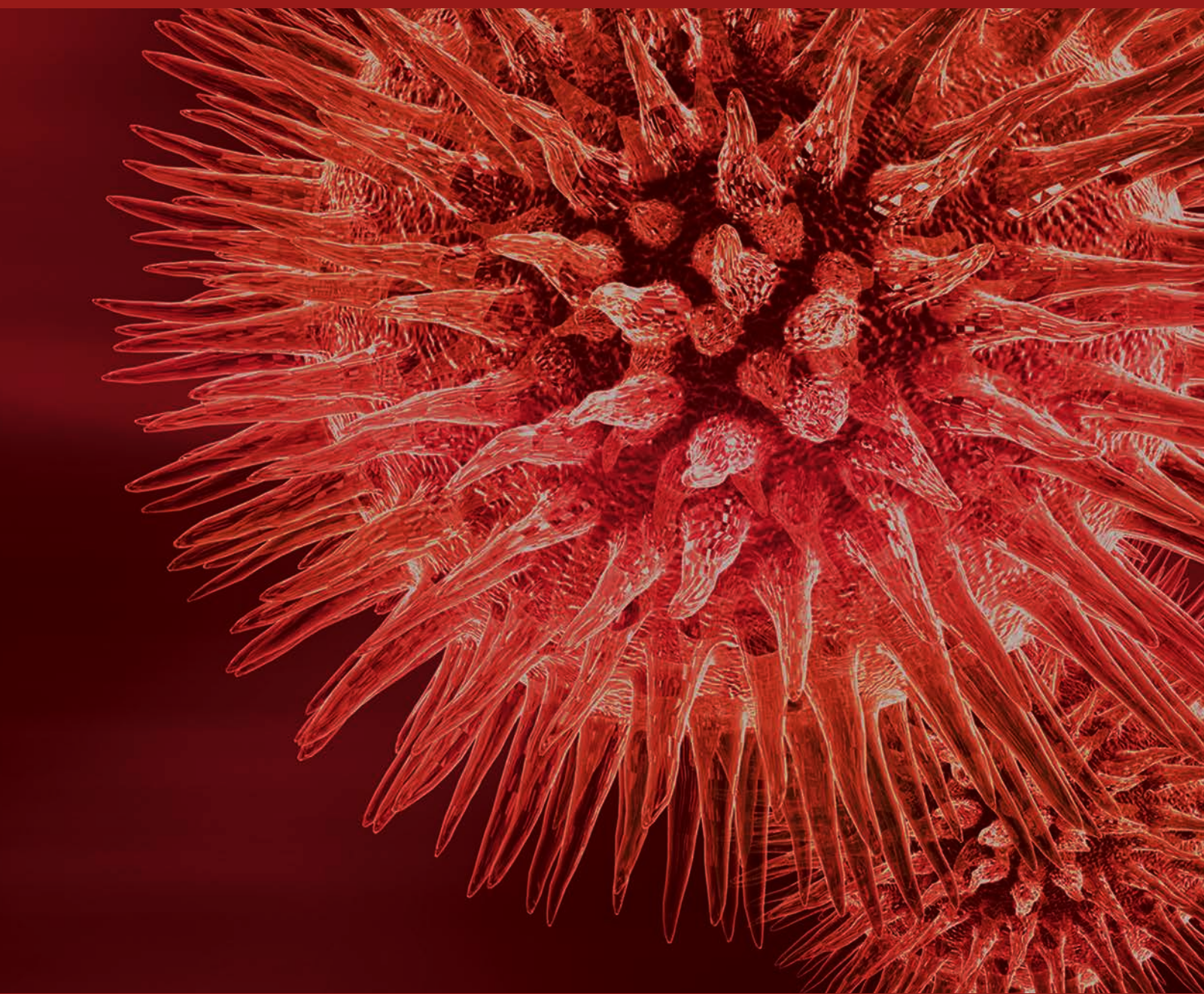
BioMed Research International

Current Computational Models for Prediction of the Varied Interactions Related to Noncoding RNAs

Guest Editors: Xing Chen, Huiming Peng, and Zheng Yin



Current Computational Models for Prediction of the Varied Interactions Related to Noncoding RNAs

BioMed Research International

Current Computational Models for Prediction of the Varied Interactions Related to Noncoding RNAs

Guest Editors: Xing Chen, Huiming Peng, and Zheng Yin

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 4183574, 2 pages
http://dx.doi.org/10.1155/2016/4183574

Editorial

Current Computational Models for Prediction of the Varied Interactions Related to Noncoding RNAs

Xing Chen,[1] Huiming Peng,[2] and Zheng Yin[3]

[1]*School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China*
[2]*Department of Math, Science & Technologies, Forsyth Technical Community College, Winston-Salem, NC, USA*
[3]*Department of Systems Medicine and Bioengineering, Houston Methodist Research Institute, Houston, TX, USA*

Correspondence should be addressed to Xing Chen; xingchen@amss.ac.cn

Received 19 October 2016; Accepted 20 October 2016

Noncoding RNA (ncRNA) refers to a kind of endogenous small RNA molecules that have no protein coding capacity. In recent years, extensive studies have been conducted to study the roles of ncRNAs in cell biology, and accumulating evidences show that these RNA molecules do not constitute transcriptional noise but play important roles in cellular functions, such as transcriptional and posttranscriptional regulation, epigenetic regulation, organ or tissue development, cell differentiation, cell cycle control, cellular transport, metabolic processes, and chromosome dynamics. Their deregulation heavily contributes to various pathological conditions of human complex diseases, including breast cancer, hepatocellular cancer, prostate cancer, colon cancer, bladder cancer, thyroid cancer, lung cancer, ovarian cancer, leukemia, Alzheimer's disease, diabetes, and HIV. The development of computational models to predict the various complex ncRNA-related interactions could significantly benefit the inference of ncRNA function, the identification of ncRNA-disease associations, the detection of ncRNA biomarker, the identification of drug-target interactions, and drug design.

The papers that follow in the next pages describe recent findings in the field of computational models for prediction of the varied interactions related to ncRNAs. They represent only a fraction of the current research, the emerging trends, and applications of computational models for ncRNAs. The special issue consists of eight research papers and one review paper. The research papers include identification and annotation of ncRNAs (four papers), miRNA-target interaction prediction (one paper), ncRNA function prediction (one paper), miRNA-transcriptional factor interaction prediction (one paper), and protein-protein interaction prediction (one paper). A brief description of the papers follows.

G. Wan et al. analyzed RNA-seq data and ChIP-seq histone modification data to identify the relationship between lncRNA genes transcription and histone modification H3K4me3 or H3K27me3 in the paper "Transcriptional Regulation of lncRNA Genes by Histone Modification in Alzheimer's Disease."

J. Qu et al. detected 254 small noncoding RNAs in genome of *Pleurotus ostreatus* and analyzed the evolutionary conservation of them in the paper "Identification and Characterization of Small Noncoding RNAs in Genome Sequences of the Edible Fungus *Pleurotus ostreatus*."

A meta-path-based prediction method RMLM was developed in the paper "A Meta-Path-Based Prediction Method for Human miRNA-Target Association" by J. Luo et al. for predicting potential miRNA-target interactions. The authors also developed RMLMSe, in which sequence information was utilized to improve the performance of RMLM.

L. Jiang et al. employed backpropagation (BP) neural network together with 98-dimensional novel features for microRNA (miRNA) precursor identification in the paper "BP Neural Network Could Help Improve Pre-miRNA Identification in Various Species." The authors demonstrated that the total prediction accuracy of this method was nearly 13.17% greater than the state-of-the-art miRNA precursor prediction software tools.

A framework named Gene2Function to annotate Gene Reference into Functions (GeneRIFs) was given by Y. Hu et al. in the paper "Annotating the Function of the Human Genome with Gene Ontology and Disease Ontology."

M-fold, TargetScan, and GeneCoDia3 were used in the paper "Human Ribosomal RNA-Derived Resident MicroRNAs as the Transmitter of Information upon the Cytoplasmic Cancer Stress" for investigating RNA relationships between rRNA and miRNA against cellular stresses by M. Yoshikawa and Y. R. Fujii. The authors detected 17 RNA sequences identical with known miRNAs in the human rRNA and termed as rRNA-hosted miRNA analogs (rmiRNAs).

The paper "Ens-PPI: A Novel Ensemble Classifier for Predicting the Interactions of Proteins Using Autocovariance Transformation from PSSM" by Z.-G. Gao et al. provided a novel predictor based on the Rotation Forest (RF) algorithm combined with Autocovariance (AC) features extracted from the Position-Specific Scoring Matrix. The method achieved promising prediction performance when implemented on the protein-protein interaction datasets of Yeast, *H. pylori*, and independent datasets.

Q. Zhao et al. constructed a computational model of miRNA-mediated feed-forward loops (FFLs) in the paper "Effect of Dynamic Interaction between MicroRNA and Transcription Factor on Gene Expression." The authors introduced four possible structural topologies of FFLs with two gate functions (AND gate and OR gate) based on the different dynamic interactions between miRNA and TF on gene expression.

Finally, in the review paper "Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination" by S. Han et al., several popular methods for lncRNA identification such as Coding Potential Calculator (CPC), Coding-Potential Assessment Tool (CPAT), Coding-Non-Coding Index (CNCI), predictor of long noncoding RNAs and messenger RNAs based on an improved *k*-mer scheme (PLEK), Long noncoding RNA IDentification (LncRNA-ID), lncRScan-SVM, lncRNA-MFDL, and LncRNApred were summarized with their advantages, disadvantages, and application scopes.

Acknowledgments

We would like to thank the authors who submitted their work for consideration to this special issue as well as the reviewers for their efforts and constructive criticism. The work of Xing Chen was supported by National Natural Science Foundation of China under Grant nos. 11301517 and 11631014.

Xing Chen
Huiming Peng
Zheng Yin

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8496165, 14 pages
http://dx.doi.org/10.1155/2016/8496165

Review Article

Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination

Siyu Han,[1] Yanchun Liang,[1,2] Ying Li,[1] and Wei Du[1]

[1]*College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China*
[2]*Zhuhai Laboratory of Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Zhuhai College of Jilin University, Zhuhai 519041, China*

Correspondence should be addressed to Ying Li; liying@jlu.edu.cn and Wei Du; weidu@jlu.edu.cn

Received 12 August 2016; Revised 5 October 2016; Accepted 13 October 2016

Academic Editor: Xing Chen

Long noncoding RNA (lncRNA) is a kind of noncoding RNA with length more than 200 nucleotides, which aroused interest of people in recent years. Lots of studies have confirmed that human genome contains many thousands of lncRNAs which exert great influence over some critical regulators of cellular process. With the advent of high-throughput sequencing technologies, a great quantity of sequences is waiting for exploitation. Thus, many programs are developed to distinguish differences between coding and long noncoding transcripts. Different programs are generally designed to be utilised under different circumstances and it is sensible and practical to select an appropriate method according to a certain situation. In this review, several popular methods and their advantages, disadvantages, and application scopes are summarised to assist people in employing a suitable method and obtaining a more reliable result.

1. Introduction

Long noncoding RNAs (lncRNAs), one of the most poorly understood but also the most common RNA species, are those noncoding transcripts with length more than 200 nucleotides. Initially, people classified noncoding RNA (ncRNA) genes as "junk gene" or transcriptional "noise" [1]. Nonetheless, researchers found that about 70% of the genome is transcribed in various contexts and cell types [2, 3], about 80% of the genome has biochemical functions [4], and many DNAs code for RNAs as the end products instead of proteins [5]. LncRNAs are involved in a wide range of cellular mechanisms such as the regulation of genome activity [6], histone modifications [7, 8], and DNA methylation [9]. In addition, lots of studies have demonstrated that lncRNAs have a significant role in diverse biological processes; thus lncRNAs are especially important to the studies of human biology and diseases [10]. For example, in prostate cancer of human, lncRNA *SChLAP1* and chromatin remodelling complex SWI/SNF have opposing roles. *SchLAP1* has an interaction with the SNF5 subunit of SWI/SNF and inhibits binding of SWI/SNF to chromatin, which leads to genome-wide derepression of gene activity [11]. Moreover, aberrant expression of lncRNAs in cancer can be regarded as biomarkers and therapeutic targets because of its extremely specific expression [6]. The LncRNADisease database now integrates more than 1000 lncRNA-disease entries and 475 lncRNA interaction entries, which suggested that lncRNAs are associated with diseases closely [12].

Since lncRNAs so closely interact with diseases, many lncRNA-disease association detection tools are invented. Assuming that lncRNAs with similar functions tend to associate with similar diseases, a semisupervised method, Laplacian Regularized Least Squares for LncRNA-Disease Association (LRLSLDA) [13], was developed; this tool displays a satisfying result and needs no negative samples. Nonetheless, this method is facing the problems of parameter selection and classifier combination. The principal idea of

LRLSLDA, as mentioned above, is to measure the functional similarity of lncRNAs, which means that the performance of similarity calculation model largely determines the performance of association model. The similarity calculation model of LRLSLDA is LFSCM (LncRNA Functional Similarity Calculation based on the information of MiRNA) which is based on lncRNA-miRNA interactions and miRNA-disease associations. In 2015, novel lncRNA functional similarity calculation models (LNCSIM) [14] were provided by Chen et al. By integrating LRLSLDA and LNCSIM, the performance was enhanced. Recently, a new lncRNA functional similarity calculation model, FMLNCSIM (Fuzzy Measure-Based LncRNA Functional Similarity Calculation Model) [15], has been developed; this new model has a web interface (http://219.219.60.245/) for users' convenience. Considering that nowadays the experimentally confirmed data of miRNA-disease associations are much easier to obtain than the ones of lncRNA-disease, Chen [16] utilised the miRNA-disease association and miRNA-lncRNA interaction to identified lncRNA-disease association. This method (HGLDA) circumvents the utility of LncRNADisease database but still presents the desired results. Currently, many other tools, such as RWRlncD [17] and RWRHLD [18], were designed aiming at predicting lncRNA-disease association and obtaining more reliable results. Unfortunately, they have their own limitations [16]. As the titles of these methods implied, RWRlncD and RWRHLD mainly predict the association by utilising Random Walk with Restart (RWR). RWRlncD can only be applied to the case that lncRNAs have known related diseases and RWRHLD cannot deal with the circumstance that lncRNAs have unknown lncRNA-miRNA interactions. Another method, Improved Random Walk with Restart for LncRNA-Disease Association (IRWRLDA) [19], is also based on RWR, but IRWRLDA can predict the associations even when diseases show no known related lncRNAs.

Research [20] has illustrated the lncRNA-disease association extensively and comprehensively. Basically, there are three approaches to performing lncRNA-disease association prediction [20]: to train a model based on machine learning algorithm; to construct a heterogeneous network; or to integrate lncRNA-miRNA interactions and miRNA-disease associations. Currently, researches have acknowledged that it is imperative to analyse the role of lncRNAs in many diseases especially cancer, but the first step and fundamental work is how to discriminate lncRNAs from genes. With the rapid development of next-generation sequencing technologies, thousands and thousands of transcriptomes have been discovered, which furnished us with more and more useful information on ncRNAs. Meanwhile, many ncRNAs identification approaches have been developed to facilitate the researches and analyses. Each kind of ncRNA has its own prediction tools such as tRNAscan-SE (1997) [21] and tRNA-Predict (2015) [22] for transfer RNA (tRNA) identification; mirnaDetect (2014) [23] and imDC (2015) [24] for microRNA (miRNA) prediction; and RNAmmer (2007) [25] for ribosomal RNA (rRNA) discrimination. Both tRNA-Predict and mirnaDetect are constructed with the features of secondary structure and codon-bias. The method imDC is an algorithm of ensemble learning to deal with imbalanced data and is applied to miRNA classification. The research area of ncRNA is fast growing. However, it is still a challenge to distinguish lncRNAs from protein-coding genes in that lncRNAs share many features similar to mRNAs. Moreover, the incomplete transcripts or genes poorly annotated or containing sequencing errors also thwart the discrimination and functional inference. During the last ten years, many efforts on lncRNA identification have been made and many approaches have been developed to make a more accurate discrimination. Several studies [26, 27] have summarised and reviewed the approaches of ncRNAs identification and analysis, but a few report the discussion of lncRNAs prediction methods. Wang et al. [26] discussed several ncRNA detection methods based on homology information and common features. Different approaches aiming at detecting different kinds of ncRNAs are presented and an overview of some useful tools was given, yet no analysis on application scopes was provided. Hence, the summary of these methods is more theoretical than practical. Veneziano et al. [27] summarised some computational approaches of ncRNA analysis based on deep sequencing technology. Some lncRNA prediction tools were discussed briefly but many other helpful tools were excluded.

In this paper, we mainly focus on the tools for lncRNA identification. The aim of this paper is to summarise the popular algorithms of lncRNA identification and to assist researchers in determining which method is more appropriate for their purpose. Here, comprehensive analyses and discussions of these tools were provided. Then, we compared several popular machine learning based methods, including Coding Potential Calculator (CPC) [28], Coding Potential Assessment Tool (CPAT) [29], Coding-Non-Coding Index (CNCI) [30], predictor of long noncoding RNAs and messenger RNAs based on an improved k-mer scheme (PLEK) [31], Long noncoding RNA IDentification (LncRNA-ID) [32], and lncRScan-SVM [33]. In addition, lncRNA-MFDL [34] and LncRNApred [35], two artificial neural network- (ANN-) involved tools, are also introduced in this paper. However, the provided access link of lncRNA-MFDL has been forbidden; LncRNApred often throws errors while handling massive-scale data which can be processed by CPC and CPAT successfully. Thus, we only briefly introduce the algorithms of the classification model but omit the discussions of application scope. We expect that this review can be a practical manual when readers conduct lncRNA identification researches.

CPC (2007) is used to assess the protein-coding potential of transcripts with high accuracy and speed [28]. However, with the emergence of new programs, speed is scarcely considered as a merit. The features of CPC can be divided into two categories. The first one is based on the extent and quality of the Open Reading Frame (ORF), and the other category is derived from BLASTX research. The authors employed the LIBSVM package to train support vector machine (SVM) model with the standard radial basis function kernel [36].

CPAT (2013) is another protein-coding potential assessment tool based on the model of logistic regression. The selected features include the quality of the ORF, Fickett Score, and hexamer score. Fickett Score is used to evaluate each base's unequal content frequency and asymmetrical

Table 1: Overview of the methods concerning lncRNA identification.

	Published year	Testing datasets	Training species	Model	Query file format	Web interface
CPC	2007	ncRNA*	Eukaryotic	SVM	FASTA	Yes
CPAT	2013	lncRNA*	Human; mouse; fly; zebrafish	LR	BED; FASTA	Yes
CNCI	2013	lncRNA	Human; plant	SVM	FASTA; GTF	No
PLEK	2014	lncRNA	Human; maize	SVM	FASTA	No
lncRNA-MFDL	2015	lncRNA	Human	DL	*Unknown***	*Unknown***
LncRNA-ID	2015	lncRNA	Human; mouse	RF	BED; FSATA	No
lncRScan-SVM	2015	lncRNA	Human; mouse	SVM	GTF	No
LncRNApred	2016	lncRNA	Human	RF	FASTA	Web only

Testing datasets denote that one specific method is developed to discriminate ncRNAs or lncRNAs from protein-coding transcripts. The classification model of CPC, CNCI, PLEK, and lncRScan-SVM is support vector machine (SVM); CPAT employs logistic regression (LR); LncRNA-ID and LncRNApred utilise random forests (RF) and lncRNA-MFDL uses deep stacking networks (DSNs) of deep learning (DL) algorithm.

*Note that the most popular tool CPC is trained and tested on datasets of ncRNAs and protein-coding transcripts. The training datasets of CPAT are also ncRNAs and protein-coding transcripts, though test on lncRNAs for CPAT is conducted and achieved a higher accuracy.

**The access link of lncRNA-MFDL has expired; thus, we cannot verify the information that the original paper failed to mention.

distribution in the positions of codons in one sequence. Hexamer score is mainly based on the usage bias of adjacent amino acids in proteins.

CNCI (2013) is a classifier to differentiate protein-coding and noncoding transcripts by profiling the intrinsic composition of the sequence. According to the unequal distribution of adjoining nucleotide triplets (ANT) in two kinds of sequences, a $64 * 64$ ANT Score Matrix is constructed to evaluate the sequence and the sliding window is used as a supplement to achieve a more robust result [30]. ANT bears some similarities to the hexamer score of CPAT, but much more comprehensive and intricate analysis was conducted to facilitate the incomplete transcripts classification. The classification model of CPAT is SVM with a standard radial basis function kernel.

PLEK (2014) uses *k*-mer scheme and sliding window to analyse the transcripts. For multiple species, PLEK does not have too many advantages over CNCI on testing data of normal sequence. Nevertheless, compared with PLEK, the results of CNCI will deteriorate when the sequence contains some insert or deletion (indel) errors. These errors are very common in today's sequencing platforms. The classification model of PLEK is SVM with a radial basis function kernel.

LncRNA-ID (2015) has 11 features which can be categorized according to ORF, ribosome interaction, and the conservation of protein. The first category is similar to the ORF features in CPC and CPAT. The foundation of the second feature category is the interactions between mRNAs and ribosomes during protein translation since some studies displayed that lncRNAs can be associated with ribosomes [37, 38] but do not show the release of ribosomes [39]. The profile hidden Markov model-based alignment is used to assess the conservation of protein. The classification model of LncRNA-ID is improved using random forest which assists LncRNA-ID effectively in handling imbalanced training data.

Some tools are initially designed to predict ncRNAs but can also be applied to lncRNAs prediction, such as Phylogenetic Codon Substitution Frequencies (PhyloCSF, 2011) [40] and RNAcon (2014) [41]. Based on nucleotide substitutions and formal statistical comparison of phylogenetic codon models [40], PhyloCSF utilises multiple sequence alignments to find conserved protein-coding regions. As an alignment-based method, PhyloCSF entails high-quality alignments and suffers from low efficiency. RNAcon mainly predicts ncRNAs utilising *k*-mer scheme. Based on graph properties [41, 42], RNAcon can also perform ncRNAs classification and classify different ncRNA classes.

Some methods are especially developed for long intergenic noncoding RNAs (lincRNAs, one subgroup of lncRNAs) classification, such as iSeeRNA (2012, web server and L inux binary package available at http://137.189.133.71/iSeeRN/index.html) [43] and LincRNA Classifier based on selected features (linc-SF, 2013) [44]. iSeeRNA built a SVM model with three feature groups: ORF; adjoining nucleotides frequencies (GC, CT, TAG, TGT, ACG, and TCG); and conservation score obtained from Phast [45]. The classifier of linc-SF evaluates the sequences with the criteria of sequence length, GC content, minimum free energy (MFE), and *k*-mer scheme.

2. Details of the Methods

In this part, we will discuss the machine learning models and the selected features of each method more specifically. Firstly, for users' convenience, some brief information of each method is displayed in Table 1 and the details of using are summarised. Then the details of each method are provided in the following. Table 2 is a summary about the features selected by each method.

2.1. Details of Using. CPC can be downloaded from http://cpc.cbi.pku.edu.cn/download/. CPC has a user-friendly web interface at http://cpc.cbi.pku.edu.cn/programs/run_cpc.jsp. Documents and User Guide are provided at the website. To run CPC on a local PC, a comprehensive protein reference database is required and users can download it from ftp://ftp.ncbi.nlm.nih.gov/blast/db/ or ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref90/. About 20 gigabytes (GB) of free space is also needed for storing the protein reference database.

Table 2: Summary of the features of each method selected.

	ORF	Codon	Sequence structure	Ribosome interaction	Alignment	Protein conservation
CPC	Quality; coverage; integrity	No	No	No	BLASTX	Number and *E*-value of hits; Distribution of hits
CPAT	Length; coverage	Hexamer Frequency	Content of the bases Position of the bases	No	No	No
CNCI	No	ANT matrix; Codon-bias	MLCDS	No	No	No
PLEK	No	No	Improved *k*-mer scheme	No	No	No
lncRNA-MFDL	Length; coverage	No	*k*-mer scheme Secondary structure MLCDS	No	No	No
LncRNA-ID	Length; coverage	No	Kozak motif	Ribosome release signal Changes of binding energy	Profile HMM based alignment	Score of HMMER Length of the profile Length of aligned region
lncRScan-SVM	No	Distribution of stop codon	Score of txCdsPredict; length of transcripts; length and count of exon	No	Phylo-HMM based alignment	Average PhastCons scores
LncRNApred	Length; coverage	No	Length of the sequence; signal to noise ratio; *k*-mer scheme; G + C content	No	No	No

All features are categorized into six groups according to the similarity or basic principles. Thus, some items in the table might not be exactly in one-to-one correspondence with the feature names given in the corresponding published references.

CPAT is also available both for download and as a web-server. Users can obtain the latest resource code from https://sourceforge.net/projects/rna-cpat/files/?source=navbar. Pre-releases, tutorial files, and examples are also supplied on the pages. CPAT requires Python 2.7.x; numpy; cython; and R when running offline. The web server is available at http://lilab.research.bcm.edu/cpat/index.php.

CNCI can be downloaded at https://github.com/www-bioinfo-org/CNCI. Version 2 is updated on Feb 28, 2014. Setup and running steps are attached on the websites. Libsvm-3.0 has been enclosed in the package. Other additional files can be downloaded at http://www.bioinfo.org/np/.

PLEK was implemented by C and Python. The source code can be freely downloaded from https://sourceforge.net/projects/plek/files/. Several videos to assist user in utilising PLEK correctly are also provided. Python 2.7.x is required.

Scripts of LncRNA-ID can be obtained at https://github.com/zhangy72/LncRNA-ID.

LncRScan-SVM provided scripts, gene annotation files, and datasets. The scripts can be downloaded at https://sourceforge.net/projects/lncrscansvm/?source%20=%20directory. A Readme file is also attached on this site.

All the stand-alone versions of these tools require Linux/UNIX operating system.

The link of lncRNA-MFDL provided is https://compgenomics.utsa.edu/lncRNA_MDFL/. LncRNApred only has the web interface and is available at http://mm20132014.wicp.net:57203/LncRNApred/home.jsp. However, the link of lncRNA-MFDL expired when we did this research. And LncRNApred only provides a web server which cannot handle too many sequences at one time.

2.2. CPC in Detail. CPC [28] extracted six features to evaluate the coding potential of transcripts. Log-odds score, coverage, and integrity of ORF are used to assess the ORFs of one sequence. ORFs are predicted by *framefinder*. A high-quality

ORF tends to have a high log-odds score and a larger ORF coverage. The integrity of ORF means ORFs in protein-coding transcripts are disposed of to begin with a start codon and end with a stop codon. The other three features are number of hits, hit score, and frame score, which are derived from the output of BLASTX search. A protein-coding transcript prefers more hits in alignment with lower E-values. Then the hit score is defined as follows [28]:

$$S_i = \underset{j}{\text{mean}} \left\{-\log_{10} E_{ij}\right\}, \quad i \in 0, 1, 2,$$
$$\text{Hit Score} = \underset{i\in\{0,1,2\}}{\text{mean}} \{S_i\} = \frac{\sum_{i=0}^{2} S_i}{3}, \tag{1}$$

where E_{ij} is the E-Value of the jth hits in the ith ORF. A noncoding transcript may also contain some hits, but these hits are inclined to scatter in three frames rather than be located in one. The frame score to calculate the distribution of hits among three ORFs is defined in the following:

$$\text{Frame Score} = \underset{i\in\{0,1,2\}}{\text{variance}} \{S_i\} = \frac{\sum_{i=0}^{2} \left(S_i - \overline{S}\right)^2}{2}. \tag{2}$$

Thus, a protein-coding transcript will achieve a higher hit score and frame score because of the lower E-value and biased distribution of the hits.

The training data of CPC [46] are eukaryotic ncRNAs from the RNAdb [47] and NONCODE [48, 49] databases. CPC is designed to assess transcripts' protein-coding potential, which means it will have high accuracy of discriminating protein-coding transcripts. Moreover, CPC also has the error tolerance capacity, which owes much to *framefinder*'s accurate prediction. *Framefinder* performed well even though input transcripts may have some point mutations, indel errors, and truncations. CPC is slightly inferior in distinguishing noncoding transcripts in respect of the fact that lncRNAs may contain putative ORFs and transcript length is also familiar to protein-coding transcripts. The slow speed is another imperfection of CPC.

2.3. CPAT in Detail. CPAT [29] is an alignment-free program. CPAT uses a logistic regression model and can be trained on own data of users. Apart from the features of maximum length and coverage of ORF akin to CPC, Fickett Score is another criterion. Fickett Score can be regarded as a dependent classifier; it is mainly based on calculating the position of each base favoured and the content of each base in the sequence [50]. The base's position parameter of CPAT is defined as follows:

$$A_1 = \text{Number of As in positions } 0, 3, 6, \ldots$$
$$A_2 = \text{Number of As in positions } 1, 4, 7, \ldots$$
$$A_3 = \text{Number of As in positions } 2, 5, 8, \ldots$$
$$A\text{-position} = \frac{\max\left(A_1, A_2, A_3\right)}{\min\left(A_1, A_2, A_3\right) + 1},$$
$$A\text{-content} = \frac{\text{Occurence Number of } A}{\text{Total Number of all bases}}, \tag{3}$$

where A in the formula means the base A and the other three bases are measured in a similar way. The parameter of position calculates each base's favoured position and the parameter of content is the percentage of each base in the sequence. Then according to distributions of eight parameters' values [50], it is easy to obtain the probability that the sequence will be a protein-coding transcript. Next, each probability is multiplied by a weight to make a more accurate result. The weight is the percentage of the times that the estimate of each parameter alone is correct. Finally, according to the above descriptions, Fickett Score can be determined as follows:

$$\text{Fickett Score} = \sum_{i=1}^{8} p_i w_i. \tag{4}$$

According to Fickett [50], Fickett Score alone can correctly discriminate about 94% of the coding segments and 97% of the noncoding segments with 18% of "No Opinion."

The last feature of CPAT is hexamer score, which is the most discriminating feature. Hexamer means the adjacent amino acids in proteins. The features of the in-frame hexamer frequency of coding and noncoding transcripts are calculated and hexamer score is defined in the following:

$$\text{Hexamer Score} = \frac{1}{m} \sum_{i=1}^{m} \log\left(\frac{F\left(H_i\right)}{F'\left(H_i\right)}\right). \tag{5}$$

There are $64 * 64$ kinds of hexamers, and i denotes each hexamer. $F(H_i)$ $(i = 0, 1, 2, \ldots, 4095)$ means the in-frame hexamer frequency of protein-coding transcripts, while $F'(H_i)$ means noncoding transcripts. For a transcript containing m hexamers, a positive hexamer score indicates a protein-coding transcript.

A high-quality training dataset is constructed containing 10,000 protein-coding transcripts selected from RefSeq database with the annotations of the Consensus Coding Sequence project and 10,000 noncoding transcripts randomly collected from GENCODE database. CPAT is prebuilt hexamer tables and logit models for human, mouse, fly, and zebrafish. Meanwhile, CPAT uses pure linguistic features to facilitate discrimination of the poorly annotated transcripts. CPAT has an efficient offline program and also provides a user-friendly web interface.

2.4. CNCI in Detail. CNCI [30] is mainly based on sequence intrinsic composition, it evaluates the transcripts by calculating the usage frequency of adjoining nucleotide triplets (ANT). Firstly, two ANT matrices are constructed based on the usage frequency of ANT in noncoding sequences and coding region of the sequences (CDS). For 4,096 ANT, the

formulas to calculate each ANT usage frequency are defined as follows:

$$X_iN = \sum_{j=1}^{n} S_j(X_i),$$
$$T = \sum_{i=1}^{m} X_iN = \sum_{i=1}^{m}\sum_{j=1}^{n} S_j(X_i); \quad m = 64 \times 64;\ n = 1, \ldots, N, \tag{6}$$
$$X_iF = \frac{X_iN}{T},$$

where X means one kind of ANT; $S_j(X_i)$ is the occurrence number of X_i in one sequence S_j. Thus, X_iN denotes the total occurrence number of one kind of ANT in the dataset while T indicates the total occurrence number of all kinds of ANT in the dataset. Accordingly, X_iF is the usage frequency of ANT. Then the ANT Score Matrix is utilised, which is the $\log_2$-ratio of the two above-mentioned ANT matrices, to score a sequence and make a discrimination.

$$\text{ANT Score Matrix} = \log_2 \frac{\text{CDS Matrix}}{\text{Non-coding Matrix}}. \tag{7}$$

The distinguishing results of ANT Score Matrix are fairly well, but the matrix is constructed by computing the ANT usage frequency of coding region and noncoding region; consequently the untranslated region (UTR) of the entire sequence will interfere with the performance of discrimination. The sliding window is employed with one ANT (3 nt) in each scan step to identify the CDS of a sequence by scanning six reading frames of each sequence. The different sizes $(30, 60, 90, \ldots, 300\,\text{nt})$ of the sliding windows are examined and the size of 150 nt for this classification model is found to obtain the most robust result. For a sequence consisting of k ANT, there will be $k - 1$ segments in this sequence. Based on the ANT Score Matrix, each segment will get an S-Score, and each reading frame can obtain an array comprised of the S-Scores. The formula of S-Score is defined as follows:

$$S\text{-Score} = \sum_{i=1}^{n} \left\{H_p(X_i)\right\}, \tag{8}$$

where X means ANT, H_p is the ANT Score Matrix, and n is the total number of the ANT in one segment or the whole sequence. Hence, a correct reading frame of coding transcript tends to have a higher whole sequence S-Score and, in this array of reading frame, the region composed of consecutive high S-Scores is the CDS. For long noncoding transcripts, the Maximum Interval Sum [51] program is used to identify the most-like CDS (MLCDS) which is the region that gained the largest sum of consecutive S-Scores in each reading frame. Among those six MLCDS, the length and S-Score of the MLCDS with the highest value are selected as the features of CNCI. Furthermore, the features of the LENGTH-Percentage, SCORE-Distance, and codon-bias are also selected to improve accuracy:

$$\text{LENGTH-Percentage} = \frac{M1}{\sum_{i=0}^{n}(Y_i)},$$
$$\text{SCORE-Distance} = \frac{\sum_{j=0}^{n}\left(S - E_j\right)}{5}, \tag{9}$$

where $M1$ is the length of the MLCDS with the highest S-Score, Y_i is the length of each MLCDS, S is the highest S-Score among six MLCDS, and E_j is the S-Score of other five MLCDS. Codon-bias (3-mer frequencies) is a parameter to evaluate the usage bias of different codons in protein-coding or long noncoding transcripts. The $\log_2$-ratio of occurrence frequency of each codon (stop codons are excluded) in protein-coding genes and lncRNAs is calculated, and most codons have distinct usage bias in two kinds of sequences.

The training datasets of CNCI contain protein-coding transcripts selected from RefSeq database and long noncoding transcripts selected from GENCODE [52]. The CNCI is applied to other species with the aim of examining the scope of application. The results of vertebrates (except birds), especially mammals, can be accepted since the program was trained on human gene set. CNCI can be used to discriminate incomplete transcripts, especially those high-throughput sequencing data of poorly explored species.

