. 31–38, 2015.

[3] A. J. Link, J. Eng, D. M. Schieltz et al., "Direct analysis of protein complexes using mass spectrometry," *Nature Biotechnology*, vol. 17, no. 7, pp. 676–682, 1999.

[4] Y. Ho, A. Gruhler, A. Heilbut et al., "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[5] A. Gavin, M. Bösche, R. Krause et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[6] X. Lei, Y. Ding, H. Fujita, and A. Zhang, "Identification of dynamic protein complexes based on fruit fly optimization algorithm," *Knowledge-Based Systems*, vol. 105, pp. 270–277, 2016.

[7] M. Bertolaso, A. Giuliani, and L. De Gara, "Systems biology reveals biology of systems," *Complexity*, vol. 16, no. 6, pp. 10–16, 2011.

[8] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, no. 1, p. 2, 2003.

[9] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted PPI networks," *Bioinformatics*, vol. 25, no. 15, pp. 1891–1897, 2009.

[10] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature Methods*, vol. 9, no. 5, pp. 471-472, 2012.

[11] A.-C. Gavin, P. Aloy, P. Grandi et al., "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–636, 2006.

[12] H. C. Leung, Q. Xiang, S. M. Yiu, and F. Y. Chin, "Predicting protein complexes from PPI data: a core-attachment approach," *Journal of Computational Biology*, vol. 16, no. 2, pp. 133–144, 2009.

[13] M. Wu, X. Li, C.-K. Kwoh, and S.-K. Ng, "A core-attachment based method to detect protein complexes in PPI networks," *BMC Bioinformatics*, vol. 10, article 169, 2009.

[14] R. V. Solé, R. Ferrer-Cancho, J. M. Montoya, and S. Valverde, "Selection, tinkering, and emergence in complex networks. Crossing the land of tinkering," *Complexity*, vol. 8, no. 1, pp. 20–33, 2002.

[15] B. Chen, W. Fan, J. Liu, and F. X. Wu, "Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks," *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 177–179, 2014.

[16] J. Wang, X. Peng, M. Li, Y. Luo, and Y. Pan, "Active protein interaction network and its application on protein complex detection," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '11)*, pp. 37–42, 2011.

[17] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Li, and B. Xu, "Protein complex prediction in large ontology attributed protein-protein interaction networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 3, pp. 729–741, 2013.

[18] S. M. Van Dongen, Graph clustering by flow simulation, 2001.

[19] H. Wu, L. Gao, J. Dong, and X. Yang, "Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks," *PLoS ONE*, vol. 9, no. 3, Article ID e91856, 2014.

[20] P. Manikandan, D. Ramyachitra, and D. Banupriya, "Detection of overlapping protein complexes in gene expression, phenotype and pathways of Saccharomyces cerevisiae using Prorank based Fuzzy algorithm," *Gene*, vol. 580, no. 2, pp. 144–158, 2016.

[21] J. Zhao, X. Lei, and F. Wu, "Identifying protein complexes in dynamic protein-protein interaction networks based on Cuckoo Search algorithm," in *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1288–1295, Shenzhen, China, December 2016.

[22] X. Lei, F. Wang, F.-X. Wu, A. Zhang, and W. Pedrycz, "Protein complex identification through Markov clustering with firefly algorithm on dynamic protein-protein interaction networks," *Information Sciences*, vol. 329, pp. 303–316, 2016.

[23] J. Zeng, D. Li, Y. Wu, Q. Zou, and X. Liu, "An empirical study of features fusion techniques for protein-protein interaction prediction," *Current Bioinformatics*, vol. 11, no. 1, pp. 4–12, 2016.

[24] F. Luo, J. Liu, and J. Li, "Discovering conditional co-regulated protein complexes by integrating diverse data sources," *BMC Systems Biology*, vol. 4, no. 2, article no. 4, 2010.

[25] M. Li, X. Wu, J. Wang, and Y. Pan, "Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data," *BMC Bioinformatics*, vol. 13, no. 1, article 109, 2012.

[26] C. Tu, "A dynamical method to estimate gene regulatory networks using time-series data," *Complexity*, vol. 21, no. 2, pp. 134–144, 2015.

[27] X. Tang, J. Wang, B. Liu, M. Li, G. Chen, and Y. Pan, "A comparison of the functional modules identified from time course and static PPI network data," *BMC Bioinformatics*, vol. 12, article no. 339, 2011.

[28] J. Wang, X. Peng, M. Li, and Y. Pan, "Construction and application of dynamic protein interaction network based on time course gene expression data," *Proteomics*, vol. 13, no. 2, pp. 301–312, 2013.

[29] C. C. Friedel and R. Zimmer, "Inferring topology from clustering coefficients in protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, no. 1, article 519, 2006.

[30] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Paris, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[31] X. Shang, Y. Wang, and B. Chen, "Identifying essential proteins based on dynamic protein-protein interaction networks and RNA-Seq datasets," *Science China Information Sciences*, vol. 59, no. 7, Article ID 070106, 2016.

[32] X.-S. Yang, *Nature-inspired metaheuristic algorithms*, Luniver Press, 2010.

[33] A. M. Reynolds and C. J. Rhodes, "The Lévy flight paradigm: Random search patterns and mechanisms," *Ecology*, vol. 90, no. 4, pp. 877–887, 2009.

[34] I. Xenarios, Ł. Salwínski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.

[35] N. J. Krogan, G. Cagney, H. Yu et al., "Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[36] U. Güldener, M. Münsterkötter, M. Oesterheld et al., "MPact: the MIPS protein interaction resource on yeast.," *Nucleic acids research.*, vol. 34, pp. D436–441, 2006.

[37] B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight, "Cell biology: Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes," *Science*, vol. 310, no. 5751, pp. 1152–1158, 2005.

[38] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak, "Up-to-date catalogues of yeast protein complexes," *Nucleic Acids Research*, vol. 37, no. 3, pp. 825–831, 2009.

[39] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[40] X. Lei, Y. Ding, and F.-X. Wu, "Detecting protein complexes from DPINs by density based clustering with Pigeon-Inspired Optimization Algorithm," *Science China Information Sciences*, vol. 59, no. 7, Article ID 070103, 2016.

[41] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," *BMC Bioinformatics*, vol. 7, article 207, 2006.

Hindawi
Complexity
Volume 2017, Article ID 4826206, 10 pages
https://doi.org/10.1155/2017/4826206

Research Article

DriverFinder: A Gene Length-Based Network Method to Identify Cancer Driver Genes

Pi-Jing Wei,[1] Di Zhang,[1] Hai-Tao Li,[2] Junfeng Xia,[3] and Chun-Hou Zheng[1]

[1]*College of Computer Science and Technology, Anhui University, Hefei, Anhui 230601, China*
[2]*State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, Jiangsu 210018, China*
[3]*Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China*

Correspondence should be addressed to Chun-Hou Zheng; zhengch99@126.com

Received 7 March 2017; Revised 8 June 2017; Accepted 3 July 2017; Published 10 August 2017

Academic Editor: Haiying Wang

Integration of multi-omics data of cancer can help people to explore cancers comprehensively. However, with a large volume of different omics and functional data being generated, there is a major challenge to distinguish functional driver genes from a sea of inconsequential passenger genes that accrue stochastically but do not contribute to cancer development. In this paper, we present a gene length-based network method, named DriverFinder, to identify driver genes by integrating somatic mutations, copy number variations, gene-gene interaction network, tumor expression, and normal expression data. To illustrate the performance of DriverFinder, it is applied to four cancer types from The Cancer Genome Atlas including breast cancer, head and neck squamous cell carcinoma, thyroid carcinoma, and kidney renal clear cell carcinoma. Compared with some conventional methods, the results demonstrate that the proposed method is effective. Moreover, it can decrease the influence of gene length in identifying driver genes and identify some rare mutated driver genes.

1. Background

At present, understanding the mechanisms of cancer development and uncovering actionable target genes for cancer treatment are still difficult challenges. With rapid advances in high-throughput sequencing technologies, some large-scale cancer genomics projects, such as The Cancer Genome Atlas (TCGA) [1] and International Cancer Genome Consortium (ICGC) [2], have produced different omics data including a rich dataset of whole-exome and RNA sequence data [3, 4], which provides chances to allow us to accurately infer tumor-specific alterations [5] and help in precision medicine in cancers treatment [6, 7]. However, many of genetic changes represent neutral variations that do not contribute to cancer development which are called passenger mutations [6, 8]. Only a few alterations are causally implicated in the process of oncogenesis and provide a selection growth advantage which are referred to as driver mutations [8, 9]. Hence, it is a major challenge to distinguish pathogenic driver mutations from the so-called random mutated passenger mutations [10].

Previously, there were multiple computational methods to identify driver genes based on gene mutational frequency (termed as frequency-based method) in a large cohort of cancer patients [11–13]. However, the infrequently mutated drivers are inclined to be ignored by frequency-based methods. Also mutational heterogeneity in cancer genomes is an important factor affecting the performance of frequency-based methods [14]. In addition, further studies have realized that driver mutations or genes disrupt some cellular signaling or regulatory pathways which promote the progression of cancer [15, 16]. In fact, genes affect various biological processes by related complex networks instead of acting in isolation in cancer [17]. In addition, the cancer is a result of interplay of various types of genetic changes which form complex and dynamic networks [18]. Thus, many network-based and pathway-based approaches have been proposed to prioritize driver mutations and genes. For instance, Dendrix was a pathway-based algorithm for discovery of mutated driver pathways in cancer using somatic mutation data [19]. After that, Multi-Dendrix algorithm was proposed to extend

Dendrix method in order to guarantee yielding the optimal set of pathways [20]. MDPFinder was also a pathway-based method to solve the so-called maximum weight submatrix problem proposed in Dendrix method [19] which was aimed at identifying mutated driver pathways from mutation data in cancer [21]. And Zhang et al. proposed CoMDP method which focused on cooccurring driver pathways rather than single pathway [22]. In addition, iMCMC was a network-based method by integrating somatic mutation, CNVs, and gene expressions without any prior information [6]. Another method, DawnRank, was also a network-based algorithm to discover personalized causal driver mutations by ranking mutated genes according to their potential to be drivers based on PageRank algorithm [23]. Bashashati et al. developed a method called DriverNet which comprehensively analyzed genomes and transcriptomes datasets to identify likely driver genes in population-level by virtue of their effect on mRNA expression networks and also reveal the infrequent but important genes and patterns of pathway [10]. VarWalker was a personalized network-assisted approach to prioritize well-known, infrequently mutated genes and interpret mutation data in NGS studies [24].

Although some proposed methods can determine potential drivers, most of them do not consider the influence of gene length to the results; in other words, they may identify some likely false positive driver genes according to known driver genes datasets. And it has been indicated that driver genes are related to not only mutation frequency, but also mutation context or gene length [25] and variants tend to arise more frequently in long genes [26]. For example, *TTN*, the longest gene in human genome, accumulates many variants just due to its length [24, 26]. *TTN* may be selected in many computational methods; however, it usually serves as passenger gene [27]. This phenomenon indicates that many current methods have a strong preference towards identifying long genes [24]. So it is essential to filter those frequently mutated genes due to long length. VarWalker takes into account the gene length; however, it does not consider the influence of mutation to expression. In addition, some genomic variations in a gene may lead to extreme changes in some outlying genes expression level which are associated with the mutated gene through gene-gene interaction network or pathways and these outlying genes are often called outliers [10]. And, it has been proved that cancer-associated genes are more effectively detected by interindividual variation analysis rather than only calculating differences in the mean expression across different samples [28]. That is, the outliers are related to not only tumor expression distribution but also the corresponding normal expression distribution. Moreover, various cellular processes are often affected by genes in complex networks rather than genes acting in isolation [17] and cancer is also related to a set of genes interacting together in a molecular network [29]. So networks usually provide a convenient way to explore the context within which single gene operates [30]. It should be noted that prior knowledge such as protein-protein interaction (PPI) network can provide some useful information; however, prior knowledge is limited and may lead to discarding some important information in some instances [22]. In our previous work [31], we only consider the prior information of gene-gene interaction network. So it is essential to extend gene-gene interaction network.

In this study, we proposed an integrated framework named DriverFinder to identify driver genes by integrating somatic mutation data, copy number variations (CNVs), tumor and normal expression data, and gene-gene interaction network. Firstly, the gene length is taken into consideration to filter some frequently genes because of long length. Moreover, in this method, we integrated tumor expression and normal expression to construct outlier matrix rather than only using tumor expression. Furthermore, to increase accuracy of identifying drivers, we calculated Pearson correlation coefficients (PCC) of genes and combined them with PPI network to construct a new dynamic interaction network for each cancer type. In order to estimate the performance of DriverFinder method, we applied it to four different large-scale TCGA datasets, including breast cancer (BRCA), head and neck squamous cell carcinoma (HNSC), thyroid carcinoma (THCA), and kidney renal clear cell carcinoma (KIRC), and compared it with MUFFINN [32], DriverNet [10], and frequency-based method. The results demonstrated that DriverFinder can identify drivers effectively and decrease false positive, that is, filtering some long and frequently mutated but functionally neutral genes.

2. Materials and Methods

We proposed DriverFinder to identify cancer driver genes by integrating multiomics data (Figure 1). The detailed description of it is shown in Figure 1.

As is shown in Figure 1, the first step is to estimate the occurrence of mutation events in the genome by fitting them into a generalized additive model [24]. Then a weighted resample-based test is used to filter long passenger genes according to approximated probabilities based on benchmark of coding gene length. Secondly, for gene expression, we compare expression of tumors with normal samples to determine the outlying genes. Then a gene-gene interaction network combined with prior knowledge and PCC among expression data is used to relate mutations to their consequent effect on gene expression. The associations between mutated and outlying genes are formulated using a bipartite graph where the left nodes indicate the genes mutation status and the right nodes indicate the outlying expression status in each patient. For each patient, there is an edge between g_i and g_j if the left partition gene g_i is mutated and the right gene g_j is an outlier in some samples and they also have high correlations in gene-gene interaction network. Secondly, greedy algorithm is used to prioritize mutated genes based on the coverage. In each iteration of the greedy algorithm, the mutated gene on the left partition of the bipartite graph which relates to the most outlying expression genes is chosen. Until all the outlying expression genes are covered by the least mutated genes on the left, iteration is stopped. Then the mutated genes are ranked based on their coverage. So genes with the most outlying expression are appointed as candidate driver genes. Finally, the statistical significance test based on null distribution is applied to these putative driver genes.

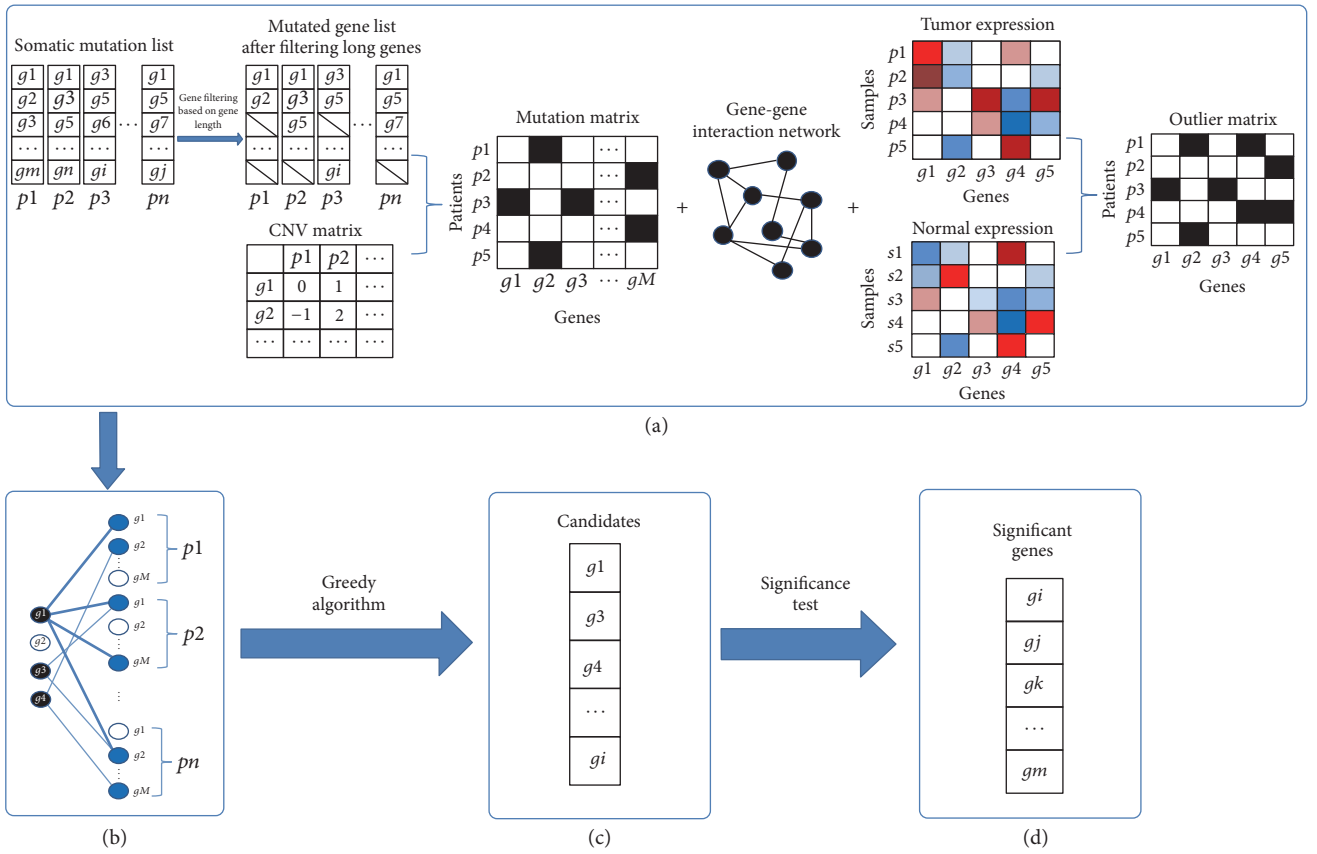


FIGURE 1: *The flowchart of DriverFinder method*. (a) Input datasets consist of somatic mutation, CNV, normal and tumor expression data, and gene-gene interaction network. A generalized additive model is performed on somatic mutation data to filter mutated genes which occurred at random due to long length. After that, the residual significant genes are combined with CNV to construct mutation data. The gene-gene interaction network is constructed by integrating prior gene-gene interaction network and Pearson correlated coefficient network. And the outlying matrix is constructed by analyzing interindividual variation in tumor and normal expression. (b) Given mutation data, outlying data, and gene-gene interaction network, the bipartite graph is obtained. The black nodes on the left indicate mutated genes and the blue nodes on the right represent outlying expression genes. (c) Candidate genes are obtained by greedy algorithm. The more outlying expression events the gene overlaps, the higher the gene ranks are. (d) Statistical test is performed on candidate genes to select important putative drivers by p value < 0.05.

2.1. Construction of Mutation Matrix M. In terms of somatic mutation, we downloaded it from TCGA data portal (https://cancergenome.nih.gov/) and only considered the data of level 2. As a matter of fact, Jia et al. explored long genes in two datasets and examined gene length effects by plotting the proportion of mutated genes versus their complementary DNA (cDNA) length [24]. They discovered that two sets of mutated genes were positively correlated with cDNA length, and longer genes were more likely to be mutated genes [24]. Hence, frequency-based methods may be inclined to select long genes as drivers. So, it is necessary to perform gene length-based filtration. In this study, in order to accurately estimate the mutation rate for each gene, generalized additive model was adopted to compute a probability weight vector (PWV) for the mutation genes of each sample [24].

Here, only somatic mutated genes mapped onto the benchmark of consensus coding sequences (CCDS) dataset [33] which contains a core set of consistently annotated and high quality human and mouse protein coding regions are reserved. And those mapped genes in this study have been allocated cDNA length based on their coding sequences [24]. Assuming the vector X as the cDNA gene length and the following model is used to assess the probability of a gene to be mutated,

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \sim S(X), \tag{1}$$

where $S(\cdot)$ represents an unspecified smooth function and π = num of mutant genes/total num of CCDS genes represents the proportion of mutant genes in the specific samples [24]. Each gene then would be assigned a PWV value. Afterwards, a resampling test based on the probability of each gene is performed 1000 times in each sample and the null distribution is that in genes mutations occur at random. Then we define mutation frequency as

$$f = \frac{\text{num of selecting the genes in 1000 times}}{1000}, \tag{2}$$

where f represents mutation frequency. Next we filter out genes with frequencies $\geq 5\%$ in random datasets unless

they are Cancer Gene Census (CGC) genes. Then a list of significant mutant genes S is obtained.

As for CNVs, which have been processed by GISTIC 2.0, they are acquired from http://gdac.broadinstitute.org/runs/(v2014_10_17). There are five types of copy number including amplification, gain, diploid, heterozygous deletion, and homozygous deletion in this dataset. Here, we only screen out amplifications and homozygous deletions to construct CNV matrix C. Finally, the significant mutated gene list S and CNV matrix C are combined to generate the patient-mutation binary matrix M, in which $M_{ij} = 1$ indicates there is genetic alteration, that is, mutation, amplification, or homozygous deletion, in the jth gene of the ith sample. Otherwise, $M_{ij} = 0$.

2.2. Construction of Expression Outlier Matrix E. Gene expression data (level 3) including tumor and normal expression data is also downloaded from TCGA data portal. Moreover, some studies have shown that assessment of interindividual variation of gene expression performs well in predicting cancer-associated genes [28]. So in this study, the outlying matrix is determined based on analysis of interindividual variation in tumor and normal expression rather than differences in mean expression levels or only tumor expression distribution [28]. For each type of cancers, there are two expression datasets $T(i, j)$ and $N(i, j)$ which indicate the real-valued expression measure of gene i in sample j of tumor and normal datasets, respectively. For each gene, the outliers in this study are defined as tumors whose expression levels are outside the four-standard deviation range of the expression values of the gene across all the normal samples [28]. It is formularized as

$$\begin{aligned} &\text{tumor expression} < m(N) - 4 * \text{sd}(N) \\ &\text{or tumor expression} > m(N) + 4 * \text{sd}(N) \end{aligned} \tag{3}$$

in which $m(N)$ is the mean expression and $\text{sd}(N)$ indicates the standard deviation of gene expression in normal samples. Then the binary patient-outlier matrix E is constructed and the value of $E(i, j)$ indicated whether gene i in patient j is an outlier among the population-level distribution for that gene. If the expression of gene i is an outlier in patient j, $E(i, j) = 1$; otherwise, $E(i, j) = 0$.

2.3. Gene-Gene Interaction Network. It is noteworthy that most prior knowledge such as PPI network or pathways is incomplete and a great deal of knowledge about biological pathways remains unclear [22]. In our previous study [31], we relied on prior knowledge about gene influence graph integrated from known gene networks [10] which often leaved out some likely important nodes. So in this study, we constructed a new dynamic gene-gene interaction network by incorporating gene-gene correlation coefficients with prior knowledge. That is, firstly, PCCs between pairwise genes are obtained by normalized tumor expression. Acceptable correlations with PCC > 0.75 are often considered high correlated and selected [34]. Here, we choose 0.8 as threshold to ensure selected pairwise genes being higher correlated and increase reliability. The edges with PCC > 0.8 are selected and set to 1, otherwise 0. In order to retrieve some important prior knowledge simultaneously, the known gene network (termed the influence graph in DriverNet [10]) is mapped onto the binary matrix obtained from correlation coefficients matrix. So a new and dynamic gene-gene interaction network (termed as G after) included prior knowledge and deduced knowledge is established. When there is a correlation, that is, PCC > 0.8 or 1 in influence graph between gene i and gene j, $G_{ij} = 1$; otherwise, $G_{ij} = 0$.

2.4. Significance Estimation. With the aim of testing the statistical significance of the driver candidates, we apply a randomization framework. The algorithm is run on the random N permuted original datasets (mutation data and outlier data). Then we assess the significance by seeing if the results on real data are significantly different from the results on random datasets and obtain the p value of each candidate drivers. The statistical significance of g is defined as follows [10]:

$$p \text{ value}(g) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(\text{COV}_{gij} > \text{COV}_g\right)}{\sum_{i=1}^{N} M_i}, \tag{4}$$

where N is permutation times and M_i is the number of candidate drivers in the ith run of the approach. COV_g is the coverage of g calculated from our method. Here we choose $N = 50$. The statistical significance of g means that the times of the observed driver genes with coverage are more than COV_g. Finally, genes with p value less than 0.05 are nominated as candidate drivers.

3. Results

3.1. Datasets. In this work, four TCGA datasets, BRCA, HNSC, KIRC, and THCA, were applied to our method. For each cancer type, four different omics data consisting of somatic mutation, tumor expression, normal expression, and CNV were used. The BRCA dataset includes copy number, 531 tumor samples' and 62 normal samples' expression data, and 962 samples' somatic mutation data. KIRC dataset contains copy number variations, 417 samples' somatic mutations data, and accompanying 534 tumor and 72 normal samples' expression data. The HNSC dataset contains 509 patients' somatic mutation data, 522 tumor and 44 normal samples' expression data, and copy number data. For THCA, it includes copy number variations, 435 patients' somatic mutation, and 513 tumor patients' and 59 normal samples' expression data. For each type of cancer, we only consider the samples which are common in tumor expression dataset and somatic mutation dataset.

3.2. Performance Evaluation. To evaluate the performance of our method on identifying known driver genes, we used annotated cancer-related genes datasets CGC database (15/7/2015) [35] and 20/20 rule [25] as approximate benchmarks. CGC is a database which catalogues 571 genes whose mutations have been causally involved in cancer [35]. 20/20

rule contains 138 driver genes in which 125 genes are affected by subtle mutations and 13 are affected by amplification or homozygous deletion [25]. We compared our method with frequency-based method, DriverNet [10], and MUFFINN [32] based on these two benchmarks.

In this work, we first counted the number of known drivers according to CGC genes of these four methods. For comparison, three measures based on top N genes including precision, recall, and $F1$ score are used which are defined as follows:

$$\begin{aligned}
\text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{(\#\text{ genes found in DriverFinder}) \cap (\#\text{ mutated genes in CGC})}{(\#\text{ genes found in DriverFinder})}, \\
\text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{(\#\text{ genes found in DriverFinder}) \cap (\#\text{ mutated genes in CGC})}{(\#\text{ genes in CGC})}, \\
F1\text{ score} &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}},
\end{aligned} \tag{5}$$

where TP indicates the number of overlapping genes found in our method and annotated genes associated with cancers in CGC. FP means the number of genes identified in our method, however not cataloged by CGC. FN is the number of genes in CGC but not contained in our method.

In general, DriverFinder almost outperforms other three methods in the top ranking genes of all the four cancer datasets (Figure 2; results of DriverFinder are shown in Supplementary File 1, in Supplementary Material available online at https://doi.org/10.1155/2017/4826206). Although, after approximately ranking top 30 genes in KIRC, MUFFINN performs comparably with DriverFinder, it has poorer performance in the top 30 genes. And the same phenomenon arises after top 60 genes in THCA. Analogous to it, the cumulative number of retrieved cancer genes annotated by 20/20 rule in KIRC by MUFFINN (Figure 3(c)) is also more than DriverFinder. A potential explanation of them may be that the total numbers of mutations in KIRC (10359 genes with 21089 mutations) and THCA (8899 genes with 16497 mutations) are remarkably less than in BRCA (16717 genes with 118098 mutations) and HNSC (14830 genes with 57164 mutations). So the gene-gene interaction network in KIRC and THCA may be simpler than in BRCA and HNSC; that is, genes may easily correlate directly with each other. And MUFFINN consider mutations only in direct neighbors [32]. On the one hand, this difference may help MUFFINN retrieve more genes; on the other hand, the number of mutations indicates that there may be more passengers (i.e., noise) in BRCA and HNSC and DriverFinder is more stable with noises.

Moreover, DriverFinder outperforms other three methods in BRCA, HNSC, and THCA by the cumulative numbers with 20/20 rule (Figure 3). In KIRC, it also has a better performance than DriverNet and frequency-based method within the top 100 genes.

3.3. DriverFinder Decreases the Effect of Gene Length. It is worth noting that, from the results of the DriverFinder, we can find that it can decrease the false positive because it has the good performance on filtering randomly mutated genes due to long length. The longest gene across human genome is *TTN* and it has been proven that higher mutation rate in it is likely to be artifacts [23, 36]. For example, in BRCA, *TTN* ranked 4 and 6 in frequency-based method and in MUFFINN, respectively, due to high mutation rate. Also it ranked 51 (p value = 0.031) as a candidate driver in DriverNet algorithm; however, it was filtered out with DriverFinder. In KIRC and HNSC, it ranked 4 and 3, respectively, based on mutation frequency and 22 and 18 in DriverNet, respectively. In addition, it is also at top 4 and 3 with MUFFINN in KIRC and HNSC, respectively, but it does not rank among the top 160 or 790 separately according to DriverFinder. Furthermore, for THCA, frequency-based method and MUFFINN ranked *TTN* as top 4 and 5, respectively. However, it is not identified by DriverFinder. These results proved the better performance of DriverFinder on filtering randomly mutated genes with long length than DriverNet, MUFFINN, and frequency-based method.

3.4. Pathway Enrichment Analysis. In order to investigate cancer-related pathways among the significant candidate drivers, Kyoto Encyclopedia of Genes and Genomes (KEGG) [37] pathway enrichment analysis was performed by statistics significantly candidate driver genes with p values less than 0.05 (see Supplementary File 2). The top 20 significant pathways are shown in Figure 4. We observed that the most enriched terms are cancer-related pathways in four cancer datasets. Moreover, ErbB signaling pathway, which is significantly enriched in BRCA (p value = $4.79E-07$) and KIRC (p value = $6.23E-07$), has been reported to play important roles in many tumors and *ErbB2/ErbB3* heterodimer functioned as an oncogenic unit in breast cancer [38]. Also, the significantly enriched VEGF signaling pathway ($4.94E-05$) in HNSC plays a pivotal role in tumor angiogenesis [39].

3.5. Discovering Rare Driver Genes. In this subsection, we exhibited that DriverFinder can identify rarely mutated but significant candidate driver genes which are defined as genes whose mutation frequency < 2% across all patients cohort. Here, we only selected highly ranked (top 30) rare genes for further analysis.

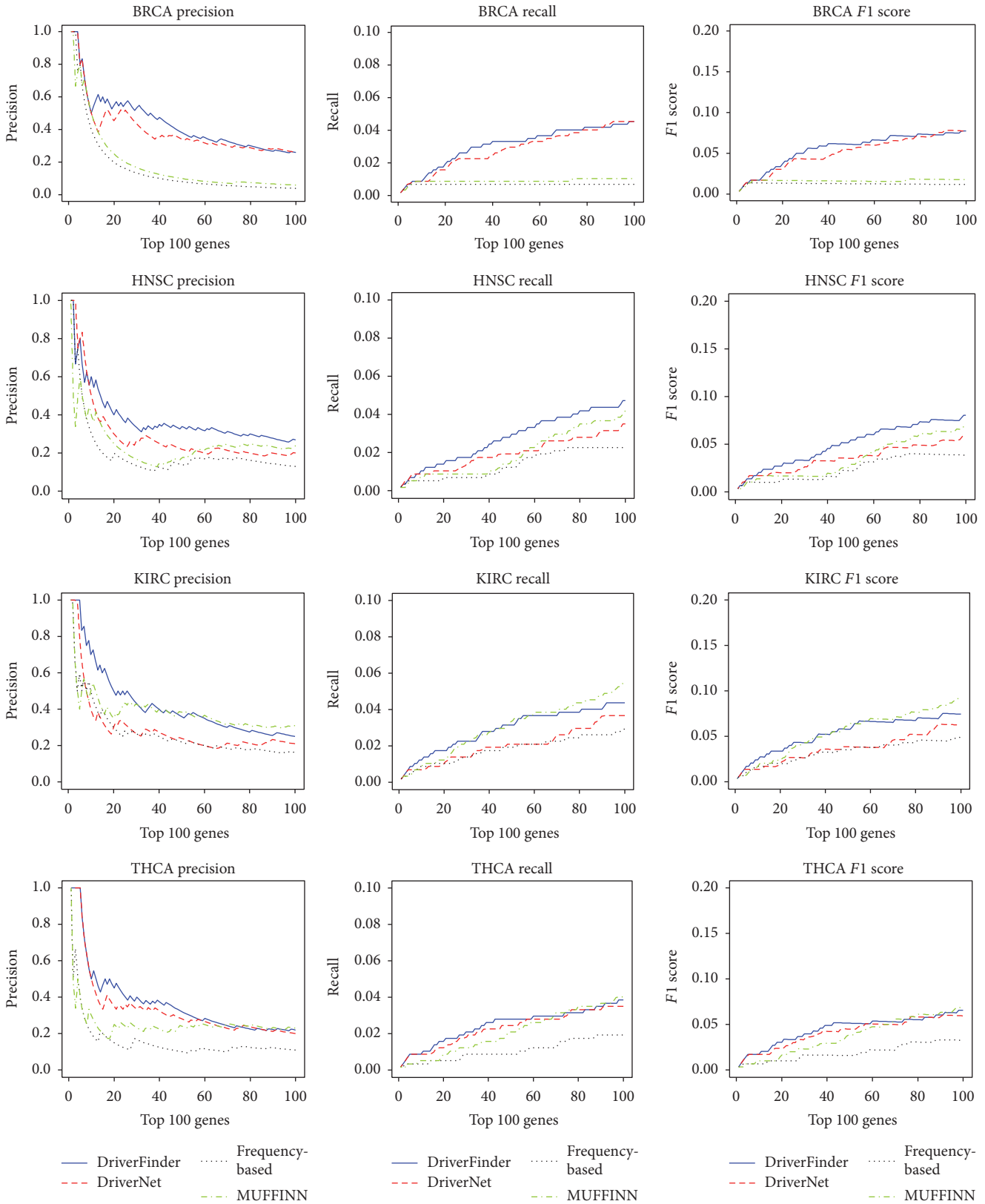


Figure 2: Performance comparison with CGC (precision, recall, and $F1$ score) of DriverFinder, DriverNet, MUFFINN, and frequency-based method on BRCA, HNSC, KIRC, and THCA datasets.

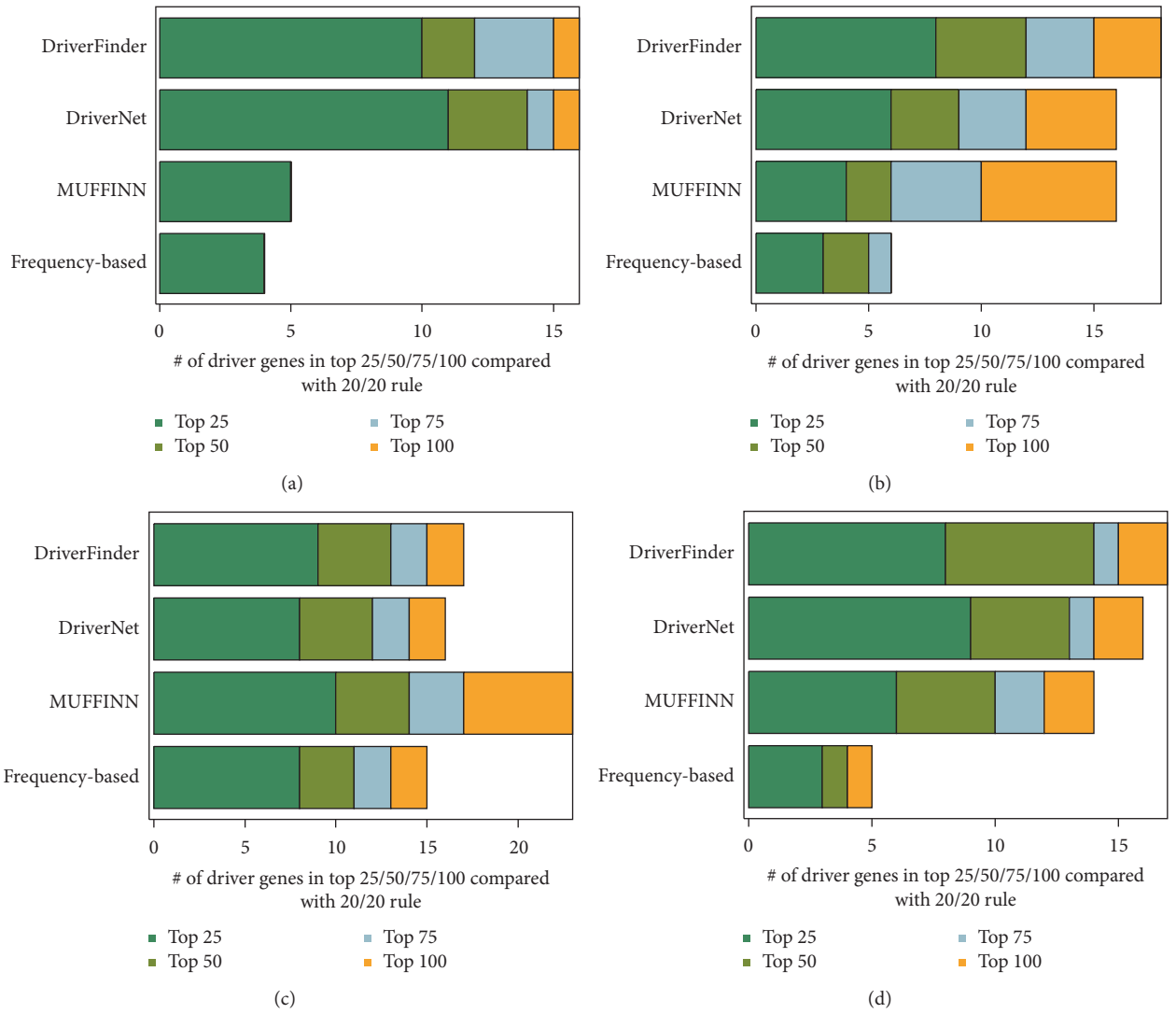


Figure 3: Cumulative numbers of retrieved cancer genes annotated by 20/20 rule within top 25, 50, 75, and 100 of (a) BRCA, (b) HNSC, (c) KIRC, and (d) THCA using four different methods.

In BRCA, 3 rare genes (*PIK3R1*, *CREBBP*, and *PRKACB*) are ranked in top 30. In them, for *PIK3R1* (ranked 23) which is not ranked top 30 in the other three methods, underexpression might possibly lead to PI3K pathway activation and confer tumor development and progression in humans and it is a clinically useful independent prognostic marker in breast cancer [40]. Due to its low frequency mutations, any further statistical analyses concerning a possible association between *PIK3R1* mutations and clinical parameters are not allowed [40] and it is easily ignored by frequency-based methods. Moreover, for *CREBBP* (ranked 25) which also was not ranked top 30 in the other three methods, it has been occasionally reported in breast cancer [41]. *PRKACB* downregulates in non-small cell lung cancer and the effect of its upregulation on cell proliferation, apoptosis, and invasion also has been investigated [42].

In HNSC, also 3 rare genes are selected and one of them (*UGT2B4*, ranked 30) has shown potential to be a novel driver. *UGT2B4* genotypes associated with decreased enzyme activities are found to increase the risk of esophageal squamous cell carcinoma [43].

For KIRC and THCA, there are some important rare genes not identified by other methods. For example, *CLTC* in KIRC, which is contained in CGC and ranked 11 with low mutated frequency 1.68% in DriverFinder, encodes a major subunit of clathrin and is a fusion partner of *TFE3*. And CLTC-TFE3 is the fifth gene fusion involving *TFE3* in pediatric renal cell carcinomas [44]. Another example is *AKT1* (0.69% of cases) in THCA, which is identified by DriverFinder (ranked 30) and contained in CGC; it is a serine/threonine protein kinase and its downstream proteins have been reported to be frequently activated in human cancers [45].

4. Discussion

Cancer is a complex disease and difficult to treat, and distinguishing driver genes from a mass of neutral passenger genes

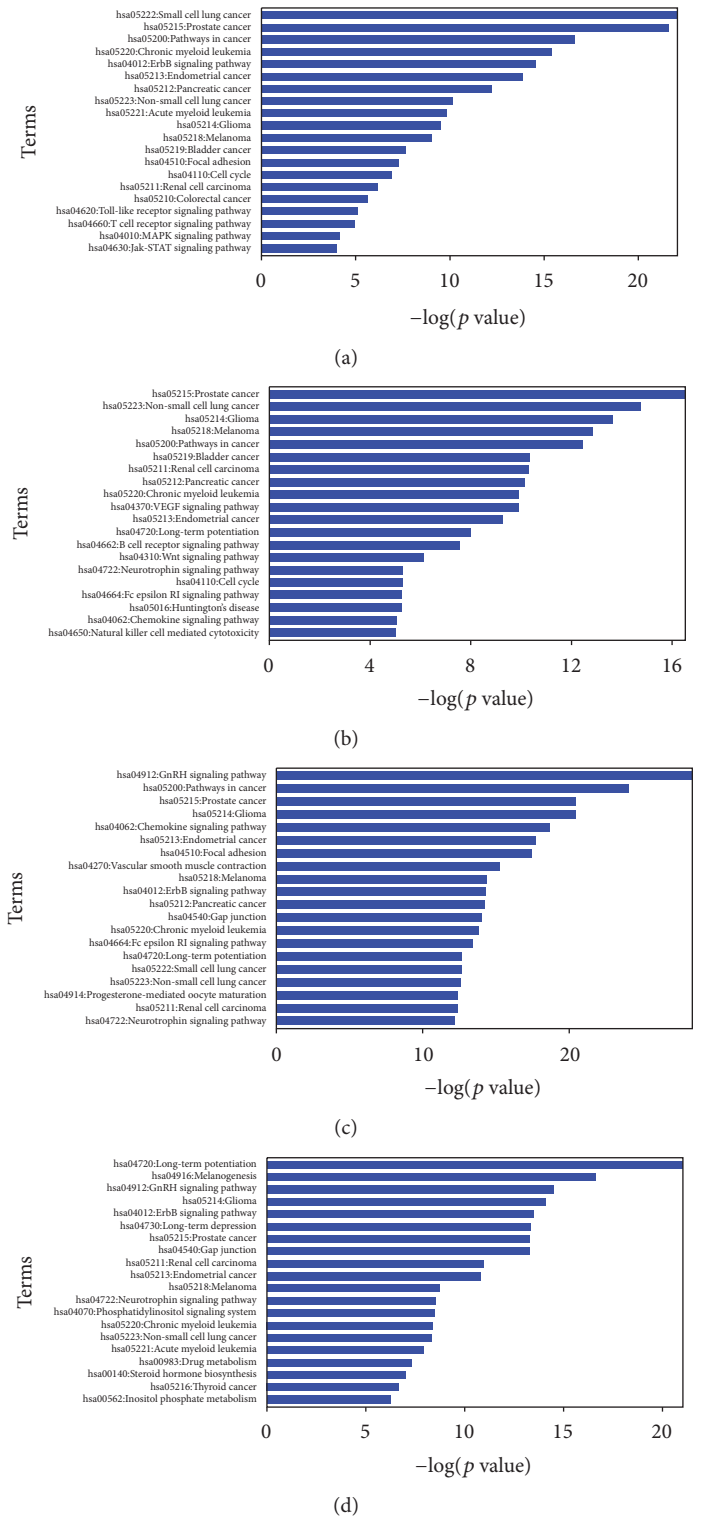


Figure 4: The KEGG pathway enrichment of (a) BRCA, (b) HNSC, (c) KIRC, and (d) THCA by significant genes with p values < 0.05 in DriverFinder.

is extremely important to understand the mechanism of the cancer and design targeted treatments. In this study, we introduced a comprehensive framework DriverFinder to identify driver genes by incorporating genomes, transcriptomes, and gene-gene interaction information. We implemented gene-based filtering model to exclude genes that were mutated largely due to random events. The method was applied to four independent cancer datasets from TCGA and the

results demonstrated that the power of it across multiple tumor types was mainly better than DriverNet, MUFFINN, and frequency-based methods. In summary, this method has advantages in both filtering random mutated genes and identifying driver genes regardless of their mutation frequencies. We expect that it can also be applied to other complex cancer types.

However, in this work, we only explained the changes of the expression by somatic mutations, although other molecular or genetic changes such as transcription factors, methylations, and microRNAs also affect expression of other genes and play important roles in the development of cancer [46]. Therefore, it is necessary to extend the method so that driver genes can be determined by not only somatic alterations but also other different types of molecular changes. Also, we can extend our aim of identifying drivers by some methods such as machine learning methods [47–51].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (no. 61672037), the Key Project of Anhui Provincial Education Department (no. KJ2017ZD01), and the Anhui Provincial Natural Science Foundation (nos. 1508085QF135 and 1608085MF136).

References

[1] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, and G. M. Mastrogianakis, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

[2] T. J. Hudson, W. Anderson, A. Aretz, A. D. Barker, C. Bell, and R. R. Bernabé, "International network of cancer genome projects," *Nature*, vol. 464, pp. 993–998, 2010.

[3] J. X. Liu, Y. Xu, Y. L. Gao, C. H. Zheng, D. Wang, and Q. Zhu, "A class-information-based sparse component analysis method to identify differentially expressed genes on RNA-seq data," *IEEE/ACM Transactions on Computational Biology Bioinformatics*, vol. 13, no. 2, pp. 392–398, 2016.

[4] Y. X. Wang, J. X. Liu, Y. L. Gao, C. H. Zheng, and J. L. Shang, "Differentially expressed genes selection via Laplacian regularized low-rank representation method," *Computational Biology Chemistry*, 2016.

[5] C. Suo, O. Hrydziuszko, D. Lee et al., "Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival," *Bioinformatics*, vol. 31, no. 16, pp. 2607–2613, 2015.

[6] J. Zhang, S. Zhang, Y. Wang, and X.-S. Zhang, "Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data," *BMC systems biology*, vol. 7, p. S4, 2013.

[7] J. Zhang and S. Zhang, "Discovery of cancer common and specific driver gene sets," *Nucleic Acids Research*, vol. 45, no. 10, article e86, 2017.

[8] J. Foo, L. L. Liu, K. Leder et al., "An evolutionary approach for identifying driver mutations in colorectal cancer," *PLoS Computational Biology*, vol. 11, no. 9, Article ID e1004350, 2015.

[9] R. Tian, M. K. Basu, and E. Capriotti, "Computational methods and resources for the interpretation of genomic variants in cancer," *BMC Genomics*, vol. 16, supplement 8, article S7, 2015.

[10] A. Bashashati, G. Haffari, J. Ding et al., "DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer," *Genome Biology*, vol. 13, article R124, 2012.

[11] Y. Ping, Y. Deng, L. Wang et al., "Identifying core gene modules in glioblastoma based on multilayer factor-mediated dysfunctional regulatory networks through integrating multi-dimensional genomic data," *Nucleic Acids Research*, vol. 43, no. 4, pp. 1997–2007, 2015.

[12] A. Youn and R. Simon, "Identifying cancer driver genes in tumor genome sequencing studies," *Bioinformatics*, vol. 27, no. 2, Article ID btq630, pp. 175–181, 2011.

[13] L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, and K. Cibulskis, "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, pp. 1069–1075, 2008.

[14] J. Zhang and S. Zhang, "The discovery of mutated driver pathways in cancer: models and algorithms," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. PP, no. 99, 2016.

[15] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature Medicine*, vol. 10, no. 8, pp. 789–799, 2004.

[16] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[17] M. Grechkin, B. A. Logsdon, A. J. Gentles, and S.-I. Lee, "Identifying network perturbation in cancer," *PLoS Computational Biology*, vol. 12, no. 5, Article ID e1004888, 2016.

[18] F. Cheng, J. Zhao, and Z. Zhao, "Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes," *Briefings in Bioinformatics*, vol. 17, no. 4, Article ID bbv068, pp. 642–656, 2016.

[19] F. Vandin, E. Upfal, and B. J. Raphael, "De novo discovery of mutated driver pathways in cancer," *Genome Research*, vol. 22, no. 2, pp. 375–385, 2012.

[20] M. D. M. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael, "Simultaneous identification of multiple driver pathways in cancer," *PLoS Computational Biology*, vol. 9, no. 5, Article ID e1003054, 2013.

[21] J. Zhao, S. Zhang, L.-Y. Wu, and X.-S. Zhang, "Efficient methods for identifying mutated driver pathways in cancer," *Bioinformatics*, vol. 28, no. 22, pp. 2940–2947, 2012.

[22] J. Zhang, L.-Y. Wu, X.-S. Zhang, and S. Zhang, "Discovery of co-occurring driver pathways in cancer," *BMC Bioinformatics*, vol. 15, article 271, 2014.

[23] J. P. Hou and J. Ma, "DawnRank: discovering personalized driver genes in cancer," *Genome Medicine*, vol. 6, article 56, 2014.

[24] P. Jia and Z. Zhao, "VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data," *PLoS Computational Biology*, vol. 10, no. 2, Article ID e1003460, 2014.

[25] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz Jr., and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 340, no. 6127, pp. 1546–1558, 2013.

[26] C. Shyr, M. Tarailo-Graovac, M. Gottlieb, J. J. Y. Lee, C. Van Karnebeek, and W. W. Wasserman, "FLAGS, frequently mutated genes in public exomes," *BMC Medical Genomics*, vol. 7, article 64, 2014.

[27] J. J. Waterfall, E. Arons, R. L. Walker et al., "High prevalence of MAP2K1 mutations in variant and IGHV4-34-expressing hairy-cell leukemias," *Nature Genetics*, vol. 46, pp. 8–10, 2014.

[28] I. P. Gorlov, J.-Y. Yang, J. Byun et al., "How to get the most from microarray data: advice from reverse genomics," *BMC Genomics*, vol. 15, article 223, 2014.

[29] C. Ma, Y. Chen, D. Wilkins, X. Chen, and J. Zhang, "An unsupervised learning approach to find ovarian cancer genes through integration of biological data," *BMC Genomics*, vol. 16, supplement 9, article S3, 2015.

[30] E. E. Schadt, "Molecular networks as sensors and drivers of common human diseases," *Nature*, vol. 461, no. 7261, pp. 218–223, 2009.

[31] P.-J. Wei, D. Zhang, C.-H. Zheng, and J. Xia, "Cancer genes discovery based on integtating transcriptomic data and the impact of gene length," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '16)*, pp. 1265–1268, Shenzhen, China, December 2016.

[32] A. Cho, J. E. Shim, E. Kim, F. Supek, B. Lehner, and I. Lee, "MUFFINN: cancer gene discovery via network analysis of somatic mutation data," *Genome Biology*, vol. 17, article 129, 2016.

[33] K. D. Pruitt, J. Harrow, R. A. Harte et al., "The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes," *Genome Research*, vol. 19, no. 7, pp. 1316–1323, 2009.

[34] T. Bammler, R. P. Beyer, and S. Bhattacharya, "Standardizing global gene expression analysis between laboratories and across platforms," *Nature Methods*, vol. 2, no. 5, pp. 351–356, 2005.

[35] P. A. Futreal, L. Coin, M. Marshall et al., "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 177–183, 2004.

[36] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nature Protocols*, vol. 4, no. 7, pp. 1073–1082, 2009.

[37] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[38] N. E. Hynes and G. MacDonald, "ErbB receptors and signaling pathways in cancer," *Current Opinion in Cell Biology*, vol. 21, no. 2, pp. 177–184, 2009.

[39] J. Ma, H. Sawai, N. Ochi et al., "PTEN regulate angiogenesis through PI3K/Akt/VEGF signaling pathway in human pancreatic cancer cells," *Molecular and Cellular Biochemistry*, vol. 331, no. 1-2, pp. 161–171, 2009.

[40] M. Cizkova, S. Vacher, D. Meseure et al., "PIK3R1 underexpression is an independent prognostic marker in breast cancer," *BMC Cancer*, vol. 13, article 545, 2013.

[41] M. P. H. M. Jansen, J. W. M. Martens, J. C. A. Helmijr et al., "Cell-free DNA mutations as biomarkers in breast cancer patients receiving tamoxifen," *Oncotarget*, vol. 7, no. 28, pp. 43412–43418, 2016.

[42] Y. Chen, Y. Gao, Y. Tian, and D.-L. Tian, "PRKACB is Downregulated in non-small cell lung cancer and exogenous PRKACB inhibits proliferation and invasion of LTEP-A2 cells," *Oncology Letters*, vol. 5, no. 6, pp. 1803–1808, 2013.

[43] P. Dura, J. Salomon, R. H. M. Te Morsche et al., "High enzyme activity UGT1A1 or low activity UGT1A8 and UGT2B4 genotypes increase esophageal cancer risk," *International Journal of Oncology*, vol. 40, no. 6, pp. 1789–1796, 2012.

[44] P. Argani, M. Y. Lui, J. Couturier, R. Bouvier, J.-C. Fournet, and M. Ladanyi, "A novel CLTC-TFE3 gene fusion in pediatric renal adenocarcinoma with t(X;17) (p11.2;q23)," *Oncogene*, vol. 22, no. 34, pp. 5374–5378, 2003.

[45] D. Lee, I. G. Do, K. Choi, O. S. Chang, K. T. Jang, and D. Choi, "The expression of phospho-AKT1 and phospho-MTOR is associated with a favorable prognosis independent of PTEN expression in intrahepatic cholangiocarcinomas," *Modern Pathology An Official Journal of the United States & Canadian Academy of Pathology Inc*, vol. 25, p. 131, 2011.

[46] B. Amgalan and H. Lee, "DEOD: uncovering dominant effects of cancer-driver genes based on a partial covariance selection method," *Bioinformatics*, vol. 31, no. 15, pp. 2452–2460, 2015.

[47] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman, and S. Li, "Incremental learning for ν-support vector regression," *Neural Networks the Official Journal of the International Neural Network Society*, vol. 67, pp. 140–150, 2015.