2.5. PLEK in Detail. PLEK [31] is an alignment-free tool based on k-mer frequencies of the sequences. For a given sequence, the sliding windows with size of k scan 1 nt as a step forward. k ranges from 1 to 5, which is a trade-off between accuracy and computational time. Thus, for a sequence consisting of A, C, G, and T, the $4^1 + 4^2 + 4^3 + 4^4 + 4^5 = 1{,}364$ patterns can be obtained. Then the following formulas can be used:

$$f_i = \frac{c_i}{s_k} w_k, \quad k = 1, 2, 3, 4, 5;\ i = 1, 2, \ldots, 1364,$$
$$s_k = l - k + 1,$$
$$w_k = \frac{1}{4^{5-k}}, \tag{10}$$
$$k = 1, 2, 3, 4, 5,$$

where i is the number of the patterns; c_i denotes the number of the segments in sliding windows matching with patterns; s_k denotes the total of the segments when sliding window slides along the sequence with the size of k. Therefore, f_i is the usage frequency multiplied by a factor w_k which is used to facilitate the discrimination.

A balanced training dataset is conducted with all 22,389 long noncoding transcripts collected from the GENCODE dataset [52–54] and 22,389 protein-coding transcripts randomly selected from the human RefSeq dataset [55, 56]. Though the training model of PLEK is human, PLEK can still be applied to other vertebrates. PLEK is particularly designed for the transcripts acquired from current sequencing platforms which consist of some indel errors commonly.

For these transcripts, the performance of PLEK is better than CPC and CNCI. PLEK can be trained with users' own datasets, but it may take a long time to be accomplished.

2.6. LncRNA-MFDL in Detail. LncRNA-MFDL [34] is based on feature fusion and deep learning algorithm. LncRNA-MFDL has four kinds of features which are integrated to build a classification model based on deep stacking networks (DSNs, one kind of deep learning algorithm) [57, 58]. Four feature groups of lncRNA-MFDL include k-mer; secondary structure; ORF, obtained by utilising txCdsPredict program (http://genome.ucsc.edu/) [59]; and MLCDS features which are inspired by CNCI [30].

The k-mer scheme employed in lncRNA-MFDL is unlike the one in PLEK. Here, the k only ranges from 1 to 3, but the frequencies are calculated on the regions of the whole sequence and ORF at the same time. Considering that the secondary structure is more conserved and stable than primary structure, a representative criterion, the minimum free energy (MFE), is used to assess the secondary structure of the transcripts. Utilising RNAfold program of ViennaRNA Package [60], the MFE, the ratio of MFE to sequence length, and the number of paired bases and unpaired bases can be easily obtained.

2.7. LncRNA-ID in Detail. LncRNA-ID [32] has three categories of features as mentioned earlier. Except for the length and coverage of ORF, the features based on translation mechanism and protein conservation are extracted.

Many studies [61–63] have demonstrated that several nucleotide sites in Kozak motif play a prominent role during the initiation of protein translation. An efficient translation indicates that the highly conserved nucleotides appear at the positions $\{-3, +4\}$ and $\{-2, -1\}$ of Kozak motif GCC*R*CCAUGG (*R* represents purine and the position of A in start codon AUG is +1). Thus, these conserved sites are more likely to exist in protein-coding transcripts. Moreover when the translation starts, the binding energy will change along with the interaction between the $3'$ end of rRNAs and mRNA transcripts. The Ribosome Coverage to calculate the changes of the binding energy is defined as follows:

$$\text{Ribosome Coverage} = \sum_{i=1}^{L} \{N_i \mid \delta_i < 0\}, \tag{11}$$

where δ_i is the free energy at position i and N_i is the number of base pairs starting at position i in a sequence with the length of L. Next, the three levels of ribosome occupancy by computing Ribosome Coverage on three regions, respectively, are obtained: the whole transcript, ORF, and $3'$UTR. Accordingly, a true protein-coding transcript tends to attain higher Ribosome Coverage on the whole transcript and the ORF region. When the translation terminates, the ribosomes will be released from protein-coding transcripts. Therefore, it is likely to capture a considerable drop of ribosome occupancy when ribosomes reach stop codons. The Ribosome Release Score to capture this change of ribosome occupancy is defined:

$$\text{Ribosome Release Score} = \frac{\text{Ribosome coverage of ORF/length (ORF)}}{\text{Ribosome coverage of } 3'\text{UTR/length}\left(3'\text{UTR}\right)}, \tag{12}$$

and a protein-coding transcript inclines to exhibit a higher Ribosome Release Score. For protein translation category, the selected features including nucleotides at two positions of Kozak motif, Ribosome Coverage on three regions, and Ribosome Release Score are selected.

The protein conservation of the sequences is evaluated according to profile hidden Markov model-based alignment scores. HMMER [64] is a software suite for sequence homology detection using probabilistic methods. LncRNA-ID employed HMMER with the E-value cutoff of 0.1 to align the transcripts against all available protein families. A protein-coding transcript is expected to get a higher score, longer aligned region, and a reasonable length of the profile in the alignment.

In human genome, although the amount of lncRNA is at least four times more than protein-coding genes [65], the majority class in training data is protein-coding transcript on account of poorly annotated lncRNA. Hence, the classification model of this method is balanced random forest [66, 67] which is derived from random forest but could utilise the sufficient protein-coding data and avoid inaccurate results caused by the imbalanced training data at the same time. The human prebuilt model of LncRNA-ID contains 15,308 protein-coding transcripts and 4586 lncRNAs from GENCODE [52]. For mouse, the training datasets are comprised of 22,033 protein-coding transcripts and 2,457 lncRNAs randomly selected from GENCODE. These two datasets were also used to draw receiver operation characteristic (ROC) curves in the next section (Figure 2). Users can train LncRNA-ID with their own dataset and apply it to various species.

2.8. LncRScan-SVM in Detail. LncRScan-SVM [33] classifies the sequences mainly by evaluating the qualities of nucleotide sequences, codon sequence, and transcripts structure. The counts and average length of exon in one sequence are calculated. The protein-coding transcripts are disposed of to include more exons, thus having a longer exon length than lncRNA. Another feature is the score of *txCdsPredict*. This third-part program from UCSC genome browser [68] can determine if a transcript is protein-coding. Conservation score is obtained by calculating the average of PhastCons scores [45] from Phast (http://compgen.cshl.edu/phast/). Transcript length and standard deviation of stop codon counts between three ORFs are the last two features.

The reliable datasets are constructed from GENCODE [54] composed of 81,814 protein-coding transcripts and 23,898 long noncoding transcripts of human. And, for mouse, 47,394 protein-coding transcripts and 6,053 long noncoding transcripts from GENCODE [52] are also contained within the dataset. After being trained on human and mouse

datasets, lncRScan-SVM obtains a good performance on lncRNA prediction.

2.9. LncRNApred in Detail. Before constructing the classifier, self-organizing feature map (SOM) clustering [69] is employed to select representative samples as the training dataset, which enhanced the performance of LncRNApred. As to the features, the length and coverage of the longest ORF, one of the classical and typical features, are selected as the criteria. In addition, G + C content, k-mer (k is from 1 to 3 just like lncRNA-MFDL), and length of the sequence are also the features of LncRNApred. The novel idea of LncRNApred is SNR, which transforms one sequence into four binary numeric sequences:

$$u_b = \begin{cases} 1, & S[n] = b, \\ 0, & S[n] \neq b, \end{cases} \quad n = 0, 1, 2, \ldots, N-1, \; b \in \{A, T, C, G\}, \tag{13}$$

where b means four kinds of bases, N is the length of one sequence, and $S[n]$ denotes a sequence of length N. Thus, there will be four binary sequences $\{u_b \mid b \in (A, T, C, G)\}$. Then applying Discrete Fourier Transform (DFT) to these four binary numeric sequences, the power spectrum $\{P[k]\}$ can be obtained:

$$U_b[k] = \sum_{n=0}^{N-1} u_b[n] \, e^{-i(2\pi nk/N)}, \quad k = 0, 1, \ldots, N-1,$$
$$P[k] = \left|U_A[k]\right|^2 + \left|U_T[k]\right|^2 + \left|U_G[k]\right|^2 + \left|U_C[k]\right|^2. \tag{14}$$

The studies of Fickett [50, 70] have presented that positions and compositions of four bases are different in lncRNAs and protein-coding RNA, and, because of this, the power spectrum of one protein-coding transcript will have a peak at $N/3$ position. Hence, the SNR is defined as follows:

$$\overline{E} = \frac{\sum_{k=0}^{N-1} P[k]}{N}, \quad \mathrm{SNR} = \frac{P[N/3]}{\overline{E}}. \tag{15}$$

Now, there are 89 features: the length and coverage of the longest ORF, the length of the sequence, SNR, G + C content, and 4 + 16 + 64 features of k-mer. Noticing that not all the features have high discriminative power, the feature selection is made and 25 high-quality features are determined from the original 84 features of k-mer. Finally, 30 features are selected to build a random forest model. The performance of random forest is largely determined by training set. Therefore, the clustering method is used to find out the most adequate sequences to form a high standard training set. The clustering method SOM [69] achieved the best result and was chosen to select characteristic sequences

An overall procedure of these eight tools is displayed in Figure 1.

3. Performance of These Methods

To quantify the classification performance under one unified standard, we first characterise lncRNAs as the positive class and protein-coding transcripts as the negative class; then the performance of these tools can be evaluated with several standard criteria defined as follows:

$$\begin{aligned} \text{Sensitivity} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \text{Specificity} &= \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \\ \text{Accuracy} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}, \\ \text{False Positive Rate} &= \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \end{aligned} \tag{16}$$

As one of the most popular methods, CPC is especially designed for assessing protein-coding potential and performed fairly well for discriminating protein-coding transcript. It enjoys the best results when screening the coding transcripts. For 10,000 protein-coding genes and 10,000 lncRNAs selected from UCSC genome browser (GRCh37/hg19), CPC picked up about 97.62% coding transcripts while CPAT distinguished 85.28% of them. CPAT also outperforms CPC with 89.94% accuracy [33]. Table 3 shows the performance of these tools on the same testing dataset. CPC picks up 99.97% of human protein-coding genes collected from GENCODE, in comparison with the latest program LncRNA-ID whose performance is 95.28%. However, the performance of CPC appears to somewhat decline when focusing on the capability of discriminating noncoding transcripts, especially long noncoding transcripts: CPC only picked up 66.48% of human's long noncoding transcripts while the results of CPAT, PLEK, and LncRNA-ID are 86.95%, 99.52%, and 96.28%.

CPC and CPAT are the programs to assess the coding potential, but CNCI is especially used to classify protein-coding and long noncoding transcripts. With the sequences becoming longer and longer, CNCI was more superior to CPC. According to Sun et al. [30], when the length of transcript is longer than 2,000 nt, the accuracy of CPC is only around 0.4 while the CNCI still has an outstanding performance. The training dataset of CNCI is human but this method still achieved more than 90% accuracy in other vertebrates apart from the birds [30]. PLEK is tested on two datasets sequenced by PacBio and 454 platforms (refer to Table 3). Among the tools being compared, CPC still picked up about 99.90% coding genes though this figure is not that useful because it can only distinguish 19.00% and 47.20% lncRNAs. CNCI displayed better performance on both datasets, but PLEK even achieved a more satisfying result.

LncRNA-ID is another method to identify the long noncoding transcripts. Compared with other programs, LncRNA-ID strikes a good balance between sensitivity and false positive rate. According to Table 3, it is noticeable that lncRNA is better than PLEK but slightly inferior to CPC and CPAT on the testing data of coding genes, and the

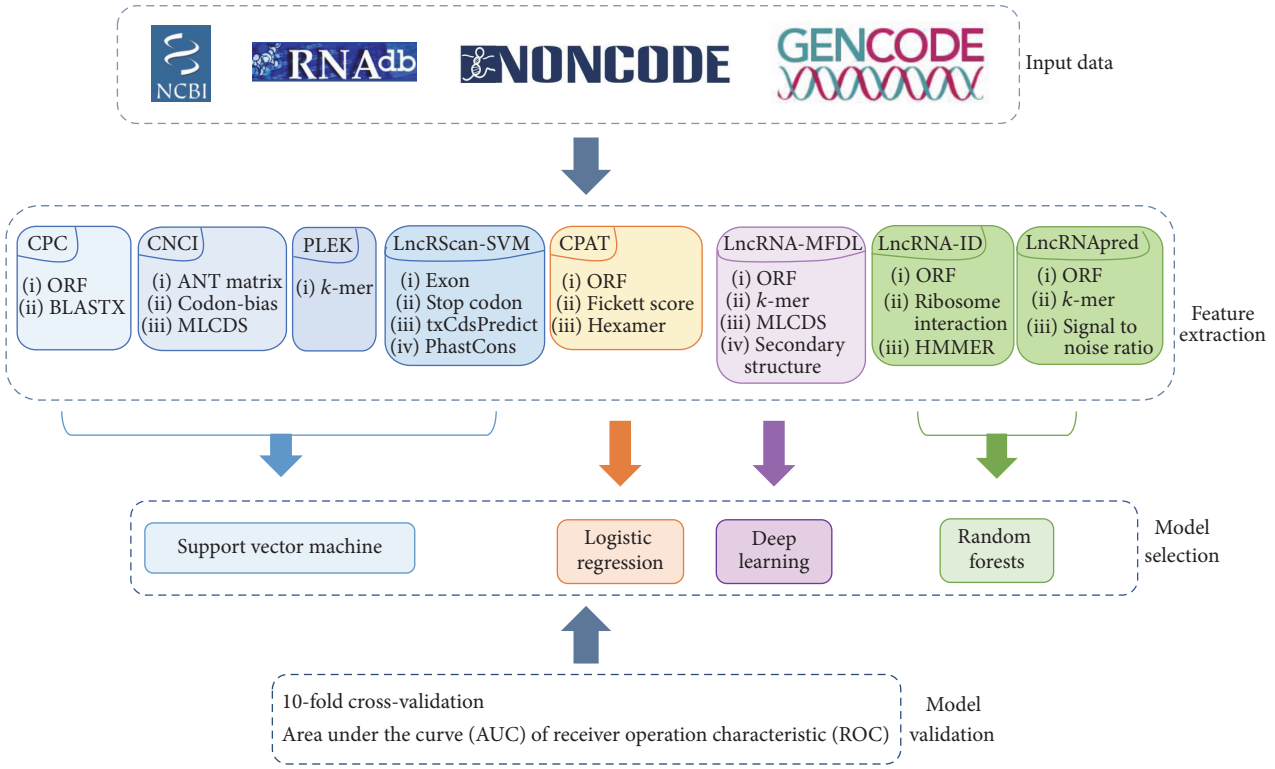


Figure 1: An overall procedure of eight tools. The features of each tool are sorted into several groups and only the categories of the features are listed in the figure.

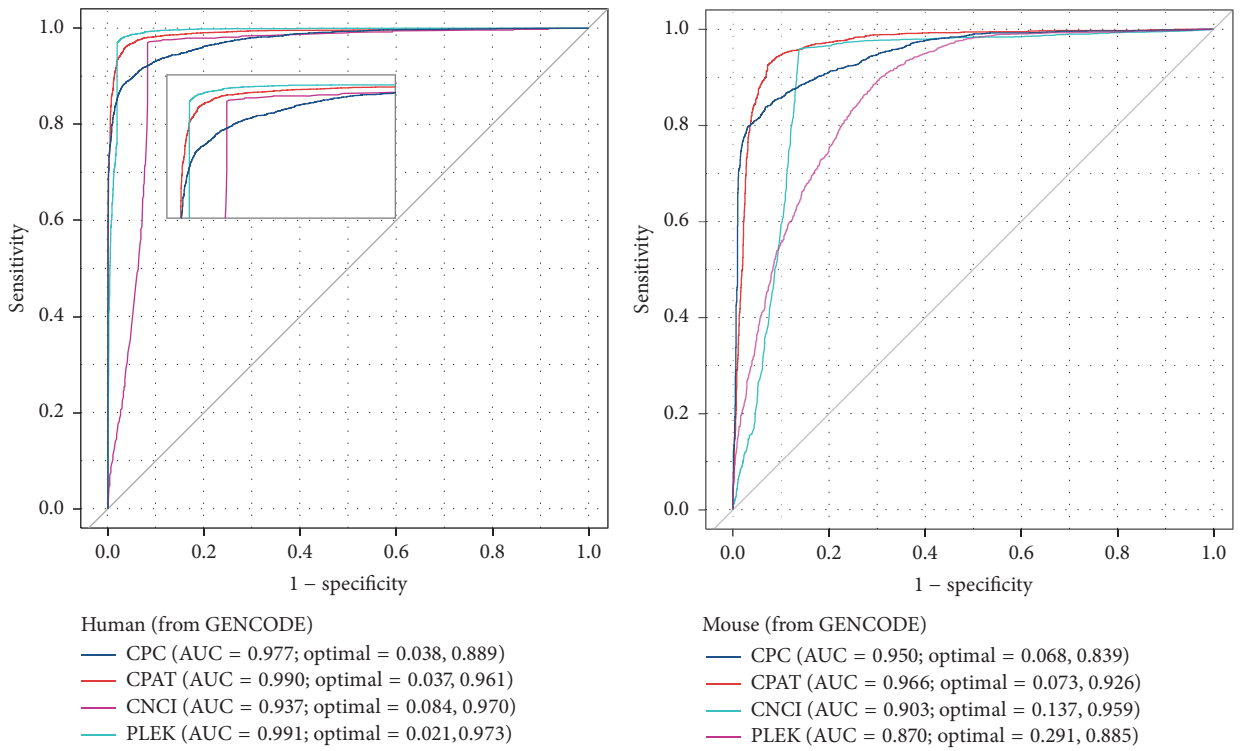


Figure 2: The ROC curves of CPC, CPAT, CNCI, and PLEK. We assessed the models using the same datasets as LncRNA-ID (selected from GENCODE) used. Both CPC and CPAT were evaluated with the latest versions.

Table 3: Overview of each tool's performance on different testing datasets.

Testing dataset	CPC	CPAT	CNCI	PLEK	LncRNA-ID	lncRScan-SVM
Human MCF-7 (PacBio)[1]						
Specificity	**99.90**		91.80	94.70		
Sensitivity	19.00		78.70	**95.80**		
Accuracy	**97.00**		91.30	94.70		
Human HelaS3 (454)[2]						
Specificity	**99.90**		93.90	95.50		
Sensitivity	47.20		81.10	**92.50**		
Accuracy	**99.00**		93.70	95.40		
Human (from GENCODE)[3]						
Specificity	**99.97**	99.55		89.18	95.28	
Sensitivity	66.48	86.95		**99.52**	96.28	
Accuracy	83.22	93.25		94.32	**95.78**	
Mouse (from GENCODE)[4]						
Specificity	98.75	**98.95**		70.94	92.10	
Sensitivity	76.55	38.80		88.11	**94.45**	
Accuracy	87.65	68.88		79.49	**93.28**	
Human (from GRCh37/hg19)[5]						
Specificity	**97.62**	85.28				89.20
Sensitivity	67.23	**94.60**				93.88
Accuracy	82.43	89.94				**91.94**
Mouse (from GRCm38/mm10)[5]						
Specificity	**98.37**	88.17				89.14
Sensitivity	75.46	**95.34**				95.29
Accuracy	86.91	91.76				**92.21**

The results of the tools being tested on the same datasets are listed above. Bold numbers denote the highest value of the metrics.
[1]MCF-7 is available at http://www.pacb.com/blog/data-release-human-mcf-7-transcriptome/; [2]dataset of HelaS3 is available at https://www.ncbi.nlm.nih.gov/sra/SRX214365; [3,4]datasets are available at https://www.dropbox.com/sh/7yvmqknartttm6k/AAAQHvLZPjgjf4dtmHM7GNCqa/H1_gencode?dl=0 and https://www.dropbox.com/sh/7yvmqknartttm6k/AACzaG-QJggvbXW6LA32oo7ba/M1_gencode?dl=0; [5]dataset of human and mouse is available at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0139654.

performance on lncRNAs is just the opposite. LncRNA-ID can be trained with users' own data; it can obtain a satisfying result even when the data is unbalanced. With the proportion of lncRNA decreasing, CPAT shows a sharp reduction from 79.51% to 54.46% on the capability of lncRNA discrimination; LncRNA-ID, by contrast, fell less than 1% [32].

ROC curves of CPC, CPAT, PLEK, and LncRNA-ID tested on human and mouse datasets were provided in [32]. Since CPC and CPAT are updated as the accumulation of gene database, it is useful to assess their performance with latest version and take CNCI into account. Here we utilise the test set of LncRNA-ID [32] (both the datasets of human and mouse are selected from GENCODE) to reevaluate CPC, CPAT, CNCI, and PLEK (Figure 2). According to [32], the area under curve (AUC) of LncRNA-ID on human dataset is 0.9829 (optimal = 0.0545, 0.9720), while on mouse it is 0.9505 (optimal = 0.0800, 0.9445) [32]. In our assessment, the performance of PLEK is identical with [32], while the performance of CPC and CPAT, as we anticipated, displayed some differences. The ROC curves were drawn in R with the package of pROC [71].

LncRScan-SVM is compared with CPC and CPAT on human and mouse datasets from UCSC (version hg19 of human and mm10 of mouse). CPC, as an excellent coding potential assessment tool, still achieves 98.37% of specificity on mouse testing dataset. CPAT, on the contrary, achieved the highest values of sensitivity both on the datasets of human and mouse. LncRScan-SVM surpasses CPAT with 89.20% and 89.14% of specificity on human and mouse datasets. For sensitivity, lncRScan-SVM obtained 93.88% and 95.29% on the same testing datasets, which are only 0.72% and around 0.1% lower than CPAT's results, respectively, but much higher than CPC's 67.23% and 75.46%. In addition, lncRScan-SVM also has the best results of accuracy and AUC [33] on these datasets.

For the same testing datasets, the running time of CPAT is the shortest and CPC shows the longest time to finish the process because of its alignment process. When being tested on a dataset containing 4,000 protein-coding and 4,000 long noncoding transcripts, CPAT takes 35.36 s and LncRNA-ID takes 65.35 s to accomplish the discrimination while PLEK and CPC need 21.47 m and 86.51 h, respectively [32]. PLEK is 8 times and 244 times faster than CNCI and CPC, respectively, on the same testing data [31], and lncRScan-SVM also needs about 10 times as much as CPAT to finish computation [33].

Table 4: Priority of employing different methods on different situations.

	CPC	CPAT	CNCI	PLEK	LncRNA-ID	lncRScan-SVM
Coding potential assessment	√	√				
Human lncRNAs	√	√	√	√	√	√
Mouse lncRNAs		√			√	√
Other Species[1]		√	√	√	√	
Testing data with sequencing errors[2]	√		√	√		
Lack of annotation		√	√	√		
Massive-scale data[3]		√		√	√	√
Trained by users[4]		√		√	√	
Web interface	√	√				

This table only presents the preferences under different situations, which means a method with a tick can achieve a better performance under a certain circumstance.
[1]Only CPAT, LncRNA-ID, and lncRScan-SVM provide the model for mouse. When analysing other species, CPAT has the model for fly and zebrafish; CNCI and PLEK can predict the sequences of vertebrata and plant. CPAT, PLEK, and LncRNA-ID can build a new model based on users' datasets. [2]Users can choose CNCI for incomplete sequences and CPC or PLEK for the transcripts with indel errors. [3]CPAT is the most efficient method. Though lncRScan-SVM needs more time than CPAT and LncRNA-ID, it is also acceptable. [4]LncRNA-ID can handle the imbalanced training data. Training PLEK with users' own datasets may be a time-consuming task.

4. Application Scopes of the Methods

All these methods have own particular scopes to exert their talents, which means an appropriate program can help us obtain a satisfying result. The priority of utilising these tools under some particular circumstances is summarised in Table 4.

CPC is based on sequence alignment which facilitates protein-coding transcripts selection but impairs the performance of noncoding transcripts in that long noncoding transcripts share more similarities with coding transcripts such as putative ORF, which could mislead CPC. Also, because of alignment process, utilising CPC to analyse massive-scale data is a time-consuming process.

CPAT is also used to evaluate the coding potential, though the performance on long noncoding transcripts is acceptable. CPAT has a compromise between coding and noncoding transcripts that is not bad. Since the model of CPAT is logistic regression and the input file is FASTA format, CPAT is markedly superior in computational time which means CPAT is more suitable for being applied to data on a large scale. Furthermore, linguistic features make CPAT be able to analyse the sequences without annotation, and allowing users to train the model with their own dataset extends CPAT's scope of application. Users can apply CPAT to other species instead of being confined to human or mouse only.

CNCI is designed to distinguish between coding and long noncoding transcripts without the annotations of sequences. Because lots of lncRNAs are poorly annotated, this quality provided a more accurate discrimination for these sequences. CNCI is trained on human dataset but can also be applied to other mammals such as mouse and orangutan. CNCI displays acceptable results on vertebrates (except fish), but, for plants and invertebrates, the result is not very satisfying. CNCI is valuable when the sequences lack annotations or users do not have training set of other species. CNCI also shows a good performance when the transcripts are incomplete.

PLEK employs a higher fault tolerance algorithm and performs better when the sequences have indel errors. It is a proper tool to analyse the *de novo* assembled transcriptome datasets such as the sequences obtained from Roche (454) and Pacific Biosciences (PacBio) sequencing platforms. In addition to human and mouse, PLEK can also be used to other vertebrates and displays comparable results with the ones of CNCI. PLEK's model can be trained by users, but it takes a long time to be completed.

LncRNA-ID has many merits and delivers better all-round performance on human and mouse datasets. Although the time LncRNA-ID spent on classifying is nearly twice of CPAT, LncRNA-ID is still more efficient than other methods, which makes it a reasonable choice when data are on a massive-scale. The model of LncRNA-ID can be trained by users, but the most excellent attribute is the competence of handling the unbalanced training data. For studying those not well-explored species, LncRNA-ID takes priority when users have training datasets.

LncRScan-SVM achieves a good trade-off between the discrimination of coding and long noncoding genes. LncRScan-SVM is slower than CPAT and LncRNA-ID, but it is still acceptable. For analysing human and mouse datasets, lncRScan-SVM can be considered as a proper approach.

5. Discussion

According to the features selected by each tool, it is apparent that different tools have their own advantages and disadvantages. CPC is developed to assess coding potential of the transcripts; moreover, CPC is trained on datasets of protein-coding and noncoding RNA which means it achieves excellent performance when analysing ncRNAs. CPC provides a stand-alone version and a web server, but both of the two programs need vast amounts of time to process the sequences. As alignment-based tools, the performance of CPC varied when using different protein reference database. CPAT can present satisfying results efficiently partly because CPAT builds the logistic model which is faster than SVM. The web server of CPAT can display the result in an instant,

which facilitates small scale prediction tasks. A minor disadvantage of CPAT is that the cutoff of CPAT varies from species to species and users have to determine the optimum cutoff value when they are training a new model. CNCI is designed to predict the transcripts assembled from whole-transcriptome sequencing data. Thus, CNCI offers a high accuracy on incomplete transcripts. CNCI did not provide result of elaborate comparison between CNCI and CPAT, but CPAT has no regard for the problem of incomplete transcripts. Meanwhile, UTRs of the transcripts may also interfere with the performance of CPAT. The features of ANT of CNCI closely resemble the hexamer of CPAT, but the distinguishing process of CNCI is more complicated and accurate than CPAT. However, the sliding window of CNCI slides 3 nt in each step, and consequently some deletion or frameshift errors may lead to a false shift and present users with a disappointing performance. In such cases, PLEK has made a considerable improvement and exhibits more flexibility when handling the indel sequencing errors. Indel errors are very common in the sequences obtained by today's sequencing platform, which means PLEK performs well for *de novo* assembled transcriptomes. With the indel error rate increasing, the accuracy of CNCI is decreasing while PLEK has no distinct fluctuation. Nonetheless, since the nucleotides compositions differ slightly among different species, the performance of PLEK on multiple species is not better than or approximately equivalent to CNCI whose performance is more stable on different species. Both LncRNA-ID and lncRScan-SVM achieve a balance between protein-coding and lncRNAs. But the capacity of lncRScan-SVM will be limited when analysing the sequences with a lack of annotation. Another point that needs to be brought up is that lncRScan-SVM and CNCI support ∗ .GTF as input file format.

It is apparent that nucleotides composition (such as *k*-mer and G + C content) and ORF are two classic and widely used feature groups. These features have strong discriminative power because protein-coding genes will finally be transcribed and translated to produce a specific amino acid chain, which requires some specified nucleotides composition and high-quality ORFs. As to the models of these tools, SVM (CPC, CNCI, and PLEK), logistic regression (CPAT), and random forest (LncRNA-ID) are more practical for lncRNA identification, though ANN or deep learning is a more popular machine learning algorithm now. Along with the protein-coding genes prediction, the annotations of lncRNA gene have been performed as well. A new tool named AnnoLnc (2015, available at http://annolnc.cbi.pku.edu.cn/index.jsp) has just been developed to annotate new discovered lncRNAs but related article has not yet been officially published. Users can access its web server for more information.

LncRNAs are receiving increasing attention and lncRNA identification has always been a challenge for researches of life science. For so many different types of sequences, various excellent tools should be developed to tackle different problems under various circumstances in the future. In this review, we summarised several tools for lncRNAs identification and concluded respective scopes. Due to their different scopes of application, using a method apposite to particular situation will be of essence to achieve convincing results. We hope this review can help researchers employ a more appropriate method in certain situations.

Additional Points

Key Points. (i) Different tools have different scopes. Users should select a proper tool according to the type of sequences. (ii) From the perspective of sequence types, CPC and CPAT are mainly used to assess coding potential. CNCI and PLEK can be applied to the sequences obtained from high-throughput sequencing platforms or the poorly annotated. LncRNA-ID and lncRScan-SVM are more accurate on human and mouse datasets. (iii) From the perspective of other functions, CPC and CPAT have web interfaces. The classification models of CPAT, LncRNA-ID, and PLEK can be trained on users' own datasets. CPAT, LncRNA-ID, and lncRScan-SVM can be utilised when the data to be analysed are on a massive-scale.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (61272207, 61472158, and 61402194) and the Science-Technology Development Project from Jilin Province (20130522118JH, 20130101070JC) and China Postdoctoral Science Foundation (2014T70291).

References

[1] A. F. Palazzo and E. S. Lee, "Non-coding RNA: what is functional and what is junk?" *Frontiers in Genetics*, vol. 5, article 2, pp. 1–11, 2015.

[2] Y. Okazaki, M. Furuno, T. Kasukawa et al., "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs," *Nature*, vol. 420, pp. 563–573, 2002.

[3] S. Djebali, C. A. Davis, A. Merkel et al., "Landscape of transcription in human cells," *Nature*, vol. 489, pp. 101–108, 2012.

[4] I. Dunham, A. Kundaje, S. F. Aldred et al., "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, 2012.

[5] E. Pennisi, "ENCODE project writes eulogy for junk DNA," *Science*, vol. 337, no. 6099, pp. 1159–1161, 2012.

[6] S. U. Schmitz, P. Grote, and B. G. Herrmann, "Mechanisms of long noncoding RNA function in development and disease," *Cellular and Molecular Life Sciences*, vol. 73, no. 13, pp. 2491–2509, 2016.

[7] K. Plath, J. Fang, S. K. Mlynarczyk-Evans et al., "Role of histone H3 lysine 27 methylation in X inactivation," *Science*, vol. 300, no. 5616, pp. 131–135, 2003.

[8] S. T. da Rocha, V. Boeva, M. Escamilla-Del-Arenal et al., "Jarid2 is implicated in the initial xist-induced targeting of PRC2 to the inactive X chromosome," *Molecular Cell*, vol. 53, no. 2, pp. 301–316, 2014.

[9] V. O'Leary, S. V. Ovsepian, L. G. Carrascosa et al., "PARTICLE, a triplex-forming long ncRNA, regulates locus-specific methylation in response to low-dose irradiation," *Cell Reports*, vol. 11, no. 3, pp. 474–485, 2015.

[10] A. C. Marques and C. P. Ponting, "Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness," *Genome Biology*, vol. 10, no. 11, article R124, 2009.

[11] J. R. Prensner, M. K. Iyer, A. Sahu et al., "The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex," *Nature Genetics*, vol. 45, no. 11, pp. 1392–1398, 2013.

[12] G. Chen, Z. Wang, D. Wang et al., "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic Acids Research*, vol. 41, no. 1, pp. D983–D986, 2013.

[13] 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.

[14] X. Chen, C. Clarence 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.

[15] X. Chen, Y.-A. Huang, X.-S. Wang, Z. You, and K. C. Chan, "FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 29, pp. 45948–45958, 2016.

[16] 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.

[17] J. Sun, H. Shi, Z. Wang et al., "Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network," *Molecular BioSystems*, vol. 10, no. 8, pp. 2074–2081, 2014.

[18] M. Zhou, X. Wang, J. Li et al., "Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network," *Molecular BioSystems*, vol. 11, no. 3, pp. 760–769, 2015.

[19] X. Chen, Z. You, G. Yan, and D. Gong, "IRWRLDA: improved random walk with restart for lncRNA-disease association prediction," *Oncotarget*, vol. 7, no. 36, pp. 57919–57931, 2016.

[20] X. Chen, C. C. Yan, X. Zhang, and Z. You, "Long non-coding RNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, 2016.

[21] T. M. Lowe and S. R. Eddy, "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence," *Nucleic Acids Research*, vol. 25, no. 5, pp. 955–964, 1997.

[22] Q. Zou, J. Guo, Y. Ju, M. Wu, X. Zeng, and Z. Hong, "Improving tRNAscan-SE annotation results via ensemble classifiers," *Molecular Informatics*, vol. 34, no. 11-12, pp. 761–770, 2015.

[23] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identificationof human microRNAs by incorporatinga high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[24] C. Y. Wang, L. L. Hu, M. Z. Guo, X. Y. Liu, and Q. Zou, "imDC: an ensemble learning method for imbalanced classification with miRNA data," *Genetics and Molecular Research*, vol. 14, no. 1, pp. 123–133, 2015.

[25] K. Lagesen, P. Hallin, E. A. Rødland, H.-H. Stærfeldt, T. Rognes, and D. W. Ussery, "RNAmmer: consistent and rapid annotation of ribosomal RNA genes," *Nucleic Acids Research*, vol. 35, no. 9, pp. 3100–3108, 2007.

[26] C. Wang, L. Wei, M. Guo, and Q. Zou, "Computational approaches in detecting non- coding RNA," *Current Genomics*, vol. 14, no. 6, pp. 371–377, 2013.

[27] D. Veneziano, G. Nigita, and A. Ferro, "Computational approaches for the analysis of ncRNA through deep sequencing techniques," *Frontiers in Bioengineering and Biotechnology*, vol. 3, article 77, 2015.

[28] L. Kong, Y. Zhang, Z.-Q. Ye et al., "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine," *Nucleic Acids Research*, vol. 35, no. 2, pp. W345–W349, 2007.

[29] L. Wang, H. J. Park, S. Dasari et al., "Coding-potential assessment tool using an alignment-free logistic regression model," *Nucleic Acids Research*, vol. 41, no. 6, article e74, 2013.