[48] Y. Zheng, J. Byeungwoo, D. Xu, Q. M. J. Wu, and Z. Hui, "Image segmentation by generalized hierarchical fuzzy C-means algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 28, pp. 4024–4028, 2015.

[49] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation," *Pattern Recognition*, vol. 40, no. 3, pp. 825–838, 2007.

[50] X. Wen, L. Shao, Y. Xue, and W. Fang, "A rapid learning algorithm for vehicle classification," *Information Sciences*, vol. 295, pp. 395–406, 2015.

[51] S. G. Ge, J. Xia, W. Sha, and C. H. Zheng, "Cancer subtype discovery based on integrative model of multigenomic data," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. PP, no. 99, p. 1, 2016.

Hindawi
Complexity
Volume 2017, Article ID 9194801, 8 pages
https://doi.org/10.1155/2017/9194801

Research Article

Identifying the Risky SNP of Osteoporosis with ID3-PEP Decision Tree Algorithm

Jincai Yang,[1] Huichao Gu,[1] Xingpeng Jiang,[1] Qingyang Huang,[2] Xiaohua Hu,[1] and Xianjun Shen[1]

[1]*School of Computer Science, Central China Normal University, Wuhan 430079, China*
[2]*School of Life Science, Central China Normal University, Wuhan 430079, China*

Correspondence should be addressed to Jincai Yang; jcyang@mail.ccnu.edu.cn

Received 31 March 2017; Revised 26 May 2017; Accepted 8 June 2017; Published 7 August 2017

Academic Editor: Fang-Xiang Wu

In the past 20 years, much progress has been made on the genetic analysis of osteoporosis. A number of genes and SNPs associated with osteoporosis have been found through GWAS method. In this paper, we intend to identify the suspected risky SNPs of osteoporosis with computational methods based on the known osteoporosis GWAS-associated SNPs. The process includes two steps. Firstly, we decided whether the genes associated with the suspected risky SNPs are associated with osteoporosis by using random walk algorithm on the PPI network of osteoporosis GWAS-associated genes and the genes associated with the suspected risky SNPs. In order to solve the overfitting problem in ID3 decision tree algorithm, we then classified the SNPs with positive results based on their features of position and function through a simplified classification decision tree which was constructed by ID3 decision tree algorithm with PEP (Pessimistic-Error Pruning). We verified the accuracy of the identification framework with the data set of GWAS-associated SNPs, and the result shows that this method is feasible. It provides a more convenient way to identify the suspected risky SNPs associated with osteoporosis.

1. Introduction

Osteoporosis is a type of systemic skeletal disease that is characterized by reduced bone mass and microarchitecture deterioration of bone tissues, thereby leading to the loss of strength and increased risk of fractures [1]. It is one of the age-related diseases with arteriosclerosis, hypertension, diabetes, and cancer. Currently, none of the medical methods is safe and effective to cure osteoporosis. Therefore, it is necessary to provide theoretical basis for developing a medical strategy to cure the disease from the pathogenesis of osteoporosis.

With the completion of the International HapMap Project and 1000 Genomes Project, about ten millions SNPs of human were annotated, among which more than 3 million are common SNPs. Genetic analysis has reached the stage of genome-wide association study (GWAS). The GWAS is applied to the study of 40 kinds of diseases that are related to more than 500 thousands SNPs [2].

Osteoporosis is a complex and polygenic disease of bone system with the heritability of bone mass is about 60–80% [3]. Much progress has been made on the genetic analysis of osteoporosis in the past 20 years and it has been found that a lot of genes and SNPs are associated with osteoporosis through GWAS [4, 5].

Computational biology refers to the development and application of data analysis, the theory of data method, mathematical modeling, and computer simulation technology, used in the study of biology, behavioral, and social group system of a discipline [6]. The rapid mass of biological data accumulation is unprecedented in the history of human science. Now, a variety of methods and tools of computational biology through the Internet have been successfully applied in every aspect in the field of biological research. They are powerful for post-GWAS studies [7] and could identify the potential and promising causal SNPs that require experimental tests for follow-up functional studies. Extensive work has

been done in this area in recent years. The performances were well validated through identifying numerous disease-associated SNPs for further study and revealing previously unknown mechanisms for complex diseases [8].

The method of computational biology can also be used to study and understand these osteoporosis-susceptible genes and the function of SNP. All the osteoporosis associated genes and SNPs (including linkage disequilibrium (LD) SNPs) sequence information were collected and aggregated from the national center for biological information (NCBI) database, and the effects of osteoporosis GWAS-associated lead SNPS and their linked SNPs to transcription factor (TF) binding affinity were studied through JASPAR database. At the same time, the osteoporosis GWAS-associated genes have also been analyzed with Protein-Protein Interaction (PPI) network analysis tool in the study of the osteoporosis GWAS-associated SNPs associated by the online PPI tool named String. Combining with GO and pathway analysis, we found that the hub proteins associated and the Wnt signaling pathway were related to the mesenchymal stem cell differentiation and hormone signaling that was related to the metabolism of osteoporosis [9]. Finally, it was found that the osteoporosis GWAS-associated SNPs in special region of genes had long-range interaction signal with other locus by analyzing the long-range interaction of osteoporosis associated SNPs on GWAS3D [10].

In the BIBM workshop paper [11], we utilized the known osteoporosis GWAS-associated SNPs and genes as the data set to identify the osteoporosis suspected risky SNPs. The process for identification was achieved by computation method. In this extension, we made some improvements on the paper. Firstly, we had achieved graphical description for the SNPs identification process. We added a flow chart for the paper to describe the process of identification method that made the method more intuitive. Secondly, we used ID3 decision tree algorithm with PEP method instead of ID3 decision tree algorithm in the second part of the method. We made the improvement to solve the overfitting problem in ID3 decision tree algorithm; we used the C4.5 algorithm to make a comparison with our ID3-PEP algorithm. Finally, we added type 2 diabetes (T2D) GWAS-associated SNPs and genes as the negative data set based on osteoporosis GWAS-associated SNPs and genes to verify the accuracy of the method comprehensively.

2. Material and Method

We identified the suspected risky SNPs associated with osteoporosis by algorithm based on the analysis of osteoporosis GWAS-associated SNPs with the method mentioned above [9]. It is assumed that the SNPs that are similar to the osteoporosis GWAS-associated SNPs are possible risky SNPs associated with osteoporosis. The identification process of the suspected risky SNPs includes two steps in general. Firstly, we constructed a Protein-Protein Interaction (PPI) network based on the Protein-Protein Interaction analysis of the osteoporosis GWAS-associated genes and the genes associated with suspected risky SNPs and identify whether the genes associated with the suspected risky SNPs are associated with osteoporosis through random walk algorithm based on Markov chain. By the algorithm, we also selected the suspected risky SNPs whose associated genes are identified to be associated with osteoporosis. We then classified those SNPs based on their characteristics of function and their loci features by a classification decision tree, and the decision tree was constructed by ID3 decision tree algorithm with Pessimistic-Error Pruning. Figure 1 describes the process to identify the osteoporosis risky SNPs.

2.1. The Identification of Genes Associated with Suspected Risky SNPs. According to the modular property of the genetic diseases, many scholars have proposed prioritization algorithms to predict the disease-causing genes based on the PPI, Human Disease Network, and DISEASOME recently [12–16]. Similarly, we obtained the scores of the genes associated with the suspected risky SNPs through the random walk algorithm based on the PPI of the osteoporosis GWAS-associated genes and the genes associated with suspected risky SNPs. Then, the result was acquired by setting up a threshold k, and the genes associated with suspected risky SNPs are probably the osteoporosis associated genes if their scores are greater than k.

2.2. The Random Walk Algorithm Based on Markov Chain. Kohler proposed a method for the problem of candidate-gene prioritization by random walk algorithm based on the global network distance of PPI. The results indicate that the algorithm is more effective than the local network distance algorithm [17]. The random walk algorithm was applied to Protein-Protein Interaction network of all associated genes.

An undirected graph $G = (V, E)$ is defined for the Protein-Protein Interaction network of all associated genes. In the undirected graph G, V is the set of vertices for the interactors of the network. And V is defined as $V = \{v_1, v_2, \ldots, v_n\}$; E is the set of edges; and E is defined as $E = \{\langle v_i, v_j\rangle \mid v_i,\ v_j \in V\}$. Every edge in the set of edges corresponds to two nodes of the set of vertices for the interaction between the interactors. Moreover, it is assumed that a random process meets the condition of Markov chain. The random process should be as follows:

(a) The probability distribution of time $t+1$ is only related to the state of time t, and it is not related to the state before time t.

(b) The state transition is not related to the value of t from the time t to time $t + 1$. Therefore, the Markov chain model is defined as

$$(S, P, Q)\,. \tag{1}$$

S is a nonempty set that consists of all the possible states of the system. It is a state space that can be a limited and denumerable set or a nonempty set. $P = [P_{ij}]_{n\times n}$ is the state transfer-probability matrix, P_{ij} is the probability that the system is in the state i at time t to the state j at time $t + 1$. N is the number of system states. $Q = \{q_0, q_1, \ldots, q_{n-1}\}$ is the initial probability distribution of the system, q_i is the

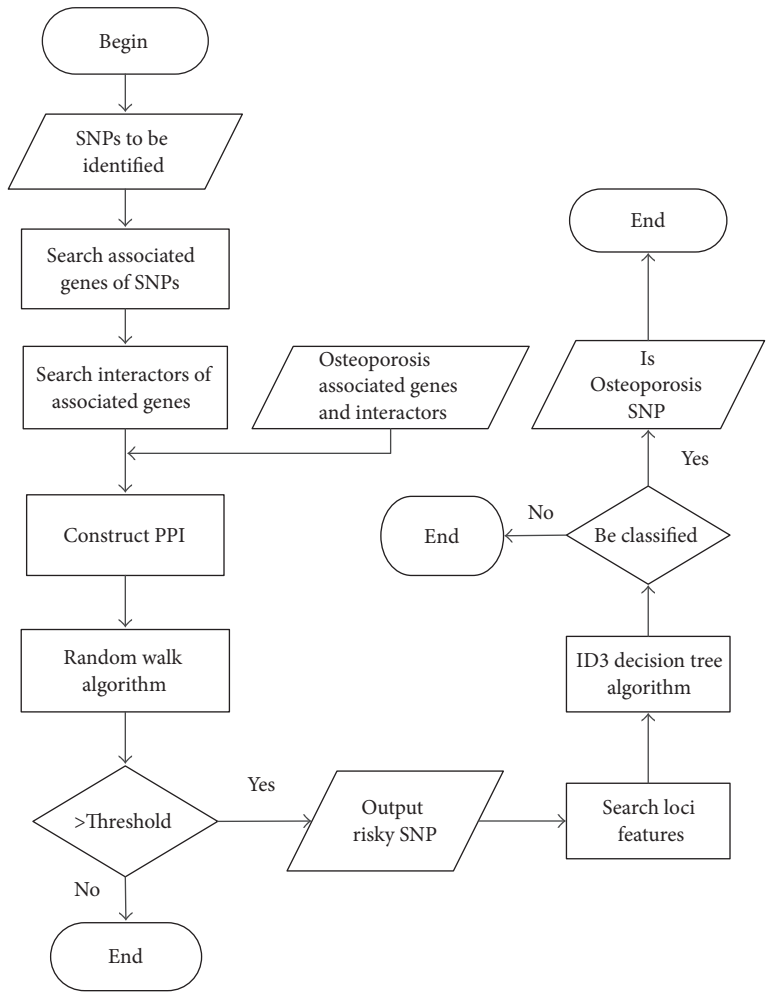


Figure 1: Process to identify the suspected risky SNPs associated with osteoporosis.

probability that the system in state i at the initial time, and $\sum_{i=0}^{N} q_i = 1$.

Based on the above theory model, the random walk on graphs is defined as an iterative walk's transition from its current node to a randomly selected neighbor starting at given source node [17]. The random walk is defined as

$$P^{t+1} = (1 - \alpha) P^t W + \alpha P^0. \quad (2)$$

P^t is a vector in which the ith element holds the probability of being at node i at time step t. α is a constant between 0 and 1 that it is the restart of the walk in every step at the node i with probability α, and $\alpha \in (0, 1]$ [17]. P^0 is a row vector of $1 \times n$ which is the initial state of the system, and n is the element number of V. The value of known elements of P^0 is equal, and the sum of them is 1. And the value of other elements is 0. W is the transition probability matrix which is defined as

$$W = D^{-1} A. \quad (3)$$

A is an adjacency matrix of undirected graph V. Every element a_{ij} of A is defined as follows: if there is interaction between v_i and v_j in the network, the element $a_{ij} = 1$; otherwise, $a_{ij} = 0$ the formula is defined as

$$a_{ij} = \begin{cases} 1, & \langle v_i, v_j \rangle \in E \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

D is a diagonal matrix. Each element d_{ij} of D is defined as follows: if $i = j$ then it should have $d_{ij} = d_{ii}$; otherwise $d_{ij} = 0$. d_{ij} is the degree of v_i in the network. The formula is defined as

$$d_{ij} = \begin{cases} \sum_{k=1}^{n} a_{ik}, & i = j \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

The transition probability matrix W is also a row-normalized adjacency matrix of the graph. Formula (2) meets the state of stationary distribution of Markov train model obviously, so the central point of random walk algorithm is evaluating the stationary distribution state of the probability of the nodes in the undirected network G which consists of PPI. Firstly, the transition probability matrix W should be

obtained and the initial value is set for P^0. Then, process t times iteration based on formula (2) until $\lim(p^{t+1} - p^t) = 0$, p^{t+1} is a convergence vector. A threshold is set for the probability value, and if the probability value of the nodes (or genes) is greater than the threshold, they are osteoporosis associated genes.

2.3. Classify the Suspected Risky SNPs by ID3 Decision Tree Algorithm. ID3 decision tree algorithm is a classification algorithm for tree structure [18, 19]. The goal of the algorithm is to predict target variable based on multiple input variables and deduce a classification rule with decision tree form from a group of irregular samples. We assume that all input characteristic elements have a limited discrete domain and need an individual characteristic element as a category. The nonleaf nodes of a classification decision tree classify the samples by characteristics of samples, and each leaf node of the tree is a class or classes of probability distribution. Therefore, we chose decision tree to classify the SNP based on the condition of the training set and algorithm characteristics.

SNPs located within the promoter or distant enhancer region of genes may alter the binding of TFs with DNA and subsequently regulate gene expression [20]. The suspected risky SNPs are classified by ID3 decision tree algorithm based on four features of significant position on genes, mapping on putative enhancer region, mapping on distal interaction, and the region where the SNPs are located [21].

The decision tree algorithm chooses the attribute with the maximum information gain after it is split, and the algorithm searches the decision-space by way of top-down greedy algorithm. S is defined as the training set of SNPs with their loci features, and the training set is divided into n classes. That is, $C = \{S_1, S_2, \ldots, S_n\}$. The number of the training instances in ith class is defined as $|S_i| = C_i$. The number of the training instances in S is $|S|$. The probability that a training instance belongs to the ith class is $P(S_i)$. And a formula is defined as

$$P\left(S_i\right) = \frac{C_i}{|S|}. \tag{6}$$

For the training set S, $H(S)$ is defined as the information entropy of C, and we have the formula

$$H\left(S\right) = -\sum_{i=1}^{n} P\left(S_i\right) \log_2 P\left(S_i\right). \tag{7}$$

The greater the value of information entropy $H(S)$ is, the smaller the degree of uncertainty for the division of C is. The attribute T is selected as the test attribute which is the loci features of the training set SNPs, and the value set for attribute T is $T = \{t_1, t_2, \ldots, t_m\}$. The probability of the attribute belongs to ith class when $T = t_j$ can be formulated as

$$P\left(S_i \mid T = t_j\right) = \frac{C_{ij}}{|S|}. \tag{8}$$

C_{ij} is the number of training instances which belongs to ith class.

When the attribute $T = t_j$, a formula is used to define the conditional entropy of the attribute T as

$$H\left(X_j\right) = -\sum_{i=1}^{n} P\left(S_i \mid T = t_j\right) \log_2 P\left(S_i \mid T = t_j\right). \tag{9}$$

X_j is the training instances set of training set S.

The information entropy of attribute T is defined as

$$\mathrm{IG}\left(T\right) = H\left(S\right) - \sum_{j=1}^{m} P\left(S_i \mid T = t_j\right) H\left(X_j\right). \tag{10}$$

We built a top-down decision tree and classified the training instances by choosing the attribute with the maximum information entropy based on the formulas above.

However, the overfitting problem could not be avoided if there were many noise samples in the training set, because of a complicated classification decision tree constructed by ID3 decision tree algorithm with a fair amount of noise samples in the training set. To solve the problem, a PEP (Pessimistic-Error Pruning) algorithm was exerted on the ID3 decision tree classification algorithm. PEP is the most accurate top-down pruning strategy which deals with the pruning problem without separating the training set.

We define a decision tree T which grows on a large scale based on the training set of SNPs with their loci features. T_1 is a nonleaf node set, T_2 is a leaf node set, and T_3 is for all nodes of T. The formula is $T_3 = T_1 \cup T_2$.

Before pruning, we define $r(t)$ as the error rate of node t in the decision tree. The formula is

$$r\left(t\right) = \frac{e\left(t\right)}{n\left(t\right)}. \tag{11}$$

$n(t)$ is the number of samples in node t, and $e(t)$ is the number of samples that does not belong to node t actually.

We define T_t as a subtree of the decision tree T, and t is the root node of T_t. So the error rate of the subtree T_t is

$$r\left(T_t\right) = \frac{\sum_{s \in S_t} e\left(s\right)}{\sum_{s \in S_t} n\left(s\right)}. \tag{12}$$

S_t is the leaf node set of subtree T_t, and we define $S_t = \{s_1, s_2, \ldots, s_n\}$.

Apparently, the formula for error rate of the subtree T_t is binomial distribution. We define a continuity correction factor $r'(t)$ in order to make the binomial distribution approach the normal distribution. And the formula is

$$r'\left(t\right) = \frac{e\left(t\right) + 1/2}{n\left(t\right)}. \tag{13}$$

Therefore, we deduce the continuity correction factor for the subtree T_t. The formula is

$$r'\left(T_t\right) = \frac{\sum_{s \in S_t} \left[e\left(s\right) + 1/2\right]}{\sum_{s \in S_t} n\left(s\right)} = \frac{\sum_{s \in S_t} e\left(s\right) + \left|S_t\right|/2}{\sum_{s \in S_t} n\left(s\right)}. \tag{14}$$

In order to simplify the formula, we define $e'(t)$ as the error sample number instead of error rate. So the error sample number of node t in the decision tree T is

$$e'(t) = e(t) + \frac{1}{2}. \tag{15}$$

Therefore, the error sample number of the subtree T_t is

$$e'(T_t) = \sum_{s \in S_t} e(s) + \frac{|S_t|}{2}. \tag{16}$$

Similarly, the formula for the error sample number of subtree T_t is binomial distribution. And the standard deviation for $e'(T_t)$ is defined as

$$\mathrm{SE}\left(e'(T_t)\right) = \left[\frac{e'(T_t) \times \left(n(t) - e'(T_t)\right)}{n(t)}\right]^{1/2}. \tag{17}$$

Finally, we deduce from formulas above that the subtree T_t will be cut if the node t meets the condition:

$$e'(t) \leq e'(T_t) + \mathrm{SE}\left(e'(T_t)\right). \tag{18}$$

The process of the PEP algorithm is as follows:

Algorithm: PEP
Begin
Input: decision tree T before pruned
Output: decision tree T after pruned
(1) Get the nonleaf node set T_1 of the decision tree T
(2) For $k = 1$ to length (T_1)
(3) Do get a subtree T_t whose root node is $t[k]$ $(t[k] \in T_1)$
(4) If $(e'(t) \leq e'(T_t) + SE(e'(T_t)))$
(5) Then delete T_t
(6) Else $k++$
(7) End
End

We classified the suspected risky SNPs effectively based on their loci characteristics and studied their functions according the ID3 decision tree algorithm and PEP.

3. Results

By the end of 2014, nine GWAS and nine meta-analyses had reported 107 genes and 129 SNPs (lead SNP) that were associated with BMD, osteoporosis, or fractures with a significant threshold of 5×10^{-8}. 222 SNPs linked to osteoporosis GWAS-associated lead SNPs had also been identified by using LD information in the Caucasians population via HapMap website [9]. Moreover, we obtained 107 known osteoporosis GWAS-associated genes which showed significant connectivity among proteins. And there were interactions between osteoporosis GWAS-associated genes and interactors. We used the common Protein-Protein Interaction databases, such as Human Protein Interaction database (HPID) and General Repository for Interaction Data (GRID), to find the interactors which had interactions with the osteoporosis GWAS-associated genes and their interactions. Then, we obtained the interaction network graph by Cytoscape v3.4.0. Figure 2 is the PPI of osteoporosis GWAS-associated genes.

Table 1: Part of the classification of training set.

SNP	bda	td	Enhancer	Gene region	Class
rs7524102	Y	Y	Y	Intergenic	C
rs34920465	Y	Y	Y	Control region	D
rs6426749	Y	N	Y	Control region	G
rs1430742	N	N	N	Cording sequence	B
rs6929137	Y	Y	Y	Missense	A
rs479336	Y	Y	N	Cording sequence	K
rs11898505	Y	N	Y	Intergenic	F
rs17040773	Y	Y	Y	Cording sequence	E
rs344081	Y	N	Y	Cording sequence	H
rs6909279	Y	Y	N	Intergenic	I

(a) The first column is part of osteoporosis GWAS-associated SNPs; (b) the column of "bda," "td," and "enhancer" means whether the SNP is on significant TFs binding affinity, mapping on distal interaction, and mapping on putative enhancer region; (c) the last column is the category the SNP belong to.

The result was verified by 10-fold cross-validation based on the data set of osteoporosis GWAS-associated genes and SNPs. We divided the data set of 129 osteoporosis GWAS-associated lead SNPs and 222 SNPs linked with them into 10 samples. One sample was then randomly chosen and saved as the validation set to verify the model from the 10 samples, and the other 9 samples were saved as training set. The verification process was repeated 10 times so that each sample was the validation set once, and the accuracy was calculated every time. A 10-fold cross-validation was completed by the process above.

We set a threshold k ($k > 10^{-3}$) as a result of the validation. The recall was calculated, which was the true positive result to positive result ratio. The 10-fold cross-validation was repeated for ten times and the average recall rate of every validation was calculated. The result was shown in Figure 3.

The classification result was also verified by 10-fold cross-validation. The osteoporosis GWAS-associated SNPs were used as the data set. The SNPs of training set were classified based on their loci features. Part of classification of the training set was shown in Table 1. We classified the SNPs of validation set through ID3 decision tree algorithm and recorded the accuracy of classification, which was the proportion of classification accurate samples to all the samples.

Then, the process of validation was repeated for ten times and calculated the average accuracy rate and average classification reliability. The result was shown in Figure 4.

We also used genome-wide association studies (GWAS) of type 2 diabetes (T2D) data as negative data to verify our method [22]. 50 lead SNPs of T2D were obtained with

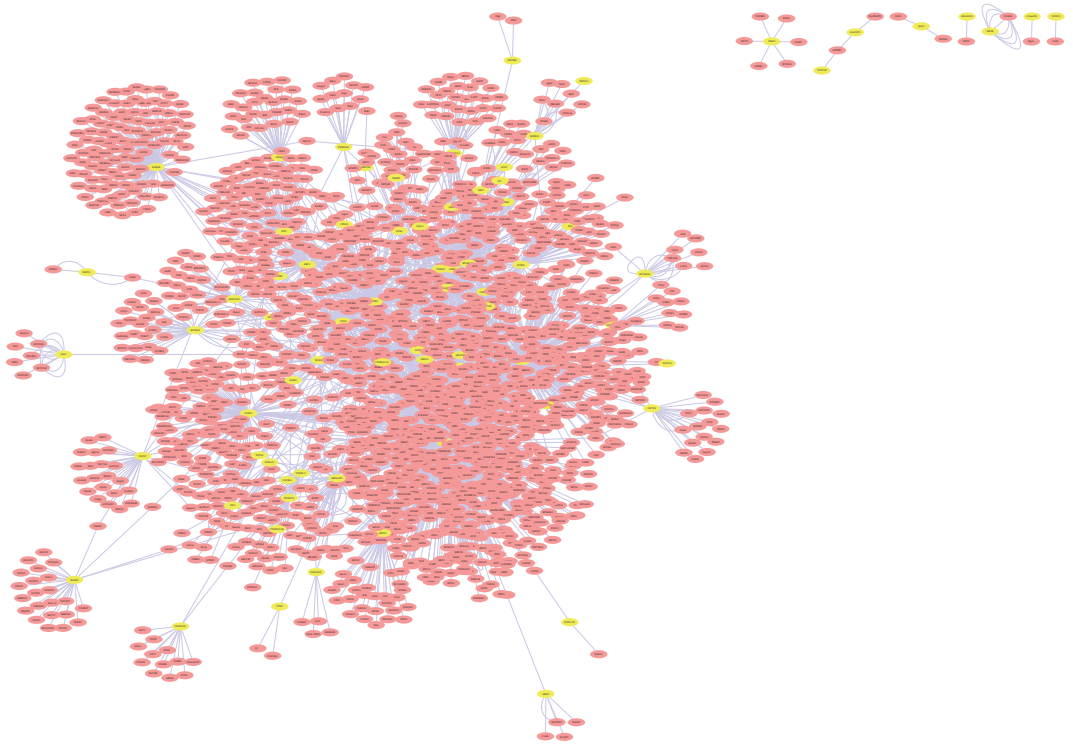


Figure 2: PPI of osteoporosis GWAS-associated genes (the pink nodes indicated those which had interactions with the osteoporosis GWAS-associated genes, and the yellow nodes indicated the osteoporosis GWAS-associated genes).

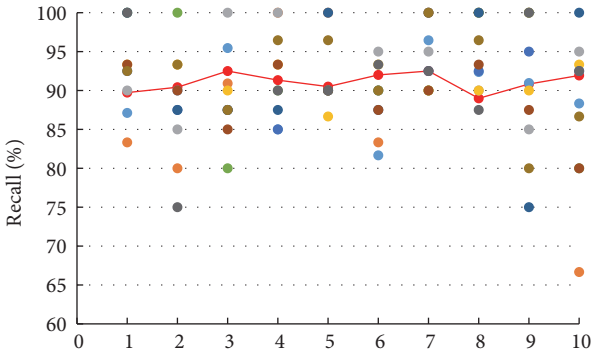


Figure 3: Result of random walk (the ten colors of the points indicated ten 10-fold cross-validation, and the same color of points indicated the validation process. The points connected by a line were the average recall value of ten experiments. The x-axis was the 10-step verification of the 10-fold cross-validation process).