[30] L. Sun, H. Luo, D. Bu et al., "Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts," *Nucleic Acids Research*, vol. 41, no. 17, article e166, 2013.

[31] A. Li, J. Zhang, and Z. Zhou, "PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved *k*-mer scheme," *BMC Bioinformatics*, vol. 15, no. 1, article 311, 2014.

[32] R. Achawanantakun, J. Chen, Y. Sun, and Y. Zhang, "LncRNA-ID: long non-coding RNA IDentification using balanced random forests," *Bioinformatics*, vol. 31, no. 24, pp. 3897–3905, 2015.

[33] L. Sun, H. Liu, L. Zhang, and J. Meng, "lncRScan-SVM: a tool for predicting long non-coding RNAs using support vector machine," *PLoS ONE*, vol. 10, no. 10, Article ID e0139654, 2015.

[34] X.-N. Fan and S.-W. Zhang, "LncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning," *Molecular BioSystems*, vol. 11, no. 3, pp. 892–897, 2015.

[35] C. Pian, G. Zhang, Z. Chen et al., "LncRNApred: classification of long non-coding RNAs and protein-coding transcripts by the ensemble algorithm with a new hybrid feature," *PLoS ONE*, vol. 11, no. 5, Article ID e0154567, 2016.

[36] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[37] M. C. Frith, A. R. Forrest, E. Nourbakhsh et al., "The abundance of short proteins in the mammalian proteome," *PLoS Genetics*, vol. 2, no. 4, pp. 515–528, 2006.

[38] D. M. Anderson, K. M. Anderson, C.-L. Chang et al., "A micropeptide encoded by a putative long noncoding RNA regulates muscle performance," *Cell*, vol. 160, no. 4, pp. 595–606, 2015.

[39] M. Guttman, P. Russell, N. T. Ingolia, J. S. Weissman, and E. S. Lander, "Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins," *Cell*, vol. 154, no. 1, pp. 240–251, 2013.

[40] M. F. Lin, I. Jungreis, and M. Kellis, "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions," *Bioinformatics*, vol. 27, no. 13, Article ID btr209, pp. i275–i282, 2011.

[41] B. Panwar, A. Arora, and G. P. S. Raghava, "Prediction and classification of ncRNAs using structural information," *BMC Genomics*, vol. 15, no. 1, article 127, 2014.

[42] L. Childs, Z. Nikoloski, P. May, and D. Walther, "Identification and classification of ncRNA molecules using graph properties," *Nucleic Acids Research*, vol. 37, no. 9, article e66, 2009.

[43] K. Sun, X. Chen, P. Jiang, X. Song, H. Wang, and H. Sun, "iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data," *BMC Genomics*, vol. 14, article S7, 2013.

[44] Y. Wang, Y. Li, Q. Wang et al., "Computational identification of human long intergenic non-coding RNAs using a GA-SVM algorithm," *Gene*, vol. 533, no. 1, pp. 94–99, 2014.

[45] A. Siepel, G. Bejerano, J. S. Pedersen et al., "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Research*, vol. 15, no. 8, pp. 1034–1050, 2005.

[46] J. Liu, J. Gough, and B. Rost, "Distinguishing protein-coding from non-coding RNAs through support vector machines," *PLoS Genetics*, vol. 2, no. 4, pp. 529–536, 2006.

[47] K. C. Pang, S. Stephen, P. G. Engström et al., "RNAdb-a comprehensive mammalian noncoding RNA database," *Nucleic Acids Research*, vol. 33, pp. D125–D130, 2005.

[48] C. Liu, B. Bai, G. Skogerbø et al., "NONCODE: an integrated knowledge database of non-coding RNAs," *Nucleic Acids Research*, vol. 33, pp. D112–D115, 2005.

[49] Y. Zhao, H. Li, S. Fang et al., "NONCODE 2016: an informative and valuable data source of long non-coding RNAs," *Nucleic Acids Research*, vol. 44, no. 1, pp. D203–D208, 2016.

[50] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Research*, vol. 10, no. 17, pp. 5303–5318, 1982.

[51] S. Mukherjee and Y. Zhang, "MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming," *Nucleic Acids Research*, vol. 37, no. 11, 2009.

[52] T. Derrien, R. Johnson, G. Bussotti et al., "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression," *Genome Research*, vol. 22, no. 9, pp. 1775–1789, 2012.

[53] J. Harrow, F. Denoeud, A. Frankish et al., "GENCODE: producing a reference annotation for ENCODE," *Genome Biology*, vol. 7, supplement 1, pp. S4.1–S4.9, 2006.

[54] J. Harrow, A. Frankish, J. M. Gonzalez et al., "GENCODE: the reference human genome annotation for the ENCODE project," *Genome Research*, vol. 22, no. 9, pp. 1760–1774, 2012.

[55] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott, "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy," *Nucleic Acids Research*, vol. 40, no. 1, pp. D130–D135, 2012.

[56] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 35, no. 1, pp. D61–D65, 2007.

[57] L. Deng, D. Yu, and J. Platt, "Scalable stacking and learning for building deep architectures," in *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '12)*, pp. 2133–2136, Kyoto, Japan, March 2012.

[58] B. Hutchinson, L. Deng, and D. Yu, "Tensor deep stacking networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1944–1957, 2013.

[59] W. J. Kent, C. W. Sugnet, T. S. Furey et al., "The human genome browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002.

[60] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen et al., "ViennaRNA Package 2.0," *Algorithms for Molecular Biology*, vol. 6, no. 1, article 26, 2011.

[61] M. Kozak, "Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6," *The EMBO Journal*, vol. 16, no. 9, pp. 2482–2492, 1997.

[62] M. Kozak, "Initiation of translation in prokaryotes and eukaryotes," *Gene*, vol. 234, no. 2, pp. 187–208, 1999.

[63] H. Xu, P. Wang, Y. Fu et al., "Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts," *Cell Research*, vol. 20, no. 4, pp. 445–457, 2010.

[64] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Research*, vol. 39, no. 2, pp. W29–W37, 2011.

[65] P. Kapranov, J. Cheng, S. Dike et al., "RNA maps reveal new RNA classes and a possible function for pervasive transcription," *Science*, vol. 316, no. 5830, pp. 1484–1488, 2007.

[66] C. Chen, A. Liaw, and L. Breiman, *Using Random Forest to Learn Imbalanced Data*, University of California, Berkeley, Calif, USA, 2004.

[67] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[68] J. A. Sogn, "Structure of the peptide antibiotic polypeptin," *Journal of Medicinal Chemistry*, vol. 19, no. 10, pp. 1228–1231, 1976.

[69] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[70] J. W. Fickett and C.-S. Tung, "Assessment of protein coding measures," *Nucleic Acids Research*, vol. 20, no. 24, pp. 6441–6450, 1992.

[71] X. Robin, N. Turck, A. Hainard et al., "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, article 77, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 2676282, 10 pages
http://dx.doi.org/10.1155/2016/2676282

Research Article

Effect of Dynamic Interaction between microRNA and Transcription Factor on Gene Expression

Qi Zhao,[1,2,3] Hongsheng Liu,[3,4] Chenggui Yao,[5] Jianwei Shuai,[1] and Xiaoqiang Sun[6,7,8]

[1] *Department of Physics, College of Physics Science and Technology, Xiamen University, Xiamen 361005, China*
[2] *School of Mathematics, Liaoning University, Shenyang 110036, China*
[3] *Research Center for Computer Simulating and Information Processing of Bio-Macromolecules of Liaoning Province, Shenyang 110036, China*
[4] *School of life science, Liaoning University, Shenyang 110036, China*
[5] *Department of Mathematics, Shaoxing University, Shaoxing 312000, China*
[6] *Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou 510080, China*
[7] *Guangdong Provincial Key Laboratory of Orthopedics and Traumatology, The First Affiliated Hospital of Sun Yat-Sen University, Guangzhou 510000, China*
[8] *School of Mathematical and Computational Science, Sun Yat-Sen University, Guangzhou 510275, China*

Correspondence should be addressed to Qi Zhao; zhaoqi@lnu.edu.cn and Xiaoqiang Sun; sunxq6@mail.sysu.edu.cn

Received 11 August 2016; Accepted 10 October 2016

Academic Editor: Huiming Peng

MicroRNAs (miRNAs) are endogenous noncoding RNAs which participate in diverse biological processes in animals and plants. They are known to join together with transcription factors and downstream gene, forming a complex and highly interconnected regulatory network. To recognize a few overrepresented motifs which are expected to perform important elementary regulatory functions, we constructed a computational model of miRNA-mediated feedforward loops (FFLs) in which a transcription factor (TF) regulates miRNA and targets gene. Based on the different dynamic interactions between miRNA and TF on gene expression, four possible structural topologies of FFLs with two gate functions (AND gate and OR gate) are introduced. We studied the dynamic behaviors of these different motifs. Furthermore, the relationship between the response time and maximal activation velocity of miRNA was investigated. We found that the curve of response time shows nonmonotonic behavior in Co1 loop with OR gate. This may help us to infer the mechanism of miRNA binding to the promoter region. At last we investigated the influence of important parameters on the dynamic response of system. We identified that the stationary levels of target gene in all loops were insensitive to the initial value of miRNA.

1. Introduction

MicroRNAs (miRNAs) [1, 2] are a class of endogenous small noncoding RNAs that bind to partially complementary sequences in target mRNAs, negatively regulating their protein production in higher eukaryotes, plants, and animals [1, 3–5]. Many experimental studies have revealed that miRNAs can regulate various biological functions [6, 7], for instance, development and metabolisms [8]. Also, they have been demonstrated to be involved in many cellular signaling regulation processes, including apoptosis, proliferation, and differentiation [9–11]. Moreover, a lot of biological and clinical experiments have shown that miRNAs are involved in the initiation and development of many diseases [12, 13], such as cancers [14] and HIV [15]. More and more attention has been focused on the molecular mechanisms related to miRNAs and their functions [16].

The production of miRNA is regulated by certain transcription factors (TFs) that are also key regulators in gene expression. It has been demonstrated that miRNAs and TFs are often highly interacted in a dependent or independent manner [17]. Therefore, miRNA functions can be understood more clearly only in the context of regulatory interactions between TF and miRNA. Experimental data have

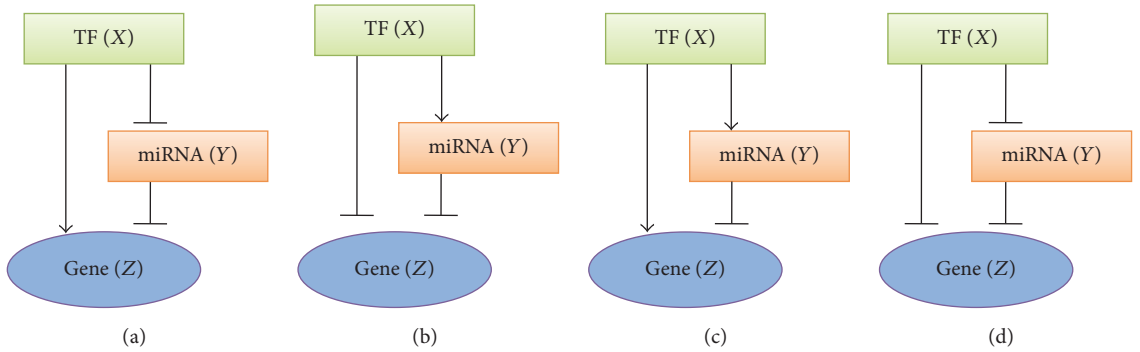


FIGURE 1: The coherent and incoherent feedforward loops. Arrows mean activation, the turned-over T-bars indicate repression. (a) Type 1 coherent FFL, TF activates target gene and represses miRNA synthesis. (b) Type 2 coherent FFL, TF represses target gene and activates miRNA synthesis. (c) Type 1 incoherent FFL, TF activates both target gene and miRNA synthesis. (d) Type 2 incoherent FFL, TF represses both target gene and miRNA synthesis.

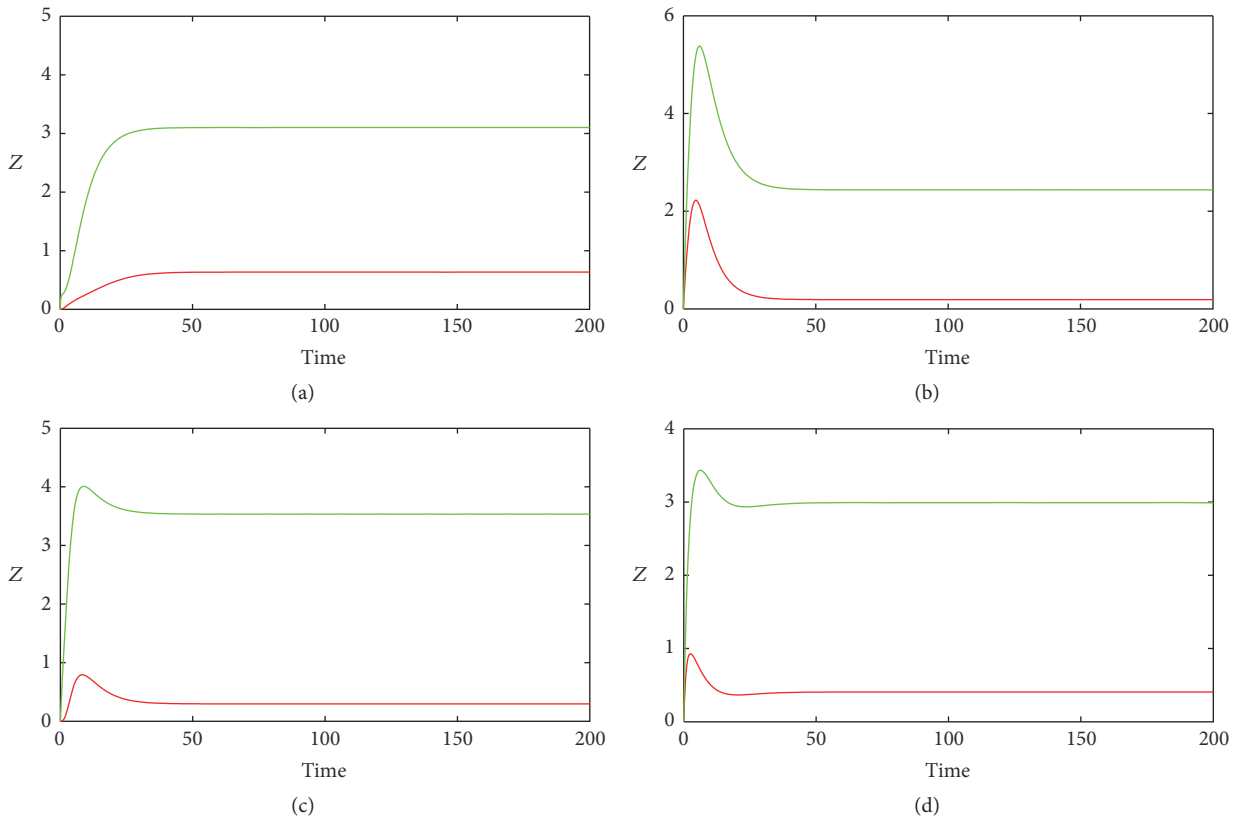


FIGURE 2: The time evolutions of Z in various FFLs with different gate functions when k_1 is constant input. Types 1-2 coherent FFLs are shown in (a)-(b), while types 1-2 incoherent FFLs are given in (c)-(d). The red line corresponds to AND gate function, and the green line represents OR gate function. Here we fix $k_1 = 0.25$.

demonstrated that gene regulatory networks are often constituted of some basic subcircuits involving feedforward or feedback loops [18], which are often called motif [19]. Feedforward loops (FFLs) have been shown to be a major member of biological network motifs. Many theoretical works [20–22] and experimental studies [23] have been conducted to investigate their structure and functions within the context of gene expression regulation. These studies focused on FFLs at the transcriptional level, in which gene expression is controlled by two regulatory TFs. Moreover, certain miRNA-containing motifs are often embedded in a lot of gene regulatory networks (GRNs). It has been known that all miRNAs operate through a repressive action on target mRNA. However, considering the interaction between miRNA and TFs, the role of miRNA in gene regulatory network is not simply repressive. Therefore, the investigation

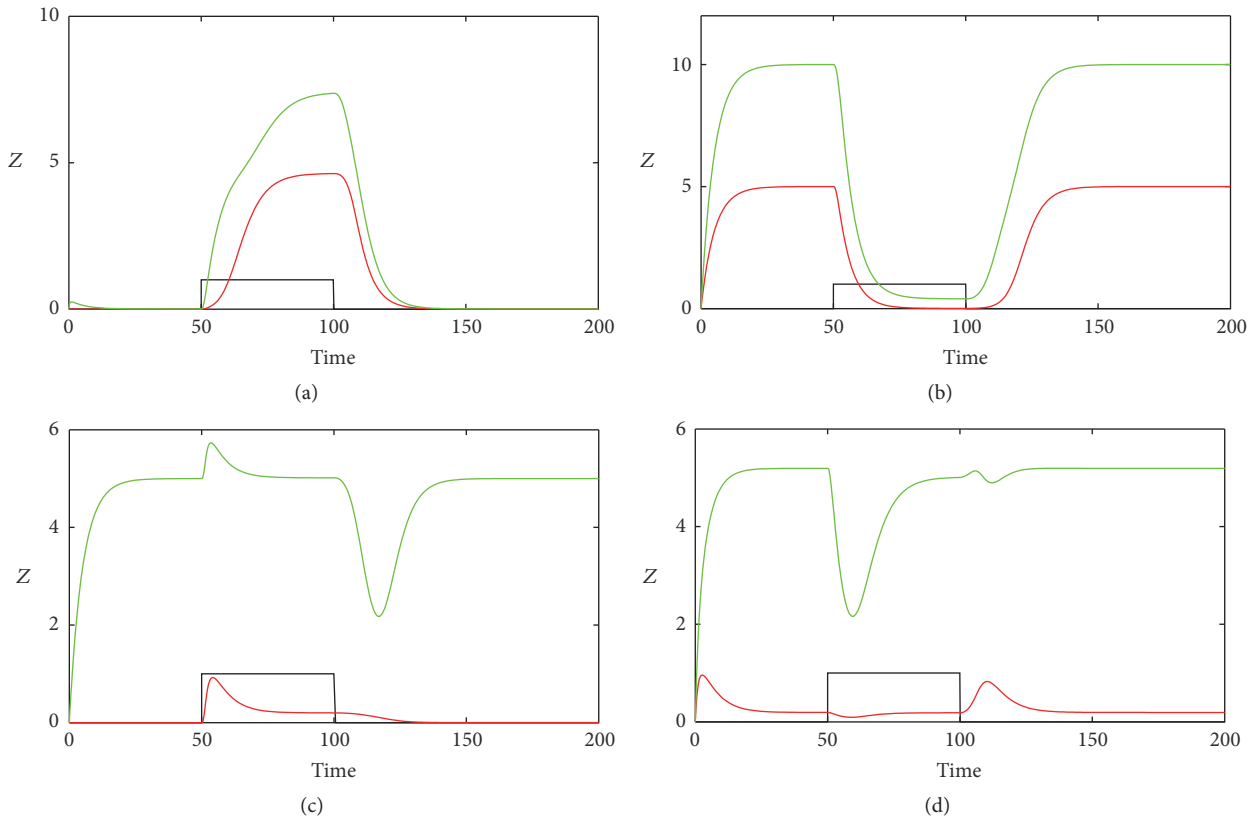


FIGURE 3: The time evolutions of Z in various FFLs with different gate functions in response to on and off steps of k_1. Types 1-2 coherent FFLs are shown in (a)-(b), while types 1-2 incoherent FFLs are given in (c)-(d). The red line corresponds to AND gate function, and the green line represents OR gate function. k_1 is set to 1 during the time between 50 and 100 and 0 in other time ranges (the black line).

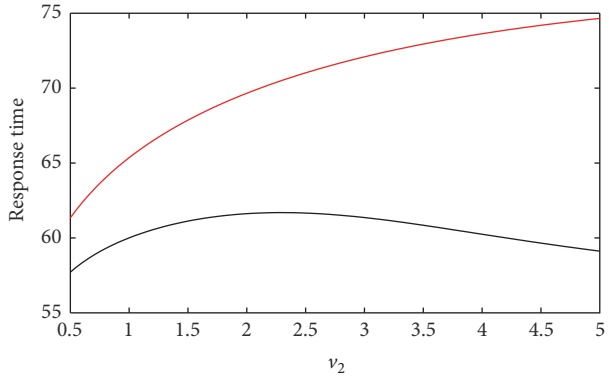


FIGURE 4: The response time is plotted against the variation of v_2 in Co1 loop with different gate regulations. The red line corresponds to AND gate function, and the black line represents OR gate function. k_1 is set to 1 during the time between 50 and 100 and 0 in other time ranges.

of the effect of interaction between TF and miRNA on gene expression is very important to help us understand the role of miRNAs in the GRN and disease.

Mathematical model is a powerful tool used to describe the biological systems and discriminate between different tentative mechanisms [24–36]. Several studies have examined the mechanisms of miRNA-containing motifs using mathematical models. Osella et al. [37] used a detailed analytical model and simulations to investigate the function of the miRNA-mediated FFL. Their analysis demonstrated that the incoherent version of such FFL motif can provide precision and stability to the overall gene expression program with an efficient noise control, given the existence of fluctuations in upstream regulators. Morozova et al. [38] developed a mathematical model containing nine known mechanisms of miRNA action and discriminated among different possible individual mechanisms based on the kinetic signatures. Duk et al. [39] analyzed three mathematical models, in which miRNA either represses translation of its target or promotes target mRNA degradation or is not reused but degrades along with target mRNA. They showed that different mechanisms of miRNA action lead to a variety of types of dynamical behavior of feedforward loops. However, none of previous studies examined the effects of dependence (AND gate) or independence (OR gate) between miRNA and TFs on gene expression.

In this paper, we developed a mathematical model to quantitatively analyze the dynamics of miRNA-containing FFLs and investigate the interaction between miRNA and TF on gene expression. We examined four FFLs, in which each

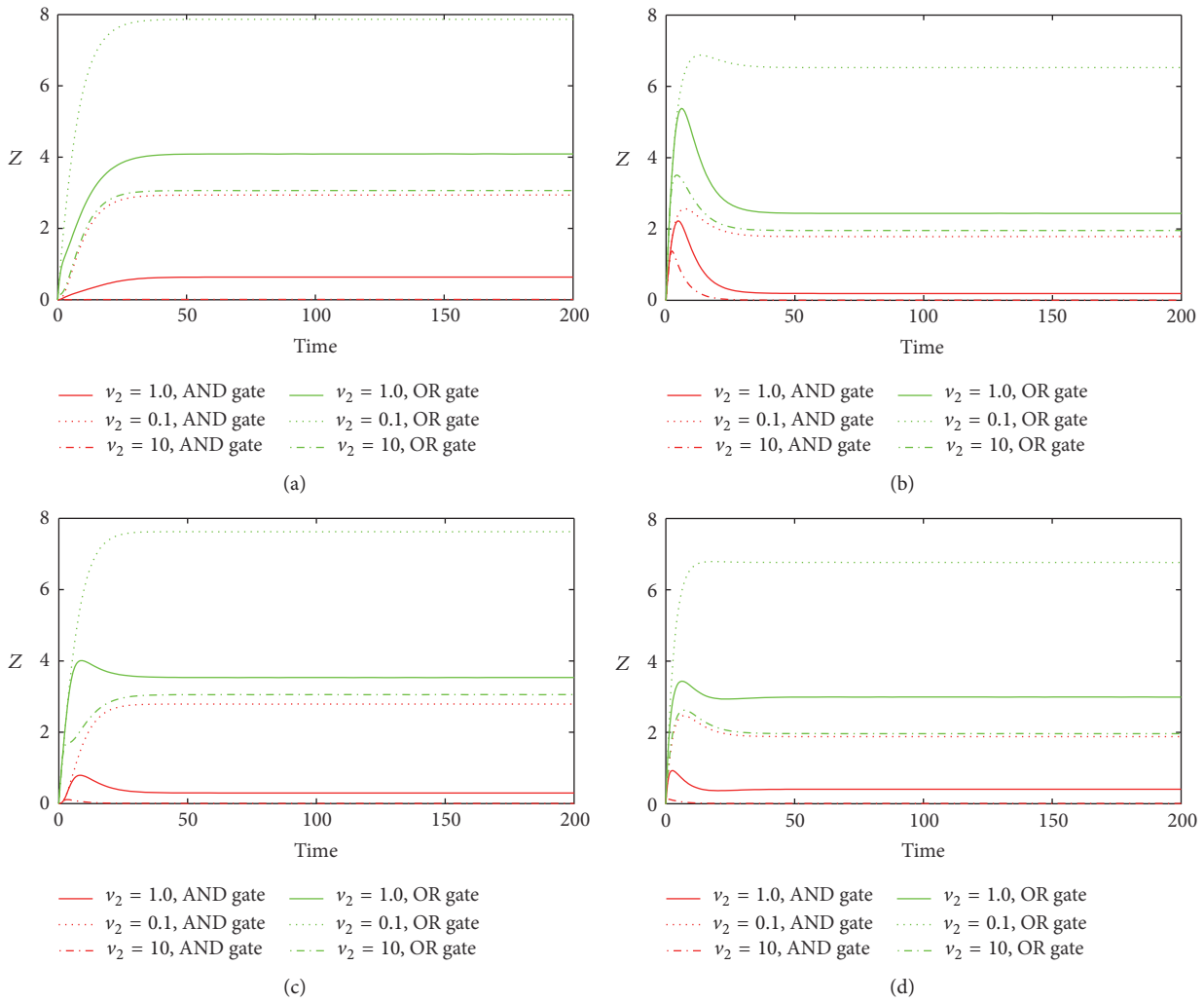


FIGURE 5: The time evolutions of Z in various FFLs with different gate functions in response to variation of v_2. Types 1-2 coherent FFLs are shown in (a)-(b), while types 1-2 incoherent FFLs are given in (c)-(d). The red line corresponds to AND gate function, and the green line represents OR gate function. Here we fix $k_1 = 0.25$.

contains AND gate or OR gate. We analyzed the different dynamical behaviors between AND gate and OR gate for each of these four FFLs. Our results showed that different mechanisms with respect to AND or OR gate might produce distinct dynamics of the GRN. In addition, we examined the relationship between response time of gene expression and certain parameters in the model. Finally we investigated the influence of important parameters on the response of system. Our study advances our quantitative understanding on the dynamic interaction between TF and miRNA, particularly, with AND or OR gate in the GRN, and provides some implications on the miRNA-mediated dieses.

2. Results

2.1. Mathematical Model of FFLs. Figure 1 illustrates the general structure of FFLs in miRNA-mediated gene transcription network, similar to that in [24–27]. The upstream transcription factor (TF) regulates the target gene via two parallel pathways: directly and by interaction with miRNA, which also regulates the target gene. Therefore, regulatory interactions in FFL create four possible structural topologies (Figure 1). Two of these configurations are named "coherent": the sign of the direct regulation path from TF to gene is the same as the overall sign of the indirect regulation path from TF via miRNA to gene. The other two structures are termed "incoherent": the sign of the direct regulation path is opposite to that of indirect path. We specify these configurations as type 1 or 2 coherent FFLs and type 1 or 2 incoherent FFLs, respectively. The biological network motif under investigation is described by 3 variables, the concentrations of transcription factor (X), miRNA (Y), and target gene (Z). The dynamical behavior of the FFLs is governed by the following equations:

$$\frac{dX}{dt} = k_1 - d_1 X,$$

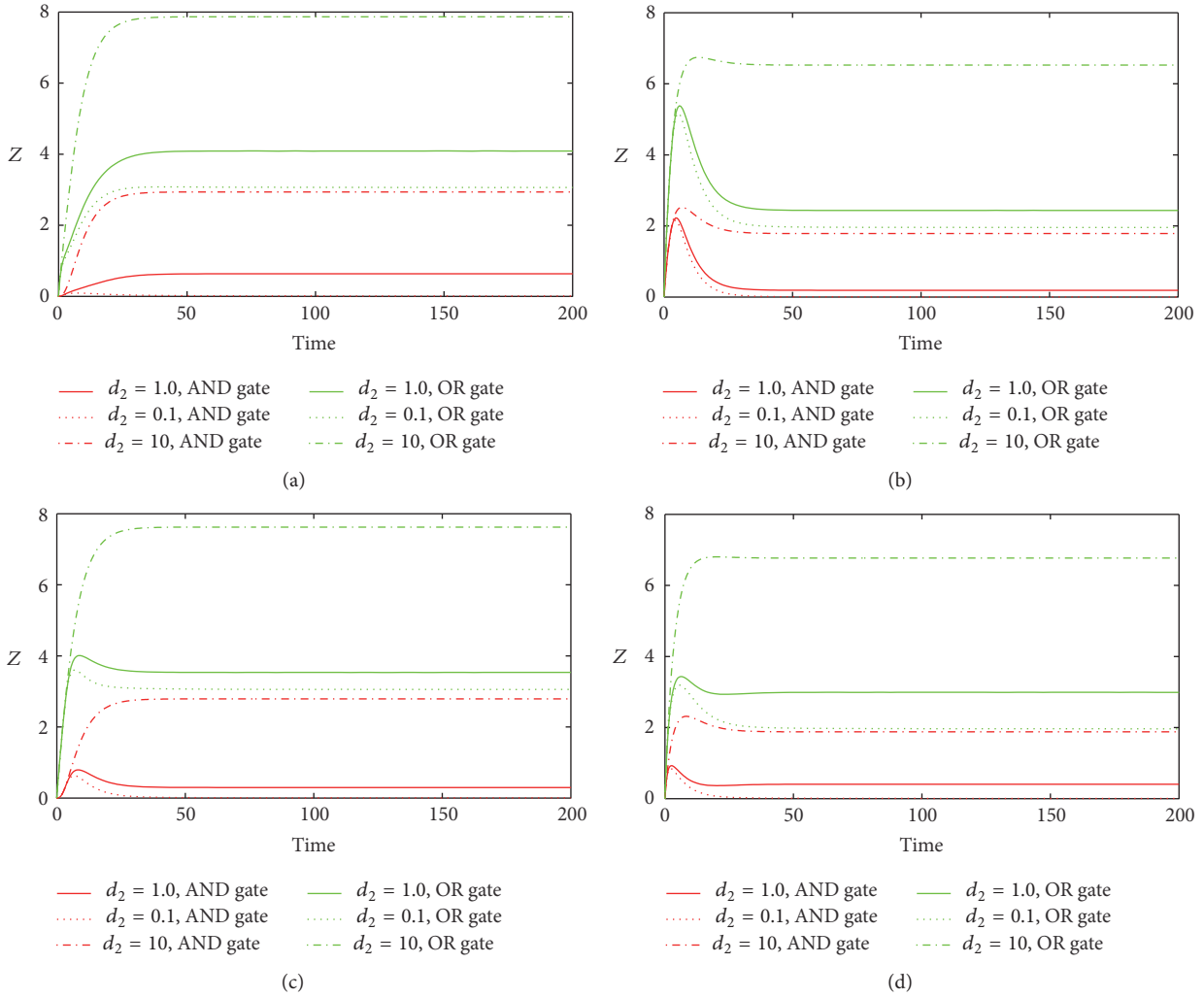


Figure 6: The time evolutions of Z in various FFLs with different gate functions in response to variation of d_2. Types 1-2 coherent FFLs are shown in (a)-(b), while types 1-2 incoherent FFLs are given in (c)-(d). The red line corresponds to AND gate function, and the green line represents OR gate function. Here we fix $k_1 = 0.25$.

$$\frac{dY}{dt} = v_2 f\left(X, k_{12}\right) - d_2 Y,$$
$$\frac{dZ}{dt} = v_3 g\left(X, k_{13}; Y, k_{23}\right) - d_3 Z. \tag{1}$$

The regulation function for an activator is $f(u, k_{ij}) = (u/k_{ij})^n/(1 + (u/k_{ij})^n)$ and for a repressor is $f(u, k_{ij}) = 1/(1 + (u/k_{ij})^n)$, similar to that we used before in [40, 41]. $g(X, k_{13}; Y, k_{23})$ is the gate function, the mechanisms underlying miRNA-mediated repression are not clear so far, and for this reason we consider that the gate function has two forms. The gate function for an AND gate is $g(X, k_{13}; Y, k_{23}) = f(X, k_{13}) * f(Y, k_{23})$, while for an OR gate we have $g(X, k_{13}; Y, k_{23}) = f(X, k_{13}) + f(Y, k_{23})$. For more details about the values of parameters and initial concentrations we use, see Tables 1 and 2.

2.2. Comparative Analysis of FFLs' Temporal Behavior under Different Gate Functions. We shall use for brevity the following abbreviations for the FFL identification: Co1 will mean type 1 coherent FFL, Co2 type 2 coherent FFL, In1 type 1 incoherent FFL, and In2 type 2 incoherent FFL, respectively.

Figure 2 shows the time courses of Z in various FFLs with different gate functions when k_1 is constant number. Here k_1 represents the basal synthesis rate of TF. The dynamics of target gene in Co1 loop has a form of increasing function and then tends to a constant vale (Figure 2(a)). The target gene profiles in Co2, In1, and In2 loops show pulse-like behavior due to repression mediated by miRNA (Figures 2(b), 2(c), and 2(d)). At the steady state, the concentrations of target gene in all the loops with AND gate are much lower than those with OR gate function. It is easy to understand this, because OR gate function makes the synthesis rate bigger than that of AND gate.

Table 1: The values of parameters in the mathematical model.