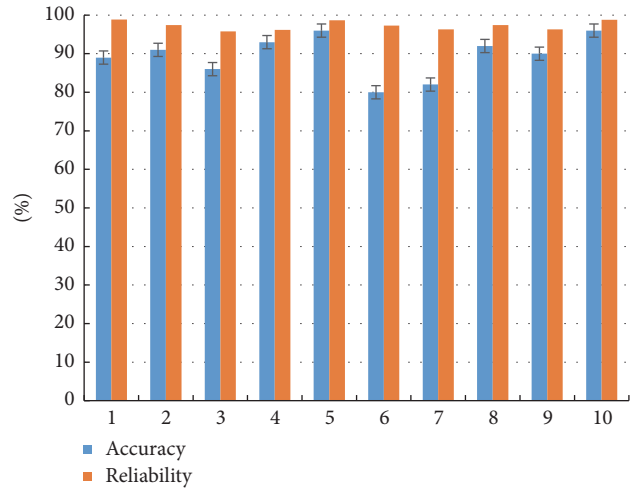


Figure 4: Result of ID3 decision tree (the blue credibility refers to the average accuracy values of 10-fold cross-validation, and the orange credibility refers to the average reliability value).

their position features and associated genes. We searched the interactors of the associate genes from the PPI database and constructed the PPI network with the known osteoporosis GWAS-associated genes. The random walk algorithm was used on the PPI network.

We then used PEP for ID3 decision tree to construct a simplified classification decision tree. We combined the two steps of the risky SNPs identification method and verified the method by 10-fold cross-validation. Finally, we found that not only was the computation efficiency improved, but also the accuracy rate of the result by using ID3 decision tree algorithm with PEP in the identification method was higher. The improvement is due to the fact that we had cut the subtrees which were constructed by the noise samples and solved the overfitting problem. While we defined ID3

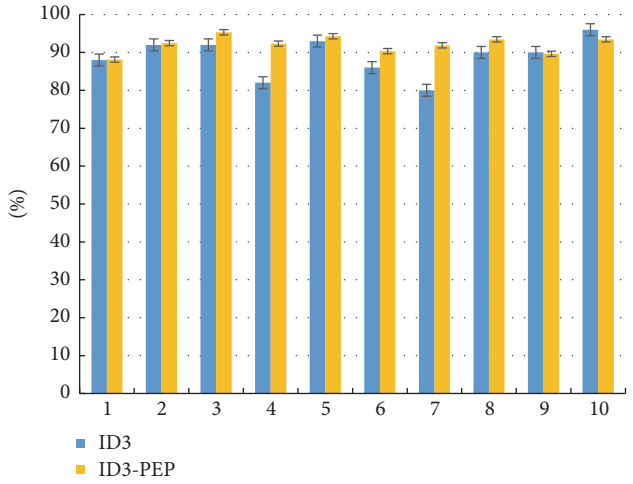


Figure 5: Comparison of two classification algorithm (the blue credibility refers to the classification accuracy by ID3 algorithm, and the yellow credibility refers to the classification accuracy by ID3 decision tree algorithm with PEP).

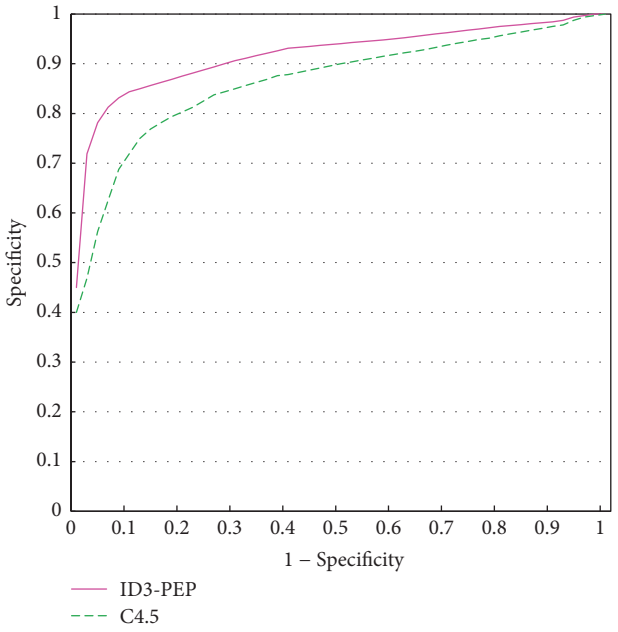


Figure 6: The comparison of ID3-PEP and C4.5.

decision tree algorithm with PEP in the identification method as ID3-PEP and ID3 decision tree algorithm as ID3, the result comparison of these two classification algorithm in the identification method was described by Figure 5. According to the result, we concluded that the ID3-PEP in the identification method was more stable than ID3 algorithm, and it had better effect for the classification problem.

C4.5 is the optimization of ID3. They have the same way to learn training set and build a classification decision tree, but the difference of them is the way of choosing split attribute. C4.5 algorithm chooses the maximum attribute with information gain ratio to split. In order to solve the problem of overfitting in ID3 decision tree algorithm, C4.5 algorithm needs to scan the data set and rank them in every step. This calculation method and process of the algorithm have low operational efficiency. ID3-PEP algorithm solved the problem and was more accurate than C4.5. We made a comparison of these two algorithms through ROC curve, which is shown in Figure 6. Result shows that ID3-PEP is better than C4.5 in our classification.

4. Discussion and Conclusion

Since SNP plays a key role in the process of pathology and susceptibility of osteoporosis [23], it is necessary to find the unknown risky SNPs. Using the data set of known osteoporosis GWAS-associated SNPs and genes [8], we identified the genes of suspected risky SNPs associated with osteoporosis by random walk algorithm on the PPI network constructed by osteoporosis GWAS-associated genes and the genes associated with suspected risky SNPs. The suspected risky SNPs were classified based on the features of their loci position and function. We used 10-fold cross-validation to verify our method.

The result of the experiment above showed that the identification method for risky SNPs of osteoporosis was correct and effective. Our method efficiently achieved the process of identifying osteoporosis suspected risky SNPs.

However, there is still a need to perfect the identification method. First of all, we need to search the loci features of suspected risky SNPs associated with osteoporosis and the interactors of associated genes manually. The training set for our method is the known osteoporosis GWAS-associated SNPs, which is not large enough to identify the risky SNPs accurately. Therefore, further research is needed. Firstly, a workflow can be constructed to improve the identification process, aiming to automatically identify the suspected risky SNPs' features. In order to improve the accuracy of our method, more features of the SNPs should be examined, such as the conservation of SNPs and the influence of the SNPs on miRNA binding site. Finally, we use our method to predict risky SNPs associated with osteoporosis by constructing the PPI network of all the human genes.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper and the received funding did not lead to any conflicts of interest regarding the publication of this manuscript.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grants nos. 61532008 and 31371275), the National Social Science Foundation of China (no. 14BYY093), and the Fundamental Research Funds for the Central Universities (no. CCNU17TS0003).

References

[1] F. Rivadeneira, U. Styrkársdottir, K. Estrada, B. V. Halldórsson, Y. H. Hsu, J. B. Richards et al., "Twenty bonemineral-density loci identified by large-scale meta-analysis of genome-wide association studies," *Nature Genetics*, vol. 41, no. 11, pp. 1199-206, 2009.

[2] D. Welter, J. MacArthur, J. Morales et al., "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1001–D1006, 2014.

[3] Q.-Y. Huang, R. R. Recker, and H.-W. Deng, "Searching for osteoporosis genes in the post-genome era: Progress and challenges," *Osteoporosis International*, vol. 14, no. 9, pp. 701–715, 2003.

[4] Q. Huang and A. W. C. Kung, "Genetics of osteoporosis," *Molecular Genetics and Metabolism*, vol. 88, no. 4, pp. 295–306, 2006.

[5] J. B. Richards, H. F. Zheng, and T. D. Spector, "Genetics of osteoporosis from genome-wide association studies: advances and challenges," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 576–588, 2012.

[6] K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak, "VISTA: computational tools for comparative genomics," *Nucleic Acids Research*, vol. 32, pp. W273–W279, 2004.

[7] Q. Y. Huang, "Genetic study of complex diseases in the post-GWAS era," *Journal of Genetics and Genomics*, vol. 42, no. 3, pp. 87–98, 2015.

[8] S. L. Edwards, J. Beesley, J. D. French, and A. M. Dunning, "Beyond GWASs: illuminating the dark road from association to function," *American Journal of Human Genetics*, vol. 93, no. 5, pp. 779–797, 2013.

[9] L. Qin, Y. Liu, Y. Wang et al., "Computational characterization of osteoporosis associated SNPs and genes identified by genome-wide association studies," *Plos One*, vol. 11, no. 3, Article ID e0150070, pp. 1–14, 2016.

[10] M. J. Li, L. Y. Wang, Z. Xia, P. C. Sham, and J. Wang, "GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications," *Nucleic acids research*, vol. 41, pp. W150–W158, 2013.

[11] J. Yang, H. Gu, X. Jiang, Q. Huang, X. Hu, and X. Shen, "Walking in the PPI network to predict the risky SNP of osteoporosis with decision tree algorithm," in *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM '16)*, pp. 1283–1287, Shenzhen, China, 2016.

[12] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[13] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, M. A. Durbin, and R. E. Handsaker, "An integrated map of genetic variation from 1, 092 human genomes," *Nature*, vol. 467l, no. 7311, pp. 52–58, 2010.

[14] K. Lage, E. O. Karlberg, Z. M. Storling et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[15] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular Systems Biology*, vol. 4, no. 1, 2008.

[16] R. K. Nibbe, S. A. Chowdhury, M. Koyuturk, R. Ewing, and M. R. Chance, "Protein-protein interaction networks and subnetworks in the biology of disease," *Systems Biology and Medicine*, vol. 3, no. 3, pp. 357–367, 2010.

[17] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[18] M. J. Blow, D. J. McCulley, Z. Li et al., "ChIP-seq identification of weakly conserved heart enhancers," *Nature Genetics*, vol. 42, no. 9, pp. 806–812, 2010.

[19] J. R. Quinlan, "Generating production rules from decision trees," in *Proceedings of the IJCAI-87*, Milan, Italy, 1987.

[20] L. A. Hindorff, P. Sethupathy, H. A. Junkins et al., "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 23, pp. 9362–9367, 2009.

[21] A. Visel, E. M. Rubin, and L. A. Pennacchio, "Genomic views of distant-acting enhancers," *Nature*, vol. 461, no. 7261, pp. 199–205, 2009.

[22] M. Cheng, X. Liu, M. Yang, L. Han, A. Xu, and Q. Huang, "Computational analyses of type 2 diabetes-associated loci identified by genome-wide association studies," *Journal of Diabetes*, vol. 9, no. 4, pp. 362–377, 2016.

[23] E. T. Dermitzakis and A. G. Clark, "Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover," *Molecular Biology and Evolution*, vol. 19, no. 7, pp. 1114–1121, 2002.

Hindawi
Complexity
Volume 2017, Article ID 9514369, 11 pages
https://doi.org/10.1155/2017/9514369

Research Article

A Resting-State Brain Functional Network Study in MDD Based on Minimum Spanning Tree Analysis and the Hierarchical Clustering

Xiaowei Li,[1] Zhuang Jing,[1] Bin Hu,[1] Jing Zhu,[1] Ning Zhong,[2] Mi Li,[2] Zhijie Ding,[3] Jing Yang,[4] Lan Zhang,[4] Lei Feng,[5] and Dennis Majoe[6]

[1]*School of Information Science & Engineering, Lanzhou University, Lanzhou, China*
[2]*International WIC Institute, Beijing University of Technology, Beijing, China*
[3]*The Third People's Hospital of Tianshui City, Tianshui, China*
[4]*Lanzhou University Second Hospital, Lanzhou, China*
[5]*Beijing Anding Hospital of Capital Medical University, Beijing, China*
[6]*Computer Systems Institute, ETH Zürich, Zürich, Switzerland*

Correspondence should be addressed to Bin Hu; bh@lzu.edu.cn

Received 31 March 2017; Revised 3 June 2017; Accepted 12 June 2017; Published 25 July 2017

Academic Editor: Jianxin Wang

A large number of studies demonstrated that major depressive disorder (MDD) is characterized by the alterations in brain functional connections which is also identifiable during the brain's "resting-state." But, in the present study, the approach of constructing functional connectivity is often biased by the choice of the threshold. Besides, more attention was paid to the number and length of links in brain networks, and the clustering partitioning of nodes was unclear. Therefore, minimum spanning tree (MST) analysis and the hierarchical clustering were first used for the depression disease in this study. Resting-state electroencephalogram (EEG) sources were assessed from 15 healthy and 23 major depressive subjects. Then the coherence, MST, and the hierarchical clustering were obtained. In the theta band, coherence analysis showed that the EEG coherence of the MDD patients was significantly higher than that of the healthy controls especially in the left temporal region. The MST results indicated the higher leaf fraction in the depressed group. Compared with the normal group, the major depressive patients lost clustering in frontal regions. Our findings suggested that there was a stronger brain interaction in the MDD group and a left-right functional imbalance in the frontal regions for MDD controls.

1. Introduction

Major depressive disorder is a global mental disorder and has an unfavourable influence on physical and psychological health [1]. In addition to profound personal suffering, MDD patients lack the necessary social and occupational functioning [2]. Moreover, The World Health Organization predicted that depression would become the second leading cause of illness by the year 2020 [3]. In this light, exploring the neurobiological signature of MDD from multiple imaging modalities was considered to sharpen the reach of depression and develop treatments, including electroencephalogram (EEG), magnetoencephalogram (MEG), functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT) [4]. In recent years, the research results of MDD based on different approaches had been presented substantially such as frontal EEG asymmetry, "small-word" network characteristics, and increased/disrupted cognition connectivity network [5–8]. These results revealed neurophysiology characteristics in different aspects for depression disease and made a great contribution to the study of the depression. However, there were disputes and contradictions in these results due to the differences of subjects, experimental environment, methods, and other restrictions. So, more methods and techniques are expected for exploring MDD.

The human brain is a complex system, characterized by its dynamical neural communications and mutual interactions based on synchronous oscillations among different brain areas [9, 10]. In the human brain, oscillatory patterns reflect the activity of the brain and provide reliable markers of the brain function or dysfunction [11, 12]. Therefore it is a direct and effective way to explore the brain activity based on the brain oscillatory nature. In present imaging modalities, electroencephalogram (EEG) is regarded as a convenient technology to reflect the comprehensive electrophysiological activity of neuron populations [13]. In recent years, it has been widely used in biomedical fields to estimate the oscillatory patterns on account of higher temporal resolution and noninvasiveness in mental disorders such as major depressive disorder (MDD), Alzheimer's disease, and schizophrenia disease [14, 15]. EEG oscillations are rhythmic electrical events coming from the brain and can be used to define the interaction of different brain regions [16]. Because of this feature, it is suggested that the information processing of the brain can be reflected in characteristic EEG oscillation rhythms [17]. Using this approach, a large number of findings were presented in the study of the depression. Some results based on EEG oscillations demonstrated that patients with MDD had more frontal theta, alpha, and beta oscillations [18–20]. Using EEG oscillations, Lee et al. [21] suggested that the brain affected by a major depressive disorder showed a slower decay of the long-range temporal (auto)correlations (LRTC). Recently, an EEG oscillations study on MDD reported that depressive brain was manifested in the superposition of distributed multiple oscillations [22].

At present, studying functional interactions between brain regions plays a vital role in understanding the dynamic interactions between the neural systems [23]. For defining the interaction of different brain regions, the synchronization of EEG oscillations is an important and effective indicator which can be estimated by EEG coherence [24]. And EEG coherence is conceptualised as the correlation in the time domain between two signals in a given frequency band [25, 26]. The high EEG coherence reflects synchronized neuronal oscillations between different brain areas, whereas low EEG coherence represents independent active neuronal [25]. This approach has been applied for evaluation and auxiliary diagnosis of various mental disorders, including Attention-Deficit/Hyperactivity Disorder, Autism Spectrum Disorder, MDD, and Alzheimer's disease. In one study of MDD, EEG coherence was used to estimate the sleep EEG rhythms, which suggested that low temporal coherence in depression reflects a breakdown in the organization of sleep EEG rhythms within and between two hemispheres [27]. Li et al. [7] found that the global EEG coherence of patients with MDD was significantly higher than that of healthy controls in both gamma bands. Prior EEG coherence based on discriminant function analysis (DFA) rules was used to explore possible neurophysiological differences between Asperger's Syndrome (ASP) and the Autism Spectrum Disorders (ASD) and successfully distinguished ASP and ASD populations [28]. Using EEG source-based coherence in Alzheimer's disease (AD) showed increased delta coherences between the bilateral precentral, left supplementary motor area (SMA), and right precentral [29].

In recent years, with the analysis of the connectivity based on EEG data, graph theoretical analysis has been widely applied. With the rising interest in graph theoretical studies of brain networks, the problem of the determination of the connectivity structure in the brain has been a subject of the intense research. The approaches standing behind the connectivity have a crucial impact on the results of the study and complicate a comparison of results across different studies and different brain imaging techniques [30]. Conventionally, a threshold, or a range of thresholds, is used to confirm whether connections exist or do not exist. Currently, this method is most widely used in the study of the brain functional networks [31–34]. But it might be biased by the choice of the threshold due to the number of links being decided by the size of the threshold values. Importantly, most network characteristics depending on the number of links in the network would be also influenced such as clustering coefficient, characteristic path length, and node degree. It may be an assignable cause that there were some controversial findings in the study of the brain functional network. In order to avoid this bias, some of the studies used the weighted brain functional network to avoid the choice of threshold [35–37]. Although the weighted network overcomes the problem of this subjective factor and provides a more realistic representation of functional networks, there are also problems in the weighted network. Spurious weak connections are also taken into account, potentially influencing the brain functional network.

In this context, using minimum spanning tree (MST) to represent brain networks may be one promising unbiased solution to this problem [38]. The MST is a subgraph without forming cycles that connects all the nodes in the original weighted network [39]. In this way, MST will obtain the same number of nodes and links, therefore enabling the direct comparison of network properties between groups and avoiding the aforementioned methodological biases and defect. Due to this advantage, the MST has been wildly used in a variety of mental diseases. Stam et al. [30] illustrated MST characterization allowed the representation of the observed brain networks and may simplify the construction of simple generative models of normal and abnormal brain network organizations. MST appeared in a variety of studies. MST was used as an elegant and sensitive method to capture subtle developmental organization changes in the brain networks of children [40]. In a study of Multiple Sclerosis, findings indicate that MST network analyses were able to detect network changes in the Multiple Sclerosis (MS) patients [41]. In the study of Alzheimer's disease, MST was regarded as an effective method for analyzing cortical networks [42].

The clustering and community structures have been regarded as one of the most significant features of complex networks [43]. In the brain networks, clustering or community structure was defined as a subset of highly interconnected nodes which had similar characteristics [44]. Moreover, we have known that brain networks demonstrate the property of the hierarchical modularity. Constructing the brain hierarchical modularity, the hierarchical clustering

is a special method which characterizes major building blocks or the hierarchical modularity of brain networks, corresponding to specialized brain functions [45]. Among the multitudinous hierarchical clustering algorithms, a tree agglomerative hierarchical clustering (TAHC) method can successfully detect clusters in both artificial trees and the MSTs of weighted social networks. This method was raised in [43]. Moreover, the hierarchical clustering has regarded the MST as a foundation of the complex networks [30]. And the hierarchical clustering in combination with the MST has been used in different cognitive domains. Hierarchical clustering analysis of the MST showed that the connections between the parahippocampal gyrus and posterior cingulate gyrus were disrupted in Alzheimer patients [42]. Impaired communication between functional clusters in AD was found in one MEG study [46]. Alexander-Bloch et al. [47] found disrupted modularity and local connectivity of the brain functional networks in childhood-onset schizophrenia using hierarchical clustering analysis of the MST.

This study used eyes-closed resting-state EEG recordings involving the patients with MDD and the healthy subjects. The aim of this study was using an unbiased method to construct the brain network and then explore the network characters and clustering of nodes for both MDD and healthy groups. Although the study of "resting-state" is rather exploratory, we believe that the approach used in this study is feasible. The choice of threshold has a great impact on the properties of the brain functional network. So we used an unbiased MST method to build the main network of the brain. As far as we know, this method was first used for the depression disease. Then we estimated the leaf fraction, mean link weight, and node degree fraction of MST. The hierarchical partitioning of the brain functional network in resting-state for MDD and healthy groups was not clear. So, in this study, TAHC method was used to characterize major building blocks and hierarchical partitioning of the brain networks. In the end, we discriminated the differences between the MDD group and the healthy group and then summarized and discussed our findings.

2. Subjects and Methods

2.1. Subjects. The study was approved by the local ethics committee. Written informed consent was obtained before the study began. Twenty-three patients with MDD (13 males and 10 females, right-handed) were recruited from the Lanzhou University Second Hospital. The mean age of the MDD group was 33.17 ± 19.83 years. All patients had no history of the manic episode. 14 healthy subjects (7 males and 7 females, right-handed) were recruited from the society with the mean age of 31.29 ± 21.71. To ensure the effectiveness of this study, the participants were aged between 18 and 55. Besides, primary or higher education level was required. Strict exclusion criteria were enforced before the experiment, and exclusion criteria for all participants included past history or the presence of any medical or neurological disorders, presence of drug or alcohol abuse, and past head trauma with loss of consciousness. Before the experiment, all participants participated in an interview in which the Mini and PHQ-9 [48] were administered with the help of an experienced clinical psychiatrist. Mini was used to ensure the correctness of the classification. The score of PHQ-9 was used to evaluate state anxiety, general anxiety, and depression levels. The mean PHQ-9 score of MDD group was 17.10, and the healthy group was 2.57 ($F = 11.504$, $p < 0.001$). In addition, all participants gave informed consent and were rewarded for their participation.

2.2. EEG Data Processing. The experiment was conducted in a quiet and dim light room kept away from electromagnetic interference. Participants were required to have a seat on a wooden chair comfortably with eyes closed but keeping awake during the EEG recording. They were also asked to avoid blinking and making movements. Five minutes' resting-state EEGs were recorded with a 128-channel HydroCel Geodesic Sensor Net and Net Station software, version 4.5.4. The MATLAB R2013b software package was used to process the data, and artifacts from vertical and horizontal eye movements and blinks were removed offline by an ocular correction algorithm. All channels were referenced to Cz during the acquisition, and electrode impedances were below 70 kΩ. The continuous EEG signals were recorded at sampling rates of 250 Hz, 0.3–70 Hz frequency band. We chose the following 72 electrodes: 1-F10, 2-AF8, 3-AF4, 4-F2, 5-FCz, 6-FP2, 7-Fz, 8-FC1, 9-FPz, 10-AFz, 11-Cz, 12-F1, 13-FP1, 14-AF3, 15-F3, 16-AF7, 17-F5, 18-FC5, 19-FC3, 20-C1, 21-F9, 22-F7, 23-FT7, 24-C3, 25-CP1, 26-FT9, 27-T7, 28-C5, 29-CP3, 30-T9, 31-T3, 32-TP7, 33-CP5, 34-P5, 35-P3, 36-TP9, 37-T5, 38-P7, 39-P1, 40-Pz, 41-PO7, 42-PO3, 43-O1, 44-POz, 45-Oz, 46-PO4, 47-O2, 48-P2, 49CP2, 50-PO8, 51-P8, 52-P4, 53-CP4, 54-T6, 55-P6, 56-CP6, 57-TP10, 58-TP8, 59-C6, 60-C4, 61-C2, 62-T4, 63-T8, 64-FC4, 65-FC2, 66-T10, 67-FT8, 68-FC6, 69-FT10, 70-F8, 71-F6, and 72-F4.

2.3. Coherence. The coherence is defined as the spectral cross-correlation between two signals normalized by their power spectra [7]. There are different measuring methods that analyze the coherence from different pairs of electrodes per frequency. In this study, the magnitude-squared coherence (MSC) was calculated for a particular frequency f between two given EEG signals x and y.

$$C_{xy} = \frac{\left|S_{xy}(f)\right|}{S_{xx}(f)\,S_{yy}(f)}. \tag{1}$$

$S_{xx}(f)$ is the power spectral density (PSD) estimate of x at the frequency of f and $S_{xy}(f)$ is the cross PSD estimate of x and y at the frequency of f, using Welch's averaged, modified periodogram method. The value of the MSC ranges between 0 and 1, where 0 represents no coherence and 1 indicates maximum linear interdependence between two signals.

In this study, we calculated the coherence between each possible pair of 72 EEG channels with the respect to each single frequency. A square 72 ∗ 72 coherence matrix was obtained for each participants (72 was the number of the chosen EEG channels), and each element in the coherence

matrix indicated the coherence between the corresponding two electrodes. The MDD coherence matrix was defined as the mean of all coherence matrices of the patients with MDD, and the definition of coherence matrix for healthy group was the mean of all coherence matrices of the healthy participants. The global coherence was defined as the mean value of all elements in the coherence matrix. It indicated the level of interdependence of the whole brain. In order to confirm which frequency bands have significant difference rapidly, we calculated the global coherence between two groups in each single frequency within 3–35 Hz.

2.4. MST. The minimum spanning tree is a simple acyclic connected subgraph of the original weighted network that can be used to direct comparison of networks with the same number of nodes and simplifies the network characterization. Although it is not a priori given, it will still capture most of the important topological information in the original network [30]. We constructed the MST based on the aforementioned EEG coherence matrix by employing Kruskal's algorithm [49]. The details of this algorithm used in this study were as follows: (1) ordering the elements of the EEG coherence matrix in a degressive order; (2) linking the N nodes with maximal EEG coherence until all the nodes being linked in a loopless subgraph consisting $N - 1$ edges; (3) skipping the link, if adding this link leading a circle.

In this study, we used 72 channels. So the number of the nodes in the topology of MST was 72 and the number of edges was 71. We also analyzed the topology properties of the MSTs for the two groups. Leaf fraction and the mean weight of all links included in the MST were calculated to evaluation of the MST topology. After that, statistical analyses were used to assess the credibility of differences.

2.5. Hierarchical Clustering. The clustering is an effective method to explore nontrivial information in the network. MSTs contain most of the information about the underlying clusters of the original weighted networks. In this study, TAHC method was used for the detection of the clusters in the aforementioned MSTs. The summary of TAHC algorithm [43] is as follows: first we use geodesic distances between all possible pairs of nodes of the given graph as an input to the agglomerative hierarchical clustering algorithm. Next, compute the similarity between every node pairs in an MST based on geodesic distances. Then find the most similar pair of clusters and merge them into a single cluster. Finally, recalculate similarities between the new cluster and each of the old clusters based on average-linkage clustering and remerge clusters until all nodes are merged into a single cluster.

The geodesic distance between two nodes in a tree is equal to the number of links in the shortest path. So, geodesic distances were calculated using Dijkstra Shortest Path algorithm [50]. Geodesic distances between all possible pairs of nodes in a graph constituted the geodesic distance matrix C which was a weighted matrix. In the geodesic distance matrix C, each node corresponds to a row vector. Based on C, we calculated vector similarities using Spearman's rank correlation between all row pairs of C.

$$\rho = 1 - \frac{6 \sum d_i^2}{n^3 - n}. \quad (2)$$

ρ is Spearman's rank correlation, and the value of the MSC ranges between 0 and 1. The larger the value is, the higher the similarity of the two nodes is. n is the number of the entries in a row. d_i is the difference between the two rows of each observation.

In our study, we analyzed the hierarchical clustering organization of the group-level MST for each group in interested frequency bands. Thus, each group-level dendrogram obtained by the TAHC method corresponds to each group-level MST. The distribution of nodes in the hierarchical clustering was described based on a global electrode graph.

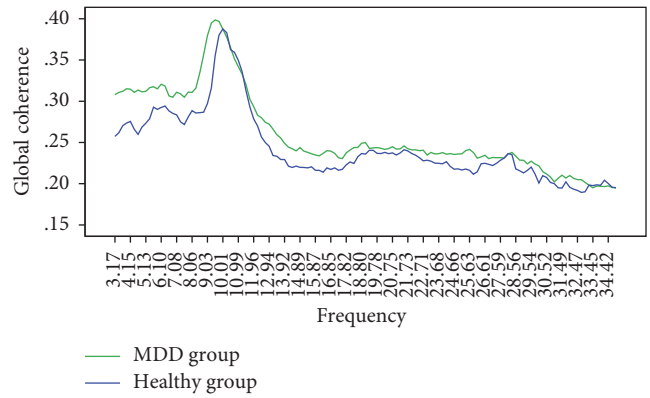


Figure 1: The coherence-frequency graph of MDD group and healthy group. The horizontal axis denoted the frequency and the vertical axis denoted the global coherence. There was the significant difference in the theta band (4–8 Hz). In alpha (8–13 Hz) and beta bands (13–30 Hz), there was no diacritic difference in both groups.

3. Results

3.1. Global Coherence. A large number of researches indicated that most of the EEG signals were at low frequencies when the brain was in the resting-state. In order to explore which frequency bands contain significant differences between the MDD group and the healthy group rapidly, we calculated global coherence in each single frequency within range of 3–35 Hz and used a frequency-coherence line chart to describe the relationship between the global coherence and the frequency, as seen in Figure 1. In this graph, we can see that the higher coherence appeared in the alpha band (8–13 Hz) for both groups. And the global coherence of the MDD group is significantly higher than that of the healthy group in the theta band (4–8 Hz). In alpha (8–13 Hz) and beta bands (13–30 Hz), there is no diacritic difference in both groups.

3.2. Difference in Coherence. Because the difference obtained from the global coherence is not obvious in the alpha and beta bands, we paid more attention to the theta band.

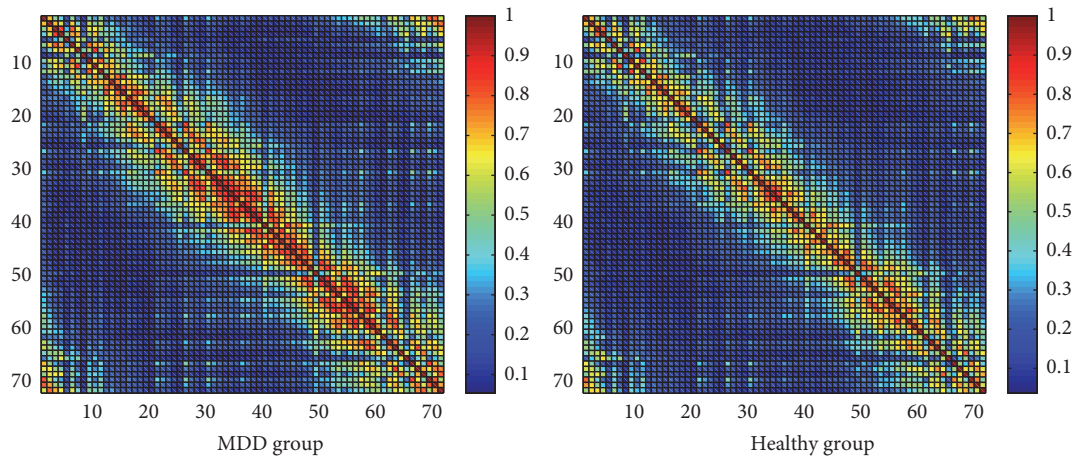


Figure 2: The coherence of MDD group and healthy group in the theta band. The size of the coherence matrix was 72 ∗ 72. In the matrix map, each chromatic point represented the coherence of two corresponding channels. The horizontal and vertical axes denoted 72 channels.

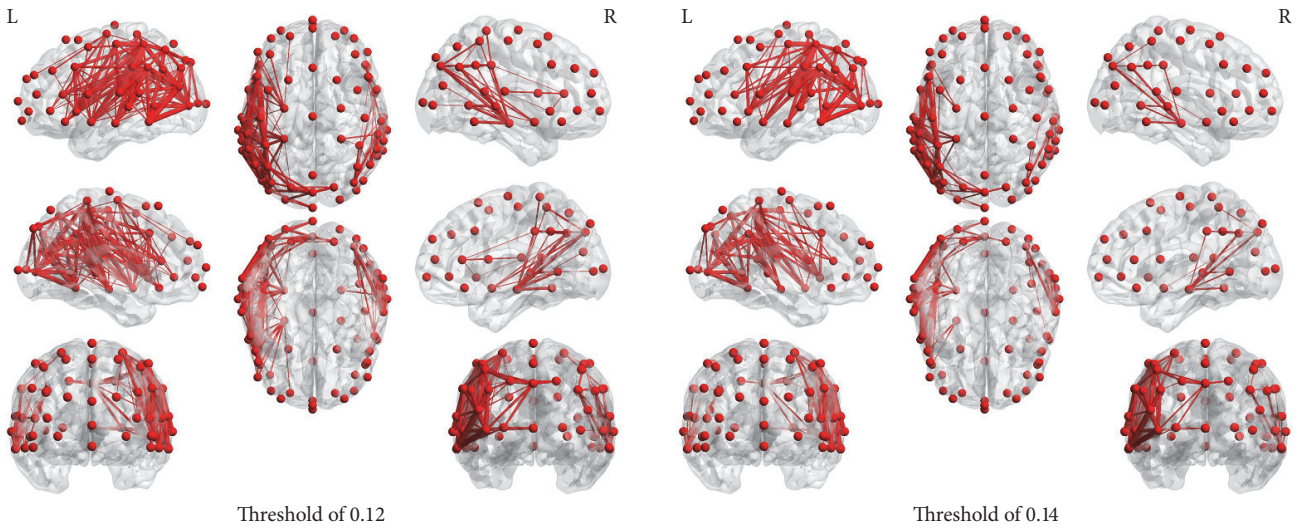


Figure 3: The results of the distribution which was the difference of the coherence in the theta band. Red nodes represent 72 electrodes, and the red lines between nodes show the difference in coherence at the threshold of 0.12 and 0.14. The thicker the red line is, the more the threshold is exceeded.

We calculated the coherence of each pair of the channels. The coherence matrices were plotted in Figure 2. The coherence matrices of both groups showed a complex but rather similar pattern, with various regions of high and low levels of the interdependence. There was no evident difference of the distribution for highlight areas in both groups. Compared with the healthy controls, the patients with MDD had more strengthened coherence values in some regions corresponding to the red areas. In order to explore the distribution of this difference, we obtained a difference matrix using the coherence matrix of the MDD group to subtract the coherence matrix of the healthy group. We plotted this matrix in a 3D graph as seen in Figure 3. The topographic maps were plotted by the threshold of 0.12 and 0.14, which could clearly indicate the difference between two groups in the theta band according to our above results. At the threshold of 0.12, most of the links were distributed in the left hemisphere of the brain except the front regions. There were also some links in the right-temporal region. At the threshold of 0.14, most of the links were distributed in the left hemisphere of the brain especially in the parietal and temporal regions. There were few links in the right hemisphere of the brain. Coherence analysis results showed that the depressed group had significantly higher coherence in the left hemisphere of the brain especially in parietal and temporal regions compared with the healthy controls in the theta band.

3.3. Characteristics of MST in Both Groups. In order to avoid the biase coming from the choice of the threshold, MST was used to construct the brain functional network. In one extreme, all nodes are connected to two other nodes, with the exception of the two nodes at either end, which have

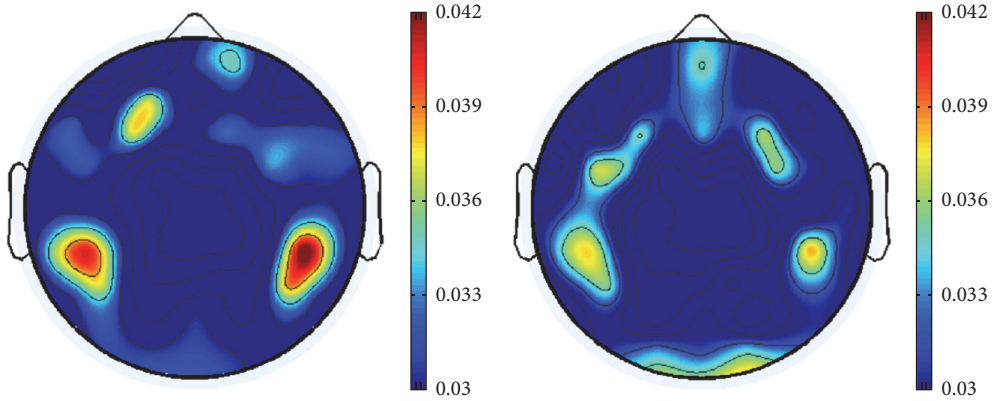


Figure 4: The brain topographic mappings of the degree fraction. The depth of the color represents the degree size of the nodes.

only one link. The other extreme is a star. In this star, there is one central node to which all other nodes are connected with one link. This tree configuration would have more leaf nodes. Between the two extremes of one path and one star, many different types of the tree configuration are possible. We calculated the leaf fraction for both groups. The leaf fraction of the MDD group was 0.5984 and for the healthy group was 0.5437 ($F = 3.1368$, $p = 0.0035$). Compared with the healthy group, MST of the MDD group tended to a star-network which had more leaf nodes. We also calculated the mean weight of all links. The mean weight of the MDD group was 0.8464 and for the healthy group was 0.7967 ($F = 2.2168$, $p = 0.0335$). Then the degree fraction of nodes was presented in the brain topographic mappings. Figure 4 showed the results. In the topographic map, we could see that clearly the degree fraction of the MDD group in the temporal regions in two hemispheres of the brain was higher than that of the healthy group. But the number of center nodes in the MDD group was less than that of the healthy group.

3.4. Hierarchical Clustering Analysis. TAHC algorithm was used to construct the hierarchical clustering of the brain functional network for both groups. Figure 5 showed the results. In the hierarchical clustering of the MDD group, it tended to cluster according to the physical structure of the brain, and the clusters corresponded to the left and right hemispheres. In the healthy group, the clustering tended to the functional structure of the brain, corresponding to the different functional areas. In order to explore the clusters detailedly, we plotted the distribution graph of the clusters in Figure 6. The hierarchical clustering was integrated into 6 clusters for the MDD group and the healthy group. We found a significant difference in the front region. In this region, there was a cluster which contained a large number of nodes in the distribution graph of the healthy group. However the MDD group had not.

4. Discussion

The results showed an increased global coherence of the MDD controls compared with the healthy controls in the theta band. In the left hemisphere of the brain, the MDD group had higher coherence, especially in the parietal and temporal regions. Several studies have reported that there was abnormal EEG absolute power, hemispheric asymmetry, or coherence with depression patients in some specific frequencies [51]. Hinrikus [52] found that the controls with MDD had the increased coherence between some brain regions. Tucker and Dawson [53] also reported that the parietal and temporal cortical regions were more activated in the left hemisphere of the depressive patients based on the coherence measure. In other frequency ranges, the increased coherence was also found. Fingelkurts et al. [54] reported the increased synchronization was observed in the EEG alpha and theta bands in the patients with MDD. And an increased topographic EEG coherence in the frontal brain areas in the MDD group in the EEG alpha, beta, and theta bands was reported by Leuchter's team [55]. However, the findings in the previous studies are inconsistent and sometimes even contradictory to each other. Knott et al. [19] found the decreased coherence in MDD subjects compared to normal controls. It may be due to the limited number of individual electrodes or the improper measurement method. For the increased global coherence of the MDD group, some researchers interpreted it as adaptive and compensatory mechanisms aimed to overcome the deficient semantic integration [55]. Our findings further support this adaptive and compensatory mechanism. However, it is not clear why the higher coherence was obtained in the left hemisphere of the brain. But the hemispheric asymmetry in the emotion processing has been observed, and the functional complementation of left and right hemispheres is important for adaptive emotion regulation [56]. Deficit of emotion regulation ability is the main manifestation of the patients with MDD [57]. So we guess it may be related to the dysfunction of the left hemisphere. Future studies about function and activation of the left and right hemispheres are needed to reach more definitive conclusions.

The functional network of the patients with MDD has been widely explored and studied. And a large number of great achievements have already been obtained. But, due to the differences in environments, subjects, and methods during the experiment and data analysis process, different

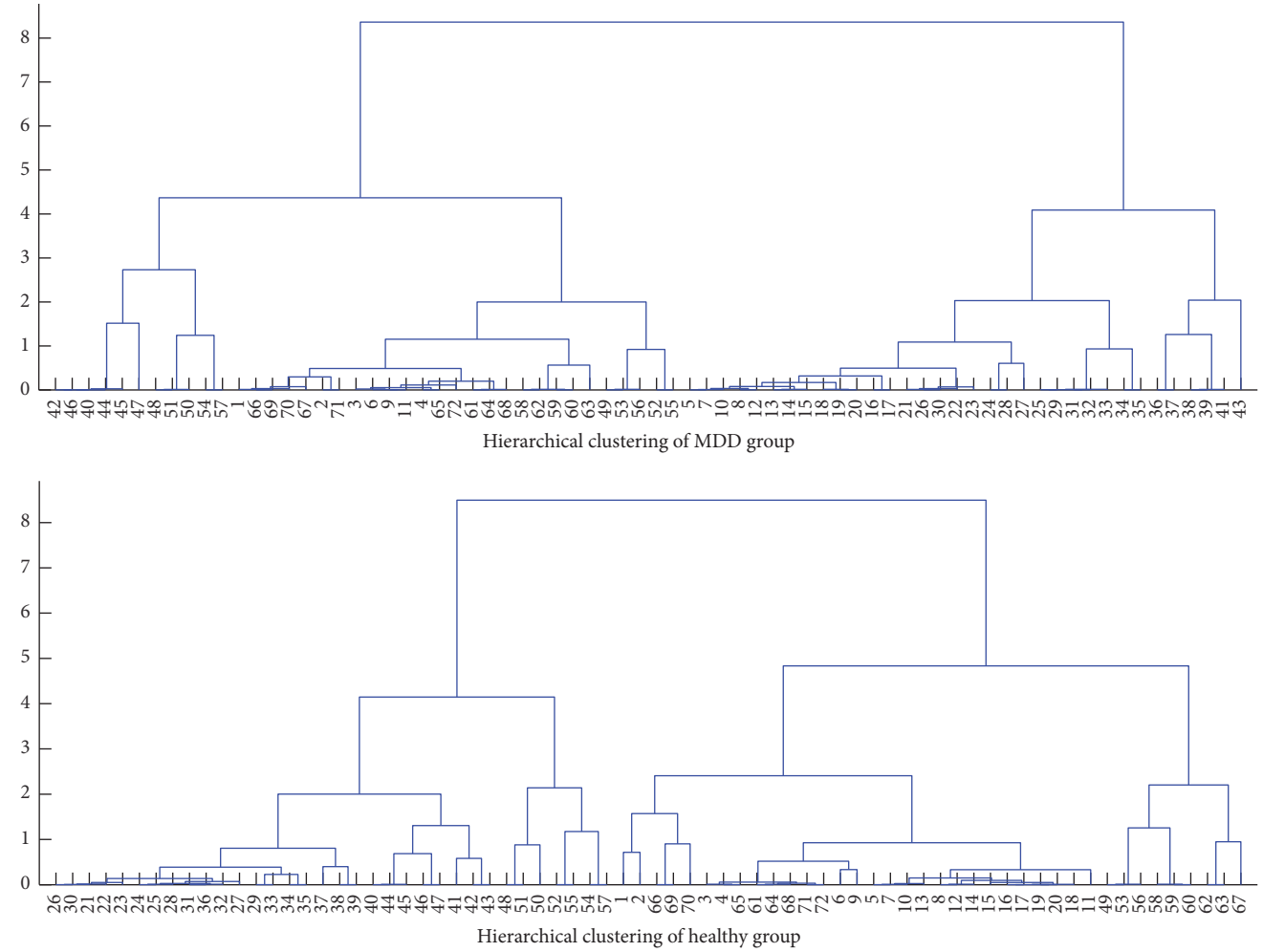


Figure 5: The hierarchical clustering graphs of the MDD group and the healthy group. The number in the graphs corresponds to the name of channels: 1-F10, 2-AF8, 3-AF4, 4-F2, 5-FCz, 6-FP2, 7-Fz, 8-FC1, 9-FPz, 10-AFz, 11-Cz, 12-F1, 13-FP1, 14-AF3, 15-F3, 16-AF7, 17-F5, 18-FC5, 19-FC3, 20-C1, 21-F9, 22-F7, 23-FT7, 24-C3, 25-CP1, 26-FT9, 27-T7, 28-C5, 29-CP3, 30-T9, 31-T3, 32-TP7, 33-CP5, 34-P5, 35-P3, 36-TP9, 37-T5, 38-P7, 39-P1, 40-Pz, 41-PO7, 42-PO3, 43-O1, 44-POz, 45-Oz, 46-PO4, 47-O2, 48-P2, 49CP2, 50-PO8, 51-P8, 52-P4, 53-CP4, 54-T6, 55-P6, 56-CP6, 57-TP10, 58-TP8, 59-C6, 60-C4, 61-C2, 62-T4, 63-T8, 64-FC4, 65-FC2, 66-T10, 67-FT8, 68-FC6, 69-FT10, 70-F8, 71-F6, and 72-F4.

or even opposite results were obtained in the research of the MDD brain functional networks. Some teams discovered increased brain functional connectivity in the patients with MDD [54, 58–60], and decreased brain functional connection was also found in Yang et al. [61] and Wang et al. [62] teams. For the functional connection, the choice of the threshold was commonly used to convert the coherence into functional connection. It is a reason that could not be ignored for emergence of the different conclusions. In this study, we used an unbiased method MST to construct the main brain functional connection. This method has been used in multiple mental disorders. As far as we know this was the first time for this method to be used for the depression disease. In a previous study, it suggested that more random networks showed low clustering and a short path length, corresponding to MSTs' shorter diameters and higher leaf numbers [40]. Our finding that leaf fraction of the MDD group was higher than the healthy group indicated a shift toward randomization in the brain networks of the MDD group. Similar conclusions have been mentioned in the study of brain functional networks. In sleep neuronal functional networks of depressed patients, Leistedt et al. [63] indicated the functional reorganization of depressed patients lost "small-world network" (SWN) characteristics. Zhang et al. [64] and Li et al. [7] teams also indicated the MDD patients showed a shift toward randomization in MDD brain networks compared with the healthy controls. For the MST, the higher the degree of a node is, the more important this node is during the brain information processing. Our finding about degree fraction indicated that temporal regions played an important role in information processing of the MDD group. Previous studies had the analogous conclusion. An EEG source location study suggested that larger degree fraction in temporal regions of the MDD subjects may be related to the dysregulated temporal pole activity [14].

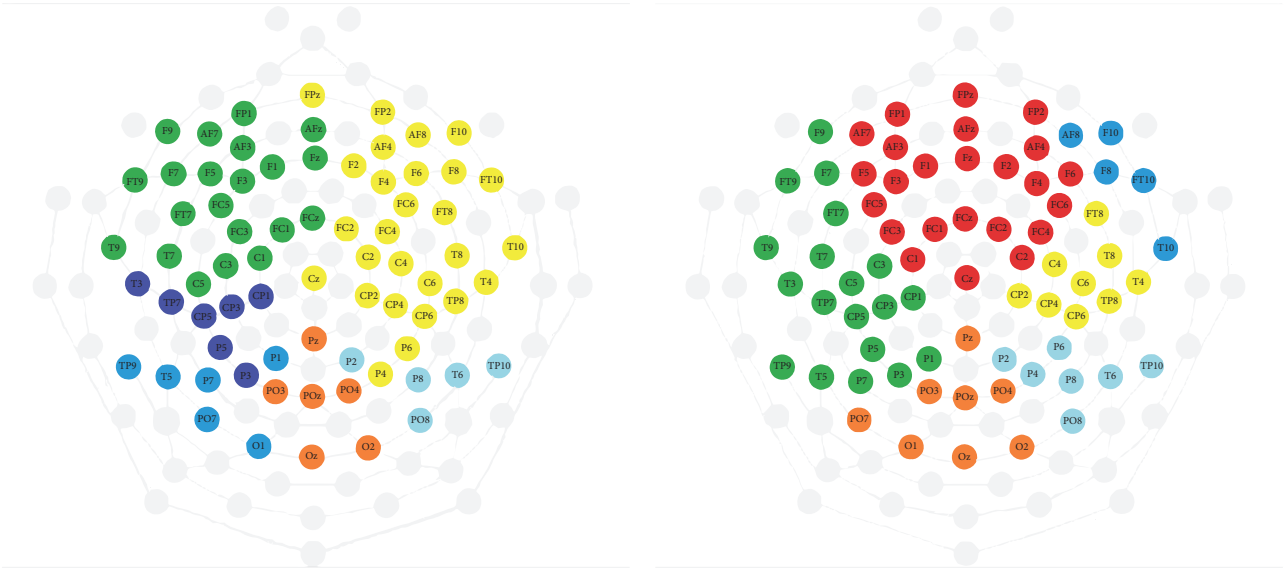


Figure 6: The distribution of the hierarchical clustering graphs under the condition of six clusters.

The prefrontal cortex mediates the control of high-level cognitive functions and is associated with the regulation of many aspects of the affective system [65]. This idea had been supported in neuroimaging studies. For the study of depression, the depression is known to involve a disturbance of mood and impaired cognitive functions [66, 67]. In previous studies, the frontal regions obtained more attention and exploration and played an important role in the development of the depression [68]. So far, many results reveal that major depression can be distinguished by specific histopathology of both neurons and glial cells in the prefrontal cortex [69]. Biver et al. [70] found frontal metabolic disturbances in the unipolar depression. Disruption of paralimbic pathways linking frontal cortex in secondary depression was indicated in [71]. But the most mentioned finding was the frontal brain asymmetry [72–75], which was described with greater activation in the right compared to the left frontal lobes [76]. The asymmetric frontal cortical activity in the MDD group had been widely presented not only in alpha band but also in the theta band [24, 77]. Besides EEG study, a combined MEG, PET, and rTMS Study also pointed out that prefrontal left-right functional imbalance and disrupted prefrontothalamic circuitry were plausible mechanisms for the depression [78]. Hierarchical clustering was a useful method to divide the brain functional network into several submodules. The nodes in the same submodule had a strong similarity, and the information processing among these nodes was very efficient. Importantly, the submodules typically corresponded to functional systems of the brain. In our study, we found that there was a cluster which contained a large number of nodes in the healthy group. However the MDD group had not, and the nodes in frontal regions of the MDD group were divided into two clusters in the left forehead and the right forehead, respectively. It was the performance of the forehead imbalance in patients with depression. This result indicated that there was a left-right functional imbalance in the frontal regions for MDD controls.

5. Conclusion

In conclusion, abnormally increased EEG coherence of the MDD group was found in the theta band, and the higher coherence was described in the left hemisphere of the brain especially in the parietal and temporal regions. An unbiased method of MST was used to construct the brain functional networks for the MDD group and the healthy group. The higher leaf fraction and mean weight were found in the MDD group. This finding indicated a shift toward randomization in the brain networks of the MDD group. Additionally, the hierarchical clustering opened up a new way for obtaining the characteristics of the brain functional network. The results that the MDD controls lose a frontal clustering indicated that the MDD group lacked the coordination in the forehead. A possible disadvantage of the MST approach is that it may miss the information about the network topology and it may contain the weaker connections in this brain functional network. Then, in the further study, we will try to use two levels' MST to construct the brain functional networks to overcome the above problems.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Basic Research Program of China (973 Program) (no. 2014CB744600), the National Natural Science Foundation of China (Grant nos.

61632014, 61210010, and 61402211), and the International Cooperation Project of Ministry of Science and Technology (no. 2013DFA11140).

References

[1] Z. I. Santini, A. Koyanagi, S. Tyrovolas, C. Mason, and J. M. Haro, "The association between social relationships and depression: a systematic review," *Journal of Affective Disorders*, vol. 175, pp. 53–65, 2015.

[2] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," *The Lancet*, vol. 367, no. 9524, pp. 1747–1757, 2006.

[3] G. H. Brundtland, "From the World Health Organization. Mental health: new understanding, new hope.," *Journal of the American Medical Association*, vol. 286, no. 19, p. 2391, 2001.

[4] D. A. Pizzagalli, "Frontocingulate dysfunction in depression: toward biomarkers of treatment response," *Neuropsychopharmacology*, vol. 36, no. 1, pp. 183–206, 2011.

[5] T. Shen, C. Li, B. Wang et al., "Increased cognition connectivity network in major depression disorder: a FMRI study," *Psychiatry Investigation*, vol. 12, no. 2, pp. 227–234, 2015.

[6] J. J. B. Allen and S. J. Reznik, "Frontal EEG asymmetry as a promising marker of depression vulnerability: Summary and methodological considerations," *Current Opinion in Psychology*, vol. 4, pp. 93–97, 2015.

[7] Y. Li, D. Cao, L. Wei, Y. Tang, and J. Wang, "Abnormal functional connectivity of EEG gamma band in patients with depression during emotional face processing," *Clinical Neurophysiology*, vol. 126, no. 11, pp. 2078–2089, 2015.

[8] F. Sambataro, N. D. Wolf, M. Pennuto, N. Vasic, and R. C. Wolf, "Revisiting default mode network function in major depression: Evidence for disrupted subsystem connectivity," *Psychological Medicine*, vol. 44, no. 10, pp. 2041–2051, 2014.

[9] O. Sporns, "The human connectome: a complex network," *Annals of the New York Academy of Sciences*, vol. 1224, no. 1, pp. 109–125, 2011.

[10] B. Güntekin and E. Başar, "A review of brain oscillations in perception of faces and emotional pictures," *Neuropsychologia*, vol. 58, no. 1, pp. 33–51, 2014.

[11] R. W. Thatcher and J. F. Lubar, "History of the scientific standards of QEEG normative databases," in *Introduction to Quantitative EEG and Neurofeedback: Advanced Theory and Applications*, pp. 29–59, 2009.

[12] J. R. Hughes and E. R. John, "Conventional and quantitative electroencephalography in psychiatry," *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 11, no. 2, pp. 190–208, 1999.