Parameter number	Symbol	Value	Description
1	d_1	0.2	Degradation rate of TF
2	v_2	1.0	Maximal activation velocity of miRNA by TF
3	d_2	0.2	Degradation rate of miRNA
4	v_3	1.0	Maximal activation velocity of target gene by TF and miRNA
5	d_3	0.2	Degradation rate of target gene
6	k_{12}	1.0	Michaelis constant of miRNA by TF
7	k_{13}	1.0	Michaelis constant of target gene by TF
8	k_{23}	1.0	Michaelis constant of target gene by miRNA
9	n	2	Hill coefficient

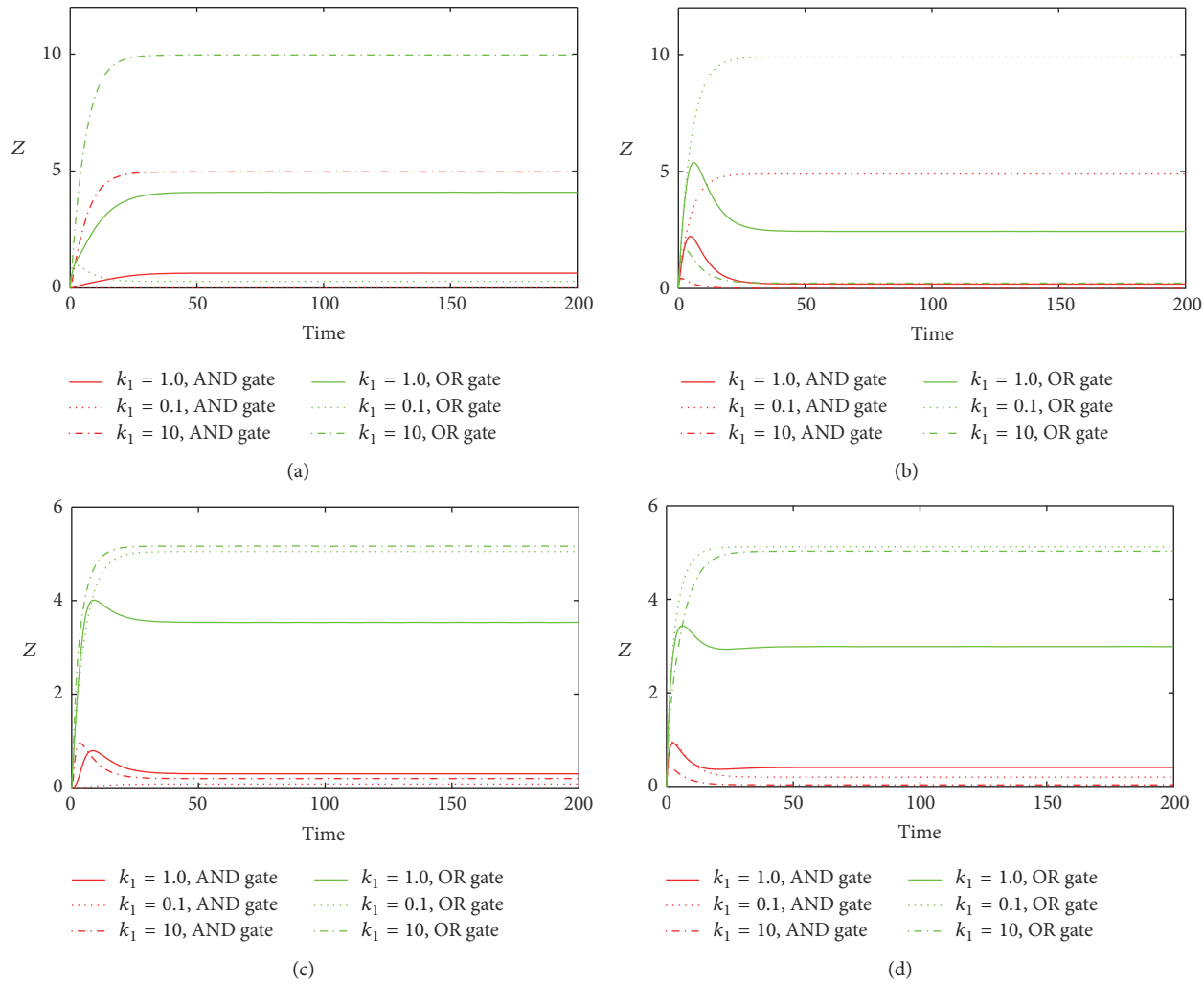


Figure 7: The time evolutions of Z in various FFLs with different gate functions in response to variation of k_1. Types 1-2 coherent FFLs are shown in (a)-(b), while type 1-2 incoherent FFLs are given in (c)-(d). The red line corresponds to AND gate function, and the green line represents OR gate function. Here we fix $k_1 = 0.25$.

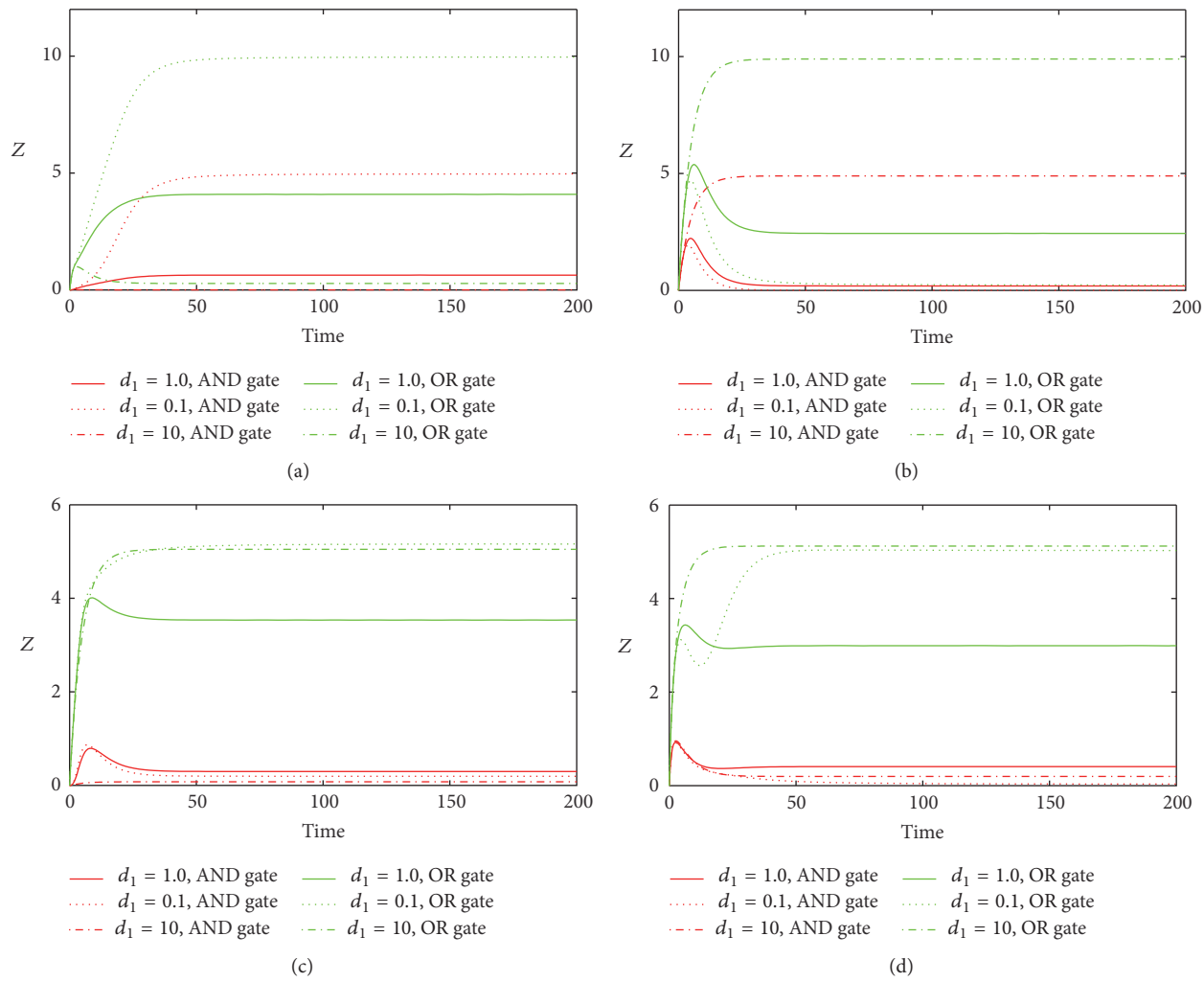


Figure 8: The time evolutions of Z in various FFLs with different gate functions in response to variation of d_1. Types 1-2 coherent FFLs are shown in (a)-(b), while types 1-2 incoherent FFLs are given in (c)-(d). The red line corresponds to AND gate function, and the green line represents OR gate function. Here we fix $k_1 = 0.25$.

Table 2: Initial values of the mathematical model.

Parameter number	Symbol	Value	Description
1	X	0	Initial value of TF
2	Y	0	Initial value of miRNA
3	Z	0	Initial value of target gene

Living cells constantly have to respond to a changing environment. To understand how cells deal with a fluctuating environment, we need to know how cells transduce time varying signals. Next we consider the effect of providing the system with simultaneous pulse, a biological scenario which corresponds to continued exposure to environmental stimuli within a certain time range. Accordingly, we set k_1 to be a piecewise constant function

$$k_1 = \begin{cases} 1 & 50 \le t \le 100, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Figure 3 shows the variations in the response of the output in the motifs. We first compare the kinetics of Z in Co1 and In1 loops (Figures 3(a) and 3(c)). When k_1 turns on, we find out only the steady states of Z in Co1 and In1 loops with both gate functions rising up due to the direct activation of Z by TF (Figures 3(a) and 3(c)). But in In1 loop, Z first rises slightly and then falls down because TF inhibits Z by promoting miRNA. When k_1 turns off, both the concentrations of Z in Co1 and In1 loops decrease, but Z in In1 loop with OR gate eventually grows again to the stationary level. We then compare the kinetics of Z in Co2 and In2 loops (Figures 3(b) and 3(d)); we observe that the concentration of Z in Co2 loop decreases as k_1 turns on and increases as k_1 turns off (Figure 3(b)). But Z in In2 loop with OR gate rises up again to the steady state level after Z falls down, as k_1 turns on (Figure 3(d)), while Z in In2 loop with AND gate just slightly decreases when k_1 changes to 1. Z in In2 loop with two types of gate functions shows pulse-like behavior after k_1 turns to 0; however, the amplitude of Z in In2 loop with OR gate is

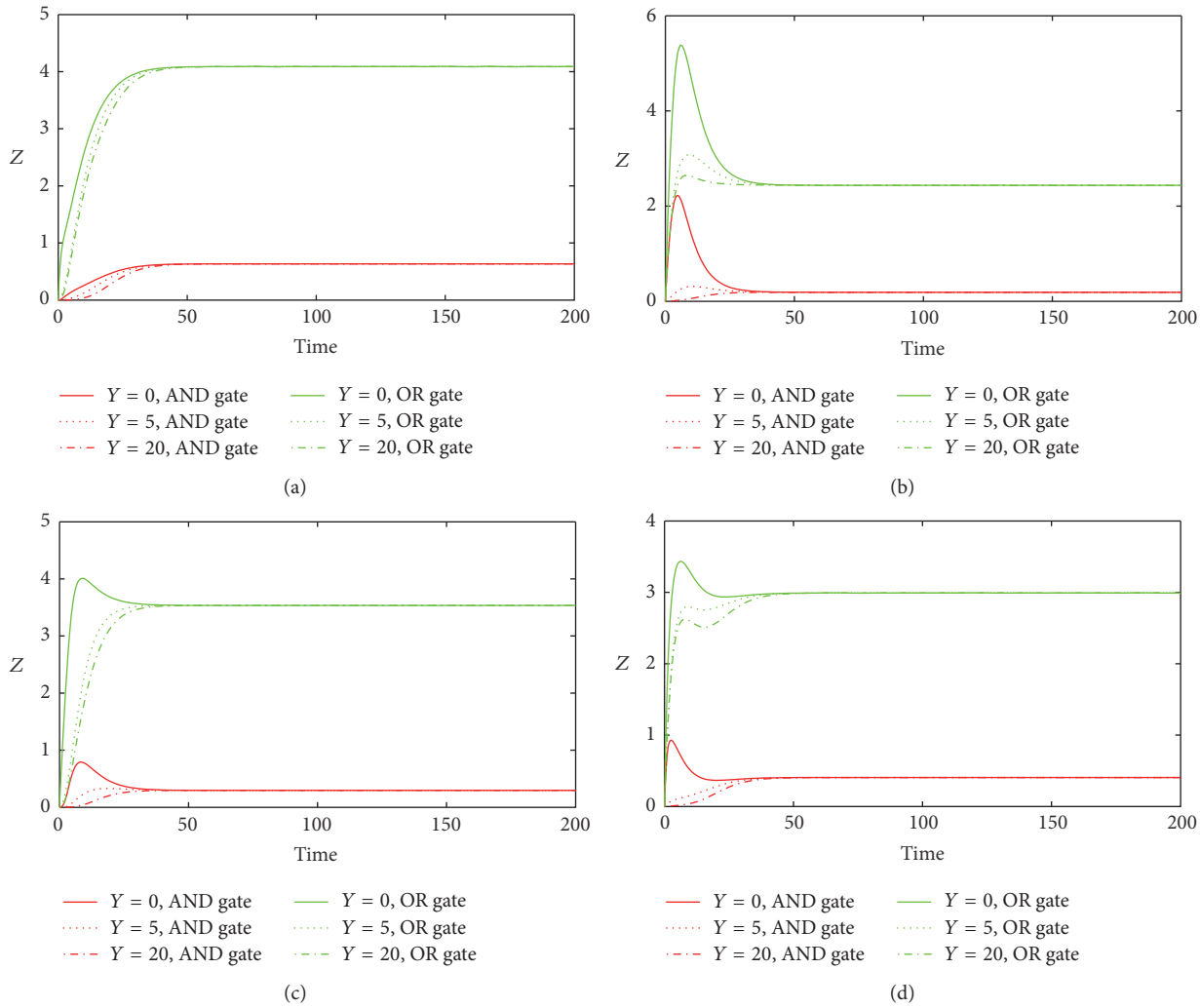


FIGURE 9: The time evolutions of Z in various FFLs with different gate functions in response to the different initial values of miRNA. Types 1-2 coherent FFLs are shown in (a)-(b), while types 1-2 incoherent FFLs are given in (c)-(d). The red line corresponds to AND gate function, and the green line represents OR gate function. Here we choose three different initial values for Y, $Y = 0$, $Y = 5$, and $Y = 20$. We fix $k_1 = 0.25$.

much smaller than that with AND gate. From the subfigures in Figure 3, we can find that Z in In2 loop with AND gate is more robust in the presence of k_1 addition, and Z in In1 loop with AND gate is more stable after an off step of k_1.

The response time is a measure of the time which a gene product takes to reach its physiologically determined steady state level. The speed of the response is characterized by the response time, which Z takes to reach half of its steady state level. Here v_2 is the maximal activation velocity of miRNA by TF. In Figure 4, we study the relationship between the response time and v_2 in Co1 loop with both gate regulations when providing the system with simultaneous pulse. We can observe that the response time has a form of increasing function as v_2 turns bigger in Co1 loop with AND gate, which means the system responses more slowly as v_2 increases. This is easy to understand; larger v_2 induces more miRNA generation which further represses target gene synthesis, so the response time turns slowly. But for the case in Co1 loop with OR gate, the response time shows nonmonotonic behavior, which first climbs and then damps as further increasing v_2. This indicates that there exists a value of v_2 such that the system responses most slowly. To understand this, we need to refer to OR gate function we use. It is a nonmonotonic function as v_2 increases, so the form of function decides the speed of the response of the system. Our result here might be useful to infer the mechanism of miRNA binding to the promoter region, whether or not the TF and miRNA compete for binding to the target gene. Also, we obtain that the response of gene expression in Co1 loop with OR gate is faster than that in Co1 loop with AND gate during the period of v_2 changing.

2.3. Variations of Parameters on the Response of System. It is known that the model coefficients might affect the dynamical behavior of FFLs. Therefore, we further examine how the changes in parameters affect the temporal behavior of the target gene. We investigate the effect of changes in v_2, d_2, k_1, and d_1 on the dynamical behavior of Z.

Figure 5 shows the time course of Z in various FFLs with different gate functions in response to variation of v_2. We choose three typical values of v_2: the original value, 10-fold, and 0.1-fold of v_2. We find that bigger v_2 induces less expression of target gene when Z reaches the steady state. We can understand this from the interaction relationship in Figure 1. Larger v_2 results in more miRNA generation which further represses target gene synthesis, so at last less target gene was observed. Parameter d_2 is the degradation rate of miRNA. For the influence of d_2, the situation is opposite, in which bigger d_2 results in higher level of gene expression after it gets to the stationary level (Figure 6). This is because that larger d_2 induces less miRNA generation, which results in less inhibition of miRNA on Z synthesis.

We also investigate the effect of changes in k_1 and d_1 on the dynamical behavior of Z (Figures 7 and 8). In Co1 loop, bigger k_1 induces more Z with both gate functions, while, in Co2 loop, the situation is opposite; lager k_1 makes less Z with both gate functions. This is due to the fact that TF activates target gene directly and promotes it indirectly in Co1 loop, while, in Co2 loop, TF inhibits target gene directly and represses it indirectly. For the cases in In1 and In2 loops with OR gate, both lager k_1 and small k_1 generate nearly the same stationary level of Z which is higher than what the original value makes. For the cases in In1 and In2 loops with AND gate, both lager k_1 and small k_1 induce nearly the same stationary level of Z which is slightly lower than that induced by the original value. For the variations of d_1 (Figure 8), we get similar results in In1 and In2 loops with both gates, but with the opposite results in Co1 and Co2 loops. Furthermore, we study the effect of different initial values of miRNA on the response of the system (Figure 9). We find that the different initial values of miRNA have no significant influence on the steady state of target gene after it passes the transient state.

3. Conclusions

In summary, there are multiple variations of the feedforward loops occurring in the nature based on different types of feedback. Hence, we constructed a mathematical model of FFLs in miRNA-mediated gene transcription network. We introduced four possible structural topologies of FFLs associated with two different gate functions which describe the dynamic interaction between miRNA and TF on gene expression. Dynamical behaviors of model component were investigated by computational simulation. Furthermore, the different features of system's response to simultaneous pulse were investigated. The influence of important parameters on the response of system was also considered. We first identified that only the dynamics of target gene in Co1 loop does not show pulse-like behavior when the synthesis rate of TF is constant. While providing the system with simultaneous pulse, we found that target gene in In2 loop with AND gate is more robust in the presence of stimulus addition, and target gene in In1 loop with AND gate is more stable after an off step of stimulus. Furthermore, we studied the relationship between the response time and maximal activation velocity of miRNA when providing the system with simultaneous pulse. We found that the curve of response time shows nonmonotonic behavior in Co1 loop with OR gate. We further showed that the stationary levels of target gene in all loops were insensitive to the initial value of miRNA.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Qi Zhao and Xiaoqiang Sun conceived the study, built the model, performed the simulations, interpreted the results, and wrote the paper. Hongsheng Liu and Chenggui Yao participated in discussions. Jianwei Shuai improved and revised the manuscript. All authors have read and approved the final version of the manuscript.

Acknowledgments

The researches of Qi Zhao and Hongsheng Liu were supported by Innovation Team Project (no. LT2015011) from the Education Department of Liaoning Province. The research of Chenggui Yao was supported by grants from National Natural Science Foundation of China (11675112) and Natural Science Foundation of Zhejiang Province (LY16A050001). The research of Jianwei Shuai was supported by grants from National Natural Science Foundation of China (31370830). The research of Xiaoqiang Sun was supported by grants from National Natural Science Foundation of China (61503419), Guangdong Nature Science Foundation (2014A030310355 and 2016A030313234), the fund for Guangdong Provincial Key Laboratory of Orthopedics and Traumatology (2016B030301002), and "985 project" of Sun Yat-Sen University (no. 50000-31101302).

References

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

[2] L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," *Nature Reviews Genetics*, vol. 5, no. 7, pp. 522–531, 2004.

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

[4] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel, "Most mammalian mRNAs are conserved targets of microRNAs," *Genome Research*, vol. 19, no. 1, pp. 92–105, 2009.

[5] O. Voinnet, "Origin, biogenesis, and activity of plant MicroRNAs," *Cell*, vol. 136, no. 4, pp. 669–687, 2009.

[6] X. Yan, H. Liang, T. Deng et al., "The identification of novel targets of miR-16 and characterization of their biological functions in cancer cells," *Molecular Cancer*, vol. 12, no. 1, pp. 1–11, 2013.

[7] C. B. Santos-Rebouças and M. M. G. Pimentel, "MicroRNAs: macro challenges on understanding human biological functions and neurological diseases," *Current Molecular Medicine*, vol. 10, no. 8, pp. 692–704, 2010.

[8] J. Krützfeldt, M. N. Poy, and M. Stoffel, "Strategies to determine the biological function of microRNAs," *Nature Genetics*, vol. 38, no. 1, pp. S14–S19, 2006.

[9] G. Suffert, G. Malterer, J. Hausser et al., "Kaposi's sarcoma herpesvirus microRNAs target caspase 3 and regulate apoptosis," *PLoS Pathogens*, vol. 7, no. 12, Article ID e1002405, 2011.

[10] H. Liang, X. Li, L. Wang et al., "MicroRNAs contribute to promyelocyte apoptosis in As_2O_3-treated APL cells," *Cellular Physiology & Biochemistry*, vol. 32, no. 6, pp. 1818–1829, 2013.

[11] A. Hata and H. Kang, "Functions of the bone morphogenetic protein signaling pathway through microRNAs (review)," *International Journal of Molecular Medicine*, vol. 35, no. 3, pp. 563–568, 2015.

[12] V. Pileczki, R. Cojocneanu-Petric, M. Maralani, I. B. Neagoe, and R. Sandulescu, "MicroRNAs as regulators of apoptosis mechanisms in cancer," *BMJ Clinical Research Journal*, vol. 322, no. 7301, pp. 1528–1532, 2016.

[13] S. Donzelli, M. Cioce, P. Muti, S. Strano, Y. Yarden, and G. Blandino, "MicroRNAs: non-coding fine tuners of receptor tyrosine kinase signalling in cancer," *Seminars in Cell & Developmental Biology*, vol. 50, pp. 133–142, 2016.

[14] S. L. Yong and A. Dutta, "MicroRNAs in cancer," *Annual Review of Pathology: Mechanisms of Disease*, vol. 4, pp. 199–227, 2009.

[15] J. Huang, F. Wang, E. Argyris et al., "Cellular microRNAs contribute to HIV-1 latency in resting primary CD4+ T lymphocytes," *Nature Medicine*, vol. 13, no. 10, pp. 1241–1247, 2007.

[16] R. Schickel, B. Boyerinas, S.-M. Park, and M. E. Peter, "MicroRNAs: key players in the immune system, differentiation, tumorigenesis and cell death," *Oncogene*, vol. 27, no. 45, pp. 5959–5974, 2008.

[17] N. J. Martinez and A. J. M. Walhout, "The interplay between transcription factors and microRNAs in genome-scale regulatory networks," *BioEssays*, vol. 31, no. 4, pp. 435–445, 2009.

[18] A. Re, D. Corá, D. Taverna, and M. Caselle, "Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human," *Molecular BioSystems*, vol. 5, no. 8, pp. 854–867, 2009.

[19] S. Kagale, M. G. Links, and K. Rozwadowski, "Genome-wide analysis of ethylene-responsive element binding factor-associated amphiphilic repression motif-containing transcriptional regulators in arabidopsis," *Plant Physiology*, vol. 152, no. 3, pp. 1109–1134, 2010.

[20] B. Ghosh, R. Karmakar, and I. Bose, "Noise characteristics of feed forward loops," *Physical Biology*, vol. 2, no. 1, pp. 36–45, 2005.

[21] S. Mangan and U. Alon, "Structure and function of the feed-forward loop network motif," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 11980–11985, 2003.

[22] S. Mangan, A. Zaslaver, and U. Alon, "The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks," *Journal of Molecular Biology*, vol. 334, no. 2, pp. 197–204, 2003.

[23] S. Mangan, S. Itzkovitz, A. Zaslaver, and U. Alon, "The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*," *Journal of Molecular Biology*, vol. 356, no. 5, pp. 1073–1081, 2006.

[24] X. Sun, J. Bao, K. C. Nelson, K. C. Li, G. Kulik, and X. Zhou, "Systems modeling of anti-apoptotic pathways in prostate cancer: psychological stress triggers a synergism pattern switch in drug combination therapy," *PLoS Computational Biology*, vol. 9, no. 12, article e1003358, 2013.

[25] X. Sun, X. Zheng, J. Zhang, T. Zhou, G. Yan, and W. Zhu, "Mathematical modeling reveals a critical role for cyclin D1 dynamics in phenotype switching during glioma differentiation," *FEBS Letters*, vol. 589, no. 18, pp. 2304–2311, 2015.

[26] X. Sun, J. Bao, and Y. Shao, "Mathematical modeling of therapy-induced cancer drug resistance: connecting cancer mechanisms to population survival rates," *Scientific Reports*, vol. 6, Article ID 22498, 2016.

[27] X. Sun, H. Xian, S. Tian et al., "A hierarchical mechanism of RIG-I ubiquitination provides sensitivity, robustness and synergy in antiviral immune responses," *Scientific Reports*, vol. 6, Article ID 29263, 2016.

[28] X. Chen, C. C. Yan, X. Zhang, and Z. You, "Long non-coding RNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, 2016.

[29] X. Chen, Y.-A. Huang, X.-S. Wang, Z.-H. You, and K. C. C. Chan, "FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 29, pp. 45948–45958, 2016.

[30] Y. A. Huang, X. Chen, Z. You, D. Huang, and K. C. Chan, "ILNCSIM: improved lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 18, pp. 25902–25914, 2016.

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

[32] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Scientific Reports*, vol. 5, Article ID 16840, 2015.

[33] 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.

[34] Z. Zhu, L. Li, Y. Zhang, Y. Yang, and X. Yang, "Comp Map: a reference-based compression program to speed up read mapping to related reference sequences," *Bioinformatics*, vol. 31, no. 3, pp. 426–428, 2014.

[35] Z. Zhu, Y. Zhang, Z. Ji, S. He, and X. Yang, "High-throughput DNA sequence data compression," *Briefings in Bioinformatics*, vol. 16, no. 1, Article ID bbt087, pp. 1–15, 2013.

[36] Y. Zhang, L. Li, Y. Yang, X. Yang, S. He, and Z. Zhu, "Light-weight reference-based compression of FASTQ data," *BMC Bioinformatics*, vol. 16, no. 1, article 188, 2015.

[37] M. Osella, C. Bosia, D. Corá, and M. Caselle, "The role of incoherent microRNA-mediated feedforward loops in noise buffering," *PLoS Computational Biology*, vol. 7, no. 3, article e1001101, 2011.

[38] N. Morozova, A. Zinovyev, N. Nonne, L.-L. Pritchard, A. N. Gorban, and A. Harel-Bellan, "Kinetic signatures of microRNA modes of action," *RNA*, vol. 18, no. 9, pp. 1635–1655, 2012.

[39] M. A. Duk, M. G. Samsonova, and A. M. Samsonov, "Dynamics of miRNA driven feed-forward loop depends upon miRNA action mechanisms," *BMC Genomics*, vol. 15, supplement 12, pp. 1–18, 2014.

[40] Q. Zhao, M. Yi, and Y. Liu, "Spatial distribution and dose-response relationship for different operation modes in a reaction-diffusion model of the MAPK cascade," *Physical Biology*, vol. 8, no. 5, article 055004, 2011.

[41] Q. Zhao, C. Yao, J. Tang, and L. Liu, "Study of spatial signal transduction in bistable switches," *Frontiers of Physics*, vol. 11, no. 5, Article ID 110501, 2016.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 3164238, 4 pages
http://dx.doi.org/10.1155/2016/3164238

Research Article

Transcriptional Regulation of lncRNA Genes by Histone Modification in Alzheimer's Disease

Guoqiang Wan,[1] Wenyang Zhou,[1] Yang Hu,[1] Rui Ma,[1] Shuilin Jin,[2] Guiyou Liu,[1] and Qinghua Jiang[1]

[1]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*
[2]*Department of Mathematics, Harbin Institute of Technology, Harbin 150001, China*

Correspondence should be addressed to Qinghua Jiang; qhjiang@hit.edu.cn

Received 11 August 2016; Accepted 27 September 2016

Academic Editor: Xing Chen

Increasing studies have revealed that long noncoding RNAs (lncRNAs) are not transcriptional noise but play important roles in the regulation of a wide range of biological processes, and the dysregulation of lncRNA genes is associated with disease development. Alzheimer's disease (AD) is a chronic neurodegenerative disease that usually starts slowly and gets worse over time. However, little is known about the roles of lncRNA genes in AD and how the lncRNA genes are transcriptionally regulated. Herein, we analyzed RNA-seq data and ChIP-seq histone modification data from CK-p25 AD model and control mice and identified 72 differentially expressed lncRNA genes, 4,917 differential peaks of H3K4me3, and 1,624 differential peaks of H3K27me3 between AD and control samples, respectively. Furthermore, we found 92 differential peaks of histone modification H3K4me3 are located in the promoter of 39 differentially expressed lncRNA genes and 8 differential peaks of histone modification H3K27me3 are located upstream of 7 differentially expressed lncRNA genes, which suggest that the majority of lncRNA genes may be transcriptionally regulated by histone modification in AD.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease with unknown etiology [1–3]. The main clinical manifestation is intelligence damage. In addition, it is the cause of 60% to 70% of cases of dementia. AD often begins in people over 65 years of age, and it affects approximately 6% of people aged 65 years and older [4]. There are about 48 million persons suffering from AD around the world in 2015, and dementia resulted in about 486,000 deaths in 2010 [5].

Long noncoding RNAs (lncRNAs) are non-protein-coding transcripts longer than 200 nucleotides in length. Thousands of human and mouse lncRNAs have been identified and emerging studies have revealed that lncRNAs play important roles in a wide range of biological processes and diseases [6–12]. Many studies have demonstrated that lncRNAs play crucial roles in the regulation of gene expression at epigenetic, transcriptional, and posttranscriptional level [13]. However, little is known about how lncRNA genes are transcriptionally regulated [14] in disease such as AD.

In this paper, we analyzed RNA-seq data and ChIP-seq histone modification data from control mice and CK-p25 AD model at 2 weeks after induction of neurodegeneration and checked whether lncRNA genes are transcriptionally regulated by histone modification in AD.

2. Materials and Methods

2.1. RNA-seq and ChIP-seq Data in AD and Control. The RNA-seq and ChIP-seq data were downloaded from GEO database with ID GSE65159 [15]. There are three control samples and three AD mice model samples at 2 weeks after induction of neurodegeneration. The histone modification marks include H3K4me3 and H3K27me3.

2.2. Identifying Differentially Expressed lncRNA Genes between AD and Control. We used RNA-seq data to evaluate gene expression on control mice and CK-p25 Alzheimer's disease model. We used the mm10 reference sequence to build an index by Bowtie2-build [16]; the mm10 reference sequence

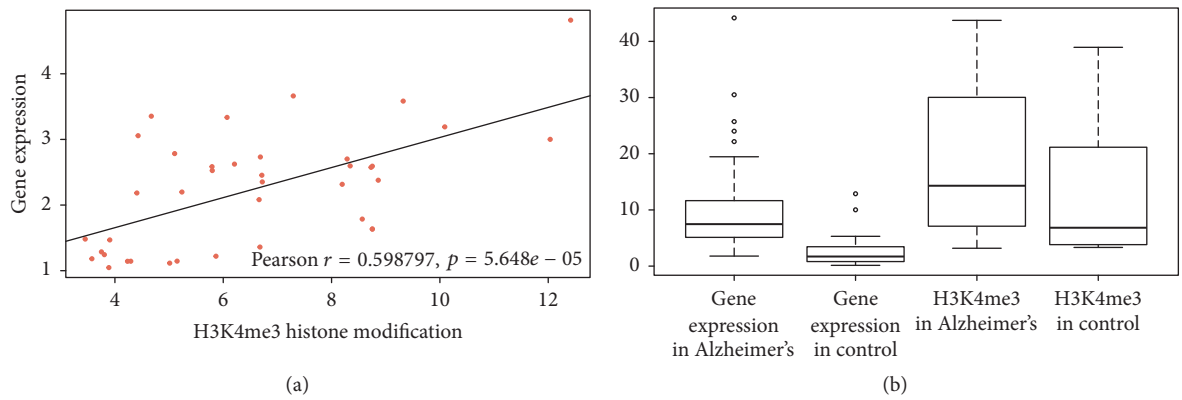


Figure 1: Positive association between expression level of differential lncRNA genes and H3K4me3 modification level in promoters of the differential lncRNA genes. (a) Scatter diagram and a fitting line show the positive association between fold change of lncRNA gene expression and fold change of H3K4me3 modification level in the promoter. (b) Boxplot of expression level of differential lncRNA genes and H3K4me3 modification level in the promoters of the differential lncRNA genes in AD and control samples, respectively, which shows that lncRNA genes with high H3K4me3 level in the promoters have high expression level. The circle in (b) refers to a singular point in statistics, differential from other points. But the singular point has statistical significance, showing the accuracy and objectivity of this article.

was downloaded from UCSC. Next the RNA-seq data are mapped to the mm10 reference sequence with TopHat2 [17] by default parameters. Cufflinks [18] was used to assemble the outcome of mapping and evaluate gene expression index. The lncRNA annotation was downloaded from GENCODE database, and differentially expressed lncRNA genes were identified by Cuffdiff with default parameters, a component of Cufflinks software.

2.3. Identifying Differential Histone Modification Peaks. To explore whether differentially expressed lncRNAs between AD and control are regulated by histone modification or not, we identified differential histone modification regions by analyzing the ChIP-seq data of histone marks H3K4me3 and H3K27me3 in AD and control. We firstly mapped the ChIP-seq data to the mm10 reference sequence by Bowtie2 software with default parameters. Then we used MACS2-callpeak [19] to identify the peaks of histone modification regions in the control mice and CKp25 Alzheimer's disease model [20], respectively. Finally, MACS2-bdgdiff is used to identify significantly differential histone modification regions between the control and AD.

2.4. Linking the Differential lncRNA Genes with the Differential Histone Modification Peaks Based on the Genomic Position. After identifying differential histone modification regions and differentially expressed lncRNA genes, we investigated whether the differential histone modification regions are located in the regulatory regions of the differential lncRNA genes. Herein, the regulatory regions are defined as 10 kbp upstream to 1 kbp downstream of transcriptional start site (TSS) of each differentially expressed lncRNA gene.

3. Results

3.1. Differentially Expressed lncRNA Genes between AD and Control Samples. By analyzing three AD and control RNA-seq data, we identified 72 significantly differentially expressed lncRNA genes with the BH-adjusted p value < 0.05 and fold change >2 (Supplementary Table 1, in Supplementary Material available online at http://dx.doi.org/10.1155/2016/3164238).

3.2. Differential Histone Modification Peaks between AD and Control Samples. We analyzed ChIP-seq histone modification data from CK-p25 AD model and control mice and identified 4,917 differential peaks of H3K4me3 and 1,624 differential peaks of H3K27me3 between AD and control samples, respectively.

3.3. Differential Histone Modification Peaks Are Located Upstream of Differentially Expressed lncRNA Genes. We found that there are 92 H3K4me3 differential histone modification peaks located in the promoters (2 kbp upstream to −1 kbp downstream) of 39 differentially expressed lncRNA genes (Supplementary Table 2) and 8 differential H3K27me3 histone modification peaks located in the region from 10 kb upstream to −1 kb downstream of 7 differentially expressed lncRNA genes. A positive association between histone modification level of H3K4me3 and lncRNA gene expression level is shown in Figure 1, and a negative association between histone modification level of H3K27me3 and lncRNA gene expression level is shown in Figure 2. A case study for the lncRNA gene named Gm20559 was shown in Figure 3, where the lncRNA Gm20559 had differential histone modification of H3K4me3 between AD and control in its promoter region, and exon 1 and exon 3 of Gm20559 are differentially expressed between AD and control. These results suggest that the majority of lncRNA genes (39 + 7)/72 may be transcriptionally regulated by histone modification in AD.