[13] Y. Li, S. Tong, D. Liu et al., "Abnormal EEG complexity in patients with schizophrenia and depression," *Clinical Neurophysiology*, vol. 119, no. 6, pp. 1232–1241, 2008.

[14] X. Li, B. Hu, T. Xu, J. Shen, and M. Ratcliffe, "A study on EEG-based brain electrical source of mild depressed subjects," *Computer Methods and Programs in Biomedicine*, vol. 120, no. 3, pp. 135–141, 2015.

[15] A. A. Fingelkurts and A. A. Fingelkurts, "Altered structure of dynamic electroencephalogram oscillatory pattern in major depression," *Biological Psychiatry*, vol. 77, no. 12, pp. 1050–1060, 2015.

[16] W. Klimesch, "Interindividual differences in oscillatory EEG activity and cognitive performance," in *The Cognitive Neuroscience of Individual Differences*, BIS, Amsterdam, Netherlands, 2003.

[17] F. L. Da Silva, "The generation of electric and magnetic signals of the brain by local networks," in *Comprehensive Human Physiology*, pp. 509–531, Springer, 1996.

[18] N. Jaworska, P. Blier, W. Fusee, and V. Knott, "Alpha power, alpha asymmetry and anterior cingulate cortex activity in depressed males and females," *Journal of Psychiatric Research*, vol. 46, no. 11, pp. 1483–1491, 2012.

[19] V. Knott, C. Mahoney, S. Kennedy, and K. Evans, "EEG power, frequency, asymmetry and coherence in male depression," *Psychiatry Research—Neuroimaging*, vol. 106, no. 2, pp. 123–140, 2001.

[20] C. Nystrom, M. Matousek, and T. Hallstrom, "Relationships between EEG and clinical characteristics in major depressive disorder," *Acta Psychiatrica Scandinavica*, vol. 73, no. 4, pp. 390–394, 1986.

[21] J.-S. Lee, B.-H. Yang, J.-H. Lee, J.-H. Choi, I.-G. Choi, and S.-B. Kim, "Detrended fluctuation analysis of resting EEG in depressed outpatients and healthy controls," *Clinical Neurophysiology*, vol. 118, no. 11, pp. 2489–2496, 2007.

[22] A. A. Fingelkurts, A. A. Fingelkurts, H. Rytsälä, K. Suominen, E. Isometsä, and S. Kähkönen, "Composition of brain oscillations in ongoing EEG during major depression disorder," *Neuroscience Research*, vol. 56, no. 2, pp. 133–144, 2006.

[23] K. E. Stephan, J. J. Riera, G. Deco, and B. Horwitz, "The brain connectivity workshops: moving the frontiers of computational systems neuroscience," *NeuroImage*, vol. 42, no. 1, pp. 1–9, 2008.

[24] M. Ertl, M. Hildebrandt, K. Ourina, G. Leicht, and C. Mulert, "Emotion regulation by cognitive reappraisal—the role of frontal theta oscillations," *NeuroImage*, vol. 81, pp. 412–421, 2013.

[25] M. Murias, S. J. Webb, J. Greenson, and G. Dawson, "Resting state cortical connectivity reflected in EEG coherence in individuals with autism," *Biological Psychiatry*, vol. 62, no. 3, pp. 270–273, 2007.

[26] A. R. Clarke, R. J. Barry, P. C. L. Heaven, R. McCarthy, M. Selikowitz, and M. K. Byrne, "EEG coherence in adults with attention-deficit/hyperactivity disorder," *International Journal of Psychophysiology*, vol. 67, no. 1, pp. 35–40, 2008.

[27] R. Armitage, R. F. Hoffmann, and A. J. Rush, "Biological rhythm disturbance in depression: temporal coherence of ultradian sleep EEG rhythms," *Psychological Medicine*, vol. 29, no. 6, pp. 1435–1448, 1999.

[28] F. H. Duffy, A. Shankardass, G. B. McAnulty, and H. Als, "The relationship of Asperger's syndrome to autism: a preliminary EEG coherence study," *BMC Medicine*, vol. 11, no. 175, 2013.

[29] F.-J. Hsiao, W.-T. Chen, Y.-J. Wang, S.-H. Yan, and Y.-Y. Lin, "Altered source-based EEG coherence of resting-state sensorimotor network in early-stage Alzheimer's disease compared to mild cognitive impairment," *Neuroscience Letters*, vol. 558, pp. 47–52, 2014.

[30] C. J. Stam, P. Tewarie, E. Van Dellen, E. C. W. van Straaten, A. Hillebrand, and P. Van Mieghem, "The trees and the forest: characterization of complex brain networks with minimum spanning trees," *International Journal of Psychophysiology*, vol. 92, no. 3, pp. 129–138, 2014.

[31] D. J. A. Smit, C. J. Stam, D. Posthuma, D. I. Boomsma, and E. J. C. De Geus, "Heritability of "small-world" networks in the

brain: a graph theoretical analysis of resting-state EEG functional connectivity," *Human Brain Mapping*, vol. 29, no. 12, pp. 1368–1378, 2008.

[32] W. de Haan, Y. A. L. Pijnenburg, R. L. M. Strijers et al., "Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory," *BMC Neuroscience*, vol. 10, no. 101, 2009.

[33] C. J. Stam, B. F. Jones, G. Nolte, M. Breakspear, and P. Scheltens, "Small-world networks and functional connectivity in Alzheimer's disease," *Cerebral Cortex*, vol. 17, no. 1, pp. 92–99, 2007.

[34] J. Zhang, W. Cheng, Z. Wang et al., "Pattern classification of large-scale functional brain networks: identification of informative neuroimaging markers for epilepsy," *PLoS ONE*, vol. 7, no. 5, Article ID e36733, 2012.

[35] E. van Diessen, W. M. Otte, K. P. J. Braun, C. J. Stam, and F. E. Jansen, "Improved diagnosis in children with partial epilepsy using a multivariable prediction model based on EEG network characteristics," *PLoS ONE*, vol. 8, no. 4, Article ID e59764, 2013.

[36] M.-T. Kuhnert, C. Geier, C. E. Elger, and K. Lehnertz, "Identifying important nodes in weighted functional brain networks: a comparison of different centrality approaches," *Chaos. An Interdisciplinary Journal of Nonlinear Science*, vol. 22, no. 2, 023142, 7 pages, 2012.

[37] G. Ansmann and K. Lehnertz, "Surrogate-assisted analysis of weighted functional brain networks," *Journal of Neuroscience Methods*, vol. 208, no. 2, pp. 165–172, 2012.

[38] B. C. M. van Wijk, C. J. Stam, and A. Daffertshofer, "Comparing brain networks of different size and connectivity density using graph theory," *PLoS ONE*, vol. 5, no. 10, Article ID e13701, 2010.

[39] M. Mareš, "The saga of minimum spanning trees," *Computer Science Review*, vol. 2, no. 3, pp. 165–221, 2008.

[40] M. Boersma, D. J. Smit, D. I. Boomsma, E. J. De Geus, H. A. Delemarre-van de Waal, and C. J. Stam, "Growing trees in child brains: graph theoretical analysis of electroencephalography-derived minimum spanning tree in 5- and 7-year-old children reflects brain maturation," *Brain Connectivity*, vol. 3, no. 1, pp. 50–60, 2013.

[41] P. Tewarie, A. Hillebrand, M. M. Schoonheim et al., "Functional brain network analysis using minimum spanning trees in Multiple Sclerosis: an MEG source-space study," *NeuroImage*, vol. 88, pp. 308–318, 2014.

[42] K. Çiftçi, "Minimum spanning tree reflects the alterations of the default mode network during alzheimer's disease," *Annals of Biomedical Engineering*, vol. 39, no. 5, pp. 1493–1504, 2011.

[43] M. Yu, A. Hillebrand, P. Tewarie et al., "Hierarchical clustering in minimum spanning trees," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 2, p. 023107, 2015.

[44] D. Meunier, R. Lambiotte, and E. T. Bullmore, "Modular and hierarchically modular organization of brain networks," *Frontiers in Neuroscience*, vol. 4, no. 200, 2010.

[45] O. Sporns and R. F. Betzel, "Modular brain networks," *Annual Review of Psychology*, vol. 67, pp. 613–640, 2016.

[46] W. De Haan, W. M. Van der Flier, T. Koene, L. L. Smits, P. Scheltens, and C. J. Stam, "Disrupted modular brain dynamics reflect cognitive dysfunction in Alzheimer's disease," *NeuroImage*, vol. 59, no. 4, pp. 3085–3093, 2012.

[47] A. F. Alexander-Bloch, N. Gogtay, D. Meunier et al., "Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia," *Frontiers in Systems Neuroscience*, vol. 4, article no. 147, 2010.

[48] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[49] J. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical Society*, vol. 7, pp. 48–50, 1956.

[50] J.-C. Chen, "Dijkstras shortest path algorithm," *Journal of Formalized Mathematics*, vol. 15, pp. 144–157, 2003.

[51] C. Tas, M. Cebi, O. Tan, G. Hizli-Sayar, N. Tarhan, and E. C. Brown, "EEG power, cordance and coherence differences between unipolar and bipolar depression," *Journal of Affective Disorders*, vol. 172, pp. 184–190, 2014.

[52] H. Hinrikus, "Electroencephalographic spectral asymmetry index for detection of depression," *Medical & Biological Engineering & Computing*, vol. 47, no. 12, p. 1291, 2009.

[53] D. M. Tucker and S. L. Dawson, "Asymmetric EEG changes as method actors generated emotions," *Biological Psychology*, vol. 19, no. 1, pp. 63–75, 1984.

[54] A. A. Fingelkurts, A. A. Fingelkurts, H. Rytsälä, K. Suominen, E. Isometsä, and S. Kähkönen, "Impaired functional connectivity at EEG alpha and theta frequency bands in major depression," *Human Brain Mapping*, vol. 28, no. 3, pp. 247–261, 2007.

[55] A. F. Leuchter, I. A. Cook, A. M. Hunter, C. Cai, and S. Horvath, "Resting-state quantitative electroencephalography reveals increased neurophysiologic connectivity in depression," *PLoS ONE*, vol. 7, no. 2, Article ID e32508, 2012.

[56] R. J. Davidson, "Anxiety and affective style: role of prefrontal cortex and amygdala," *Biological Psychiatry*, vol. 51, no. 1, pp. 68–80, 2002.

[57] J. Joormann and I. H. Gotlib, "Emotion regulation in depression: relation to cognitive inhibition," *Cognition and Emotion*, vol. 24, no. 2, pp. 281–298, 2010.

[58] Y. I. Sheline, J. L. Price, Z. Yan, and M. A. Mintun, "Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 24, pp. 11020–11025, 2010.

[59] Y. Zhou, C. Yu, H. Zheng et al., "Increased neural resources recruitment in the intrinsic organization in major depression," *Journal of Affective Disorders*, vol. 121, no. 3, pp. 220–230, 2010.

[60] M. D. Greicius, B. H. Flores, and V. Menon, "Resting-state functional connectivity in major depressionn: abnormally increased contributions from subgenual cingulate cortex and thalamus," *Biological Psychiatry*, vol. 62, no. 5, pp. 429–437, 2007.

[61] C. G. Connolly, J. Wu, T. C. Ho et al., "Resting-state functional connectivity of subgenual anterior cingulate cortex in depressed adolescents," *Biological Psychiatry*, vol. 74, no. 12, pp. 898–907, 2013.

[62] E. Alalade, K. Denny, G. Potter, D. Steffens, and L. Wang, "Altered cerebellar-cerebral functional connectivity in geriatric depression," *PLoS ONE*, vol. 6, no. 5, Article ID e20035, 2011.

[63] S. J. J. Leistedt, N. Coumans, M. Dumont, J.-P. Lanquart, C. J. Stam, and P. Linkowski, "Altered sleep brain functional connectivity in acutely depressed patients," *Human Brain Mapping*, vol. 30, no. 7, pp. 2207–2219, 2009.

[64] J. Zhang, J. Wang, Q. Wu et al., "Disrupted brain connectivity networks in drug-naive, first-episode major depressive disorder," *Biological Psychiatry*, vol. 70, no. 4, pp. 334–342, 2011.

[65] N. T. Alves, S. S. Fukusima, and J. A. Aznar-Casanova, "Models of brain asymmetry in emotional processing," *Psychology & Neuroscience*, vol. 1, no. 1, p. 63, 2008.

[66] V. K. Sharma, S. Das, S. Mondal, U. Goswami, and A. Gandhi, "Effect of Sahaj Yoga on neuro-cognitive functions in patients suffering from major depression," *Indian Journal of Physiology and Pharmacology*, vol. 50, no. 4, pp. 375–383, 2006.

[67] M.-P. Austin, P. Mitchell, and G. M. Goodwin, "Cognitive deficits in depression: possible implications for functional neuropathology," *British Journal of Psychiatry*, vol. 178, pp. 200–206, 2001.

[68] T.-J. Lai, M. E. Payne, C. E. Byrum, D. C. Steffens, and K. R. R. Krishnan, "Reduction of orbital frontal cortex volume in geriatric depression," *Biological Psychiatry*, vol. 48, no. 10, pp. 971–975, 2000.

[69] G. Rajkowska, J. J. Miguel-Hidalgo, J. Wei et al., "Morphometric evidence for neuronal and glial prefrontal cell pathology in major depression," *Biological Psychiatry*, vol. 45, no. 9, pp. 1085–1098, 1999.

[70] F. Biver, S. Goldman, V. Delvenne et al., "Frontal and parietal metabolic disturbances in unipolar depression," *Biological Psychiatry*, vol. 36, no. 6, pp. 381–388, 1994.

[71] H. S. Mayberg, "Frontal lobe dysfunction in secondary depression," *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 6, no. 4, pp. 428–442, 1994.

[72] J. Tomarkenand and A. A. D. Keener, "Frontal brain asymmetry and depression: a self-regulatory perspective," *Cognition & Emotion*, vol. 12, no. 3, pp. 387–420, 1998.

[73] J. J. B. Allen, H. L. Urry, S. K. Hitt, and J. A. Coan, "The stability of resting frontal electroencephalographic asymmetry in depression," *Psychophysiology*, vol. 41, no. 2, pp. 269–280, 2004.

[74] C. E. Schaffer, R. J. Davidson, and C. Saron, "Frontal and parietal electroencephalogram asymmetry in depressed and nondepressed subjects," *Biological Psychiatry*, vol. 18, no. 7, pp. 753–762, 1983.

[75] A. J. Tomarken, G. S. Dichter, J. Garber, and C. Simien, "Resting frontal brain activity: Linkages to maternal depression and socio-economic status among adolescents," *Biological Psychology*, vol. 67, no. 1-2, pp. 77–102, 2004.

[76] E. Jesulola, C. F. Sharpley, V. Bitsika, L. L. Agnew, and P. Wilson, "Frontal alpha asymmetry as a pathway to behavioural withdrawal in depression: research findings and issues," *Behavioural Brain Research*, vol. 292, pp. 56–67, 2015.

[77] L. I. Aftanas and S. A. Golocheikine, "Human anterior and frontal midline theta and lower alpha reflect emotionally positive state and internalized attention: high-resolution EEG investigation of meditation," *Neuroscience Letters*, vol. 310, no. 1, pp. 57–60, 2001.

[78] C.-T. Li, L.-F. Chen, P.-C. Tu et al., "Impaired prefronto-thalamic functional connectivity as a key feature of treatment-resistant depression: a combined MEG, PET and rTMS study," *PLoS ONE*, vol. 8, no. 8, Article ID e70089, 2013.

Hindawi
Complexity
Volume 2017, Article ID 4216797, 11 pages
https://doi.org/10.1155/2017/4216797

Research Article

Robust Nonnegative Matrix Factorization via Joint Graph Laplacian and Discriminative Information for Identifying Differentially Expressed Genes

Ling-Yun Dai, Chun-Mei Feng, Jin-Xing Liu, Chun-Hou Zheng, Jiguo Yu, and Mi-Xiao Hou

School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China

Correspondence should be addressed to Jin-Xing Liu; sdcavell@126.com and Chun-Hou Zheng; zhengch99@126.com

Received 17 January 2017; Accepted 6 March 2017; Published 6 April 2017

Academic Editor: Fang X. Wu

Differential expression plays an important role in cancer diagnosis and classification. In recent years, many methods have been used to identify differentially expressed genes. However, the recognition rate and reliability of gene selection still need to be improved. In this paper, a novel constrained method named robust nonnegative matrix factorization via joint graph Laplacian and discriminative information (GLD-RNMF) is proposed for identifying differentially expressed genes, in which manifold learning and the discriminative label information are incorporated into the traditional nonnegative matrix factorization model to train the objective matrix. Specifically, $L_{2,1}$-norm minimization is enforced on both the error function and the regularization term which is robust to outliers and noise in gene data. Furthermore, the multiplicative update rules and the details of convergence proof are shown for the new model. The experimental results on two publicly available cancer datasets demonstrate that GLD-RNMF is an effective method for identifying differentially expressed genes.

1. Introduction

Cancer is one of the most serious diseases that endanger the health of human being. Millions of people die of cancer every year. With the development of gene sequencing technology and other gene detection technologies, huge gene data have been generated [1, 2]. Therefore, it is important and challenging for scientists to find pathogenic genes from a large number of gene expression data. Microarray datasets on each chip usually contain many gene expression data, and the number of samples is far less than that of genes, which makes the identification of differentially expressed genes difficult [3]. In addition, irrelevant or noisy variables may reduce the accuracy of the results. In recent years, many effective mathematical methods have been applied to identify differentially expressed genes. For example, principal component analysis (PCA) [4, 5] and penalized matrix decomposition (PMD) [6] have been used to analyze gene expression data. Liu et al. used robust principal component analysis (RPCA) to discover differentially expressed genes [7]. Zheng et al. employed nonnegative matrix factorization (NMF) on the selection of tumor genes [8]. Cai et al. proposed an algorithm named graph regularized nonnegative matrix factorization (GNMF) for data representation [9]. Wang et al. used robust graph regularized nonnegative matrix factorization (RGNMF) for identifying differentially expressed genes [10]. A CIPMD (Class-Information-Based Penalized Matrix Decomposition) algorithm was proposed to identify the differentially expressed genes on RNA-Seq data, which introduced the class information via a total scatter matrix [11]. The Consensus Clustering methodology was proposed for microarray data analysis by Giancarlo and Utro [12].

However, two characteristics of gene expression data pose a serious challenge to the existing methods. Firstly, a large number of researchers hold that gene expression data probably reside in a low dimensional manifold embedded in a high dimensional ambient space. Therefore it is critical to consider the geometrical structure in the original gene expression data. Manifold learning is clearly an effective method to preserve the data geometric structure embedded in the original gene expression data [13, 14]. Cai et al. proposed GNMF [9], in which the geometrical structure of data

was constructed by an affinity graph. Another variant of NMF called manifold regularized discriminative nonnegative matrix factorization (MD-NMF) was also introduced [15]. MD-NMF considered both the local geometry of data and the discriminative information of different classes simultaneously. Long et al. proposed a method called graph regularized discriminative nonnegative matrix factorization (GDNMF) [16], in which both the geometrical structure and discriminative label information were considered in the objective function. Secondly, gene expression data often contain a lot of outliers and noise. However, existing methods cannot effectively eliminate outliers and noise. For example, least squares methods are sensitive to outliers and noise. In recent years, many researchers have been devoted to improving the robustness to outliers and noise. Zheng et al. proposed an algorithm named generalized hierarchical fuzzy C-means [17], which is robust to noise and outliers. Wang et al. used $L_{2,1}$-norm to reduce the effect of outliers and noise [10].

A novel algorithm, which we call robust nonnegative matrix factorization via joint graph Laplacian and discriminative information (GLD-RNMF), is proposed to overcome the aforementioned problems together. The proposed algorithm preserves the geometric structure of data space by constructing an affinity graph and improves the discriminative ability by the supervised label information. To do so, a new matrix decomposition objective function by integrating the geometric structure and label information is constructed. In addition, we employ $L_{2,1}$-norm instead of L_2-norm on the error function and the regularization term to reduce the influence of outliers and noise. For completeness, we present that the convergence proof of our iterative scheme is also shown in the Appendix. Experimental results indicate that the GLD-RNMF algorithm has better results than other existing algorithms for identifying differentially expressed genes.

The remainder of the paper is arranged as follows. In Section 2, we briefly introduce some relevant mathematical foundation and propose the GLD-RNMF algorithm in detail. In Section 3, the results of differentially expressed gene selection using our GLD-RNMF method and the other four methods (GNMF, NMFSC, RGNMF, and GDNMF) are shown for comparison. Finally, we conclude this paper in Section 4.

2. Materials and Methods

2.1. Mathematical Definition of $L_{2,1}$. The mathematical definition of $L_{2,1}$-norm [18] is

$$\|\mathbf{Y}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{s} \mathbf{y}_{ij}^2} = \sum_{i=1}^{n} \|\mathbf{y}_i\|_2, \tag{1}$$

where $\mathbf{y}_i$ is the ith row of $\mathbf{Y}$ and $\mathbf{Y}$ is $n \times s$ matrix. $L_{2,1}$-norm is interpreted as follows. Firstly, we compute L_2-norm of the vector $\mathbf{y}_i$ and then compute L_1-norm of vector $\mathbf{p}(\mathbf{Y}) = (\|\mathbf{Y}_1\|_2, \|\mathbf{Y}_2\|_2, \ldots, \|\mathbf{Y}_s\|_2)$. The value of the elements of vector $\mathbf{p}(\mathbf{Y})$ represents the importance of each dimension. $L_{2,1}$-norm enables the vector $\mathbf{p}(\mathbf{Y})$ sparse to achieve the purpose of dimension reduction.

2.2. Manifold Learning. The purpose of this work is to get the best approximation of the original data. We also hope that the new representation can respect the intrinsic Riemannian structure. Recently, many researchers hold that high dimensional data often reside on a much lower dimensional manifold. The "manifold assumption" could be that data points nearby in the intrinsic geometry structure are also close under the new basis. Therefore, they usually have similar characteristics and can be categorized into the same class. In this paper, we employ manifold learning to achieve the aforementioned goal.

For a graph with N vertices, each vertex corresponds to a data point. For each data point, we can find its k nearest neighbors and connect it with the neighbors. There are many ways to define the weight matrix $\mathbf{W}$ on the graph, for example, 0-1 weighting, heat kernel weighting, and dot-product weighting. Considering that 0-1 weighting is the simplest and easy to compute, we choose 0-1 weighting as the measure in this paper.

0-1 Weight. $\mathbf{W}_{ij} = 1$, if and only if two nodes i and j are connected by an edge. That is,

$$\mathbf{W}_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathbf{N}_k\left(x_j\right) \text{ or } \mathbf{x}_j \in \mathbf{N}_k\left(x_i\right), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\mathbf{N}_k(\mathbf{x}_j)$ consists of k nearest neighbors of $\mathbf{x}_j$ and the neighbors have the same label with $\mathbf{x}_j$.

Therefore, the smoothness of the dimensional representation can be measured as follows:

$$\begin{aligned} R &= \frac{1}{2}\sum_{i,j=1}^{n} \left\|\mathbf{s}_i - \mathbf{s}_j\right\|^2 \mathbf{W}_{ij} = \sum_{i=1}^{n} \mathbf{s}_i^T \mathbf{s}_i \mathbf{B}_{ii} - \sum_{i,j=1}^{n} \mathbf{s}_i^T \mathbf{s}_j \mathbf{W}_{ij} \\ &= \operatorname{tr}\left(\mathbf{G}^T \mathbf{B}\mathbf{G}\right) - \operatorname{tr}\left(\mathbf{G}^T \mathbf{W}\mathbf{G}\right) = \operatorname{tr}\left(\mathbf{G}^T \mathbf{L}\mathbf{G}\right), \end{aligned} \tag{3}$$

where $\operatorname{tr}(\cdot)$ represents the trace of a matrix. $\mathbf{B}$ is a diagonal matrix and $\mathbf{B}_{ii}$ is the row sum (or column, because $\mathbf{W}$ is symmetric, $\mathbf{W} \in R^{n\times n}$) of $\mathbf{W}$; that is, $\mathbf{B}_{ii} = \sum_{j=1}^{n} \mathbf{W}_{ij}$. $\mathbf{L} = \mathbf{B} - \mathbf{W}$ is graph Laplacian matrix and $\mathbf{L} \in R^{n\times n}$. We measure the distance of two points in the low dimensional space by the Euclidean distance $R(\mathbf{s}_i, \mathbf{s}_j) = \|\mathbf{s}_i - \mathbf{s}_j\|^2$.

2.3. Nonnegative Matrix Factorization (NMF). We review the standard NMF in this section. Although the algorithm has been widely used in many aspects, there are still many shortcomings.

Given n nonnegative samples $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ in R^m, arranged in columns of a matrix $\mathbf{X} \in R^{m\times n}$, in this paper, each row of $\mathbf{X}$ represents the transcriptional response of the n genes in one sample and each column of $\mathbf{X}$ represents the expression level of a gene across all samples. Letting matrices $\mathbf{X} \in R^{m\times n}$, $\mathbf{V} \in R^{m\times k}$, and $\mathbf{H} \in R^{k\times n}$, NMF decomposes $\mathbf{X}$ into the product of $\mathbf{V}$ and $\mathbf{H}$; that is, $\mathbf{X} \approx \mathbf{V}\mathbf{H}$.

To ensure an approximate factorization $\mathbf{X} \approx \mathbf{V}\mathbf{H}$, two update rules are introduced [19]. One of the objective functions is constructed by minimizing the square of the

Euclidean distance between $\mathbf{X}$ and $\mathbf{VH}$. The optimization problem is described as follows:

$$\begin{aligned} \min_{\mathbf{V},\mathbf{H}} \quad & \|\mathbf{X} - \mathbf{VH}\|_F^2 \\ \text{s.t.} \quad & \mathbf{V} \geq 0, \\ & \mathbf{H} \geq 0, \end{aligned} \tag{4}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm. The corresponding optimization rules are as follows:

$$\begin{aligned} \mathbf{H}_{qj} &\longleftarrow \mathbf{H}_{qj} \frac{\left(\mathbf{V}^T\mathbf{X}\right)_{qj}}{\left(\mathbf{V}^T\mathbf{VH}\right)_{qj}}, \\ \mathbf{V}_{iq} &\longleftarrow \mathbf{V}_{iq} \frac{\left(\mathbf{XH}^T\right)_{iq}}{\left(\mathbf{VHH}^T\right)_{iq}}. \end{aligned} \tag{5}$$

The convergence of the above optimization rules has been proven [19].

2.4. Graph Regularized Discriminative Nonnegative Matrix Factorization (GDNMF). Supervised label information is added to the objective function of GNMF [16]. The definition and iterative rules of GDNMF are presented below.

Class indicator matrix $\mathbf{S} \in R^{c\times n}$ is defined as follows:

$$\mathbf{S}_{ij} = \begin{cases} 1, & \text{if } \mathbf{y}_j = i,\ j = 1, 2, \ldots, n;\ i = 1, 2, \ldots, c, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $\mathbf{y}_j \in \{1, 2, \ldots, c\}$ is the class label of $\mathbf{x}_j$ and c is the total number of classes in $\mathbf{X}$.