4. Discussion

lncRNA is a type of important regulatory RNAs that play critical roles in a wide range of biological processes. However, how the lncRNA genes themselves are transcriptionally regulated

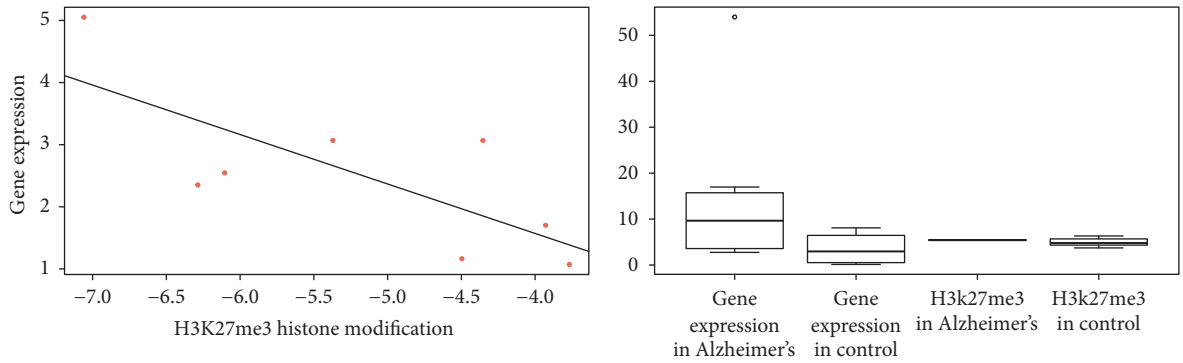


Figure 2: Negative association between expression level of differential lncRNA genes and H3K27me3 modification level in promoters of the differential lncRNA genes. (a) Scatter diagram and a fitting line show the negative association between fold change of expression level of differential lncRNA genes and fold change of H3K27me3 modification level in the promoters. (b) Boxplot of expression level of differential lncRNA genes and H3K27me3 modification level in the promoters of the differential lncRNA genes in AD and control samples, respectively, which shows that lncRNA genes with high H3K27me3 level in the promoters have low expression level. The circle in (b) refers to a singular point in statistics, differential from other points. But the singular point has statistical significance, showing the accuracy and objectivity of this article.

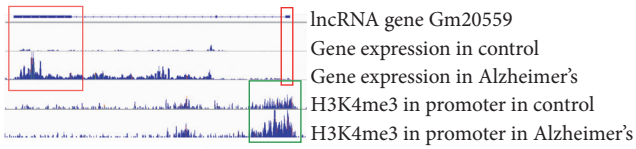


Figure 3: A lncRNA gene named Gm20559 with differential H3K4me3 modification level in the promoter between AD and control samples shows differential expression in exon 1 and exon 3. The red rectangle shows exon 1 and exon 3 regions of differentially expressed lncRNA gene Gm20559. And the green rectangle shows differential H3K4me3 histone modification in the promoter region, which suggests transcriptional regulation of Gm20559 by H3K4me3.

remains to be elucidated. In this paper, we used RNA-seq and ChIP-seq data from AD model and control to demonstrate that the majority of lncRNA genes are transcriptionally regulated by histone modification in AD.

As known, a protein-coding gene or lncRNA gene is regulated by many types of factors rather than one factor. Therefore, it sounds reasonable to integrate kinds of factors such as transcription factor, microRNA [21–24], DNA methylation, and histone modification to investigate the transcriptional regulation of lncRNAs in a specific condition such as AD, which will improve our understanding of lncRNA genes in AD.

Competing Interests

The authors declare no competing financial interests.

Authors' Contributions

Qinghua Jiang conceived the study. Guoqiang Wan collected the data, designed computational experiments, carried out statistical analysis, and wrote the manuscript. Wenyang Zhou participated in the analysis of data. Yang Hu, Shuilin Jin, and Rui Ma participated in the revision of this manuscript. Guiyou Liu gave comments and revisions to the final version of this manuscript. All authors read and approved the final manuscript. Guoqiang Wan and Wenyang Zhou equally contributed to this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61571152), the National High-Tech R&D Program of China (863 Program) [nos. 2014AA021505, 2015AA020101, and 2015AA020108], the National Science and Technology Major Project [no. 2013ZX03005012], and the Major State Research Development Program of China [no. 2016YFC1202302].

References

[1] A. Burns and S. Iliffe, "Alzheimer's disease," *British Medical Journal*, vol. 338, article b158, 2009.

[2] G. Liu and Q. Jiang, "Alzheimer's disease *CD33* rs3865444 variant does not contribute to cognitive performance," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 12, pp. E1589–E1590, 2016.

[3] G. Liu, Y. Xu, Y. Jiang, L. Zhang, R. Feng, and Q. Jiang, "PICALM rs3851179 variant confers susceptibility to alzheimer's disease in chinese population," *Molecular Neurobiology*, 2016.

[4] M. F. Mendez, "Early-onset Alzheimer's disease: nonamnestic subtypes and type 2 AD," *Archives of Medical Research*, vol. 43, no. 8, pp. 677–685, 2012.

[5] R. Lozano, M. Naghavi, K. Foreman et al., "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," *The Lancet*, vol. 380, no. 9859, pp. 2095–2128, 2012.

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

[7] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, 2016.

[8] 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.

[9] Q. Jiang, R. Ma, J. Wang et al., "LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data," *BMC Genomics*, vol. 16, supplement 3, p. S2, 2015.

[10] Q. Jiang, J. Wang, X. Wu et al., "LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression," *Nucleic Acids Research*, vol. 43, no. 1, pp. D193–D196, 2015.

[11] G. Chen, Z. Wang, D. Wang et al., "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic Acids Research*, vol. 41, no. 1, pp. D983–D986, 2013.

[12] 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.

[13] I. Ulitsky and D. P. Bartel, "LincRNAs: genomics, evolution, and mechanisms," *Cell*, vol. 154, no. 1, pp. 26–46, 2013.

[14] Q. Jiang, J. Wang, Y. Wang, R. Ma, X. Wu, and Y. Li, "TF2LncRNA: identifying common transcription factors for a list of lncRNA genes from ChIP-seq data," *BioMed Research International*, vol. 2014, Article ID 317642, 5 pages, 2014.

[15] E. Gjoneska, A. R. Pfenning, H. Mathys et al., "Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease," *Nature*, vol. 518, no. 7539, pp. 365–369, 2015.

[16] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[17] S. Ghosh and C. K. Chan, "Analysis of RNA-seq data using TopHat and cufflinks," in *Plant Bioinformatics: Methods and Protocols*, vol. 1374 of *Methods in Molecular Biology*, pp. 339–361, Springer, Berlin, Germany, 2016.

[18] C. Trapnell, A. Roberts, L. Goff et al., "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nature Protocols*, vol. 7, no. 3, pp. 562–578, 2012.

[19] Y. Zhang, T. Liu, C. A. Meyer et al., "Model-based analysis of ChIP-Seq (MACS)," *Genome Biology*, vol. 9, article R137, 2008.

[20] T. Liu, "Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells," in *Stem Cell Transcriptional Networks: Methods and Protocols*, B. L. Kidder, Ed., vol. 1150 of *Methods in Molecular Biology*, pp. 81–95, 2014.

[21] 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.

[22] 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.

[23] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *eLife*, vol. 4, Article ID e05005, 2015.

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

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 2503023, 9 pages
http://dx.doi.org/10.1155/2016/2503023

Research Article

Identification and Characterization of Small Noncoding RNAs in Genome Sequences of the Edible Fungus *Pleurotus ostreatus*

Jibin Qu,[1,2] Mengran Zhao,[1,2] Tom Hsiang,[3] Xiaoxing Feng,[4,5] Jinxia Zhang,[1,2] and Chenyang Huang[1,2]

[1] *Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing, China*
[2] *Key Laboratory of Microbial Resources, Ministry of Agriculture, Beijing, China*
[3] *School of Environmental Sciences, University of Guelph, Guelph, ON, Canada*
[4] *College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong, China*
[5] *Shenzhen Micro & Nano Research Institute of IC and System Applications, Shenzhen, Guangdong, China*

Correspondence should be addressed to Chenyang Huang; huangchenyang@caas.cn

Received 30 June 2016; Revised 21 August 2016; Accepted 24 August 2016

Academic Editor: Xing Chen

Noncoding RNAs (ncRNAs) have been identified in many fungi. However, no genome-scale identification of ncRNAs has been inventoried for basidiomycetes. In this research, we detected 254 small noncoding RNAs (sncRNAs) in a genome assembly of an isolate (CCEF00389) of *Pleurotus ostreatus*, which is a widely cultivated edible basidiomycetous fungus worldwide. The identified sncRNAs include snRNAs, snoRNAs, tRNAs, and miRNAs. SnRNA U1 was not found in CCEF00389 genome assembly and some other basidiomycetous genomes by BLASTn. This implies that if snRNA U1 of basidiomycetes exists, it has a sequence that varies significantly from other organisms. By analyzing the distribution of sncRNA loci, we found that snRNAs and most tRNAs (88.6%) were located in pseudo-UTR regions, while miRNAs are commonly found in introns. To analyze the evolutionary conservation of the sncRNAs in *P. ostreatus*, we aligned all 254 sncRNAs to the genome assemblies of some other Agaricomycotina fungi. The results suggest that most sncRNAs (77.56%) were highly conserved in *P. ostreatus*, and 20% were conserved in Agaricomycotina fungi. These findings indicate that most sncRNAs of *P. ostreatus* were not conserved across Agaricomycotina fungi.

1. Introduction

Pleurotus ostreatus (Jacq.: Fr.) Kumm. (Dikarya; Basidiomycota; Agaricomycotina; Agaricales) is an important commercially available edible fungus worldwide, and it is the most popular edible mushroom in Northern China. This fungus can grow easily on a variety of organic substrates, including agricultural wastes [1, 2]. In addition to its delicious taste and nutritional value [3], this mushroom also has health-promoting effects [4]. Furthermore, it is tolerant of a wide temperature range during the cultivation [5]. Because of its wide substrate utilization, it is a good model for the study of lignin biodegradation [6] and environmental adaptation.

Noncoding RNAs (ncRNAs) producing functional RNA products instead of proteins [7] are widely expressed in both prokaryotes and eukaryotes [8–10]. For example, around 98% of transcriptional output in human is ncRNA [11]. NcRNA families are grouped into structural ncRNA and regulatory ncRNA based on their structure and function [9]. The structural ncRNA includes transfer RNA (tRNA) and ribosomal RNA (rRNA), as well as other small but stable noncoding RNAs, such as small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), Ribonuclease P (RNase P), mitochondrial RNA processing (MRP) RNA, signal recognition particle (SRP) RNA, and telomerase RNA. Regulatory ncRNAs include microRNAs (miRNAs) and long ncRNAs (lncRNAs) [12]. These ncRNAs play important roles in splicing [13], transcription [14], translation [15], and chromatin architecture [16], and many ncRNAs are associated with diseases [17–24].

Recently, ncRNAs have been identified by experimental and computational methods in several fungi [10, 25–28].

But so far, there have been few studies related to ncRNA in basidiomycetes and even fewer for edible mushroom. Apart from rRNAs and a few tRNAs, no other ncRNAs have been annotated and characterized in the *P. ostreatus* genome. In this research, we sequenced the genome of a strain of *P. ostreatus* and identified small ncRNAs (sncRNAs) in the genome assembly. Then the distribution of genomic loci of these sncRNAs was characterized to describe the preferential locations of different sncRNAs. Lastly, we analyzed the evolutionary conservation of these sncRNAs among basidiomycetous fungi.

2. Materials and Methods

2.1. Strains and Culture Conditions. The *Pleurotus ostreatus* dikaryotic strain, CCMSSC00389, is widely cultivated in China and is preserved in the China Center for Mushroom Spawn Standards and Control, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences. From this strain, the two nuclear types were separated to constituent monokaryons by dedikaryotisation as follows: the dikaryon was grown in 10 cm diameter Petri dishes containing 25 mL of potato dextrose agar (PDA) at 25°C for 6-7 days. Mycelia (1 g) were collected from the growing margins of the plate and suspended in 2% lytic enzyme (Guangdong Institute of Microbiology, China) and 0.6 mol/L mannitol and incubated at 30°C for 4 h. The resulting protoplasts were washed twice with 0.6 mol/L mannitol and placed (dissolved = broken up) in mannitol solution. The protoplast suspension was spread onto malt-yeast-glucose (MYG) medium and incubated at 25°C for 4-5 days. Monokaryons were identified by microscopy among the regeneration clones by lack of clamp connections and further confirmed by mating to produce dikaryotic hyphae with clamps connections. A single monokaryon of each nuclear type was randomly selected and sequenced and named CCEF00389 and CCEF00389_9.

2.2. Isolation of Genomic DNA. Genomic DNAs of the two monokaryons were extracted using a DP305-Plant Genome Extraction Kit (Tianjin, China). The purity and quality of the genomic DNA were determined through spectrophotometry and electrophoresis on a 1.0% agarose gel and sequenced using the Illumina HiSeq 2500. The raw data were generated by paired-end and mate-pair sequencing with different insert sizes. Strain CCEF00389 used a whole genome *de novo* sequencing strategy with average coverage of over 300x. Three libraries were constructed for 100 bp paired-end (300 bp insert size) and mate-pair sequencing (3 kbp and 8 kbp insert length).

2.3. Transcriptomic Data. Mycelia of the same strain were inoculated on the DifcoTM Potato Dextrose Agar plates with cellophane at 25°C for four days and were subjected to heat stress at 37 centigrade for different time (0, 0.5, 1, and 1.5 h). The mycelia through different treatment were collected, respectively, for RNA extraction. The RNA samples were then sequenced with Illumina HiSeq 2500. One library for each time point was constructed for 100 bp paired-end (300 bp insert size) sequencing. The raw data were assembled to the transcriptome with *de novo* assembler TRINITY [29].

2.4. Genome Assembly and Annotation. Raw reads were first trimmed by stripping the adaptor sequences and ambiguous nucleotides using SeqPrep (https://github.com/jstjohn/SeqPrep) and Sickle (https://github.com/najoshi/sickle). Reads with quality scores less than 20 or "N" more than 10% or lengths below 25 bp were removed. The cleaned reads were assembled using the tools PLATANUS [30] and L_RNA_Scaffolder [31] with *de novo* assembly guided by the assembled transcriptome. Gene models in the genome assembly of *P. ostreatus* were predicted using BRAKER1 [32]. The protein-coding genes were then confirmed using BLAST+ (version 2.2.31) against public databases, including the NCBI nonredundant database (NR) database, the Refseq database of fungi, ESTs of *P. ostreatus* PC15 (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PleosPC15_2), the predicted protein models of 134 basidiomycetous species in JGI website (http://genome.jgi.doe.gov/basidiomycota/basidiomycota.info.html), and the transcriptome of CCEF00389. The predicted gene models were then classified according to Gene Ontology (GO) [33] with homologous sequences in the NR database and also annotated by their protein domains using InterProScan [34].

2.5. SncRNA Detection. Small ncRNAs were first identified by aligning Rfam sequences to our genome assembly using BLAST+ and Infernal (version 1.0.3). These sncRNAs included snRNAs and snoRNAs. tRNAs were predicted with tRNAscan-SE (version 1.3.1) [35]. miRNAs were detected by alignment of Rfam miRNA sequences (RF00003) to our genome assembly with BLASTn, with the *e*-value cutoff $1e-3$ and the word size 19.

2.6. Nucleotide Sequence Accession Number. This whole genome shortgun project has been deposited at DDBJ/EMBL/GenBank (http://www.ncbi.nlm.nih.gov/) under the accession number MAYC00000000 (the project accession number PRJNA327267).

3. Results and Discussion

3.1. Genome Information of CCEF00389. A 34.9-Mb genome assembly was obtained by assembling approximately 81 million Illumina reads (~300x coverage) (Table 1 and Figure 1). Gene prediction from all scaffolds of the assembled genome and transcriptomic data generated 13,438 gene models. The genome size, number of predicted genes, and the basic information of predicted genes are very similar as those of related edible Agaricales, such as *Volvariella volvacea* [36], *Agaricus bisporus* [37], and *Flammulina velutipes* [38] (see Supplement Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/2503023). Gene Ontology (GO) annotations were found for 6,566 proteins (48.9%) with homologous sequences in the NR database. In addition, 9,931 (73.9%) of all predicted genes can be annotated by their protein domains by InterProScan.

Table 1: General features of the *P. ostreatus* CCEF00389 genome assembly.

Number of scaffolds	2,529
Length of all scaffolds combined (Mb)	35.8
GC content (%)	49.54
Scaffold N50 value (bp)	394,787
Number of large scaffolds (>1000)	794
Length of large scaffolds (Mb)	34.9
Number of protein-coding genes	13,438

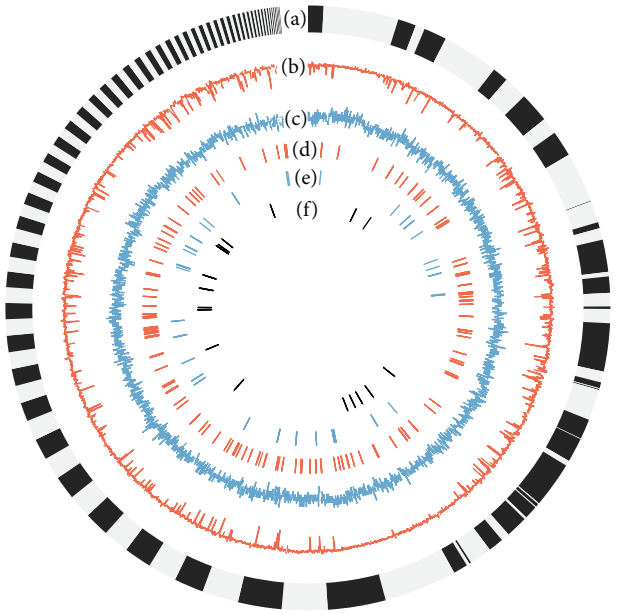


Figure 1: The ideogram showing the genomic features of *P. ostreatus*. (a) Scaffolds longer than 10 kbp. (b) GC content: the percentage of G+C in 10 kbp nonoverlapping windows. (c) Gene density: the number of genes in 10 kbp nonoverlapping windows. (d) Distribution of tRNAs. (e) Distribution of miRNAs. (f) Distribution of snRNAs, snoRNAs, and others.

3.2. Identification of sncRNAs in P. ostreatus

3.2.1. sncRNAs from Rfam 11: snRNAs, snoRNAs, and Other sncRNAs. The spliceosome contains five essential snRNAs: U1, U2, U4, U5, and U6 [39]. Four of them were identified in the CCEF00389 genome assembly: U4 and U5 exhibited a precise genomic location, while U2 and U6 had several candidate locations in the genome assembly (Table 2). To find the U1 genomic locus in the genome, we downloaded the U1 sequences of all species from Rfam and the U1 sequences of fungi from NCBI to be used as query sequences to search for homologues in the CCEF00389 genome using BLASTn. Interestingly, U1 was not found in this genome assembly, even after extensive searching with sequences from other fungi. Furthermore, U1 was not found in genome assemblies of other basidiomycetous fungi including *Agaricus bisporus* [37], *Coprinopsis cinerea* [40], *Flammulina velutipes* [38], *Schizophyllum commune* [41], *Pleurotus ostreatus PC15* [42], *Volvariella volvacea* [36], *Laccaria bicolor* [43], *and Ustilago maydis* [44]. This implies that if snRNA U1 of basidiomycetes exists, it has a sequence that varies significantly from other organisms.

Small nucleolar RNAs (snoRNAs) guide chemical modifications of other cellular RNAs, including rRNAs, tRNAs, and snRNAs. There are two major classes of snoRNA in eukaryotic cells: the C/D box snoRNAs, which are associated with methylation, and the H/ACA box snoRNAs, which are associated with pseudouridylation [10, 45]. Seven snoRNAs were identified in the CCEF00389 genome assembly: three of them were of Rfam class snoZ13_snr52, and each of the others is belonging to Rfam class snosnR60_Z15, SNORD24, Afu_455, and SNORD46, respectively. There were also six other sncRNAs in the Rfam searching result: one RNase_MRP RNA and five Hammerhead ribozymes (type 3).

3.2.2. tRNA. A transfer RNA (tRNA) is adaptor RNA molecule that serves as the physical link between the mRNA and protein [46], so it is a necessary component of translation and essential for life. However, the number of tRNAs in the genome assemblies of different organisms varies tremendously [47–49]. In the genome assembly of CCEF00389, we identified 185 tRNAs with length from 71 to 144 nt with their loci and anticodons shown in Supplement Table 2.

3.2.3. miRNA. A micro-RNA (miRNA) is a small noncoding RNA molecule about 22 nucleotides in length, which functions in RNA silencing and posttranscriptional regulation [50]. The miRNAs have been identified in the genome assemblies of most eukaryotic organisms and are very abundant in many of them [51–54]. There were only 46 mature miRNAs identified in the CCEF00389 genome assembly by BLASTn, with lengths from 19 to 23. The most important factor in uncovering putative miRNAs was the parameter "word size" of BLASTn. If this parameter was set to 20, many fewer matches (only 10) were found. As we know, some miRNAs have a variation of 1-2 nt at the end (often 3′ end) [51]. And a probable reason of the lack of miRNAs in this genome assembly is that there is no currently available miRNA database for basidiomycetes. Compared with the known miRNAs, the sequences are not evolutionarily conserved.

3.3. Distribution of snRNAs in the CCEF00389 Genome Assembly. Most sncRNAs are located in noncoding regions of the genome, including introns, UTRs, and intergenic regions. The location of ncRNA might be associated with its function. For example, the ncRNAs in UTRs and intergenic regions may play cisregulatory roles regulating their adjacent genes, and/or transregulatory roles elsewhere in the genome [55]. And the ncRNAs in introns could regulate gene expression through transcriptional gene silencing (TGS) pathways [56, 57] and posttranscriptional gene silencing pathways [58–60].

We wanted to identify the sncRNAs to locus and characterize their distribution. The UTR regions of the CCEF00389 genome assembly were not identified, so the distribution of sncRNAs located outside the ORFs (from the start codon to the stop codon) was defined quantifiably as the distance to the nearer gene boundary (start/stop codon). The UTR regions

Table 2: Genomic loci, distance to gene boundary, and neighboring gene of sncRNAs identified using Rfam.

ID	Location	Strand	Distance to gene boundary	Neighbor
U2.1	scaffold37_107912–108102	−	470	g12507
U2.2	scaffold43_54422–54612	−	487	g12899
U2.3	scaffold43_63239–63429	+	536	g12905
U2.4	scaffold37_118109–118299	+	329	g12510
RNase_MRP.1	scaffold1_325685–326108	+	172	g8330
U6.1	scaffold_59_650694–650810	+	315	g7427
U6.2	scaffold_80_192511–192627	+	150	g8186
U6.3	scaffold63_39552–39668	+	399	g4097
U5.1	scaffold16_469793–469909	−	959	g10739
snoZ13_snr52.1	scaffold46_213325–213430	−	394	g13123
U4.1	scaffold113_43830–43975	+	459	g5145
snoZ13_snr52.2	scaffold1_729232–729339	−	Intron	g8490
snoZ13_snr52.3	scaffold46_213071–213178	−	646	g13123
snosnR60_Z15.1	scaffold_75_65986–66075	−	Intron	g7941
Hammerhead_3.1	scaffold59_142237–142291	+	277	g3917
Hammerhead_3.2	scaffold59_152597–152651	−	257	g3922
Hammerhead_3.3	scaffold_12_734778–734832	+	262	g1296
Hammerhead_3.4	scaffold59_146486–146540	+	685	g3919
SNORD24.1	scaffold60_67979–68061	−	Intron	g3964
Afu_455.1	scaffold_11_316838–316924	−	Intron	g963
Hammerhead_3.5	scaffold25_333837–333891	+	Intron	g11689
SNORD46.1	scaffold37_225673–225759	+	1031	g12558

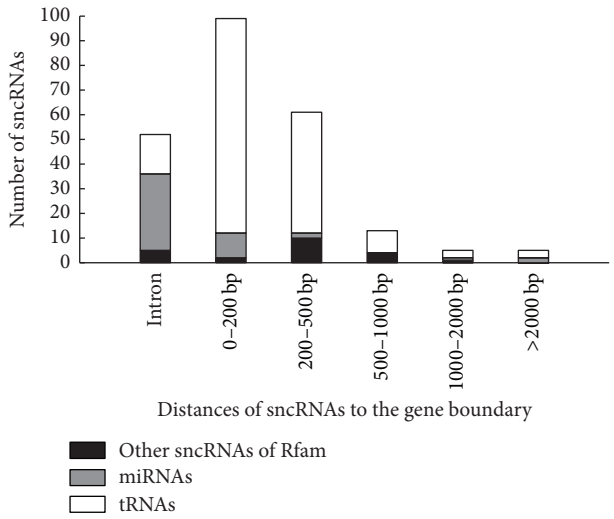


Figure 2: Distance of sncRNAs to the gene boundary (outside a start or stop codon).

usually lie within 2000 bp of gene boundary [61], and it can be assumed that the less the distance to gene boundary, the greater the possibility to be located in UTR [62].

For the three kinds of sncRNAs (tRNAs, miRNAs, and other sncRNAs from Rfam), the distribution is shown in Figure 2.

All detected snRNAs were located within 1,000 bp of a gene boundary. Among them, U5 was located 959 bp from a gene boundary, and the other snRNAs were located within 536 bp of gene boundaries. It is highly likely that all the snRNAs are located in pseudo-UTR regions of this genome. A similar distribution of snRNAs was found in the filamentous fungus *Trichophyton rubrum* [25]. For the snoRNAs, Hammerhead RNAs, and the RNase MRP, they were located diversely in the genome assembly: within 1031 bp of the gene boundary and in introns (5 out of 13) (see Table 2).

Most tRNAs (136 out of 167, 81.44%) located within 500 bp of gene boundary; this means that tRNAs distributed mainly in pseudo-UTR regions. There are also 16 tRNAs (9.58%) located in introns. For details, see Supplement Table 3.

As many as 67% (31 out of 46) of miRNAs in the CCEF00389 genome assembly located in introns, which are usually regulated together with their host genes [63, 64]. Two miRNAs, miR1171 and miR3948, located at a distance of more than 2000 bp away from an ORF, were intergenic (see Table 3).

3.4. Evolutionary Conservation of sncRNAs in P. ostreatus. In order to analyze the evolutionary conservation of sncRNAs in *P. ostreatus*, all identified sncRNAs were then aligned to the genomes of other fungi, including six *P. ostreatus*-related Agaricomycotina fungi: *Agaricus bisporus* [37], *Coprinopsis cinerea* [40], *Flammulina velutipes* [38], *Schizophyllum commune* [41], *Pleurotus ostreatus* PC15 [42], *Volvariella volvacea* [36], and finally *Ustilago maydis* [44] which is a basidiomycete, but basal to the Agaricomycotina.

Only 10 of these sncRNAs were also identified in all selected basidiomycetes, and all these conserved sncRNAs

Table 3: Genomic loci, distance to gene boundary, and neighboring gene of miRNAs.

ID	Location	Strand	Distance to gene boundary	Neighbor
miR-124-5p	scaffold_29_13029–13047	+	Intron	g3043
miR-190a-3p	scaffold48_55792–55810	+	53	g13155
miR-788-5p	scaffold8_754188–754207	−	Intron	g9721
miR-383-3p	scaffold_80_111549–111567	−	Intron	g8156
miR-466g	scaffold8_19655–19673	−	94	g9475
miR1171	scaffold_70_119604–119626	+	2094	g7800
miR-467g	scaffold_26_54067–54085	−	147	g2750
miR-190a-3p	scaffold48_55792–55810	+	53	g13155
miR1427	scaffold85_42007–42025	−	110	g4652
miR2095-5p	scaffold_29_43092–43110	+	1121	g3054
miR-788	scaffold8_754188–754207	−	Intron	g9721
miR2673a	scaffold32_244540–244559	+	71	g12059
miR2673a	scaffold25_190886–190904	+	Intron	g11628
miR2673a	scaffold_15_223801–223819	−	Intron	g1786
miR2673b	scaffold32_244540–244559	+	71	g12059
miR2673b	scaffold25_190886–190904	+	Intron	g11628
miR2673b	scaffold_15_223801–223819	−	Intron	g1786
miR-2709	scaffold64_62205–62223	+	Intron	g4162
miR-2783	scaffold392_645–663	+	Intron	g5776
miR-190a-3p	scaffold48_55792–55810	+	53	g13155
miR156h-3p	scaffold151_19717–19735	−	Intron	g5461
miR4243	scaffold_15_464570–464589	−	Intron	g1895
miR-3677-5p	scaffold61_171654–171672	−	363	g4080
miR-3775	scaffold1170_174–192	−	Intron	g6228
miR3948	scaffold155_16079–16098	+	3715	g5479
miR3948	scaffold8_7775–7794	−	Intron	g9470
miR-4459	scaffold7_14624–14642	+	Intron	g9172
miR-4968-3p	scaffold_14_64946–64966	−	Intron	g1539
miR-4968-3p	scaffold34_263134–263153	−	Intron	g12318
miR-4968-3p	scaffold_24_6521–6540	−	Intron	g2572
miR-4968-3p	scaffold3_287008–287026	+	Intron	g8829
miR-5352-5p	scaffold53_49526–49544	−	Intron	g13380
miR-5455-3p	scaffold54_13491–13509	−	Intron	g3643
miR-6012-5p	scaffold14_172648–172666	−	Intron	g10243
miR6214	scaffold58_68922–68940	−	Intron	g3823
miR-6606-5p	scaffold_19_158806–158825	+	Intron	g2084
miR-7426-5p	scaffold46_187269–187287	+	Intron	g13115
miR7734-3p	scaffold21_198083–198101	+	Intron	g11335
miR-190a-3p	scaffold48_55792–55810	+	53	g13155
miR-8481-5p	scaffold1214_35–53	+	Intron	g6248
miR-8922	scaffold21_9675–9693	+	Intron	g11273
miR-8986a	scaffold_4_231055–231073	−	Intron	g184
miR-9189b	scaffold_24_289739–289757	−	Intron	g2662
miR-9400-5p	scaffold3_354824–354842	+	Intron	g8861
miR-190a-3p	scaffold48_55792–55810	+	53	g13155
miR9773	scaffold390_1391–1409	−	292	g5774

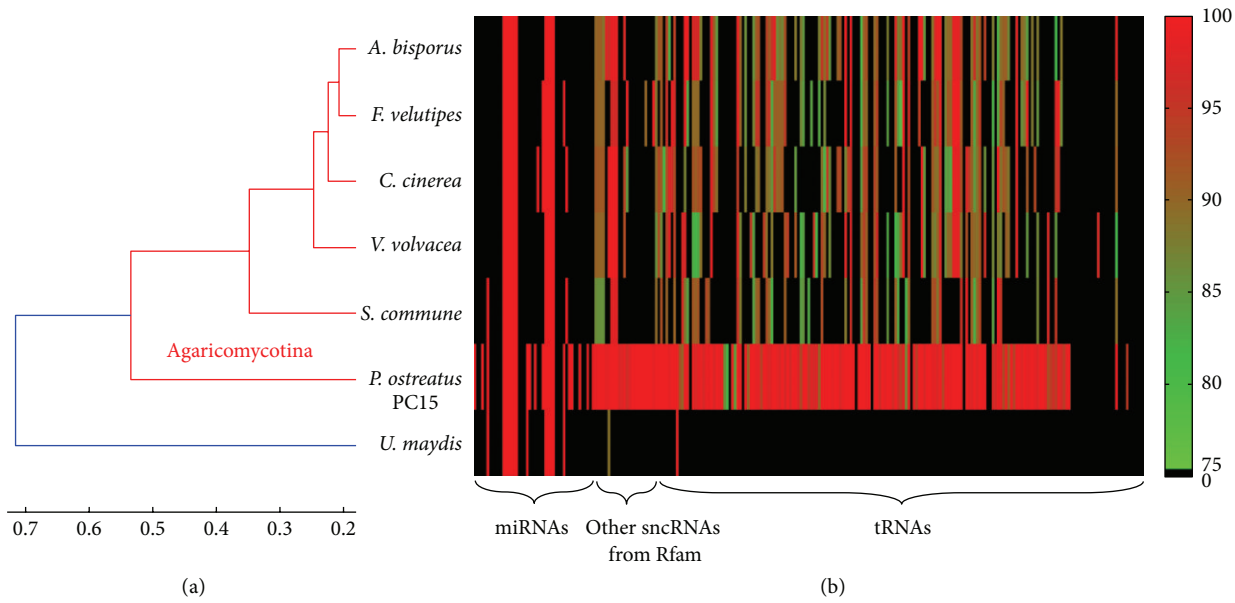


Figure 3: Similarity clustering based on sequence identities of sncRNAs between CCEF00389 and seven other basidiomycetous fungi: (a) the hierarchical clustering tree; (b) the heat map of identities; the black color means no matches at the cutoff $1e-3$ ($1e-1$ for miRNAs, because the length of their sequences is short).

were miRNAs. This means that only a small part of miRNAs are conserved because of the low rate of evolution [65]. Only 5.9% (15 out of 254) of sncRNAs in the CCEF00389 genome assembly had homologues in *Ustilago maydis*. Most sncRNAs identified in Agaricomycotina fungi do not have homologues in other groups of fungi. There were 51 of these sncRNAs also identified in all six genomes of Agaricomycotina fungi, and 74 of these sncRNAs were also identified in at least five genomes of Agaricomycotina species. Moreover, the sequence identities of the matches were above 80%. This suggests that many sncRNAs are highly conserved among Agaricomycotina fungi. These conserved sncRNAs included the snRNA U2 (4 candidates) and U6 (3 candidates), 10 miRNAs (miR2673a, miR2673b, and miR-4968-3p), and 57 tRNAs (see Supplement Table 4). In some previous researches, the microRNAs miR2673 and miR-4968-3p were found to have many target genes in many species [66, 67] and may regulate some targets [68, 69]. MicroRNA miR2673 was also be found to have stable structure and be conserved across plant species [70].

To compare the sequence similarity between sncRNAs of CCEF00389 and their homologues in selected fungi, a hierarchical clustering was performed to partition the different fungi based on the sequence identities. In the hierarchical clustering method, the Spearman correlation coefficient of sequence identities of all sncRNAs (if no matches were found, the identity was set to zero) was selected to define the dissimilarity between organisms. Figure 3 shows the result of clustering. It is clear that the homologues of sncRNAs of *Pleurotus ostreatus* PC15 were most similar to sncRNAs of CCEF00389, because they belong to the same species. There were 77.56% (197 out of 254) of matched sncRNAs with sequence identities above 81.65%. For the other five organisms, the clustering results basically reflected the currently accepted phylogenetic placement of these species [71].

4. Conclusions

The CCEF00389 genome assembly is the first released draft genome of a strain of *P. ostreatus* in China. The genome size, number of genes, and some protein families were in accordance with the released genome of PC15, which is a North American strain of *P. ostreatus*. In the CCEF00389 genome assembly, we detected 254 sncRNAs which were not reported before. This was the first study of genome-scale identification of sncRNAs for a basidiomycete. The sequence length of sncRNAs accounted for 0.054% of CCEF00389 genome, and the identified sncRNAs included most classes of known sncRNAs. However, the snRNA U1 was not identified not only in CCEF00389, but also in other basidiomycetous genomes. This implies that if snRNA U1 of basidiomycetes exists, it has a sequence that varies significantly from other organisms.

For some sncRNAs, the position of loci may be associated with some potential functions. The UTR regions of the CCEF00389 genome assembly were not precisely determined, so we calculated the distances of sncRNAs to the gene boundary (start/stop codon) for possible location in pseudo-UTR regions. The snRNAs and tRNAs had a higher possibility to be located in pseudo-UTR regions, while the miRNAs were more common in introns.

There were 197 sncRNAs in CCEF00389 genome, which had detectable homologues in another strain of *P. ostreatus*, and 74 sncRNAs in CCEF00389 genome which were also found in some other Agaricomycotina fungi. However, only 15 sncRNAs in CCEF00389 genome had homologues in *Ustilago maydis*, which does not belong to Agaricomycotina. It suggests that most sncRNAs of *P. ostreatus* were not conserved across Agaricomycotina fungi.

Long ncRNA (lncRNA) is also a kind of impressive ncRNA which plays critical roles in multiple biological processes based on diverse underlying mechanisms [17, 22]. And

prediction of the interaction between ncRNAs and proteins has attracted much attention because the ncRNAs function mediated with proteins. In the future work, we will focus on identification and analysis of lncRNAs [12] and prediction of the interactions between ncRNAs and proteins [72–74].

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was supported by National Basic Research Program of China (2014CB138303) and China Agriculture Research System (CARS24).

References

[1] C. Sánchez, "Cultivation of *Pleurotus ostreatus* and other edible mushrooms," *Applied Microbiology and Biotechnology*, vol. 85, no. 5, pp. 1321–1337, 2010.

[2] G. Aggelis, D. Iconomou, M. Christou et al., "Phenolic removal in a model olive oil mill wastewater using *Pleurotus ostreatus* in bioreactor cultures and biological evaluation of the process," *Water Research*, vol. 37, no. 16, pp. 3897–3904, 2003.

[3] P. Mattila, K. Könkö, M. Eurola et al., "Contents of vitamins, mineral elements, and some phenolic compounds in cultivated mushrooms," *Journal of Agricultural and Food Chemistry*, vol. 49, no. 5, pp. 2343–2348, 2001.

[4] A. Jedinak, S. Dudhgaonkar, Q.-L. Wu, J. Simon, and D. Sliva, "Anti-inflammatory activity of edible oyster mushroom is mediated through the inhibition of NF-κB and AP-1 signaling," *Nutrition Journal*, vol. 10, no. 1, article 52, 2011.

[5] G. Eger, G. Eden, and E. Wissig, "*Pleurotus ostreatus*—breeding potential of a new cultivated mushroom," *Theoretical and Applied Genetics*, vol. 47, no. 4, pp. 155–163, 1976.

[6] T. M. Salame, D. Knop, D. Levinson, S. J. Mabjeesh, O. Yarden, and Y. Hadar, "Release of *Pleurotus ostreatus* versatile-peroxidase from Mn^{2+} repression enhances anthropogenic and natural substrate degradation," *PLoS ONE*, vol. 7, no. 12, Article ID e52446, 2012.

[7] S. R. Eddy, "Non-coding RNA genes and the modern RNA world," *Nature Reviews Genetics*, vol. 2, no. 12, pp. 919–929, 2001.

[8] M. Guttman, I. Amit, M. Garber et al., "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals," *Nature*, vol. 458, no. 7235, pp. 223–227, 2009.

[9] J. S. Mattick and I. V. Makunin, "Non-coding RNA," *Human Molecular Genetics*, vol. 15, pp. R17–R29, 2006.

[10] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, "Rfam: annotating non-coding RNAs in complete genomes," *Nucleic Acids Research*, vol. 33, pp. D121–D124, 2005.

[11] J. S. Mattick, "Non-coding RNAs: the architects of eukaryotic complexity," *EMBO Reports*, vol. 2, no. 11, pp. 986–991, 2001.

[12] J. Li, B. Wu, J. Xu, and C. Liu, "Genome-wide identification and characterization of long intergenic non-coding RNAs in *Ganoderma lucidum*," *PLoS ONE*, vol. 9, no. 6, Article ID e99442, 2014.

[13] J. W. S. Brown, D. F. Marshall, and M. Echeverria, "Intronic noncoding RNAs and splicing," *Trends in Plant Science*, vol. 13, no. 7, pp. 335–342, 2008.

[14] J. S. Mattick and I. V. Makunin, "Small regulatory RNAs in mammals," *Human Molecular Genetics*, vol. 14, no. 1, pp. R121–R132, 2005.

[15] T. M. T. Hall, "Structure and function of argonaute proteins," *Structure*, vol. 13, no. 10, pp. 1403–1408, 2005.

[16] E. Bernstein and C. D. Allis, "RNA meets chromatin," *Genes & Development*, vol. 19, no. 14, pp. 1635–1655, 2005.

[17] X. Chen, C. C. Yan, X. Zhang, and Z. H. You, "Long non-coding RNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, 2016.

[18] X. Chen, Y. A. Huang, X. Wang, Z. H. You, and K. C. Chan, "FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 29, pp. 45948–45958, 2016.

[19] Y.-A. Huang, X. Chen, Z.-H. You, D.-S. Huang, and K. C. C. Chan, "ILNCSIM: improved lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 18, pp. 25902–25914, 2016.

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

[21] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Scientific Reports*, vol. 5, Article ID 16840, 2015.

[22] 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.

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

[24] 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.

[25] T. Liu, X. Ren, T. Xiao et al., "Identification and characterisation of non-coding small RNAs in the pathogenic filamentous fungus *Trichophyton rubrum*," *BMC Genomics*, vol. 14, article 931, 2013.

[26] S.-S. Chang, Z. Zhang, and Y. Liu, "RNA interference pathways in fungi: mechanisms and functions," *Annual Review of Microbiology*, vol. 66, pp. 305–323, 2012.

[27] N. Liu, Z. D. Xiao, C. H. Yu et al., "SnoRNAs from the filamentous fungus *Neurospora crassa*: structural, functional and evolutionary insights," *BMC Genomics*, vol. 10, article 515, 2009.

[28] F. J. Van Werven, G. Neuert, N. Hendrick et al., "Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast," *Cell*, vol. 150, no. 6, pp. 1170–1181, 2012.

[29] M. G. Grabherr, B. J. Haas, M. Yassour et al., "Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[30] R. Kajitani, K. Toshimoto, H. Noguchi et al., "Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads," *Genome Research*, vol. 24, no. 8, pp. 1384–1395, 2014.

[31] W. Xue, J.-T. Li, Y.-P. Zhu et al., "L_RNA_scaffolder: scaffolding genomes with transcripts," *BMC Genomics*, vol. 14, article 604, 2013.

[32] K. J. Hoff, S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, "BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS," *Bioinformatics*, vol. 32, no. 5, pp. 767–769, 2016.

[33] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[34] P. Jones, D. Binns, H.-Y. Chang et al., "InterProScan 5: genome-scale protein function classification," *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, 2014.

[35] T. M. Lowe and S. R. Eddy, "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence," *Nucleic Acids Research*, vol. 25, no. 5, pp. 955–964, 1997.

[36] D. Bao, M. Gong, H. Zheng et al., "Sequencing and comparative analysis of the straw mushroom (*Volvariella volvacea*) genome," *PLoS ONE*, vol. 8, no. 3, Article ID e58294, 2013.

[37] E. Morin, A. Kohler, A. R. Baker et al., "Genome sequence of the button mushroom *Agaricus bisporus* reveals mechanisms governing adaptation to a humic-rich ecological niche," *Proceedings of the National Academy of Sciences*, vol. 109, no. 43, pp. 17501–17506, 2012.

[38] Y.-J. Park, J. H. Baek, S. Lee et al., "Whole genome and global gene expression analyses of the model mushroom *Flammulina velutipes* reveal a high capacity for lignocellulose degradation," *PLoS ONE*, vol. 9, no. 4, Article ID e93560, 2014.

[39] P. Schattner, W. A. Decatur, C. A. Davis et al., "Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome," *Nucleic Acids Research*, vol. 32, no. 14, pp. 4281–4296, 2004.

[40] J. E. Stajich, S. K. Wilke, D. Ahrén et al., "Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 26, pp. 11889–11894, 2010.

[41] R. A. Ohm, J. F. De Jong, L. G. Lugones et al., "Genome sequence of the model mushroom *Schizophyllum commune*," *Nature Biotechnology*, vol. 28, no. 9, pp. 957–963, 2010.

[42] R. Riley, A. A. Salamov, D. W. Brown et al., "Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 27, pp. 9923–9928, 2014.

[43] F. Martin, A. Aerts, D. Ahrén et al., "The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis," *Nature*, vol. 452, no. 7183, pp. 88–92, 2008.

[44] J. Kämper, R. Kahmann, M. Bölker et al., "Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*," *Nature*, vol. 444, no. 7115, pp. 97–101, 2006.

[45] J.-P. Bachellerie, J. Cavaillé, and A. Hüttenhofer, "The expanding snoRNA world," *Biochimie*, vol. 84, no. 8, pp. 775–790, 2002.

[46] S. J. Sharp, J. Schaack, L. Cooley, D. J. Burke, and D. Söll, "Structure and transcription of eukaryotic tRNA genes," *Critical Reviews In Biochemistry*, vol. 19, no. 2, pp. 107–144, 1985.

[47] J. Spieth, D. Lawson, P. Davis, G. Williams, and K. Howe, *Overview of Gene Structure in C. elegans*, 2005.

[48] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[49] J. M. Kim, S. Vanguri, J. D. Boeke, A. Gabriel, and D. F. Voytas, "Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence," *Genome Research*, vol. 8, no. 5, pp. 464–478, 1998.

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

[51] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic Acids Research*, vol. 36, no. 1, pp. D154–D158, 2008.

[52] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl, "Identification of tissue-specific microRNAs from mouse," *Current Biology*, vol. 12, no. 9, pp. 735–739, 2002.

[53] L. P. Lim, N. C. Lau, E. G. Weinstein et al., "The microRNAs of *Caenorhabditis elegans*," *Genes & Development*, vol. 17, no. 8, pp. 991–1008, 2003.

[54] J. M. Cock, L. Sterck, P. Rouzé et al., "The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae," *Nature*, vol. 465, no. 7298, pp. 617–621, 2010.

[55] Z. Qu and D. L. Adelson, "Bovine ncRNAs are abundant, primarily intergenic, conserved and associated with regulatory genes," *PLoS ONE*, vol. 7, no. 8, Article ID e42638, 2012.

[56] S. Mi, T. Cai, Y. Hu et al., "Sorting of small RNAs into *Arabidopsis* argonaute complexes is directed by the 5′ terminal nucleotide," *Cell*, vol. 133, pp. 116–127, 2008.

[57] K. Hirota, T. Miyoshi, K. Kugou, C. S. Hoffman, T. Shibata, and K. Ohta, "Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs," *Nature*, vol. 456, no. 7218, pp. 130–134, 2008.

[58] D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel, "The impact of microRNAs on protein output," *Nature*, vol. 455, no. 7209, pp. 64–71, 2008.

[59] Z. Xie, K. D. Kasschau, and J. C. Carrington, "Negative feedback regulation of Dicer-Like1 in *Arabidopsis* by microRNA-guided mRNA degradation," *Current Biology*, vol. 13, no. 9, pp. 784–789, 2003.

[60] D. Rearick, A. Prakash, A. McSweeny, S. S. Shepard, L. Fedorova, and A. Fedorov, "Critical association of ncRNA with introns," *Nucleic Acids Research*, vol. 39, no. 6, pp. 2357–2366, 2011.

[61] H. Lodish, *Molecular Cell Biology*, Macmillan, London, UK, 2008.

[62] F. Mignone and G. Pesole, *mRNA Untranslated Regions (UTRs)*, eLS, 2011.

[63] Y.-K. Kim and V. N. Kim, "Processing of intronic microRNAs," *The EMBO Journal*, vol. 26, no. 3, pp. 775–783, 2007.

[64] S. Baskerville and D. P. Bartel, "Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes," *RNA*, vol. 11, no. 3, pp. 241–247, 2005.

[65] B. M. Wheeler, A. M. Heimberg, V. N. Moy et al., "The deep evolution of metazoan microRNAs," *Evolution & Development*, vol. 11, no. 1, pp. 50–68, 2009.

[66] C. U. Rahul and M. K. Rajesh, "Conserved miRNA detection in the ESTs of *Ganoderma lucidum*," *Research Journal of Biotechnology*, vol. 11, pp. 34–42, 2016.

[67] Y. Lin and Z. Lai, "Comparative analysis reveals dynamic changes in miRNAs and their targets and expression during somatic embryogenesis in Longan (*Dimocarpus longan* Lour.)," *PLoS ONE*, vol. 8, no. 4, Article ID e60337, 2013.

[68] J. Yang, N. Zhang, C. Ma, Y. Qu, H. Si, and D. Wang, "Prediction and verification of microRNAs related to proline accumulation

under drought stress in potato," *Computational Biology & Chemistry*, vol. 46, pp. 48–54, 2013.

[69] C. S. Sureshan and S. K. M. Habeeb, "Identification and conformational analysis of putative microRNAs in *Maruca vitrata* (Lepidoptera: pyralidae)," *Applied & Translational Genomics*, vol. 7, pp. 2–12, 2015.

[70] N. H. M. Yusuf, W. D. Ong, R. M. Redwan, M. A. Latip, and S. V. Kumar, "Discovery of precursor and mature microRNAs and their putative gene targets using high-throughput sequencing in pineapple (*Ananas comosus* var. *comosus*)," *Gene*, vol. 571, no. 1, pp. 71–80, 2015.

[71] R. W. Riley, "Phylogeny and comparative genome analysis of a Basidiomycete fungi," LBNL Paper LBNL-4662E-Poster, Lawrence Berkeley National Laboratory, Berkeley, Calif, USA, 2011, http://www.escholarship.org/uc/item/0066t273.

[72] 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, article 21, 2015.

[73] Y. A. Huang, Z. H. You, X. Chen, K. Chan, and X. Luo, "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC Bioinformatics*, vol. 17, no. 1, article 184, pp. 1–11, 2016.

[74] X. Luo, Z. Ming, Z. You, S. Li, Y. Xia, and H. Leung, "Improving network topology-based protein interactome mapping via collaborative filtering," *Knowledge-Based Systems*, vol. 90, pp. 23–32, 2015.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 7460740, 9 pages
http://dx.doi.org/10.1155/2016/7460740

Research Article

A Meta-Path-Based Prediction Method for Human miRNA-Target Association

Jiawei Luo,[1] Cong Huang,[2] and Pingjian Ding[2]

[1]*College of Information Science and Electronic Engineering & Collaboration and Innovation Center for Digital Chinese Medicine of 2011 Project of Colleges and Universities in Hunan Province, Hunan University, Changsha, Hunan 410082, China*
[2]*College of Information Science and Electronic Engineering, Hunan University, Changsha, Hunan 410082, China*

Correspondence should be addressed to Jiawei Luo; luojiawei@hnu.edu.cn

Received 30 June 2016; Revised 14 August 2016; Accepted 21 August 2016

Academic Editor: Xing Chen

MicroRNAs (miRNAs) are short noncoding RNAs that play important roles in regulating gene expressing, and the perturbed miRNAs are often associated with development and tumorigenesis as they have effects on their target mRNA. Predicting potential miRNA-target associations from multiple types of genomic data is a considerable problem in the bioinformatics research. However, most of the existing methods did not fully use the experimentally validated miRNA-mRNA interactions. Here, we developed RMLM and RMLMSe to predict the relationship between miRNAs and their targets. RMLM and RMLMSe are global approaches as they can reconstruct the missing associations for all the miRNA-target simultaneously and RMLMSe demonstrates that the integration of sequence information can improve the performance of RMLM. In RMLM, we use RM measure to evaluate different relatedness between miRNA and its target based on different meta-paths; logistic regression and MLE method are employed to estimate the weight of different meta-paths. In RMLMSe, sequence information is utilized to improve the performance of RMLM. Here, we carry on fivefold cross validation and pathway enrichment analysis to prove the performance of our methods. The fivefold experiments show that our methods have higher AUC scores compared with other methods and the integration of sequence information can improve the performance of miRNA-target association prediction.

1. Introduction

MicroRNAs (miRNAs) are important endogenous 21-22 nt RNAs that play important regulatory roles in gene expression. Several studies have shown that miRNAs participate in the regulation of amount cellular process, such as cell proliferation and differentiation [1], development [2], and disease [3, 4]. Considering the importance of miRNAs, it is critical to identify and decipher miRNA-target interactions at a genome level.

All the time, scientists and academics have made great efforts in uncovering the associations between miRNA and its targets by using biological experiments [5–8]. However, it is impossible to depict a complete picture of miRNA regulation mechanisms only relying on biological experiments due to the high expenses on time and cost [9]. Therefore, computational approaches must be designed to be a cost-effective choice to describe the complete mechanism of miRNA regulatory. Now, many computational approaches show great advantage in predicting putative miRNA targets [10–13].

Over the past decade, plenty of miRNA-mRNA pairs prediction approaches have been developed to identify miRNA targets by using sequence data, including TargetScanS/TargetScan [14, 15], miRanda [16], Pictar [17], DITAT-MicroT [18], and PITA [19]. The majority of these prediction algorithms were built on specific binding rules, including the degree of site conservation, thermodynamic stability, sequence complementarity, energy, target site context, secondary structure, and site accessibility. Because of the complex character of miRNA-target interactions, these sequence-based methods have relatively high false-positive rate [20]. Furthermore, those predictions methods were mostly only at static sequence level, leading to those exact interactions that are specific to certain conditions or diseases. More importantly, sequence-based methods do not support statistically significant predictions as the miRNA binding

sites are small, causing the results by different methods to be inconsistent.

To identify condition-specific interactions, many methods integrating expression profiles information into sequence-based predictions have been proposed to study miRNA-mRNA regulatory mechanism. These methods are based on the assumption that gene has negative correlations with the miRNA because of the downregulation effect that miRNAs have on their targets. These methods can be divided into four categories including simple correlation analysis [21, 22], simple/regularized regression models [23–25], Bayesian inference [19, 26], and causally inference between miRNAs and their targets [27]. Pearson correlation, one of the typical simple correlation methods, is commonly used in computing the strength of the association between a pair of miRNA and mRNA. However, Pearson correlation has high false-positive rate as the simplicity of it. Furthermore, Pearson correlation is mainly used in predicting linear associations. Lasso regression [24, 25], one of the regression models, is a high-dimensional method used to extract more reliable association as they usually optimize the network provided by sequence-based method and retain the relatively reliable edges. GenMir++ [19], the first and well-cited Bayesian inference method, calculates the existence probabilities of the relationship between a miRNA and its target based on a Bayesian model. However, this method needs prior information, such as sequence information. In general, methods in Bayesian category assume different priors [28] and are difficult in learning parameters. MCMG (joint analysis of multiple cancer for MiRNA-gene interactions), based on empirical Bayesian model [29], identifies miRNA-target associations that are either specific to a cancer type or common to several cancers by jointly analyzed across cancers. Muniategui et al. use do-calculus to estimate the causal effects the miRNA have on all the target mRNAs. The four categories methods can improve prediction performance as they integrate expression profiles information into sequence-based prediction methods [30]. But, most of the existing approaches cannot effectively use the valuable experimentally validated information [31–34]. Besides, the lack of miRNA expression profile may cause the unreliability of the predicted miRNA-target associations.

On the whole, the limitations of existing methods are summarized as follows. Firstly, sequenced-based prediction algorithms suffer from a high false-positive rate; second, the methods integrating expression profile data can only analyse one cancer every time; third, some methods cannot effectively utilize validated knowledge. To solve these problems, we propose two network-based approaches, RMLM and RMLMSe, to identify miRNA-target interactions based on meta-path. Meta-path is a good measuring method to compute the relatedness between the same or different types of objects in heterogeneous information network, as it contains a certain sequence of different link types [35]. Different meta-paths have different semantic meaning corresponding to different relationships between connected objects. In RMLM, we first utilize RM (a meta-path related measure proposed by Cao et al. [36]) to evaluate the existence probability of a link between miRNA and its targets. As different meta-path corresponds to different relation graphs, we may improve the final performance when integrating these different graphs by appropriate weights corresponding to different meta-paths. Thus, we then employ logistic regression and maximum-likelihood estimation (MLE) method to estimate the weight of different meta-path. Here, the issue of relationship prediction can be regarded as a two-class classification problem by using Bayesian analysis and logistic regression and then the MLE method can be employed to estimate the parameter vector. In RMLMSe, sequence information is integrated to improve the performance of the RMLM. Furthermore, as global approaches, RMLM and RMLMSe can remodel the missing relationship for all the diseases-associated miRNAs at the same time. Fivefold cross validations, pathway enrichment analysis about global network, and three important diseases network show that our proposed methods work well in predicting the relationship between miRNA and its target.

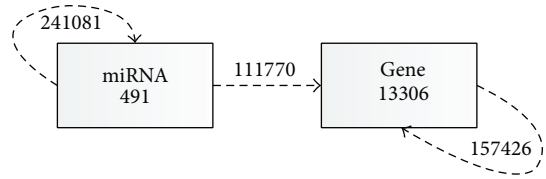


Figure 1: Network schema of the miRNA-target network. The network contains two types of objects, miRNA and its targets. Each box represents one type of nodes, and each dashed line represents one type of links. The numbers in the figure represent the numbers of nodes/links of different types.

2. Problem Definition

In this part, we describe the concepts of Heterogeneous Information Network and meta-path used in this paper.

2.1. Heterogeneous Information Network. A heterogeneous information network is an important type of information network with multiple types of nodes and multiple types of links [36–38]. It can be represented as $G = (V, E)$. V is the set of nodes, which involves n types of nodes: $V_1 = \{v_1^1, v_1^2, \ldots, v_1^x\}, \ldots, V_n = \{v_n^1, v_n^2, \ldots, v_n^y\}$, where v_i^j is jth node of type i. $E \subseteq V \times V$ is the set of links between the nodes in V, which involves m types of links.

Each type of links between source node of type i and target node of type j corresponds to a binary relation R_{ij}. More specifically, $R_{ij}^{st} = 1$ if v_i^s (sth nodes of type i) and v_j^t (tth nodes of type j) are connected by a link of type R^{ij}. For example, in Figure 1, the relation between miRNA and gene is "regulate." Particularly, R_{ij}^{st} equals 1 if sth miRNA regulates tth gene.

Moreover, a weighted matrix $W_{ij} = |V_i| \times |V_j|$ can be used to describe the relation R_{ij}, where $W_{ij}^{st} \in [0, 1]$ is the existence probability of link between nodes v_i^s and v_j^t. Particularly, $W_{ij}^{st} = 1$, if there exists an edge between v_i^s and v_j^t. Otherwise, W_{ij}^{st} is set as 0 in initialization for the unknown links.

2.2. Meta-Path. In heterogeneous information network, meta-path is defined on network schema. A meta-path P is

described in the form $A_1 \rightarrow A_2 \rightarrow \cdots A_{n-1} \rightarrow A_n$, where A_i is ith type of object and a relation must exist from A_{i-1} to A_i, $i = 2, 3, \ldots, n$. Similarly, we define the inverse path of P as P^{-1}, denoted as $A_n \rightarrow A_{n-1} \rightarrow \cdots A_2 \rightarrow A_1$. Specifically, relation $A_{i-1} \rightarrow A_i$ is the inverse relation of $A_i \rightarrow A_{i-1}$. For example, in Figure 1, a meta-path "gene → miRNA → gene" is a composite sequence between genes. The relation from miRNA to gene is "regulate" and the relation from gene to miRNA is "regulate^{-1}"; "regulate^{-1}" is the inverse relation of "regulate." Meta-path can connect object of the same or different types; thus, they can show knowledge between homologous objects or heterologous objects. For example, in Figure 1, for gene i and gene j, they can connect through another gene k, gene $i \rightarrow$ gene $k \rightarrow$ gene j; this means gene i and gene j have relation with gene k simultaneously and there may exist relation between gene i and gene j by information transfer. However, gene i and gene j can also connect by miRNA k, gene $i \rightarrow$ miRNA $k \rightarrow$ gene j; this means gene i and gene j are regulated by a common miRNA k and there may exist relation between gene i and gene j by information transfer. Different meta-paths of different relations correspond to different relation graphs with different semantics. For example, in Figure 1, the meta-path "gene → gene" denotes that two genes are connected by "PPI" links, while the meta-path "gene → miRNA → gene" corresponds to the semantic that two genes are regulated by a common miRNA. Thus, similarity between the same or different type of nodes can be described by different meta-paths with different semantics.

In this paper, the meta-path from source node of type i to target node of type j is described as P_{ij}. Particularly, P_{ii} is the meta-path between nodes of the same type i; P_{iis} is sth meta-path of P_{ii}. P_{jj} and P_{jjt} are the same to P_{ii} and P_{iis}. P_{ijst} is a meta-path by connecting P_{iis}, R_{ij}, and P_{jjt} in sequence; it can be written as a certain sequence of relations: $R_{k_0k_1}, R_{k_1k_2}, \ldots, R_{k_{n-1}k_n}$; here $k_0 = i$, $k_n = j$ and the length of P_{ijst} is n.

3. Method

RMLM and RMLMSe consist of three steps. In the first step, we utilize MISIM (proposed by Wang et al. in [39]) to calculate the miRNA functional similarity matrix and then construct the heterogeneous network. Next, we calculate the relatedness between any miRNA and its targets and extract the feature vector of these interactions. In RMLM, the feature vector only contains different relatedness of different meta-path between miRNA and its targets. However, in RMLMSe, the feature vector not only contains different relatedness from different meta-path, but also contains feature extracted from sequence information. Finally, logistic regression and MLE method are employed to compute the different weights of different meta-paths. Sections 3.1–3.4 are the detailed introduction of RMLM. Section 3.5 is about RMLMSe.

3.1. Construction of the Heterogeneous Network

3.1.1. miRNA-miRNA Similarity Estimation. In [39], Wang et al. compute miRNA-miRNA functional similarity score based on the assumption that miRNAs with similar functions tend to be related to similar disease. To get the miRNA-miRNA similarity matrix, there contains three procedures. We take miRNA i and miRNA j as an example. First, we identify diseases that related to these two miRNAs, encoded as D_i and D_j. We can obtain the relationship between miRNAs and diseases from The Human MicroRNA Disease Database (HMDD dataset). Then, we can calculate similarity of any pair of diseases using a hierarchical structure. The semantic similarity of disease is calculated based on directed acyclic graph obtained from the US National Library of Medicine in 2015 (MeSH, https://www.nlm.nih.gov/mesh/). Finally, we utilize the similarity score between D_i and D_j to compute the relatedness score between miRNA i and miRNA j. In this paper, we use SM (a 491 × 491 matrix) to represent the miRNA-miRNA similarity matrix; SM(i, j) is the functional similarity score between miRNA i and miRNA j.

3.1.2. Construction of the Heterogeneous Network. We construct the heterogeneous network by connecting the miRNA interaction network and PPI utilizing the bipartite graph of the miRNA-target association network. The schema of the heterogeneous network used in this paper is illustrated in Figure 1. The network contains two types of objects, miRNA and its targets. A meta-path P is defined at the object type level and is denoted in the form of $A_1 \rightarrow A_2 \rightarrow \cdots A_{n-1} \rightarrow A_n$, where A_i represent the object of type.

3.2. Relatedness Measure.

The RM measure [36] is a path-constrained measure and it can calculate the relatedness of heterogeneous objects with the same or different types in a uniform framework. It has been proven that RM has some good properties, such as symmetric and self-maximum, and has shown its potential to mining valuable information in heterogeneous network. Therefore, here we use RM measure to calculate the relatedness between miRNA and its targets. RM measure is based on the Linkage Homophily Principle defined as follows.

Linkage Homophily Principle. Two nodes are more likely to be directly linked if most of their respective similar nodes are linked.

In general, the computing of nodes similarity is based on their neighbors. However, in heterogeneous networks, the same type similar nodes can be linked by heterogeneous nodes through composite paths. For example, two similar genes can be connected by a common miRNA, "gene → miRNA → gene." Thus, we can utilize meta-path to extract the generalized neighbor and define the similarity. Here, we first extract the meta-path that connects the source node and target node. We take source node v_i^p and meta-path P_{iis} as an example. The neighbors of node v_i^p based on P_{iis} are the nodes of type i that linked to v_i^p by P_{iis}, denoted as N_i^p. Similarly, we can get the generalized neighbors of target node v_j^q and meta-path P_{jjt}, denoted as N_j^q. Then, we can use the connectivity between N_i^p and N_j^q to calculate the link's existence probability between nodes v_i^p and v_j^q.

Suppose RMP_{iis} is the similarity matrix of ith type node along the meta-path P_{iis}. Similarity, RMP_{jjt} represents the similarity matrix of jth type node along the meta-path P_{jjt}. In general, similarity can be calculated by the path counts. Expected path number is the number where all of the links may exist from node of type i to node of type j. Let meta-path $P_{ijst} = \{R_{k_0k_1}, R_{k_1k_2}, \ldots, R_{k_{n-1}k_n}\}$, $k_0 = i$, and $k_n = j$; then the expected path number RMP_{ijst} is computed as follows:

$$\mathrm{RMP}_{ijst} = \prod_{p=1}^{n} w_{k_{p-1}k_p} = \mathrm{RMP}_{iis} \times W_{ij} \times \mathrm{RMP}_{jjt}. \quad (1)$$

Here, P_{ijst} is a meta-path composed of P_{iis}, R_{ij}, and P_{jjt}; RMP_{ijst} is a matrix whose size is $|V_i| \times |V_j|$. The computation of RMP_{iis} (or RMP_{jjt}) is similar to the computation of RMP_{ijst}.

Now the relatedness between nodes of type i and nodes of type j along the meta-path P_{ijst} can be formulated as follows:

$$\begin{aligned} \mathrm{RM}_{ijst} &= \frac{\mathrm{RMP}_{ijst}}{\mathrm{RMP}_{iis} \times \mathbf{1} \times \mathrm{RMP}_{jjt}} \\ &= \frac{\mathrm{RMP}_{iis} \times W_{ij} \times \mathrm{RMP}_{jjt}}{\mathrm{RMP}_{iis} \times \mathbf{1} \times \mathrm{RMP}_{jjt}}. \end{aligned} \quad (2)$$

Here $\mathbf{1}$ is a matrix in which all the elements are 1 and the size of is $|V_i| \times |V_j|$. Similarly, RM_{ijst} is also a $|V_i| \times |V_j|$ matrix and RM_{ijst}^{pq} is the relatedness measured between v_i^p and v_j^q following P_{ijst}.

3.3. Construction of the Feature Vector. We can get the relatedness between miRNAs and their targets as described in Section 3.2. Now we get the feature vector as follows:

(1) Extract meta-path P_{ii} of ith type node and P_{jj} of jth type node.

(2) Compute the similarity based on any pair of meta-paths P^{ii} and P^{jj} and then get the feature vector.

In RMLM, the feature vector between miRNA i and gene j is defined as

$$\phi_{ij} = (f_1, f_2, \ldots, f_n), \quad (3)$$

where f_1 to f_n represent the different similarities of different meta-paths with different semantic meaning.

3.4. Parameter Estimation. As different meta-path corresponds to different relation graphs, the final result may be improved by combining these different graphs through different weights. Here, logistic regression and maximum-likelihood estimation (MLE) method can be employed to estimate the weight.

In this paper, we regard the issue of relationship prediction as a two-class classification problem by using Bayesian analysis and logistic regression. Based on logistic regression and under general assumption [31, 32], the posterior probability of a specific relation can be formulated as follows:

$$p(x_i = 1 \mid \varphi_i, \omega) = \frac{\exp(\omega^T \varphi_i)}{\exp(\omega^T \varphi_i) + 1}, \quad (4)$$

$$p(x_i = 0 \mid \varphi_i, \omega) = \frac{1}{\exp(\omega^T \varphi_i) + 1}. \quad (5)$$

Here ω is a weight vector served as parameters and φ_i is the feature vector of the link x_i. Then, MLE method can be employed to estimate the parameter vector ω. The likelihood function can be written as

$$L(\omega; x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i \mid \varphi_i, \omega). \quad (6)$$

Here x_i is the link to calculate and N is the number of links, φ_i is the feature vector that is calculated according to RM, and ω is the weight vector of the feature according to different meta-path. The log likelihood of (6) is

$$\begin{aligned} &\ln L(\omega; x_1, x_2, \ldots, x_N) \\ &\quad = \sum_{i=1}^{N} \left[x_i \omega^T \varphi_i - \ln\left(1 + \exp\left(\omega^T \varphi_i\right)\right) \right]. \end{aligned} \quad (7)$$

The log likelihood (7) is a convex function [40]. Hence, we can find a unique global optimal solution by solving a convex optimization problem.

3.5. Final Score. The logistic regression based algorithm returns a set of posterior probabilities. One can directly use those probabilities to make decision. However, the posterior probabilities do not always work well because it is difficult to set a threshold for a relation between miRNA and its target. Here, we utilize a percentage value as the final score to evaluate the strength of the relation between a miRNA and its target. The final score is calculated as follows:

$$q_i = \frac{\left|\left\{ j \mid p_i \geq p_j \right\}\right|}{n}, \quad i = 1, 2, \ldots, n. \quad (8)$$

Here $\{p_1, p_2, \ldots, p_n\}$ is the posterior probabilities of any association, and q_i is the top percentage value of p_i among all those posterior probabilities. The larger the final score is, the more likely the association exists.

3.6. Integration of Sequence Information. In RMLMSe, we integrate sequence information to improve the performance of the RMLM. Here, we use sequence information from database TargetScan, miRanda, and PITA. As they have a relatively high false-positive rate, we only download conserved targets information and select the data whose Pct > 0.9 from TargetScan, mirSVR > 0.6 from miRanda, and data in PITATOP to improve the reliability of the regulation relationships. Sequence information from these databases acts as new features in feature vector used in RMLMSe.

Taking interaction between miRNA i and gene j as an example, its feature vector can be written as

$$\phi_{ij} = (f_1, f_2, \ldots, f_n, f_m, f_{m+1}, f_{m+2}). \quad (9)$$

Here f_1 to f_n represent the different feature of different meta-paths and f_m, f_{m+1}, and f_{m+2} represent the feature of sequence information from TargetScan, miRanda, and PITA, respectively.

3.7. Algorithm. The process description of RMLM and RMLMSe is given as follows.

Input. The disease set d_i of each miRNA i from HMDD and DAG g_j of each disease j from MeSH, the protein interaction matrix SP, and the miRNA-protein matrix MP.

Output. The vector of final score for each unknown interaction between miRNA and its targets.

(1) Calculate the miRNA-miRNA functional similarity matrix SM as described in Section 3.1.1.

(2) Extract meta-path P_{ii} of ith type node and P_{jj} of jth type node. We set the max length of meta-path between the same type node as (3).