The objective function of GDNMF is formulated as follows:

$$\begin{aligned} \min_{\mathbf{V},\mathbf{H}} \quad & \|\mathbf{X} - \mathbf{VH}\|_F^2 + \beta \operatorname{tr}\left(\mathbf{HLH}^T\right) + \alpha \|\mathbf{S} - \mathbf{AH}\|_F^2 \\ \text{s.t.} \quad & \mathbf{V} \geq 0, \\ & \mathbf{H} \geq 0, \\ & \mathbf{A} \geq 0. \end{aligned} \tag{7}$$

The corresponding optimization rules are as follows:

$$\begin{aligned} \mathbf{H}_{qj} &\longleftarrow \mathbf{H}_{qj} \frac{\left(\mathbf{V}^T\mathbf{X} + \alpha\mathbf{A}^T\mathbf{S} + \beta\mathbf{HW}\right)_{qj}}{\left(\mathbf{V}^T\mathbf{VH} + \alpha\mathbf{A}^T\mathbf{AH} + \beta\mathbf{HB}\right)_{qj}}, \\ \mathbf{V}_{iq} &\longleftarrow \mathbf{V}_{iq} \frac{\left(\mathbf{XH}^T\right)_{iq}}{\left(\mathbf{VHH}^T\right)_{iq}}, \end{aligned} \tag{8}$$

where $\mathbf{A} \in R^{c\times k}$ is initialized to a random nonnegative matrix in the algorithm. α and β are nonnegative regularization parameters, respectively. Essentially, GDNMF incorporates the graph Laplacian and supervised label information into the objective function of NMF, which ensures the algorithm to keep consistent with the intuitive geometric structure of the data and improves the discriminative power of different classes.

2.5. Robust Nonnegative Matrix Factorization via Joint Graph Laplacian and Discriminative Information (GLD-RNMF)

2.5.1. The Objective Function. For the purpose of dimension reduction, NMF represents original data $\mathbf{X} \in R^{m\times n}$ by a product of a nonnegative matrix $\mathbf{V} \in R^{m\times k}$ and coefficient matrix $\mathbf{H} \in R^{k\times n}$. The approximation error is calculated according to the squared residuals; that is, $\|\mathbf{X} - \mathbf{VH}\|_F^2 = \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{Vh}_i\|^2$. Due to the squared term in the objective function, smaller outliers can lead to larger errors. In this paper, we enforce $L_{2,1}$-norm constraint on the objective function to reduce the impact of outliers and noise.

By employing $L_{2,1}$-norm on GDNMF model, we can formulate the objective function of GLD-RNMF as follows:

$$\begin{aligned} \min_{\mathbf{V},\mathbf{H}} \quad & \|\mathbf{X} - \mathbf{VH}\|_{2,1} + \beta \operatorname{tr}\left(\mathbf{HLH}^T\right) + \alpha \|\mathbf{S} - \mathbf{AH}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{V} \geq 0, \\ & \mathbf{H} \geq 0, \\ & \mathbf{A} \geq 0. \end{aligned} \tag{9}$$

This objective function can solve high dimensional, negative, noisy and sparse data simultaneously, keep consistent with the intuitive geometric structure of data, and improve the discriminative power of different classes.

2.5.2. The Multiplication Update Rules of GLD-RNMF. Although the objective function is not convex jointly about $(\mathbf{V}, \mathbf{H}, \mathbf{A})$, it is convex in regard to one of variables in $(\mathbf{V}, \mathbf{H}, \mathbf{A})$ when the others are fixed. The objective function can be expanded as follows:

$$\begin{aligned} J &= \operatorname{tr}\left((\mathbf{X} - \mathbf{VH})\,\mathbf{Q}\,(\mathbf{X} - \mathbf{VH})^T\right) + \beta \operatorname{tr}\left(\mathbf{HLH}^T\right) \\ &\quad + \alpha \operatorname{tr}\left((\mathbf{S} - \mathbf{AH})\,\mathbf{Q}\,(\mathbf{S} - \mathbf{AH})^T\right) \\ &= \operatorname{tr}\left(\mathbf{XQX}^T\right) - 2\operatorname{tr}\left(\mathbf{XQH}^T\mathbf{V}^T\right) + \operatorname{tr}\left(\mathbf{VHQH}^T\mathbf{V}^T\right) \\ &\quad + \beta\left(\mathbf{HLH}^T\right) + \alpha \operatorname{tr}\left(\mathbf{SGS}^T\right) - 2\alpha \operatorname{tr}\left(\mathbf{SGH}^T\mathbf{A}^T\right) \\ &\quad + \alpha \operatorname{tr}\left(\mathbf{AHGH}^T\mathbf{A}^T\right), \end{aligned} \tag{10}$$

where $\mathbf{Q}$ and $\mathbf{G}$ both are diagonal matrices and the diagonal elements are as follows:

$$\mathbf{Q}_{jj} = \frac{1}{\sqrt{\sum_{i=1}^m (\mathbf{X} - \mathbf{VH})_{ij} + \varepsilon}}, \tag{11}$$

$$\mathbf{G}_{jj} = \frac{1}{\sqrt{\sum_{i=1}^m (\mathbf{S} - \mathbf{AH})_{ij} + \varepsilon}}, \tag{12}$$

in which ε is an infinitesimal positive number.

In order to solve the optimization problem in (9), we introduce the Lagrange multipliers Φ, Ψ, and Ω for $\mathbf{V}$, $\mathbf{H}$, and

A, respectively. Firstly, we formulate the Lagrange function of GLD-RNMF as follows:

$$\begin{aligned} L_J = {} & \operatorname{tr}\left(\mathbf{XQX}^T\right) - 2\operatorname{tr}\left(\mathbf{XQH}^T\mathbf{V}^T\right) \\ & + \operatorname{tr}\left(\mathbf{VHQH}^T\mathbf{V}^T\right) + \beta\left(\mathbf{HLH}^T\right) + \alpha\operatorname{tr}\left(\mathbf{SGS}^T\right) \\ & - 2\alpha\operatorname{tr}\left(\mathbf{SGH}^T\mathbf{A}^T\right) + \alpha\operatorname{tr}\left(\mathbf{AHGH}^T\mathbf{A}^T\right) \\ & + \operatorname{tr}(\Phi\mathbf{V}) + \operatorname{tr}(\Psi\mathbf{H}) + \operatorname{tr}(\Omega\mathbf{A}). \end{aligned} \tag{13}$$

Taking the partial derivatives of L_J with respect to **V**, **A**, and **H** and setting them to zero and in view of $\operatorname{tr}(\mathbf{XY}) = \operatorname{tr}(\mathbf{YX})$ and $\operatorname{tr}(\mathbf{X}^T) = \operatorname{tr}(\mathbf{X})$, we get

$$\begin{aligned} \frac{\partial L_J}{\partial \mathbf{V}} &= -2\mathbf{XQH}^T + 2\mathbf{VHQH}^T + \Phi, \\ \frac{\partial L_J}{\partial \mathbf{A}} &= -2\alpha\mathbf{SGH}^T + 2\alpha\mathbf{AHGH}^T + \Omega, \\ \frac{\partial L_J}{\partial \mathbf{H}} &= -2\mathbf{V}^T\mathbf{XQ} + 2\mathbf{V}^T\mathbf{VHQ} + 2\beta\mathbf{HL} \\ &\quad + \alpha\left[-2\mathbf{A}^T\mathbf{SG} + 2\mathbf{A}^T\mathbf{AHG}\right] + \Psi. \end{aligned} \tag{14}$$

According to the KKT (Karush-Kuhn-Tucker) conditions [20], that is, $\Phi_{iq}\mathbf{V}_{iq} = 0$, $\Omega_{kq}\mathbf{A}_{kq} = 0$, and $\Psi_{qj}\mathbf{H}_{qj} = 0$, we can obtain the following equations:

$$\begin{aligned} & \left[-2\mathbf{XQH}^T + 2\mathbf{VHQH}^T\right]_{iq}\mathbf{V}_{iq} + \Phi_{iq}\mathbf{V}_{iq} = 0, \\ & \left[-2\alpha\mathbf{SGH}^T + 2\alpha\mathbf{AHGH}^T\right]_{kq}\mathbf{A}_{kq} + \Omega_{kq}\mathbf{A}_{kq} = 0, \\ & \left[-2\mathbf{V}^T\mathbf{XQ} + 2\mathbf{V}^T\mathbf{VHQ} + 2\beta\mathbf{HL}\right. \\ & \quad \left. + \alpha\left(-2\mathbf{A}^T\mathbf{SG} + 2\mathbf{A}^T\mathbf{AHG}\right)\right]_{qj}\mathbf{H}_{qj} + \Psi_{qj}\mathbf{H}_{qj} = 0. \end{aligned} \tag{15}$$

Then we can get the multivariate updating rules as follows:

$$\mathbf{V}_{iq} \longleftarrow \mathbf{V}_{iq}\frac{\left(\mathbf{XQH}^T\right)_{iq}}{\left(\mathbf{VHQH}^T\right)_{iq}}, \tag{16}$$

$$\mathbf{A}_{kq} \longleftarrow \mathbf{A}_{kq}\frac{\left(\mathbf{SGH}^T\right)_{kq}}{\left(\mathbf{AHGH}^T\right)_{kq}}, \tag{17}$$

$$\mathbf{H}_{qj} \longleftarrow \mathbf{H}_{qj}\frac{\left(\mathbf{V}^T\mathbf{XQ} + \alpha\mathbf{A}^T\mathbf{SG} + \beta\mathbf{HW}\right)_{qj}}{\left(\mathbf{V}^T\mathbf{VHQ} + \alpha\mathbf{A}^T\mathbf{AHG} + \beta\mathbf{HB}\right)_{qj}}. \tag{18}$$

The details of our method are described in Algorithm 1. The iterative procedure is performed until the algorithm converges.

Considering the three update rules above, we ensure the convergence of the algorithm by the following theorem.

Theorem 1. *The objective function* $\mathrm{O} = \|\mathbf{X} - \mathbf{VH}\|_{2,1} + \beta\operatorname{tr}(\mathbf{HLH}^T) + \alpha\|\mathbf{S} - \mathbf{AH}\|_{2,1}$ *is nonincreasing under the iterative rules in (16), (17), and (18).*

The detailed proof of the theorem is shown in the Appendix.

3. Results and Discussion

In order to verify the effectiveness of GLD-RNMF algorithm for identifying differentially expressed genes, we perform experiments on real gene expression datasets to compare our algorithm with the other four feature extraction algorithms: (a) GNMF algorithm (Cai et al. [9]); (b) NMFSC algorithm (Hoyer [21]); (c) RGNMF algorithm (Wang et al. [10]); (d) GDNMF algorithm (Long et al. [16]). We conduct these experiments on two publicly available cancer datasets: pancreatic cancer dataset (PAAD) and cholangiocarcinoma dataset (CHOL).

3.1. Identifying Differentially Expressed Genes by GLD-RNMF. In this section, we use GLD-RNMF to identify differentially expressed genes. The matrix **X** with size $m \times n$ is the original gene expression data. Each row of **X** indicates the transcriptional response of n genes in a sample. Each column of **X** indicates the expression level of a gene in all samples. Therefore, **X** can be written as follows:

$$\mathbf{X} \approx \mathbf{VH}, \tag{19}$$

where **V** is the basis matrix with size $m \times k$ and **H** is the coefficient matrix with size $k \times n$ and $k \ll \min(m, n)$. Since the matrix **V** contains all of the genes, the differentially expressed genes can be identified from the matrix **V** [10]. By GLD-RNMF, the evaluating vector $\overline{\mathbf{V}}$ is obtained in which elements are sorted in descending order:

$$\overline{\mathbf{V}} = \left[\sum_{i=1}^{k}\left|\mathbf{V}_{1i}\right|, \ldots, \sum_{i=1}^{k}\left|\mathbf{V}_{ni}\right|\right]^T. \tag{20}$$

Generally, the larger the entry in $\overline{\mathbf{V}}$ is, the more differential this gene is. Therefore, the differentially expressed genes can be obtained by the first num (num $\leq j$) largest elements in $\overline{\mathbf{V}}$.

The objective of the experiment is to identify the differentially expressed genes by GLD-RNMF algorithm. The identifying process is described below.

(1) Obtain the nonnegative matrix **X** according to the genomic dataset.

(2) Construct the label matrix **S** and the diagonal matrixes **G** and **Q**.

(3) Gain the Laplacian matrix **L** and basis matrix **V** via GLD-RNMF algorithm.

(4) Identify differentially expressed genes through the vector $\overline{\mathbf{V}}$.

(5) Check differentially expressed genes by gene ontology tool.

3.2. Parameters Selection. We assign parameters in our GLD-RNMF algorithm following the same way proposed by Long et al. [16]. Distinguishingly, there are two parameters, that is, β and α, in GLD-RNMF method.

Fortunately, if β and α are set in a reasonable range they have little effect on the performance of the algorithm [15, 16].

Input: Data matrix $\mathbf{X} \in R^{m\times n}$, indicator matrix $\mathbf{S} \in R^{c\times n}$, parameters α, β, k.
Output: Matrices $\mathbf{V} \in R^{m\times k}$, $\mathbf{H} \in R^{k\times n}$ and $\mathbf{A} \in R^{c\times k}$.
(1) Initialization: Randomly initialize three nonnegative matrices $\mathbf{V}_0 \in R^{m\times k}$, $\mathbf{H}_0 \in R^{k\times n}$ and $\mathbf{A}_0 \in R^{c\times k}$, initialize $\mathbf{Q}_0 \in R^{n\times n}$, $\mathbf{G}_0 \in R^{n\times n}$ to be identity matrix. Set $r = 0$
(2) Repeat
Update $\mathbf{V}_{r+1}$,$\mathbf{A}_{r+1}$ and $\mathbf{H}_{r+1}$ separately by

$$\mathbf{V}_{ir+1} \longleftarrow \mathbf{V}_r \frac{\left(\mathbf{X}\mathbf{Q}_r\mathbf{H}_r^T\right)}{\left(\mathbf{V}_r\mathbf{H}_r\mathbf{Q}_r\mathbf{H}_r^T\right)}$$

$$\mathbf{A}_{r+1} \longleftarrow \mathbf{A}_r \frac{\left(\mathbf{S}\mathbf{G}_r{\mathbf{H}_r}^T\right)}{\left(\mathbf{A}_r\mathbf{H}_r\mathbf{G}_r{\mathbf{H}_r}^T\right)}$$

$$\mathbf{H}_{r+1} \longleftarrow \mathbf{H}_r \frac{\left({\mathbf{V}_{r+1}}^T\mathbf{X}\mathbf{Q}_{r+1} + \alpha{\mathbf{A}_{r+1}}^T\mathbf{S}\mathbf{G}_{r+1} + \beta\mathbf{H}_r\mathbf{W}\right)}{\left({\mathbf{V}_{r+1}}^T\mathbf{V}\mathbf{H}_r\mathbf{Q}_{r+1} + \alpha{\mathbf{A}_{r+1}}^T\mathbf{A}\mathbf{H}_r\mathbf{G}_{r+1} + \beta\mathbf{H}_r\mathbf{B}\right)}$$

Calculate the diagonal matrices $\mathbf{Q}_{r+1}$ and $\mathbf{G}_{r+1}$ by (11) and (12), separately.
Until convergence.

Algorithm 1: GLD-RNMF.

Table 1: Comparison of p values of different methods on PAAD.

Gene ID	Gene name	NMFSC	GNMF	RGNMF	GDNMF	GLD-RNMF
GO:0030198	Extracellular matrix organization	$1.36E-15$	$9.04E-17$	$1.36E-15$	$1.36E-15$	**2.99E − 42**
GO:0043062	Extracellular structure organization	$1.44E-15$	$9.57E-17$	$1.44E-15$	$1.44E-15$	**3.35E − 42**
GO:0031012	Extracellular matrix	$5.64E-15$	$3.23E-17$	$5.64E-15$	$5.64E-15$	**3.53E − 37**
GO:0005615	Extracellular space	$3.09E-27$	$3.09E-27$	$3.09E-27$	$3.09E-27$	**9.70E − 37**
GO:0005578	Proteinaceous extracellular matrix	$6.74E-12$	$4.59E-14$	$6.74E-12$	$6.74E-12$	**2.00E − 29**
GO:0044420	Extracellular matrix component	$1.55E-10$	$7.70E-12$	$1.55E-10$	$1.55E-10$	**4.96E − 27**
GO:0030574	Collagen catabolic process	$3.93E-14$	$8.45E-16$	$5.64E-15$	$3.93E-14$	**8.03E − 27**
GO:0044243	Multicellular organism catabolic process	$1.24E-13$	$3.00E-15$	$1.24E-13$	$1.24E-13$	**6.06E − 26**
GO:0032963	Collagen metabolic process	$3.41E-11$	$1.45E-12$	$3.41E-11$	$3.41E-11$	**2.37E − 23**
GO:0098644	Complex of collagen trimers	$1.35E-09$	$1.56E-11$	$1.35E-09$	$1.35E-09$	**3.53E − 20**

In our experiments, we set $\beta = 0.9$ and $\alpha = 0.5$ in the GLD-RNMF algorithm. Another important parameter in our GLD-RNMF algorithm is k which is used to construct a k-nearest graph. Empirically, we set $k = 5$ and adopt the mode as the heat kernel in LPP [22]. Besides, we set the reduced dimension to 5 for all the methods. All the parameters in the other methods keep in line with those described in their paper [10, 16, 21, 23].

3.3. Gene Ontology Analysis. The gene ontology (GO) tool can interpret the genes that are input and discover the functions that these genes may have in common. As a web-based tool [24], GO Enrichment Analysis can find important GO items from a large number of genes and provide important information for the biological interpretation of high-throughput experiments. Another online tool that we use is ToppFun. It usually is used to interpret the differentially expressed genes.

To be fair, we extract 100 genes from the gene expression data by GNMF, NMFSC, RGNMF, GDNMF, and GLD-RNMF methods. Threshold parameters of ToppFun are set as follows: the maximum p value is set to 0.01 and the minimum number of gene products is set to 2.

3.4. Pancreatic Cancer Dataset. Pancreatic cancer is a tumor with high malignancy, which is difficult to diagnose and treat. Early diagnosis of pancreatic cancer is not difficult and the mortality rate is high. The cause of pancreatic cancer is still not clear until now. In the experiment, there are 20502 genes in 180 samples contained in the dataset.

The top 10 GO items extracted and the p values of the five methods are listed in Table 1. In this table, "ID" and "Name" represent items and their names associated with the GO in the whole genome and the lowest p value of the five methods has been marked in bold font. As we find from Table 1, the p values generated by GLD-RNMF are much smaller than those by the other four methods for the PAAD. Therefore, GLD-RNMF method is more superior than the other four methods for the PAAD. The name of "GO:0030198" is extracellular matrix organization. It contains DPT (Dermatopontin), POSTN (Periostin), sec24d (Sec24-related protein d), and

Table 2: Pancreatic cancer genes extracted by GLD-RNMF.

Gene ID	Gene name	Gene annotations	Relative diseases
4313	MMP2	Serine-type endopeptidase activity and metallopeptidase activity	Arthropathy and multicentric osteolysis of Torg
3486	IGFBP3	Fibronectin binding and insulin-like growth factor I binding.	Insulin-like growth factor I and acromegaly
3630	INS	Identical protein binding and protease binding	Diabetes mellitus
4316	MMP7	Peptidase activity and metallopeptidase activity	Spastic entropion and focal myositis
3880	KRT19	Structural molecule activity and structural constituent of cytoskeleton	Thyroid cancer and lung cancer
7057	THBS1	Calcium ion binding and heparin binding	Posterior uveal melanoma and thrombotic thrombocytopenic purpura
3320	HSP90AA1	Poly(A) RNA binding and identical protein binding	Lobular neoplasia and candidiasis
1508	CTSB	Peptidase activity and cysteine-type peptidase activity	Occlusion of gallbladder and ileum cancer
7076	TIMP1	Cytokine activity and protease binding.	Oral submucous fibrosis and lung giant cell carcinoma
2335	FN1	Heparin binding and protease binding	Glomerulopathy and plasma fibronectin deficiency

other genes which are related to pancreatic cancer [25–27]. For example, POSTN can create a tumor-supportive microenvironment in the pancreas [28]. Genes and gene products associated with extracellular structure organization (GO:0043062) can be found by GO tool, in which, DPT, POSTN, sec24d, uxs1 (UDP-Glucuronate Decarboxylase 1), fkrp (Fukutin Related Protein), and other genes have been illustrated to be associated with pancreatic cancer [29, 30]. For example, Sec24d is ubiquitously expressed but exhibits predominant expression in heart, placenta, liver, and pancreas. The other GO items can also be proven to be related to pancreatic cancer by some relevant literature material. Clearly, GLD-RNMF method is an effective method for identifying differentially expressed genes.

Comparing 100 genes extracted by GLD-RNMF with what we obtain from Gene Cards (http://www.genecards.org/) about pancreatic cancer, 82 of the 100 genes are associated with pancreatic cancer. Many genes, which were previously thought to be unrelated to clinical outcomes, are identified. We present the top 10 of 82 genes with higher relevance scores in Table 2, including their gene ID, names, function, and related diseases. Among the identified differentially expressed genes, MMP2, MMP7, IGFBP3, INS, and the other genes have been demonstrated to be related to pancreatic cancer [31–34]. For example, the effect of MMP-2 and its activators MT1-MMP, MT2-MMP, and MT3-MMP in pancreatic tumor cell invasion and the development of the desmoplastic reaction characteristic of pancreatic cancer tissues have been discussed [35]. Akihisa Fukuda et al. demonstrated that serum MMP7 level in human pancreatic ductal adenocarcinoma patients is correlated with metastatic disease and survival. Conditioned medium from Capan-1 pancreatic cancer cells which contains abundant IGFBP-3 has been mentioned [36]. Other genes identified by GLD-RNMF have been illustrated to be related to pancreatic cancer by some relevant literature materials as well.

On the other hand, we use Kyoto Encyclopedia of Genes and Genomes (KEGG) online analysis tool to analyze the differentially expressed genes identified by GLD-RNMF. In this experiment, putting the identified 100 genes into KEGG, we can obtain the corresponding disease pathway. Figure 1 is the pathway of pancreatic cancer. The genes that have been found by GLD-RNMF are marked with red. The chromosomal instability pathway records the disease progression from normal duct to pancreatic cancer. Infiltrating ductal adenocarcinoma is the most common malignancy of the pancreas. Normal duct epithelium progresses to the stage of infiltrating cancer through a series of histologically defined precursors. These differentially expressed genes contain two oncogenes: K-Ras and HER2/neu. p16, p53, BRCA2, and Smad4 are tumor suppressors. Slebos et al. assessed that K-ras oncogene mutations and p53 protein accumulation are associated with known or postulated risk factors for pancreatic cancer [37]. Gu et al. found the expression of Smad4 and p16 is significantly lower in pancreatic cancer tissue compared with normal tissue. The lower expression of the proteins may impact the development of pancreatic cancer [38]. From Figure 1, we can see that the pancreatic cancer pathway map contains six other pathways: PI3K-Akt signaling pathway, ErbB signaling pathway, MAPK signaling pathway, VEGF signaling, Jak-STAT signaling pathway, p53 signaling pathway, and TGF-β signaling pathway.

PI3K-Akt signaling pathway is presented in Figure 2. From Figure 2, we can find four proteinases from the differentially expressed genes identified by GLD-RNMF. Therefore, our algorithm achieves better results.

3.5. Cholangiocarcinoma Dataset. Cholangiocarcinoma is diagnosed in 12,000 patients in the US each year, but only 10 percent are discovered early enough to allow for successful surgical treatment. In our experiments, we apply the five methods on CHOL which contains 20502 genes on 45 samples. The 8 GO items closely related to cholangiocarcinoma

Table 3: Comparison of p values of different methods on CHOL.

Gene ID	Gene name	NMFSC	GNMF	RGNMF	GDNMF	GLD-RNMF
GO:0072562	Blood microparticle	$1.73E-20$	$1.73E-20$	$1.85E-17$	$1.73E-20$	**$1.25E-50$**
GO:0060205	Cytoplasmic membrane-bounded vesicle lumen	$4.19E-23$	$4.19E-23$	$3.90E-18$	$4.19E-23$	**$1.34E-41$**
GO:0031983	Vesicle lumen	$8.66E-23$	$8.66E-23$	$7.05E-18$	$8.66E-23$	**$4.47E-41$**
GO:0005615	Extracellular space	$3.21E-19$	$3.21E-19$	$2.12E-17$	$3.21E-19$	**$2.66E-37$**
GO:0034774	Secretory granule lumen	$1.59E-19$	$1.59E-19$	$1.35E-14$	$1.59E-19$	**$3.33E-34$**
GO:0004857	Enzyme inhibitor activity	$5.71E-06$	$5.71E-06$	$5.71E-06$	$5.71E-06$	**$3.93E-17$**
GO:0004866	Endopeptidase inhibitor activity	$2.92E-05$	$2.92E-05$	$2.92E-05$	$2.92E-05$	**$3.74E-16$**
GO:0061135	Endopeptidase regulator activity	$3.64E-05$	$3.64E-05$	$3.64E-05$	$3.64E-05$	**$6.50E-16$**

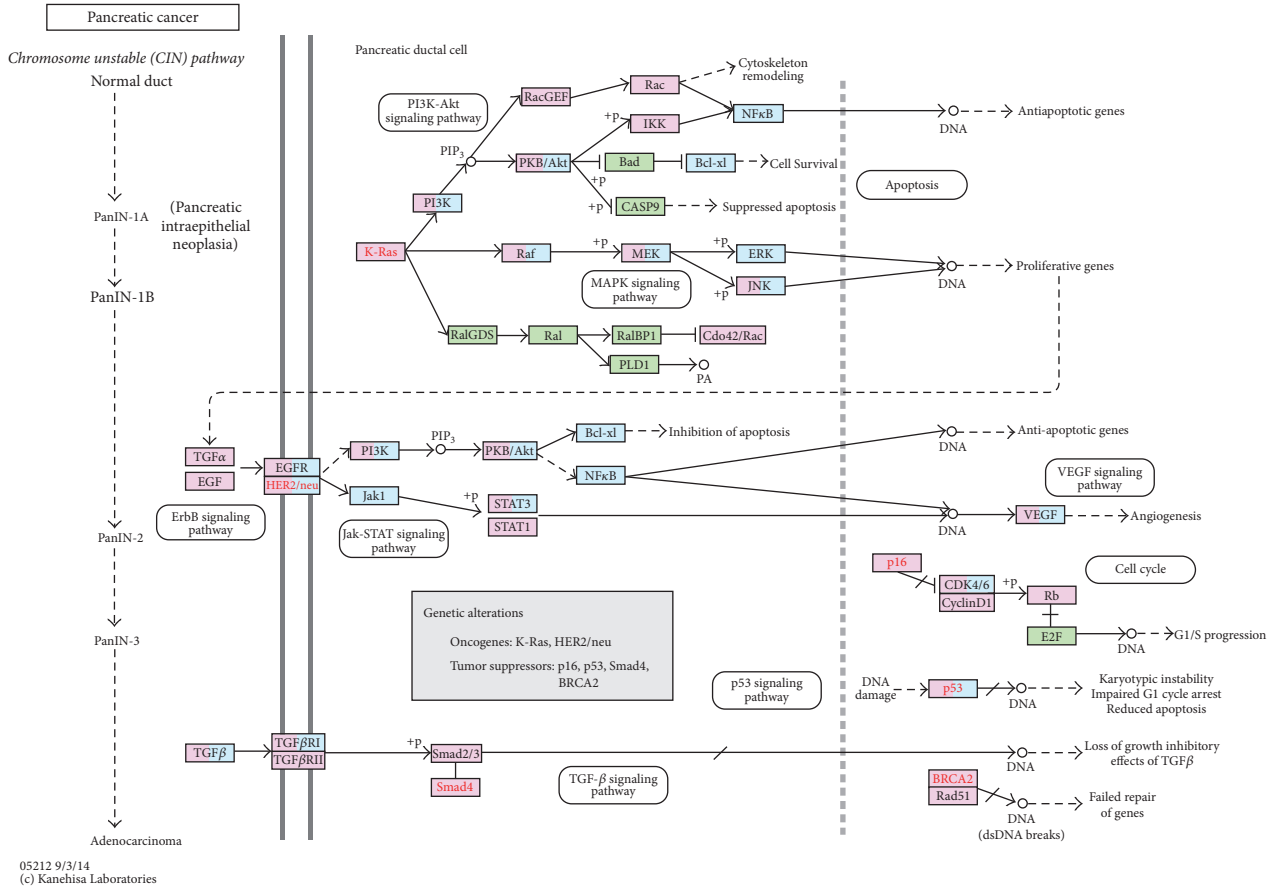


Figure 1: Pathway of pancreatic cancer, where pink background represents disease genes, light blue represents drug target genes, and light green represents human genes. Genes found by GLD-RNMF are marked with red.

and the p values of the five methods are listed in Table 3. In this table, "ID" and "Name" represent items and their names associated with the GO in the whole genome. The lowest p value of the five methods has been marked in bold font. We can see from the table that GLD-RNMF is much better than the other four methods. Genes and gene products associated with blood microparticle (GO:0072562) can be found by the GO tool, in which APOA4 (Apolipoprotein A4), kng1 (Kininogen 1), and other genes have been illustrated to be associated with cholangiocarcinoma [39–41]. The name of "GO:0060205" is cytoplasmic vesicle lumen. It contains ada (Adenosine Deaminase), DBH (Dopamine Beta-Hydroxylase), and other genes which are related to cholangiocarcinoma [42, 43]. The top 8 genes identified by GLD-RNMF are listed in Table 4 including the gene ID, names, gene annotations, and related diseases. Consistent with the previous study, ALB (Albumin), HP (Haptoglobin), SERPINC1 (Serpin Family C Member 1), C3 (Complement C3),

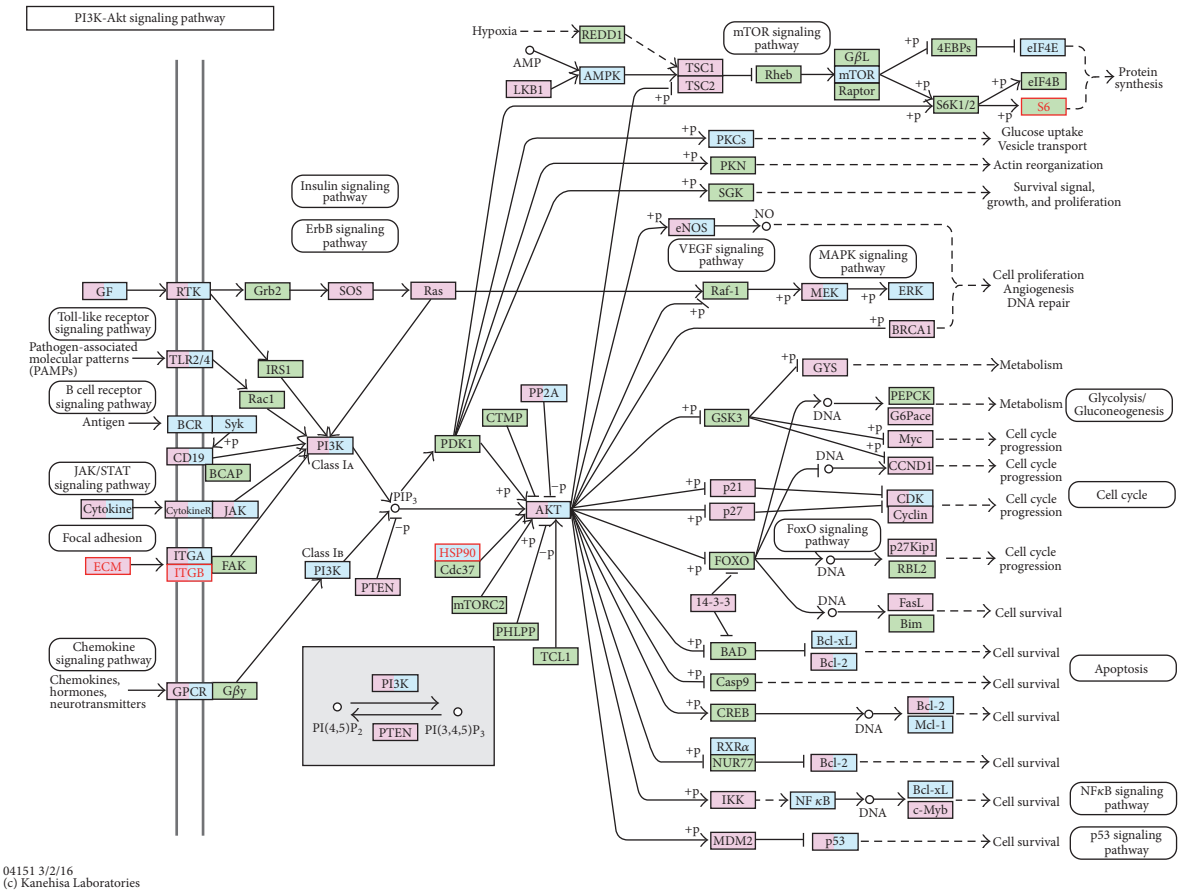


FIGURE 2: PI3K-AKT signaling pathway, where pink background represents disease genes, light blue represents drug target genes, and light green represents human genes. Genes found by GLD-RNMF are marked with red.

and other genes are successfully identified which represent potential biomarkers for cholangiocarcinoma and potential targets for clarifying the molecular mechanisms associated with cholangiocarcinoma. For example, HP was proposed as pronucleating proteins because they were highly expressed in the fast nucleating bile of patients with cholesterol stones [44]. Waghray et al. predicted survival in patients with hilar cholangiocarcinoma by serum albumin [45]. C3 and HP were identified as more abundant in cholangiocarcinoma [46]. The relative documents can illustrate that the other genes identified by GLD-RNMF are associated with cholangiocarcinoma.

4. Conclusions

By introducing $L_{2,1}$-norm, manifold graph and discriminative label information, we propose an efficient algorithm named robust nonnegative matrix factorization via joint graph Laplacian and discriminative information (GLD-RNMF) in this paper. $L_{2,1}$-norm can reduce the influence of outliers and noise, manifold graph can find the low dimensional manifold in high dimensional data space and the intrinsic law of the observed data, and the label information can increase the discriminative power of different classes. Nonnegative matrix factorization avoids the problems of high dimension and nonnegative data. As a result, GLD-RNMF can handle nonnegative, high dimension, outliers, and noise and improve the discriminative power of different classes. Experimental results on two datasets show that GLD-RNMF is superior to the state-of-the-art methods for identifying differentially expressed gene.

Appendix

Detailed Proof of Theorem

To prove Theorem, we need to show that the objective function in (9) is nonincreasing under the iterative rules in (16), (17), and (18). For the objective function O, we need to fix $\mathbf{H}$ and $\mathbf{A}$ when we update $\mathbf{V}$. Similarly, we need to fix $\mathbf{H}$ and $\mathbf{V}$ when we update $\mathbf{A}$, and we need to fix $\mathbf{A}$ and $\mathbf{V}$ when we update $\mathbf{H}$. For the reason that we have similar update rules for $\mathbf{A}$ and $\mathbf{V}$ in GLD-RNMF with those in NMF, the detailed proof can be found [27]. Hence, we just need to prove that O is nonincreasing under the iterative

Table 4: Cholangiocarcinoma genes extracted by GLD-RNMF.

Gene ID	Gene name	Gene annotations	Relative diseases
213	ALB	Enzyme binding and chaperone binding.	Analbuminemia and congenital analbuminemia
3240	HP	Serine-type endopeptidase activity and hemoglobin binding	Anhaptoglobinemia and plasmodium falciparum malaria
5265	SERPINA1	Identical protein binding and protease binding	Emphysema due to Aat deficiency and alpha 1-antitrypsin deficiency
718	C3	Receptor binding and C5L2 anaphylatoxin chemotactic receptor binding	Macular degeneration, age-related, 9, and C3 deficiency
2243	FGA	Receptor binding and protein binding, bridging	Dysfibrinogenemia, congenital, and afibrinogenemia, congenital
7448	VTN	Heparin binding and scavenger receptor activity	Glanzmann thrombasthenia and camptodactyly-arthropathy-coxa vara-pericarditis syndrome.
7018	TF	Ubiquitin protein ligase binding and ferric iron transmembrane transporter activity	Atransferrinemia and iron overload In Africa
335	APOA1	Identical protein binding and lipid binding	Amyloidosis, familial visceral, and hypoalphalipoproteinemia

rules in (18). We use an auxiliary function similar to what is used in the Expectation-Maximization algorithm [47]. In the demonstration, we present the definition of auxiliary function [16]. Definition G(h, h') is an auxiliary function of $F(h)$ if the following conditions are satisfied.

$$\begin{aligned} G\left(h, h'\right) &\geq F\left(h\right), \\ G\left(h, h\right) &= F\left(h\right). \end{aligned} \tag{A.1}$$

The auxiliary function is vital due to the following lemmas.

Lemma A.1. *If G is an auxiliary function of F, then F is nonincreasing under the update rule:*

$$h^{(t+1)} = \arg\min_{h} \quad G\left(h, h^{(t)}\right). \tag{A.2}$$

Proof. Obviously

$$\begin{aligned} F\left(h^{(t+1)}\right) &\leq G\left(h^{(t+1)}, h^{(t)}\right) \leq G\left(h^{(t)}, h^{(t)}\right) \\ &= F\left(h^{(t)}\right); \end{aligned} \tag{A.3}$$

now, we will present the update step for $\mathbf{H}$ in (18) is exactly the update in (A.2) with an appropriate auxiliary function

$$\begin{aligned} F'_{ij} &= \left(\frac{\partial O}{\partial \mathbf{H}}\right)_{ij} = \left(-2\mathbf{V}^T\mathbf{XQ} + 2\mathbf{V}^T\mathbf{VHQ} + 2\beta\mathbf{HL}\right. \\ &\quad \left. - 2\alpha\mathbf{A}^T\mathbf{SG} + 2\alpha\mathbf{A}^T\mathbf{AHG}\right)_{ij}, \\ F''_{ij} &= \left(2\mathbf{V}^T\mathbf{VQ}\right)_{ii} + 2\beta\mathbf{L}_{jj} + \left(2\alpha\mathbf{A}^T\mathbf{AG}\right)_{ii}. \end{aligned} \tag{A.4}$$

It is enough to prove that each F_{ij} is nonincreasing under the update rules because our update is essentially wised. Therefore, we present the following lemma. □

Lemma A.2. *Function,*

$$\begin{aligned} G\left(h, h_{ij}^{(t)}\right) &= F_{ij}\left(h_{ij}^{(t)}\right) + F'_{ij}\left(h_{ij}^{(t)}\right)\left(h - h_{ij}^{(t)}\right) \\ &\quad + \frac{\left(\mathbf{V}^T\mathbf{VHQ}\right)_{ij} + \beta\left(\mathbf{HB}\right)_{ij} + \alpha\left(\mathbf{A}^T\mathbf{AHG}\right)_{ij}}{h_{ij}^{(t)}}\left(h\right. \\ &\quad \left. - h_{ij}^{(t)}\right)^2, \end{aligned} \tag{A.5}$$

is an auxiliary function of F_{ij}.

Proof. We only need to demonstrate that $G(h, h_{ij}^{(t)}) \geq F_{ij}(h)$, because $G(h, h) = F_{ij}(h)$ is obvious. Consequently, comparing the Taylor series expansion of $F_{ij}(h)$,

$$\begin{aligned} &F_{ij}\left(h\right) \\ &= F_{ij}\left(h_{ij}^{(t)}\right) + F'_{ij}\left(h_{ij}^{(t)}\right)\left(h - h_{ij}^{(t)}\right) \\ &\quad + \left[\left(\mathbf{V}^T\mathbf{VQ}\right)_{ii} + \beta\mathbf{L}_{jj} + \left(\alpha\mathbf{A}^T\mathbf{AG}\right)_{ii}\right]\left(h - h_{ij}^{(t)}\right)^2, \end{aligned} \tag{A.6}$$

with (A.2), we can find that $G(h, h_{ij}^{(t)}) \geq F_{ij}(h)$ is equivalent to

$$\begin{aligned} &\frac{\left(\mathbf{V}^T\mathbf{VHQ}\right)_{ij} + \beta\left(\mathbf{HB}\right)_{ij} + \alpha\left(\mathbf{A}^T\mathbf{AHG}\right)_{ij}}{h_{ij}^{(t)}} \\ &\quad \geq \left(\mathbf{V}^T\mathbf{VQ}\right)_{ii} + \beta\mathbf{L}_{jj} + \left(\alpha\mathbf{A}^T\mathbf{AG}\right)_{ii}. \end{aligned} \tag{A.7}$$

Actually, we have

$$\left(\mathbf{V}^T\mathbf{VHQ}\right)_{ij} = \sum_{l=1}^{k} h_{il}^{(t)} \left(\mathbf{V}^T\mathbf{VQ}\right)_{lj} \geq \left(\mathbf{V}^T\mathbf{VQ}\right)_{ii} h_{ij}^{(t)},$$

$$\left(\mathbf{A}^T\mathbf{AHG}\right)_{ij} = \sum_{l=1}^{k} h_{il}^{(t)} \left(\mathbf{A}^T\mathbf{AG}\right)_{lj} \geq \left(\mathbf{A}^T\mathbf{AG}\right)_{ii} h_{ij}^{(t)}. \tag{A.8}$$

$$\begin{aligned}\beta\left(\mathbf{HB}\right)_{ij} &= \beta\sum_{l=1}^{n} h_{il}^{(t)} B_{lj} \geq \beta h_{ij}^{(t)} B_{jj} \\ &\geq \beta h_{ij}^{(t)} \left(B - W\right)_{jj} = \beta h_{ij}^{(t)} L_{jj}.\end{aligned}$$

□

Therefore, (A.7) holds and $G(h, h_{ij}^{(t)}) \geq F_{ij}(h)$. Now we can prove the convergence of theorem.

Proof of Theorem. Replacing $G(h, h_{ij}^{(t)})$ in (A.2) by (A.5), we can obtain the following update rules:

$$\begin{aligned}h_{ij}^{(t+1)} &= h_{ij}^{(t)} \\ &\quad - h_{ij}^{(t)} \frac{F_{ij}'\left(h_{ij}^{(t)}\right)}{\left(2\mathbf{V}^T\mathbf{VHQ}\right)_{ij} + 2\beta\left(\mathbf{HB}\right)_{ij} + \left(2\alpha\mathbf{A}^T\mathbf{AHG}\right)_{ij}} \\ &= h_{ij}^{(t)} \left(\frac{\mathbf{V}^T\mathbf{XQ} + \alpha\mathbf{A}^T\mathbf{SG} + \beta\mathbf{HW}}{\mathbf{V}^T\mathbf{VHQ} + \beta\mathbf{HB} + \alpha\mathbf{A}^T\mathbf{AHG}}\right)_{ij}.\end{aligned} \tag{A.9}$$

Since $G(h, h_{ij}^{(t)})$ is an auxiliary function, F_{ij} is nonincreasing under the update rule. □

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by the grants of the National Science Foundation of China, nos. 61572284, 61502272, 61373027, and 61672321.

References

[1] M. J. Heller, "DNA microarray technology: devices, systems, and applications," *Annual Review of Biomedical Engineering*, vol. 4, pp. 129–153, 2002.

[2] Y. Li and Z. Zhang, "Computational biology in microRNA," *Wiley Interdisciplinary Reviews: RNA*, vol. 6, no. 4, pp. 435–452, 2015.

[3] A. B. Tchagang and A. H. Tewfik, "DNA microarray data analysis: a novel biclustering algorithm approach," *Eurasip Journal on Applied Signal Processing*, vol. 2006, Article ID 59809, 12 pages, 2006.

[4] A. Wang and E. A. Gehan, "Gene selection for microarray data analysis using principal component analysis," *Statistics in Medicine*, vol. 24, no. 13, pp. 2069–2087, 2005.

[5] R. Luss and A. D'Aspremont, "Clustering and feature selection using sparse principal component analysis," *Optimization and Engineering*, vol. 11, no. 1, pp. 145–157, 2010.

[6] C. H. Zheng, L. Zhang, T. Y. Ng, K. S. Chi, and S. L. Wang, "Inferring the transcriptional modules using penalized matrix decomposition," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence: 6th International Conference on Intelligent Computing, ICIC 2010, Changsha, China, August 18–21, 2010. Proceedings*, vol. 6216 of *Lecture Notes in Computer Science*, pp. 35–41, Springer, Berlin, Germany, 2010.

[7] J.-X. Liu, Y. Xu, C.-H. Zheng, H. Kong, and Z.-H. Lai, "RPCA-based tumor classification using gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 964–970, 2015.

[8] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 599–607, 2009.

[9] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[10] D. Wang, J. X. Liu, Y. L. Gao, C. H. Zheng, and Y. Xu, "Characteristic gene selection based on robust graph regularized non-negative matrix factorization," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 108, p. 1, 2015.

[11] J.-X. Liu, J. Liu, Y.-L. Gao, J.-X. Mi, C.-X. Ma, and D. Wang, "A class-information-based penalized matrix decomposition for identifying plants core genes responding to abiotic stresses," *PLoS ONE*, vol. 9, no. 9, Article ID e106097, 2014.

[12] R. Giancarlo and F. Utro, "Speeding up the Consensus Clustering methodology for microarray data analysis," *Algorithms for Molecular Biology*, vol. 6, no. 1, article 1, 2011.

[13] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[14] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[15] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 2030–2048, 2011.

[16] X. Long, H. Lu, Y. Peng, and W. Li, "Graph regularized discriminative non-negative matrix factorization for face recognition," *Multimedia Tools and Applications*, vol. 723, pp. 2679–2699, 2014.

[17] Y. Zheng, B. Jeon, D. Xu, Q. M. J. Wu, and H. Zhang, "Image segmentation by generalized hierarchical fuzzy C-means algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 28, no. 2, pp. 961–973, 2015.

[18] C. Ding, D. Zhou, X. He, and H. Zha, "R 1-PCA: rotational invariant L 1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 281–288, Pittsburgh, Pa, USA, June 2006.

[19] D. D. Lee, "Algorithms for nonnegative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[20] F. Facchinei, C. Kanzow, and S. Sagratella, "Solving quasi-variational inequalities via their KKT conditions," *Mathematical Programming*, vol. 144, no. 1-2, pp. 369–412, 2014.

[21] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[22] S. Yang, C. Hou, C. Zhang, Y. Wu, and S. Weng, "Robust non-negative matrix factorization via joint sparse and graph regularization," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '13)*, pp. 1–5, August 2013.

[23] B. Yang, "Graph regularized non-negative matrix factorization with sparseness constraints," *Computer Science*, vol. 1, no. 40, pp. 218–256, 2013.

[24] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack, "Gene ontology analysis for RNA-seq: accounting for selection bias," *Genome Biology*, vol. 11, no. 2, article R14, 2010.

[25] T. Orban, J. M. Sosenko, D. Cuthbertson et al., "Pancreatic islet autoantibodies as predictors of type 1 diabetes in the diabetes prevention trial-type 1," *Diabetes Care*, vol. 32, no. 12, pp. 2269–2274, 2009.

[26] P. Baril, R. Gangeswaran, P. C. Mahon et al., "Periostin promotes invasiveness and resistance of pancreatic cancer cells to hypoxia-induced cell death: role of the β_4 integrin and the PI3k pathway," *Oncogene*, vol. 26, no. 14, pp. 2082–2094, 2007.

[27] B. L. Tang, J. Kausalya, D. Y. H. Low, M. L. Lock, and W. Hong, "A family of mammalian proteins homologous to yeast Sec24p," *Biochemical and Biophysical Research Communications*, vol. 258, no. 3, pp. 679–684, 1999.

[28] M. Erkan, J. Kleeff, A. Gorbachevski et al., "Periostin creates a tumor-supportive microenvironment in the pancreas by sustaining fibrogenic stellate cell activity," *Gastroenterology*, vol. 132, no. 4, pp. 1447–1464, 2007.

[29] Z.-Q. Ye, S. Niu, Y. Yu et al., "Analyses of copy number variation of GK rat reveal new putative type 2 diabetes susceptibility loci," *PLoS ONE*, vol. 5, no. 11, article e14077, 2010.

[30] T. Yamamoto, Y. Kato, M. Karita, M. Kawaguchi, N. Shibata, and M. Kobayashi, "Expression of genes related to muscular dystrophy with lissencephaly," *Pediatric Neurology*, vol. 31, no. 3, pp. 183–190, 2004.

[31] F. Miralles, T. Battelino, P. Czernichow, and R. Scharfmann, "TGF-β plays a key role in morphogenesis of the pancreatic islets of langerhans by controlling the activity of the matrix metalloproteinase MMP-2," *Journal of Cell Biology*, vol. 143, no. 3, pp. 827–836, 1998.

[32] S. R. Bramhall, J. P. Neoptolemos, G. W. H. Stamp, and N. R. Lemoine, "Imbalance of expression of matrix metalloproteinases (MMPs) and tissue inhibitors of the matrix metalloproteinases (TIMPs) in human pancreatic carcinoma," *Journal of Pathology*, vol. 182, no. 3, pp. 347–355, 1997.

[33] J. B. Douglas, D. T. Silverman, M. N. Pollak, Y. Tao, A. S. Soliman, and R. Z. Stolzenberg-Solomon, "Serum IGF-I, IGF-II, IGFBP-3, and IGF-I/IGFBP-3 molar ratio and risk of pancreatic cancer in the prostate, lung, colorectal, and ovarian cancer screening trial," *Cancer Epidemiology Biomarkers and Prevention*, vol. 19, no. 9, pp. 2298–2306, 2010.

[34] E. Karaskov, C. Scott, L. Zhang, T. Teodoro, M. Ravazzola, and A. Volchuk, "Chronic palmitate but not oleate exposure induces endoplasmic reticulum stress, which may contribute to INS-1 pancreatic β-cell apoptosis," *Endocrinology*, vol. 147, no. 7, pp. 3398–3407, 2006.

[35] V. Ellenrieder, B. Alber, U. Lacher et al., "Role of MT-MMPs and MMP-2 in pancreatic cancer progression," *International Journal of Cancer*, vol. 85, no. 1, pp. 14–20, 2000.

[36] X. Y. Huang, Z. L. Huang, J. H. Yang et al., "Erratum to: Pancreatic cancer cell-derived IGFBP-3 contributes to muscle wasting," *Journal of Experimental & Clinical Cancer Research*, vol. 35, pp. 1–13, 2016.

[37] R. J. C. Slebos, J. A. Hoppin, P. E. Tolbert et al., "K-ras and p53 in pancreatic cancer: association with medical history, histopathology, and environmental exposures in a population-based study," *Cancer Epidemiology Biomarkers and Prevention*, vol. 9, no. 11, pp. 1223–1232, 2000.

[38] L.-J. Gu, J. Chen, Z.-H. Lu, L. Li, W.-X. Zhou, and Y.-F. Luo, "Expression of DPC4/Smad4, p21wafI, and p16 in human pancreatic cancer," *Chinese Journal of Cancer*, vol. 21, no. 2, pp. 132–137, 2002.

[39] E.-H. Kim, J.-S. Bae, K. B. Hahm, and J.-Y. Cha, "Endogenously synthesized n-3 polyunsaturated fatty acids in fat-1 mice ameliorate high-fat diet-induced non-alcoholic fatty liver disease," *Biochemical Pharmacology*, vol. 84, no. 10, pp. 1359–1365, 2012.

[40] A. Wee and B. Nilsson, "Highly well differentiated hepatocellular carcinoma and benign hepatocellular lesions: can they be distinguished on fine needle aspiration biopsy?" *Acta Cytologica*, vol. 47, no. 1, pp. 16–26, 2003.

[41] I. Subrungruang, C. Thawornkuno, C.-P. Porntip, C. Pairojkul, S. Wongkham, and S. Petmitr, "Gene expression profiling of intrahepatic cholangiocarcinoma," *Asian Pacific Journal of Cancer Prevention*, vol. 14, no. 1, pp. 557–563, 2013.

[42] S. Tanaka, M. Iwai, Y. Harada et al., "Targeted killing of carcinoembryonic antigen (CEA)-producing cholangiocarcinoma cells by polyamidoamine dendrimer-mediated transfer of an Epstein-Barr virus (EBV)-based plasmid vector carrying the CEA promoter," *Cancer Gene Therapy*, vol. 7, no. 9, pp. 1241–1249, 2000.

[43] M. Tawfik El-Mansi, K. S. Cuschieri, R. G. Morris, and A. R. W. Williams, "Prevalence of human papillomavirus types 16 and 18 in cervical adenocarcinoma and its precursors in Scottish patients," *International Journal of Gynecological Cancer*, vol. 16, no. 3, pp. 1025–1031, 2006.

[44] A. Farina, M. Delhaye, P. Lescuyer, and J.-M. Dumonceau, "Bile proteome in health and disease," *Comprehensive Physiology*, vol. 4, no. 1, pp. 91–108, 2014.

[45] A. Waghray, A. Sobotka, C. R. Marrero, B. Estfan, F. Aucejo, and K. N. Menon, "Serum albumin predicts survival in patients with hilar cholangiocarcinoma," *Gastroenterology Report*, vol. 5, no. 1, pp. 62–66, 2017.

[46] U. Navaneethan, V. Lourdusamy, P. G. Venkatesh, B. Willard, M. R. Sanaka, and M. A. Parsi, "Bile proteomics for differentiation of malignant from benign biliary strictures: a pilot study," *Gastroenterology Report*, vol. 3, pp. 136–143, 2015.

[47] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 39, no. 1, pp. 1–38, 1977.