(3) Concatenate P_{iis} (sth meta-path of P_{ii}), R_{ij}, and P_{jjt} (tth meta-path of P_{jj}) in sequence to compose a meta-path P_{ijst} going from the source nodes of type i to target nodes of type j. Then, the relatedness between miRNA and its target based on meta-path P_{ijst} is calculated according to (2).

(4) Calculate the different similarity of different meta-path and get the feature vector of each interaction. The feature vectors used in RMLM and RMLMSe are described in Sections 3.3 and 3.5.

(5) Estimate parameters ω by maximizing the log likelihood $\ln L(\omega; x_1, x_2, \ldots, x_N)$ in (7) based on x_i and φ_i, x_i is the link to be calculated, and N is the number of links.

(6) Calculate the probability for each unknown interaction according to (4) by using ω and feature vector.

(7) Calculate the final score according to (8).

4. Results

4.1. Datasets

The Human MicroRNA Disease Database. HMDD [41] provides a comprehensive resource of experimentally verified miRNA-disease associations. We can get the information through a website at http://www.cuilab.cn/hmdd. The database (in June 2014) contains 5100 associations between 491 miRNAs and 326 diseases. In this paper, we first analyse the global network. Then, we analyse another three diseases, Ovarian Neoplasms (OV), Lung Neoplasms (Lung), and Breast Neoplasms (Breast). The miRNAs associated with OV, Lung, and Breast are 114, 132, and 202, respectively.

The Protein-Protein Interaction Database. The PPI network was constructed by combining DNA-protein data from TRANSFAC [42] and protein interaction data obtained from Bossi and Lehner [43], respectively. The database contains 13306 proteins and 157426 interactions between proteins.

Experimentally Validated miRNA-mRNA Interaction Databases. The posttranscriptional regulatory knowledge is obtained from miRNA-target database miRTarbase *v6.1*. When mapping onto our miRNA-target matrix, it retains 111770 interactions. We can get the information through a website at (http://mirtarbase.mbc.nctu.edu.tw/).

Predicted miRNA-mRNA Interaction Database. We also utilize sequence information in database TargetScan *v7.0*, miRanda released at 2010, and PITA *v6*. These databases are available online at http://www.targetscan.org/, http://www.microrna.org/, and http://genie.weizmann.ac.il/pubs/mir07/, respectively.

4.2. Comparisons with Other Methods. To compare the performance of RMLM and RMLMSe, we applied RLSMDA [44] and RM [36] to the same testing data. RLSMDA was introduced to predict disease-miRNA association. We encoded RLSMDA in MATLAB according to the derivation process of the authors. Here, we set ω used in RLSMDA as 0.5. RM was implemented in MATLAB with source code available from authors personal homepage. RM is the measurement used to calculate the similarity of objects in heterogeneous networks. Here, the sum of the different similarities corresponding to different meta-paths is utilized to predict the miRNA-gene associations. All experiments are carried on a Windows 7 professional computer (Inter(R) Xeon(R) CPU, 2.93 GHz, 56 G RAM, 64-bit OS). The performance of each method is evaluated by fivefold cross validation. First, all known miRNA-target associations were split into five sets of the same size randomly: one set was set aside as the test set and the other four sets were used as train sets. The experiment was repeated five times so that each set was hidden once and each hidden miRNA-target pair obtained a predict relevance score. The ROC (receiver operating characteristic) curve was calculated according to the various TPR (true-positive rate) and the various FPR (false-positive rate) through a varying threshold. The area under the ROC curve (AUC) is employed to show the overall performance of methods. We can see from Figure 2 that RMLM and RMLMSe always work better than RLSMDA and RM. There is only slight improvement when sequence information is employed, where the AUC score increases from 0.8919 to 0.9033. This may have two reasons. First, the performance of the RMLM already achieves a very high AUC score and there is only a little room for it to be further improved by using additional prior information. Second, the amount of the sequence information mapped onto the miRNA-target matrix is little; for example, when TargatScan, miRanda, and PITA mapped onto the miRNA-target matrix, they leave 16,7403, 10,4631, and 13,7229 interactions, about 1.6~2.6% of the entire size of the miRNA-target matrix MP (a 491 × 13306 matrix). Although the improvement of the sequence

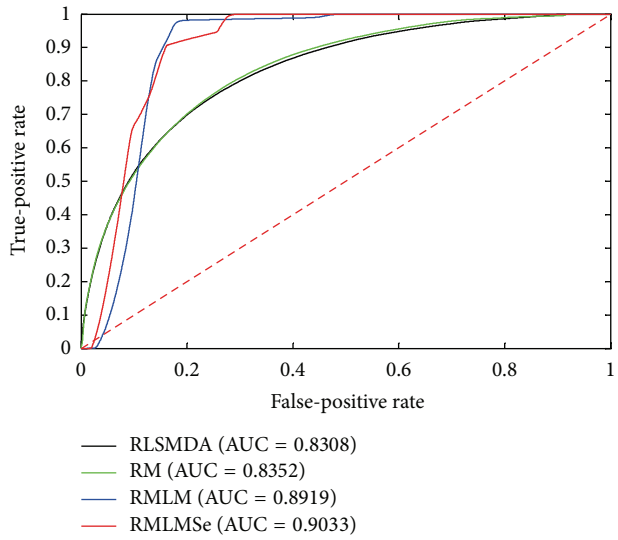


Figure 2: The ROC curve of the global network.

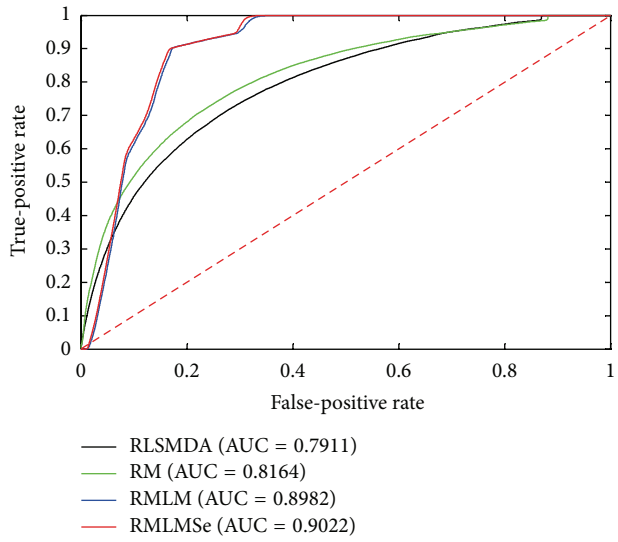


Figure 3: The ROC curve of the OV network.

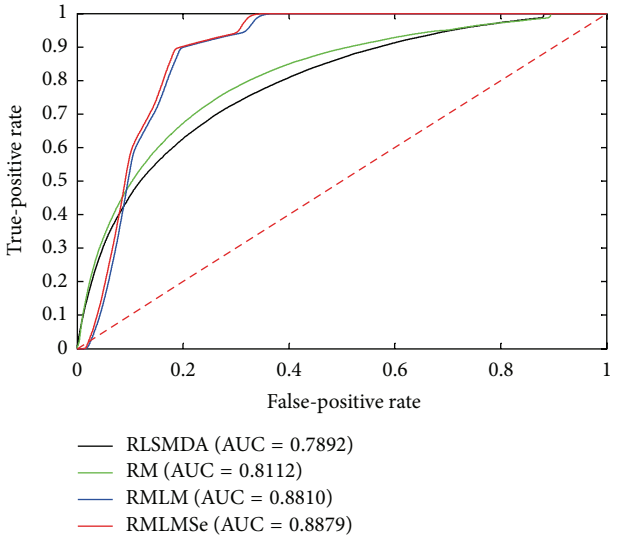


Figure 4: The ROC curve of the Lung network.

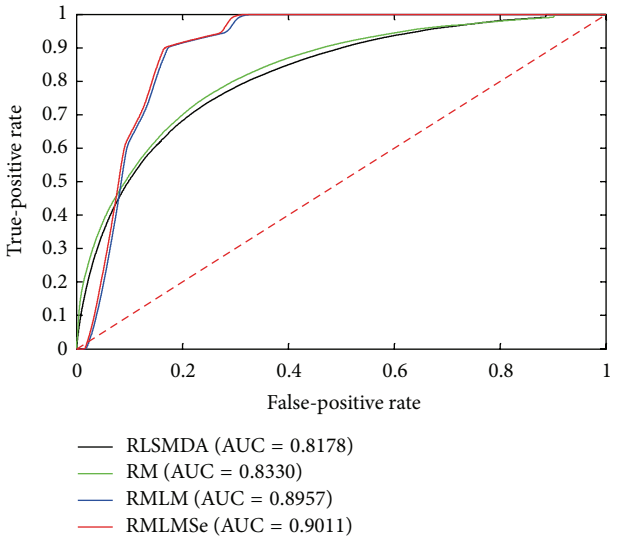


Figure 5: The ROC curve of the Breast network.

information is not significant, the increased AUC score still indicates that additional knowledge is helpful for improving the prediction performance as any prior knowledge, such as sequence information, Go Ontology annotations, gene copy numbers, and gene methylation, related to miRNA-target associations can be employed to predict associations. Figures 3, 4, and 5 are the result when we execute the methods on OV, Lung, and Breast database, respectively. The results are similar to Figure 2. RMLM and RMLMSe always work better than RLSMDA and RM, and RMLMSe only have a slight improvement than RMLM.

4.3. The Number of Links Predicted by Our Methods. Here, we present the number of interactions predicted based on different thresholds in RMLM and RMLMSe. As shown in Table 1, the numbers of interactions predicted in RMLM are higher than in RMLMSe among all of the threshold. This can further indicate the performance improvement in RMLMSe. In future, we can utilize the associations predicted by our method to construct miRNA-target regulatory network and extract regulatory modules and hub nodes.

4.4. Functional Validation of mRNAs. When we get the result of the global dataset, we compute every mRNA score and extract the top 250 mRNAs to carry on the pathway enrichment analysis with the focus on KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways (adjusted p value < 0.05). In this paper, p value calculated by hypergeometric test is a statistical value that represents the significant enrichment of pathways. The smaller the p value is, the more significant the pathway enrichment is. As shown in Table 2, many of the KEGG pathways are highly related to many cancers and

Table 1: The number of links predicted by our methods based on different thresholds.

Database	Methods	Validated	Th ≥ 0.9	Th ≥ 0.8	Th ≥ 0.7	Th ≥ 0.6	Th ≥ 0.5
Global	RMLM	11,1770	17,2912	20,4894	23,4327	26,5883	79,8049
	RMLMSe	11,1770	17,6625	21,0909	24,2946	28,1782	80,7688
OV	RMLM	4,2730	5,3683	5,9580	6,4676	6,9759	23,3784
	RMLMSe	4,2730	5,3891	5,9954	6,5526	7,1565	23,4562
Lung	RMLM	4,7764	5,8511	6,4339	6,9397	7,4816	24,5323
	RMLMSe	4,7764	5,8870	6,4881	7,0437	7,9293	24,6261
Breast	RMLM	6,4403	8,6555	9,8883	10,9659	12,0730	36,4375
	RMLMSe	6,4403	8,6690	9,9540	11,1719	12,6556	36,6573

The "validated" column is the number of links validated in database miRTarbase *v6.1* and "Th" represents the threshold.

Table 2: In RMLMSe, the enrichment KEGG pathways of global dataset.

	Enrichment KEGG pathways	p value
1	p53 signaling pathway	$4.27E-10$
2	Chronic myeloid leukemia	$8.80E-10$
3	Bladder cancer	$3.24E-09$
4	Glioma	$6.03E-09$
5	Melanoma	$1.35E-08$
6	Pathways in cancer	$2.34E-08$
7	Prostate cancer	$1.01E-07$
8	Cell cycle	$1.61E-07$
9	Small cell lung cancer	$9.71E-07$
10	Pancreatic cancer	$3.26E-06$

The p values have been obtained through hypergeometric test.

Table 3: In RMLMSe, the enrichment KEGG pathways of lung dataset.

	Enrichment KEGG pathways	p value
1	p53 signaling pathway	$5.15E-10$
2	Pathways in cancer	$3.11E-08$
3	Small cell lung cancer	$1.12E-06$
4	Non-small cell lung cancer	$1.04E-05$
5	Focal adhesion	$1.53E-05$
6	Neurotrophin signaling pathway	$1.81E-04$
7	Adherens junction	$6.05E-04$
8	ErbB signaling pathway	$1.34E-03$
9	Pathogenic *Escherichia coli* infection	$1.89E-03$
10	MAPK signaling pathway	$1.31E-02$

The p values have been obtained through hypergeometric test.

respective biological process, for instance, glioma, prostate cancer, and colorectal cancer. Furthermore, pathways in cancer are closely related to many cancers and P53 signaling pathways is proved to be related to the processes of cell division and DNA replication [45]. The result of Lung KEGG pathways is shown in Table 3. The pathway focal adhesion [46], adherens junction [47], and ErbB signaling pathway [48] are proved to be related to Lung.

5. Discussion and Conclusion

The rapid increase of various biological data provides challenges and opportunities for us to complete the global miRNA regulatory mechanism. In recent years, academics have made great efforts to predict miRNA targets. However, each method has its pros and cons, and the performance of a method varies on different datasets. Thus, how to get precise results is a long-time challenge for miRNA-target association prediction.

In this paper, two novel methods, RMLM and RMLMSe, were developed. In RMLM, we first construct miRNA-miRNA similarity matrix. Second, we use RM to evaluate the different relatedness between miRNAs and its target based on different meta-path and extract the feature vectors of links; different meta-path corresponds to different relation graphs; we can improve the performance by combining these different graphs through different weights of corresponding meta-paths. Third, logistic regression and MLE method were employed to estimate the weight. Here, the issue of relationship prediction is regarded as a two-class classification problem by using Bayesian analysis and logistic regression and then MLE method can be employed to estimate the parameter vector. Then, we estimate the posterior probabilities between miRNAs and its targets based on the feature vectors of links and the corresponding parameter vectors. Finally, the final scores are obtained by using the percentage values of individual posterior probabilities. In RMLMSe, we utilize more information such as sequence information from TargetSacn, miRanda, and PITA to improve the performance of the RMLM. The results showed that there are slight improvement when sequence information is integrated.

Compared with other methods, RMLM and RMLMSe proposed by us have higher AUC scores. Besides, we conduct pathway enrichment analysis and found many relevant pathways. These results indicate that our two methods were reasonable and credible.

The comparison results of RMLM and RMLMSe indicate that our methods have the capability to integrate more biological data, such as sequence data and gene copy number. Thus, with the rapid growth of the gene regulatory knowledge, our method can integrate more prior information to improve the prediction performance.

In addition, disease target inference [49, 50], disease-miRNA prioritization [51–54], and lncRNA-disease association prediction [55] are also the immediate areas of research

focus to further study therapeutic strategy. Due to the scalability of the proposed methods, RMLM and RMLMSe could be applied to the different constructed heterogeneous networks to infer disease target, miRNA-disease association, and lncRNA-disease association, respectively. Moreover, the performance of our methods should be further evaluated after extending.

Of course, RMLM and RMLMSe also have some limitations that need to be improved in the future. Firstly, our methods utilize the network topology and known miRNA-gene associations to calculate the relatedness between miRNA and its target. It may cause bias to miRNA-gene pair which has more neighbor nodes. Furthermore, although the better performance is obtained by our methods on the whole, the predictive results should be further improved, especially for the small output. In the future, the prediction performance will be further improved by integrating more reliable biological data and obtaining more known miRNA-gene associations.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The research is supported by National Natural Science Foundation of China (Grant no. 61572180).

References

[1] E. Wienholds and R. H. A. Plasterk, "MicroRNA function in animal development," *FEBS Letters*, vol. 579, no. 26, pp. 5911–5922, 2005.

[2] I. Alvarez-Garcia and E. A. Miska, "MicroRNA functions in animal development and human disease," *Development*, vol. 132, no. 21, pp. 4653–4662, 2005.

[3] W. C. S. Cho, "OncomiRs: the discovery and progress of microRNAs in cancers," *Molecular Cancer*, vol. 6, no. 1, article 60, pp. 1–7, 2007.

[4] F. Felicetti, M. C. Errico, L. Bottero et al., "The promyelocytic leukemia zinc finger-microRNA-221/-222 pathway controls melanoma progression through multiple oncogenic mechanisms," *Cancer Research*, vol. 68, no. 8, pp. 2745–2754, 2008.

[5] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel, "Mammalian microRNAs predominantly act to decrease target mRNA levels," *Nature*, vol. 466, no. 7308, pp. 835–840, 2010.

[6] N. Mercatelli, V. Coppola, D. Bonci et al., "The inhibition of the highly expressed mir-221 and mir-222 impairs the growth of prostate carcinoma xenografts in mice," *PLoS ONE*, vol. 3, no. 12, Article ID e4029, 2008.

[7] G. T. Huang, C. Athanassiou, and P. V. Benos, "MirConnX: condition-specific mRNA-microRNA network integrator," *Nucleic Acids Research*, vol. 39, no. 2, pp. W416–W423, 2011.

[8] B. Liu, J. Li, A. Tsykin, L. Liu, A. B. Gaur, and G. J. Goodall, "Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy," *BMC Bioinformatics*, vol. 10, article 408, 2009.

[9] S. R. A. Fisher, R. A. Fisher, S. Genetiker et al., *The Design of Experiments*, 1960.

[10] S.-D. Hsu, Y.-T. Tseng, S. Shrestha et al., "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions," *Nucleic Acids Research*, vol. 42, no. 1, pp. D78–D85, 2014.

[11] R. F. Service, "Biology's dry future," *Science*, vol. 342, no. 6155, pp. 186–189, 2013.

[12] J. C. Huang, T. Babak, T. W. Corson et al., "Using expression profiling data to identify human microRNA targets," *Nature Methods*, vol. 4, no. 12, pp. 1045–1049, 2007.

[13] T. De Bie, L.-C. Tranchevent, L. M. M. van Oeffelen, and Y. Moreau, "Kernel-based data fusion for gene prioritization," *Bioinformatics*, vol. 23, no. 13, pp. i125–i132, 2007.

[14] B. P. Lewis, I.-H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microRNA targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[15] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[16] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "MicroRNA targets in *Drosophila*," *Genome Biology*, vol. 5, no. 1, article R1, 2003.

[17] A. Krek, D. Grün, M. N. Poy et al., "Combinatorial microRNA target predictions," *Nature Genetics*, vol. 37, no. 5, pp. 495–500, 2005.

[18] M. Reczko, M. Maragkakis, P. Alexiou, I. Grosse, and A. G. Hatzigeorgiou, "Functional microRNA targets in protein coding sequences," *Bioinformatics*, vol. 28, no. 6, pp. 771–776, 2012.

[19] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microRNA target recognition," *Nature Genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[20] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou, "A guide through present computational approaches for the identification of mammalian microRNA targets," *Nature Methods*, vol. 3, no. 11, pp. 881–886, 2006.

[21] H. Liu, A. R. Brannon, A. R. Reddy et al., "Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell renal cell carcinoma," *BMC Systems Biology*, vol. 4, article 51, 2010.

[22] I. Van der Auwera, R. Limame, P. Van Dam, P. B. Vermeulen, L. Y. Dirix, and S. J. Van Laere, "Integrated miRNA and mRNA expression profiling of the inflammatory breast cancer subtype," *British Journal of Cancer*, vol. 103, no. 4, pp. 532–541, 2010.

[23] S. Kim, M. Choi, and K.-H. Cho, "Identifying the target mRNAs of microRNAs in colorectal cancer," *Computational Biology and Chemistry*, vol. 33, no. 1, pp. 94–99, 2009.

[24] Y. Lu, Y. Zhou, W. Qu, M. Deng, and C. Zhang, "A Lasso regression model for the construction of microRNA-target regulatory networks," *Bioinformatics*, vol. 27, no. 17, pp. 2406–2413, 2011.

[25] A. Muniategui, R. Nogales-Cadenas, M. Vázquez et al., "Quantification of miRNA-mRNA interactions," *PLoS ONE*, vol. 7, no. 2, Article ID e30766, 2012.

[26] N. Su, Y. Wang, M. Qian, and M. Deng, "Predicting MicroRNA targets by integrating sequence and expression data in cancer," in *Proceedings of the 5th IEEE International Conference on Systems Biology (ISB '11)*, pp. 219–224, Zhuhai, China, September 2011.

[27] T. D. Le, L. Liu, A. Tsykin et al., "Inferring microRNA-mRNA causal regulatory relationships from expression data," *Bioinformatics*, vol. 29, no. 6, pp. 765–771, 2013.

[28] F. C. Stingo, Y. A. Chen, M. Vannucci, M. Barrier, and P. E. Mirkes, "A Bayesian graphical modeling approach to microRNA regulatory network inference," *The Annals of Applied Statistics*, vol. 4, no. 4, pp. 2024–2048, 2010.

[29] X. Chen, F. J. Slack, and H. Zhao, "Joint analysis of expression profiles from multiple cancers improves the identification of microRNA-gene interactions," *Bioinformatics*, vol. 29, no. 17, pp. 2137–2145, 2013.

[30] A. Muniategui, J. Pey, F. J. Planes, and A. Rubio, "Joint analysis of miRNA andmRNA expression data," *Briefings in Bioinformatics*, vol. 14, no. 3, Article ID bbs028, pp. 263–278, 2013.

[31] F. Tai and W. Pan, "Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms," *Bioinformatics*, vol. 23, no. 14, pp. 1775–1782, 2007.

[32] Z. Tian, T. Hwang, and R. Kuang, "A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge," *Bioinformatics*, vol. 25, no. 21, pp. 2831–2838, 2009.

[33] Z. Zhao, J. Wang, H. Liu, J. Ye, and Y. Chang, "Identifying biologically relevant genes via multiple heterogeneous data sources," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*, pp. 839–847, August 2008.

[34] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D152–D157, 2010.

[35] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Pathselclus: integrating meta-path selection with userguided object clustering in heterogeneous information networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 7, no. 3, p. 11, 2013.

[36] B. Cao, X. Kong, and P. S. Yu, "Collective prediction of multiple types of links in heterogeneous information networks," in *Proceedings of the 14th IEEE International Conference on Data Mining (ICDM '14)*, pp. 50–59, Shenzhen, China, December 2014.

[37] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*, pp. 797–806, Paris, France, July 2009.

[38] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[39] 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, Article ID btq241, pp. 1644–1650, 2010.

[40] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[41] 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, no. 1, pp. D1070–D1074, 2014.

[42] V. Matys, O. V. Kel-Margoulis, E. Fricke et al., "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D108–D110, 2006.

[43] A. Bossi and B. Lehner, "Tissue specificity and the human protein interaction network," *Molecular Systems Biology*, vol. 5, article 260, 2009.

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

[45] S. L. Harris and A. J. Levine, "The p53 pathway: positive and negative feedback loops," *Oncogene*, vol. 24, no. 17, pp. 2899–2908, 2005.

[46] G. W. McLean, N. O. Carragher, E. Avizienyte, J. Evans, V. G. Brunton, and M. C. Frame, "The role of focal-adhesion kinase in cancer—a new therapeutic opportunity," *Nature Reviews Cancer*, vol. 5, no. 7, pp. 505–515, 2005.

[47] Q.-Y. Chen, D.-M. Jiao, L.-F. Wang et al., "Curcumin inhibits proliferation-migration of NSCLC by steering crosstalk between a Wnt signaling pathway and an adherens junction via EGR-1," *Molecular BioSystems*, vol. 11, no. 3, pp. 859–868, 2015.

[48] T. Yu, J. Li, M. Yan et al., "MicroRNA-193a-3p and -5p suppress the metastasis of human non-small-cell lung cancer by downregulating the ERBB4/PIK3R3/mTOR/S6K2 signaling pathway," *Oncogene*, vol. 34, no. 4, pp. 413–423, 2015.

[49] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, "Prediction and validation of gene-disease associations using methods inspired by social network analyses," *PLoS ONE*, vol. 8, no. 5, Article ID e58977, 2013.

[50] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim Scores," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.

[51] 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.

[52] P. Xuan, K. Han, M. Guo et al., "Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors," *PLoS ONE*, vol. 8, no. 8, Article ID e70204, 2013.

[53] 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.

[54] 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.

[55] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, 2016.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 9565689, 11 pages
http://dx.doi.org/10.1155/2016/9565689

Research Article

BP Neural Network Could Help Improve Pre-miRNA Identification in Various Species

Limin Jiang,[1,2] Jingjun Zhang,[2] Ping Xuan,[3] and Quan Zou[1,4]

[1] *School of Computer Science and Technology, Tianjin University, Tianjin 300350, China*
[2] *School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China*
[3] *School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China*
[4] *State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin 300074, China*

Correspondence should be addressed to Ping Xuan; 2004058@hlju.edu.cn and Quan Zou; zouquan@tju.edu.cn

Received 17 May 2016; Revised 5 July 2016; Accepted 17 July 2016

Academic Editor: Xing Chen

MicroRNAs (miRNAs) are a set of short (21–24 nt) noncoding RNAs that play significant regulatory roles in cells. In the past few years, research on miRNA-related problems has become a hot field of bioinformatics because of miRNAs' essential biological function. miRNA-related bioinformatics analysis is beneficial in several aspects, including the functions of miRNAs and other genes, the regulatory network between miRNAs and their target mRNAs, and even biological evolution. Distinguishing miRNA precursors from other hairpin-like sequences is important and is an essential procedure in detecting novel microRNAs. In this study, we employed backpropagation (BP) neural network together with 98-dimensional novel features for microRNA precursor identification. Results show that the precision and recall of our method are 95.53% and 96.67%, respectively. Results further demonstrate that the total prediction accuracy of our method is nearly 13.17% greater than the state-of-the-art microRNA precursor prediction software tools.

1. Introduction

MicroRNAs are some of the most important noncoding RNA genes with rather short length. They regulate the expression of whole organism genes at the posttranscriptional level [1]. miRNA is widely involved in the metabolic activity of the body as well as in many important life processes, including cell proliferation and apoptosis, cell differentiation, growth and development of plants and animals, and organ formation [2–4]. Recently, several studies have shown that microRNAs are related to several cancers [5–7] and other diseases [8–10]. Caligiuri et al. [11] proposed that methods and compositions involving miRNAs are useful for the treatment of various diseases and cancers. Some miRNAs are regarded as potential therapeutic targets for various diseases [12]. Recently, the target gene (cancer gene) drugs, which developed in accordance with the theory on miRNA's gene silencing, have been used for incurable disease that has become a threat to human health problems for years [13]. In addition, the viral genome can encode a large number of miRNAs by itself. Through combination with target genes and coding by viruses or host cell, these miRNAs can lead to immune escape or antiviral effect against the host cell. Therefore, the accurate prediction of miRNA and its target genes, as well as the correct understanding of miRNA mechanism, has important practical significance in medical treatments. Thus, the research on novel miRNA identification is rather essential.

Feature selection mainly dominated the performance of the prediction model in the machine learning process [14–20]. In addition, effective features can represent the characteristics of the entire sequence data, which enables easy-to-build better prediction model. To represent the microRNA precursors, Xue et al. [21] proposed 32D novel triplet features, which involved secondary structure information. Jiang et al. [22] found that random rearrangement of the sequence could help obtain significant free-energy features. However, the free-energy computation for many random rearrangement sequences is very time consuming. Wei et al. [23] combined Xue et al.'s features and triplet nucleotide frequency to 98D

features and obtained good performance result in human pre-miRNA identification. However, more features would not mean better performance because of some irrelevant and redundant features in the high dimensional or ultra-high dimensional feature set. The purpose of feature selection is to eliminate the irrelevant and redundant features of the feature set. In addition, the training time could be reduced effectively by the feature selection optimization [24]. Some studies focus on developing computational predictors by incorporating the sequence-order or structure-order effects [25, 26]. Several works indicated that proper features could improve the prediction performance of classification in a certain extent. For example, Wang et al. [27] employed the feature selection techniques to optimize the features in miR-SF. They proved that an optimized feature subset could improve the prediction performance. In addition, some popular recently proposed multiobjective optimization evolutionary algorithms can also be used as a possibly promising feature selection approach [28, 29].

Another factor that affects the performance of machine learning prediction method is the classifier algorithm. The selection of different classifiers often leads to the difference of classification results. Several different classifiers and strategies were employed for miRNA identification. Bayesian classifier algorithm was tested for predicting miRNA across different species in 2006 [30]. The method also utilized the multiple species of miRNA sequences and structural features. It proved that miRNA genes could be detected effectively in large scale of different species genomes.

MiPred classifier was tested for predicting miRNA in 2007 [22]. The method utilized random forest classifier algorithm. The prediction accuracy of MiPred is 10% higher than that of Triplet-SVM; the sensitivity and specificity of MiPred can reach to 95.09% and 98.21%. CSHMM classifier was also used for mining miRNA sequences from the genome [31], which utilized the Markov model. Overall, the accuracy of machine learning algorithm was up to 90%. The machine learning method is more accurate than the other methods.

In this study, we chose backpropagation neural network as the classifier. It has three advantages, including better generalization performance, faster learning speed, and good learning ability.

2. miRNA Identification with BP Neural Network

2.1. Pre-miRNA Features

2.1.1. n-Gram Frequency. Some studies showed that the local primary sequence is crucial to the pre-miRNA sequence [32]. Thus, the n-gram frequency is often applied for the feature map in the selection of the primary sequence feature [33, 34]. However, no good methods are still available for tuning the value of n. In general, we choose n by comparing the effect of n-gram frequency with different n-values. In our feature set, we select the different n values ($n = 2, 3, 4$) for comparison. The different frequency characteristics have almost the same effect on the classifier. Thus, consider that its base and adjacent base have practical biological significance. We chose n as 3. A total of 64 ($64 = 4^3$)-dimensional frequency features were calculated.

2.1.2. Triple Structure Sequence. In addition to high specificity of the primary sequence features, the secondary structure sequence of pre-miRNA is also a contributing factor. To analyze the contribution of the secondary structure, the secondary structure prediction software RNAfold is used to calculate the potential structures. In the secondary structure, each nucleotide of the sequence corresponds to two states, matching and nonmatching: record matching as "(" or ")" and nonmatching as "·." In the structure, three character groups are considered as a unit, and every ")" is replaced as "(." Thus, 8 ($8 = 2^3$) different combinations are available as a unit, including "(((," "((·," "(·(," "·((," "(··," "·(·," "··(," and "· · ·."

To characterize pre-miRNA sequence better, the first nucleotide of the corresponding subsequence was added to the front of each structure unit. This provides 32 ($32 = 4 \times 8$) different combinations, that is, "A(((," "U((·," . . ., "G·((," "G· · ·." For a sequence, the occurrence frequency of each combination is determined and coded into the 32D feature vector as the input of the classifier. This calculated 32D triple structure sequence feature is used to train the SVM classifier; the inclusion of the SVM classifier significantly improved the classification ability of pre-miRNA sequences [21].

2.1.3. Energy Characteristics. The real pre-miRNA sequences are generally more stable and show a lower minimum of free energy (MFE) than the randomly generated pre-miRNA. Therefore, energy characterization is often used to describe the structure pre-miRNA sequence as an aspect of feature extraction of the pre-miRNA sequence. To do this, the MFE value is obtained by using RNAfold to calculate the structure.

2.1.4. Structural Diversity Characteristics. The potential for nucleotide pairing in the sequence is a significant characteristic that can also be used to describe the pre-miRNA sequence. This includes both traditional Watson-Crick nucleotide pairing (A–U pairing and C–G pairing) and also other forms of nucleotide pairing, such as the G–U pairing that can occur in the loop of RNA hairpin structures. We included possible G–U pairing in our description of base pairing.

To summarize, we extracted 98 features for the input of the neural network, including 64-dimensional n-gram frequency characteristics, 32-dimensional triple structure sequence characteristics, one-dimensional energy feature, and one-dimensional structural diversity characteristics.

2.2. Fixing the Number of Nodes in the Hidden Layer. In general, to select the number of nodes in the hidden layer in changing the BP neural network structure is difficult. Technically, a hidden layer could facilitate operation. However, too many hidden layers can reduce the operation rate.

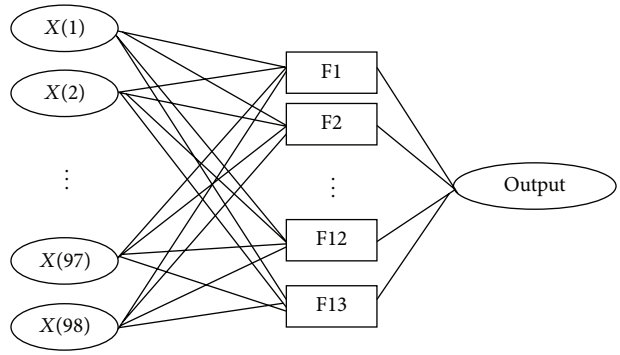


Figure 1: Topology structure of the BP neural network.

Table 1: Corresponding training results with different numbers of nodes in the hidden layers.

Hidden layers	Training times	Training errors	Hidden layers	Training times	Training errors
11	43	$9.57718e-005$	12	39	$9.88418e-005$
13	17	$9.42136e-005$	14	65	$9.92537e-005$
15	34	$9.88206e-005$	16	74	$8.38658e-005$
17	48	$7.82527e-005$	18	157	$6.63468e-005$
19	7	$9.46711e-005$	20	47	$9.3627e-005$

Currently, no theoretical methods are available to fix the number of nodes in the hidden layer. However, the number generally depends on the empirical formula, as calculated in

$$M = \sqrt{N + L} + \alpha,$$
$$M = \log_2 N, \quad (1)$$
$$M = \sqrt{NL},$$

where M represents the neuron number of the hidden layers, N is the neuron number of the input layers, L is the neuron number of the output layers, and α is a constant between 1 and 10.

In this study, $N = 98$ and $L = 1$. Therefore, (1) can be used for any values between 11 and 20. A comprehensive analysis of the training results with different numbers of nodes in the hidden layer was performed with the error set to 0.0001. A total of 621 samples were used to train the network, and one sample was used to test the network. The results are shown in Table 1.

From the data shown in Table 1, the increased number of nodes in the hidden layer did not result in better convergence. Additionally, the increased number of nodes increased the network parameters and greatly increased the amount of calculation of the classifier. Thus, keeping 13 nodes in the hidden layers required relatively less training times and less error and still produced relatively good training effects.

2.3. Fixing the Number of Nodes in the Output Layer. Two kinds of output exist, positive and negative, which are represented as 1 for a positive sample and 0 for a negative sample. The topology structure of this prediction method based on BP neural network is shown in Figure 1.

2.4. Selecting Training and Test Model Samples. The collection and organization of training samples are often limited by the objective conditions. Appropriate numbers of training samples are required to achieve sufficient precision. Therefore, it refers to the rule of experience:

$$P = (5 \sim 10) \times P_w, \quad (2)$$

where P represents the numbers of training samples and P_w is the total of network connection weight equal to the sum of nodes of the input and hidden layers. In this study, 2236 samples were used for training.

The data set used for the pre-miRNAs was downloaded from http://bioinf.sce.carleton.ca/SMIRP [35], and these data include negative and positive samples for *Arabidopsis lyrata*. The FASTA file was converted to ARFF file using a jar package written by Java converting the reference index to numerical form. We randomly selected real pre-miRNAs and pseudo pre-miRNAs to evaluate our algorithm.

2.5. Error Evaluation Steps Based on BP. The structure of the intelligent diagnosis model contains three layers of 98-13-1. First, we set the nodes of the input, output, and hidden layers as N, M, and L, respectively. Assuming the training sample set $\{\xi^p, Y\} \subset R^N \times R^L$, the weight matrix between the input and hidden layers can be written as $V = (v_{mn})_{M\times N}$, where $V_m = (v_{m1}, v_{m1}, \ldots, v_{m1})^T \in R^N$ and $m = 1, 2, \ldots, M$. We assume the connection weight matrix between the hidden and output layers as $W = (w_{lm})_{L\times M}$, where $W_l = (v_{l1}, v_{l1}, \ldots, v_{lm})^T \in R^M$, $l = 1, 2, \ldots, L$. Then, respectively, take g and f as the activation function of each node of the hidden and output layers. To simplify the derivation, we use the vector function $G(X)$ for $X = (x_1, x_2, \ldots, x_m)^T \in R^M$, where

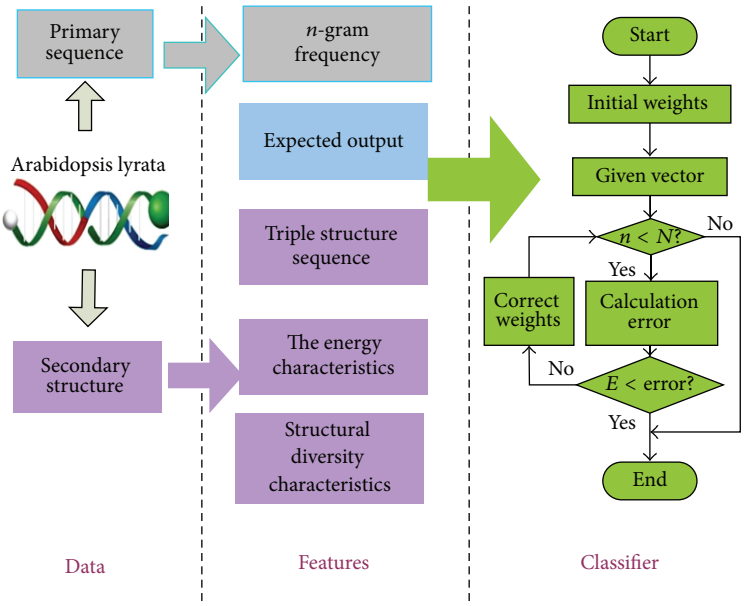


Figure 2: Process flow of model generation and training.

$G(X) = (g(x_1), g(x_2), \ldots, g(x_m))^T \in R^M$. After input of the sample $\xi^p \in R^N$, the actual output can be calculated by

$$\zeta_l^p = f\left(w_l \cdot G\left(V_l \xi^p\right)\right). \tag{3}$$

The error function is defined in

$$E(W,V) = \frac{1}{2}\sum_{p=1}^{P}\sum_{l=1}^{L}\left(O^p - f\left(w_l \cdot G\left(V_l \xi^p\right)\right)\right)^2. \tag{4}$$

Objectively, the target of BP training is to compute the W and V to minimize the solution of the error function $E(W,V)$. With this, a combination of gradient descent, common, and simple derivatives was used. To simplify the derivation process, we derive

$$f_{pl}(x) = \frac{1}{2}\left(O_l^p - f(x)\right)^2. \tag{5}$$

Then, the error function can be written as

$$E(W,V) = \sum_{p=1}^{P}\sum_{l=1}^{L} f_{pl}\left(w_l \cdot G\left(V_m \xi^p\right)\right). \tag{6}$$

The corresponding gradient function of W and V can then be expressed as

$$\begin{aligned} E_{wl}(W,V) &= \sum_{p=1}^{P} f_{pl}^{l}\left(w_l \cdot G\left(V\xi^p\right)\right) G\left(V\xi^p\right), \\ E_{vm}(W,V) &= \sum_{p=1}^{P}\sum_{l=1}^{L} f_{pl}^{l}\left(w_l \cdot G\left(V\xi^p\right)\right) w_{lm} \acute{g}\left(v_m \cdot \xi^p\right)\xi^p. \end{aligned} \tag{7}$$

Table 2: Basic parameters of the classifier based on BP neural network.

Setting items	The value set
The learning rate	0.1
Error bounds	0.0001
The number of iterations	1000
Transfer function of hidden layer nodes	Tansig
Transfer function of output nodes	Purelin
The training function	Trainlm

For arbitrary initial values of $W_0 \in R^{L\times M}$ and $V_0 \in R^{M\times N}$, gradient descent rules to modify the weight of the BP learning algorithm are applied in

$$\begin{aligned} W_l^{n+1} &= W_l^n + \Delta W_l^n, \\ \Delta W_l^n &= -\eta_n E_{wl}\left(W^n, V^n\right), \\ V_m^{n+1} &= V_m^n + \Delta V_m^n, \\ \Delta V_m^n &= -\eta_n E_{wm}\left(W^n, V^n\right), \end{aligned} \tag{8}$$

where η_n represents the learning efficiency. ΔW_l^n is the partial derivative of the error function relative to W. ΔV_m^n is the partial derivative of the error function relative to V.

2.6. Selection of Training Functions and Related Parameters. The above analysis allows fixing of the BP neural network structure. Table 2 shows the chosen training functions and the relevant parameters.

This condition allows establishment of a complete classifier based on BP neural network structure. The model generation and training are summarized in Figure 2.

Table 3: Measurements for the classification problems.

Actual result	Classification result: Forecast result P	Classification result: Forecast result N
P	TP	FN
N	FP	TN

2.7. Measurement. The use of pattern recognition and machine learning methods can be used as a two-way classification problem. Four kinds of prediction results are presented in Table 3.

The four kinds of prediction results are true positive (TP), the number of positive cases that were correctly predicted; false positive (FP), the number of positive cases represented by error prediction; true negative (TN), the number of counter negative examples that were correctly predicted; and false negative (FN), the number of negative cases represented by error prediction.

Many evaluation indicators can be used for the classification results. First, the accuracy rate (ACC) is the ratio of the correctly predicted cases for the entire data set. Precision and recall can also be used as evaluation indicators in tests of pattern recognition models. Precision is expressed as the ratio of the correctly predicted values for the entire positive data set and recall reflects the number correctly judged as positive examples in the positive example test set [36]. The above three indicators are expressed in

$$\begin{aligned} \text{ACC} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}. \end{aligned} \tag{9}$$

Additionally, sensitivity and specificity parameters may be used to evaluate the function of the model. Sensitivity record (SE) is the same as the recall and specificity record (SP) calculated in accordance with

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{10}$$

A challenge may be presented if the positive and negative test sets are unbalanced in the study of biological information. In most cases, the number of positive samples is far less than the number of negative samples. In a few cases, the number of positive samples may be much larger than the number of negative samples. We can easily obtain ACC-SP when the number of positive samples is greater than the negative samples. In this case, the classifier only reflects the classification effect of the negative samples and is unable to accurately express the prediction effect of the classifier on the entire test data set. To solve this problem, researchers typically use the geometric mean (Gm) as described in

Table 4: Comparison of classification results based on different feature sets.

Features	SP (%)	SE (%)	Gm (%)	ACC (%)
B	67.89	68.25	68.07	68.00
C	92.74	76.42	84.19	88.03
A + B	91.79	90.41	91.10	91.31
A + C	94.03	80.85	87.19	89.67
B + C	96.12	85.21	90.50	92.49
A + B + C	96.33	86.51	91.29	93.42

Notes: A: energy feature and structural diversity; B: 32-dimensional triad structure characteristic; C: 64-dimensional *n*-gram frequency characteristics.

$$\text{Gm} = \sqrt{\text{SE} \cdot \text{SP}}. \tag{11}$$

Matthew's correlation coefficient (MCC) [16, 21, 37, 38] can provide more equitable response forecast ability when a large difference exists between the number of positive samples and the number of negative samples. MCC can be expressed as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TN} + \text{FN}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP})}}. \tag{12}$$

Currently, studies on miRNA commonly use one or more of these above evaluation indices. In this work, we estimate the overall performance of the classifier by analysis of ACC, SE, SP, Gm, and MCC.

3. Results and Discussion

3.1. Analysis of Feature Set Performance. To select a better feature set for classification, we needed to determine the effect of different feature subsets on the performance of the classifier. To do this, we used the BP neural network method with the same training set (553 positive samples and 1150 samples) to test different feature sets, with the results shown in Table 4.

From Table 4, we learn that the accuracy of the entire feature sets can be as high as 93.42%. This result indicates that our feature set is more effective for processing of a more complex structure or sequence diversity. Considering that the feature sets used here are not very large and each feature subset is highly independent, reducing the dimension of the feature vector is no longer needed.

3.2. Performances of BP. V-fold cross-validation with moderate computational complexity is widely used for model selection. The selection of V is important because V not only determines the number of samples but also determines the computational complexity. Usually, a value of V between 5 and 10 is selected based on experience. Statistical

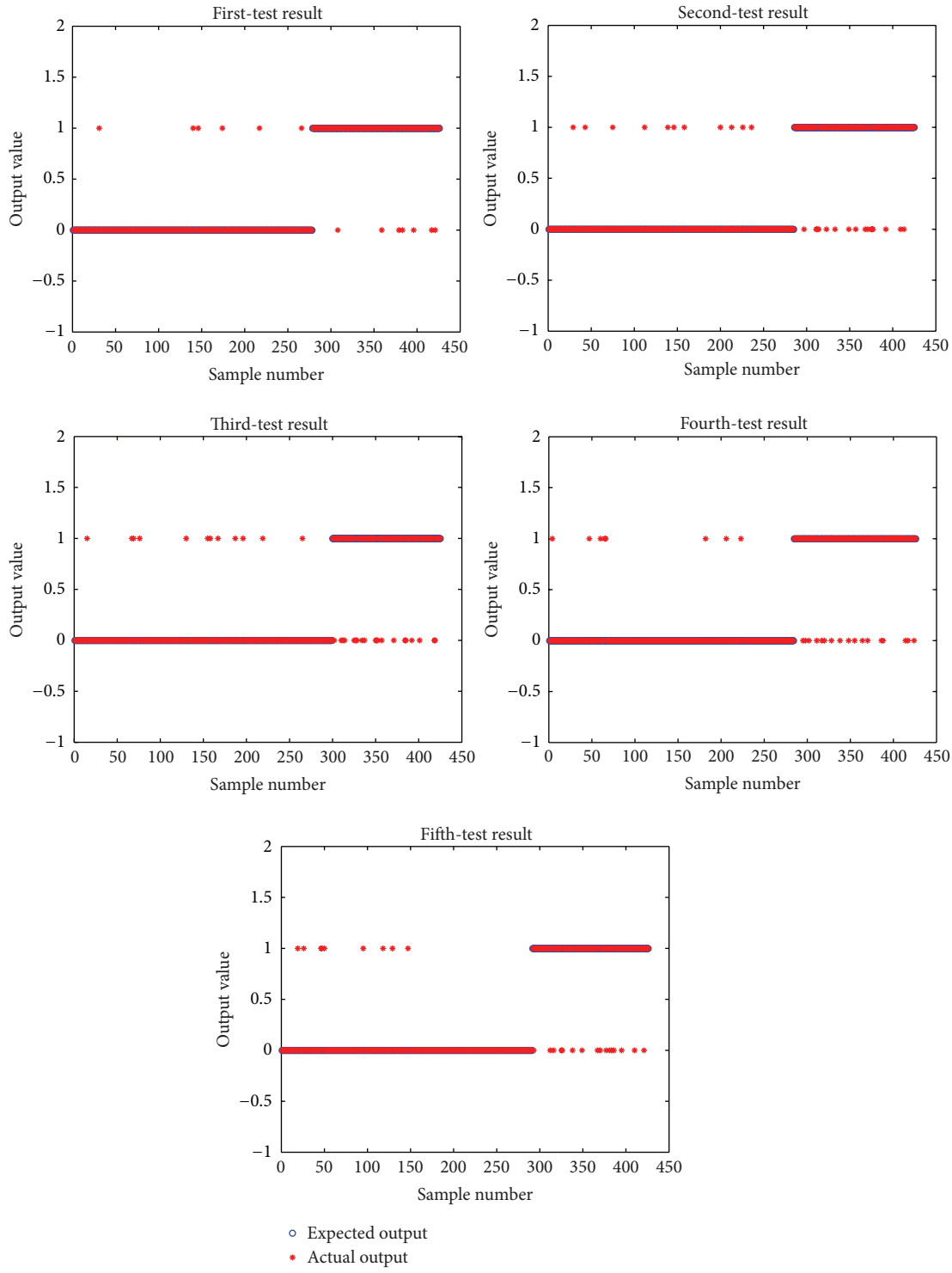


Figure 3: Different test results for varying sample quantities.

performance shows little improvement when V selection is greater than 10. Again, computational complexity must be considered; thus a value between 5 and 10 is best [32].

We divided the samples into two cases for training and testing. In the first one, a large difference was observed between the number of positive and negative samples: 518 positive samples and 1078 negative samples as the training set and 166 positive samples and 366 negative samples for the test set. The second case included equal numbers of positive and negative samples: 552 positive samples and 552 negative samples as the training set and 138 positive samples and 138 negative samples for the test set. These training and testing were repeated five times. The testing performance is shown in Figures 3 and 4.

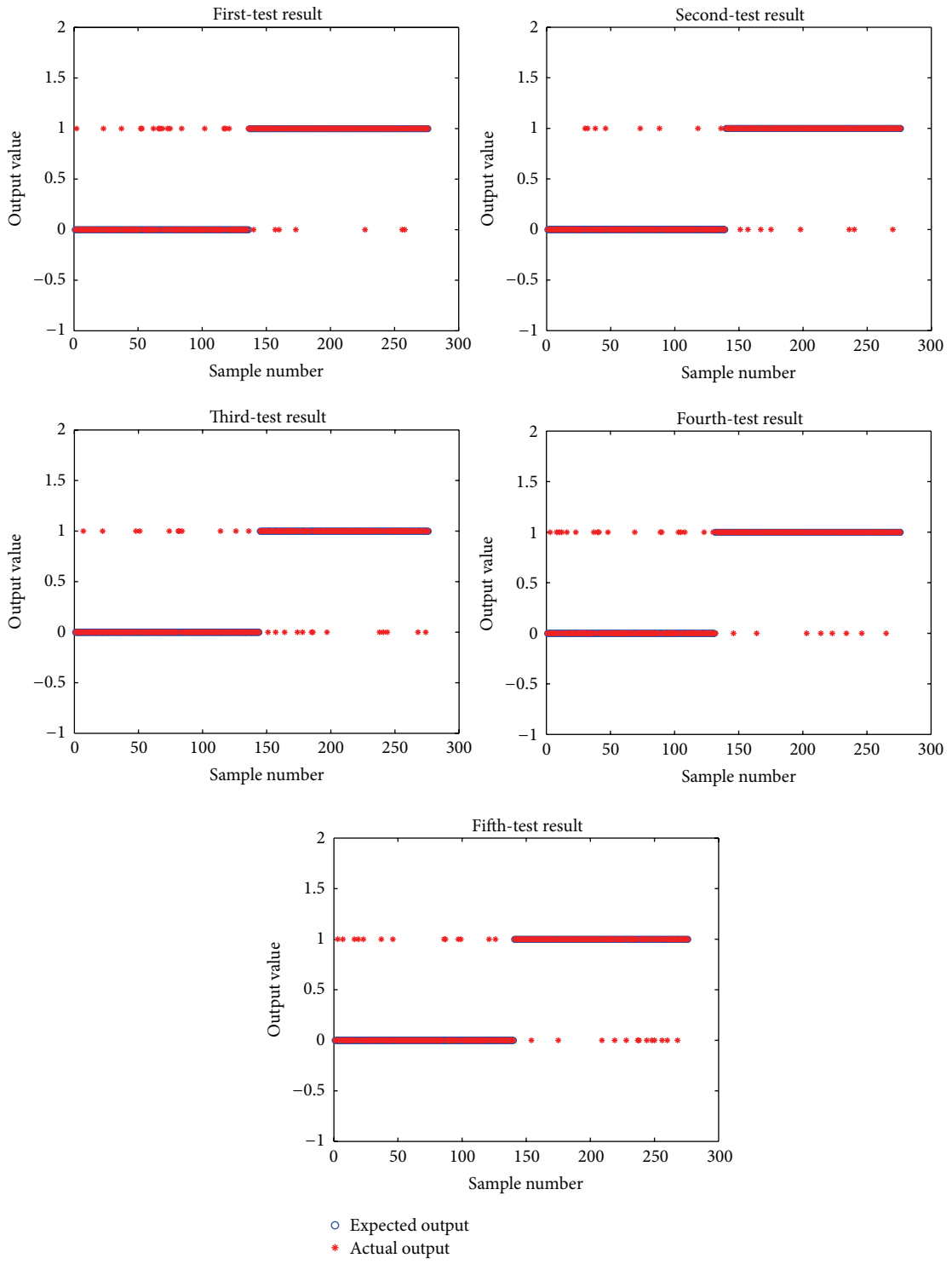


Figure 4: Different test results for same sample quantity.

From comparison of the data in Figures 3 and 4, no significant difference was observed between the actual output and the expected output of each test. As described above, the evaluation of the reference index is shown in Table 5.

From the data presented in Table 5, the number of samples affects the accuracy and recall rate of the positive samples. In particular, the precision and recall rate of the negative samples decreased with the decrease in the number of negative samples in the training set. This result indicates that

Table 5: Evaluation of the reference index.

	Training sample	Test sample	Output sample	Correct sample	Precision (%)	Recall (%)	Gm (%)
D							
Positive	553	138	128	124	96.0	90.0	93.43
Negative	1150	287	296	282	95.38	98.19	
E							
Positive	552	138	136	128	94.10	92.82	93.98
Negative	552	138	140	130	92.87	94.12	

Note: D: sample set has different numbers of positive and negative samples; E: the sample set has equal numbers of positive and negative samples; correct sample: the number of correct predictions.

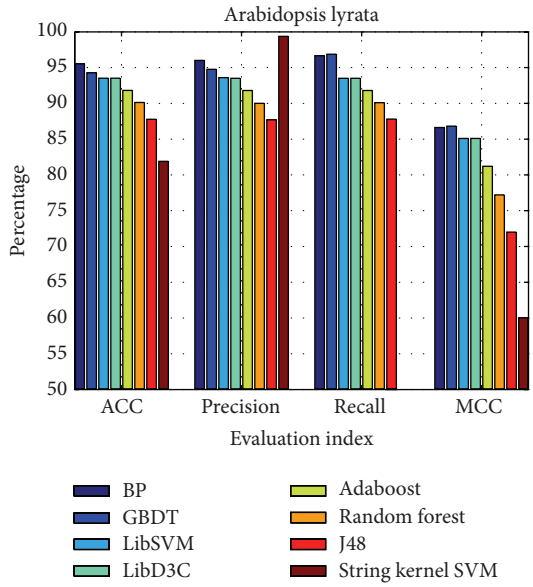


Figure 5: Comparison results of different models.

the more the samples in the training process, the better the classification effect of the classifier. At the same time, the precision and recall rate of the number of positive samples were affected. With the number of negative samples in the training set increased, the number of correct predictions increased by four and the number of error predictions was reduced by eight. This result shows that the precision and recall rate of the positive samples decreased with the increase in the number of the negative samples.

3.3. Comparison with Other Methods. The performance of our method was compared with other methods: J48, random forest, LibD3C [39], Adaboost, string kernel SVM [40], LibSVM, and GBDT, which were classified on the same data set. The data set contains 691 real pre-miRNAs and 1437 pseudo pre-miRNAs. As shown in Table 6 and Figure 5, the results demonstrate that the total prediction accuracy of our method is 13.64% greater than the string kernel SVM model and nearly 2% greater than the LibD3C and LibSVM models. The overall performance of the models as measured by MCC was in the following order: GBDT (0.8682), BP (0.8662), LibSVM (0.8510), LibD3C (0.8510), Adaboost (0.8120), random forest (0.7720), J48 (0.7200), and string kernel SVM (0.6002).

Table 6: Comparison of the BP with alternative models.

	ACC	Precision	Recall	MCC
BP	95.53%	96.00%	96.67%	0.8662
GBDT	94.27%	94.76%	96.87%	0.8682
LibSVM	93.52%	93.60%	93.50%	0.8510
LibD3C	93.52%	93.50%	93.50%	0.8510
Adaboost	91.82%	91.80%	91.80%	0.8120
Random forest	90.13%	90.00%	90.10%	0.7720
J48	87.78%	87.70%	87.80%	0.7200
String kernel SVM	81.89%	99.37%	46.31%	0.6002

Thus, we conclude that the BP method allows improved recognition accuracy.

3.4. Performance on Different Species. To demonstrate the validity and the universal applicability of the BP method, we analyzed six other species: *Anolis carolinensis*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Drosophila pseudoobscura*, Epstein-Barr virus, and *Xenopus tropicalis*. The results shown

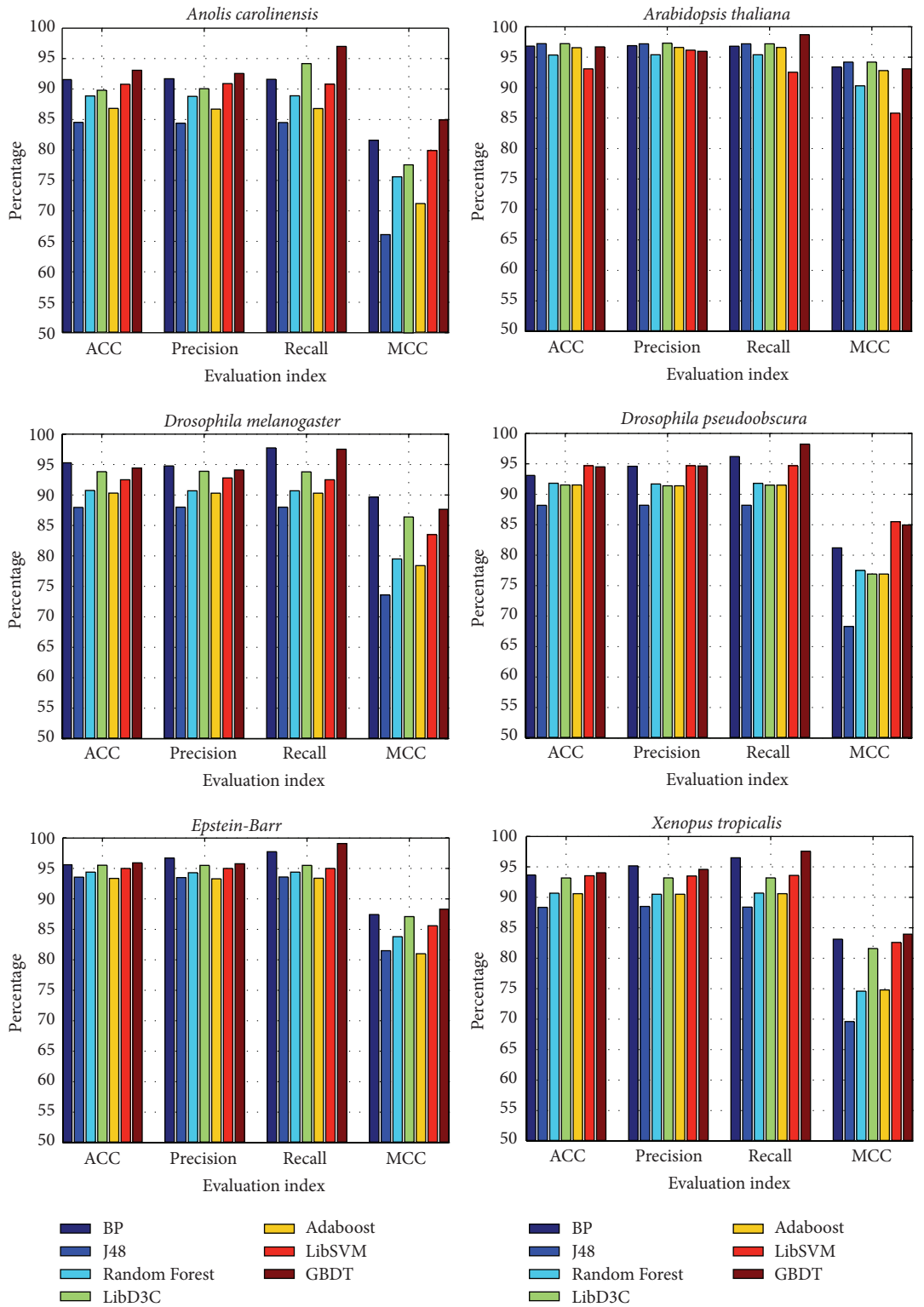


Figure 6: Test comparison results for six different species.

in Figure 6 indicate that the accuracy of the GBDT is better than BP method in some situations, but the BP method has been achieved fairly good results in terms of ACC, precision, recall, and MCC.

4. Conclusions

Identification of miRNAs is the first step toward understanding their biological characteristics. Many approaches

have been proposed to predict pre-miRNAs in recent years. However, feature extraction in these methods can result in information redundancy. To overcome this drawback, a BP neural network algorithm together with optimal 98D features was employed for this analysis. We compare our method with the existing methods of J48, random forest, LibD3C, Adaboost, GBDT, string kernel SVM, and LibSVM, which were trained on the same training data set. The results demonstrate that the total prediction accuracy of our method is 13.17% greater than the string kernel SVM model and nearly 2% greater than LibD3C and LibSVM.

After the identification step, functional analysis is also important for miRNA research. If human miRNA and diseases were focused on, two main approaches would be employed to predict the relationship. The first one is the statistical comparison analysis for the miRNA or isomiR expression [41]. The second one is the network analysis and prediction for miRNA-disease relationship [42–45]. Several advanced machine learning, network techniques, and bioinspired models can be utilized on this problem, including random forest [46], semisupervised learning [47], HeteSim Scores [48], spiking neural P systems [49–52], and membrane computing_ENREF_51 [53–57]. Functional analysis of the novel detected miRNAs would be our future works.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The work was supported by the Natural Science Foundation of China (no. 61370010 and no. 61302139) and the State Key Laboratory of Medicinal Chemical Biology in China.

References

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

[2] D. Wu, Y. Huang, J. Kang et al., "ncRDeathDB: a comprehensive bioinformatics resource for deciphering network organization of the ncRNA-mediated cell death system," *Autophagy*, vol. 11, no. 10, pp. 1917–1926, 2015.

[3] Y. Huang, N. Liu, J. P. Wang et al., "Regulatory long non-coding RNA and its functions," *Journal of Physiology and Biochemistry*, vol. 68, no. 4, pp. 611–618, 2012.

[4] X. Zhang, D. Wu, L. Chen et al., "RAID: a comprehensive resource for human RNA-associated (RNA–RNA/RNA–protein) interaction," *RNA*, vol. 20, no. 7, pp. 989–993, 2014.

[5] S. Hua, W. Yun, Z. Zhiqiang, and Q. Zou, "A discussion of micrornas in cancers," *Current Bioinformatics*, vol. 9, no. 5, pp. 453–462, 2014.

[6] Q. Wang, L. Wei, X. Guan, Y. Wu, Q. Zou, and Z. Ji, "Briefing in family characteristics of microRNAs and their applications in cancer research," *Biochimica et Biophysica Acta—Proteins and Proteomics*, vol. 1844, no. 1, pp. 191–197, 2014.

[7] C. Yang, D. Wu, L. Gao et al., "Competing endogenous RNA networks in human cancer: hypothesis, validation, and perspectives," *Oncotarget*, vol. 7, no. 12, pp. 13479–13490, 2016.

[8] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 193–203, 2016.

[9] Q. Zou, J. Li, Q. Hong et al., "Prediction of microRNA-disease associations based on social network analysis methods," *BioMed Research International*, vol. 2015, Article ID 810514, 9 pages, 2015.

[10] Y. Wang, L. Chen, B. Chen et al., "Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network," *Cell Death and Disease*, vol. 4, no. 8, article e765, 2013.

[11] M. A. Caligiuri, J. Yu, S. He, and R. Trott, "Activation of Innate Immunity by miRNA for Cancer and Infection Treatment," United States Patent, 2016.

[12] H. Zhou, X. Ge, and X. Xue, "microRNAs regulation and its role as biomarkers in diseases," *Oncology and Translational Medicine*, vol. 2, no. 1, pp. 39–46, 2016.

[13] P. S. Kelly, C. Gallagher, M. Clynes, and N. Barron, "Conserved microRNA function as a basis for Chinese hamster ovary cell engineering," *Biotechnology Letters*, vol. 37, no. 4, pp. 787–798, 2015.

[14] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016.

[15] W.-X. Liu, E.-Z. Deng, W. Chen, and H. Lin, "Identifying the subfamilies of voltage-gated potassium channels using feature selection technique," *International Journal of Molecular Sciences*, vol. 15, no. 7, pp. 12940–12951, 2014.

[16] H. Tang, W. Chen, and H. Lin, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Molecular BioSystems*, vol. 12, no. 4, pp. 1269–1275, 2016.

[17] P.-P. Zhu, W.-C. Li, Z.-J. Zhong et al., "Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition," *Molecular BioSystems*, vol. 11, no. 2, pp. 558–563, 2015.

[18] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis," *Molecular BioSystems*, vol. 10, no. 8, pp. 2229–2235, 2014.

[19] H. Ding, S.-H. Guo, E.-Z. Deng et al., "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[20] H. Ding, H. Lin, W. Chen et al., "Prediction of protein structural classes based on feature selection technique," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 3, pp. 235–240, 2014.

[21] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine," *BMC Bioinformatics*, vol. 6, article 310, 2005.

[22] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, "MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features," *Nucleic Acids Research*, vol. 35, no. 2, pp. W339–W344, 2007.

[23] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human microRNAs by incorporating a high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[24] Y. Wang, X. Chen, W. Jiang et al., "Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM," *Genomics*, vol. 98, no. 2, pp. 73–78, 2011.

[25] B. Liu, L. Fang, F. Liu, X. Wang, J. Chen, and K.-C. Chou, "Identification of real microRNA precursors with a pseudo structure status composition approach," *PLoS ONE*, vol. 10, no. 3, article e0121501, 2015.

[26] B. Liu, L. Fang, J. Chen, F. Liu, and X. Wang, "MiRNA-dis: MicroRNA precursor identification based on distance structure status pairs," *Molecular BioSystems*, vol. 11, no. 4, pp. 1194–1204, 2015.

[27] X. Wang, J. D. Laurie, T. Liu, J. Wentz, and X. S. Liu, "Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites," *Genomics*, vol. 97, no. 4, pp. 235–243, 2011.

[28] X. Zhang, Y. Tian, R. Cheng, and Y. Jin, "An efficient approach to nondominated sorting for evolutionary multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 2, pp. 201–213, 2015.

[29] X. Zhang, Y. Tian, and Y. Jin, "A knee point-driven evolutionary algorithm for many-objective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 6, pp. 761–776, 2014.

[30] M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L. C. Showe, and M. K. Showe, "Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier," *Bioinformatics*, vol. 22, no. 11, pp. 1325–1334, 2006.

[31] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[32] E. Bonnet, J. Wuyts, P. Rouzé, and Y. Van de Peer, "Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences," *Bioinformatics*, vol. 20, no. 17, pp. 2911–2917, 2004.

[33] H. Liu and L. Wong, "Data mining tools for biological sequences," *Journal of Bioinformatics and Computational Biology*, vol. 1, no. 1, pp. 139–167, 2003.

[34] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. Chou, "Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Research*, vol. 43, no. 1, pp. W65–W71, 2015.

[35] R. J. Peace, K. K. Biggar, K. B. Storey, and J. R. Green, "A framework for improving microRNA prediction in non-human genomes," *Nucleic Acids Research*, vol. 43, no. 20, article e138, 2015.

[36] S. Yang, S. Cai, F. Zheng et al., "Representation of fluctuation features in pathological knee joint vibroarthrographic signals using kernel density modeling method," *Medical Engineering and Physics*, vol. 36, no. 10, pp. 1305–1311, 2014.

[37] R. Wang, Y. Xu, and B. Liu, "Recombination spot identification Based on gapped k-mers," *Scientific Reports*, vol. 6, Article ID 23934, 2016.

[38] Y. Wu, P. Chen, X. Luo et al., "Quantification of knee vibroarthrographic signal irregularity associated with patellofemoral joint cartilage pathology based on entropy and envelope amplitude measures," *Computer Methods and Programs in Biomedicine*, vol. 130, pp. 1–12, 2016.

[39] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, and Q. Zou, "LibD3C: ensemble classifiers with a clustering and dynamic selection strategy," *Neurocomputing*, vol. 123, pp. 424–435, 2014.

[40] M. Ghandi, D. Lee, M. Mohammad-Noori, and M. A. Beer, "Enhanced regulatory sequence prediction using gapped *k*-mer features," *PLoS Computational Biology*, vol. 10, no. 7, Article ID e1003711, 2014.

[41] L. Guo, J. Yu, T. Liang, and Q. Zou, "miR-isomiRExp: a web-server for the analysis of expression of miRNA at the miRNA/isomiR levels," *Scientific Reports*, vol. 6, Article ID 23700, 2016.

[42] 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.

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

[44] 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.

[45] X. Chen, "miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method," *Molecular BioSystems*, vol. 12, no. 2, pp. 624–633, 2016.

[46] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.

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

[48] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim Scores," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.

[49] T. Song, J. Xu, and L. Pan, "On the universality and non-universality of spiking neural P systems with rules on synapses," *IEEE Transactions on NanoBioscience*, vol. 14, no. 8, pp. 960–966, 2015.

[50] X. Zhang, X. Zeng, B. Luo, and L. Pan, "On some classes of sequential spiking neural P systems," *Neural Computation*, vol. 26, no. 5, pp. 974–997, 2014.

[51] T. Song and L. Pan, "Spiking neural P systems with request rules," *Neurocomputing*, vol. 193, pp. 193–200, 2016.

[52] X. Wang, T. Song, F. Gong, and P. Zheng, "On the computational power of spiking neural P systems with self-organization," *Scientific Reports*, vol. 6, Article ID 27624, 2016.

[53] X. Zeng, L. Xu, X. Liu, and L. Pan, "On languages generated by spiking neural P systems with weights," *Information Sciences*, vol. 278, pp. 423–433, 2014.

[54] X. Zhang, L. Pan, and A. Păun, "On the universality of axon P systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 11, pp. 2816–2829, 2015.

[55] X. Chen, M. J. Pérez-Jiménez, L. Valencia-Cabrera, B. Wang, and X. Zeng, "Computing with viruses," *Theoretical Computer Science*, vol. 623, pp. 146–159, 2016.

[56] T. Wu, Z. Zhang, G. Păun, and L. Pan, "Cell-like spiking neural P systems," *Theoretical Computer Science*, vol. 623, pp. 180–189, 2016.

[57] X. Zhang, Y. Liu, B. Luo, and L. Pan, "Computational power of tissue P systems for generating control languages," *Information Sciences*, vol. 278, pp. 285–297, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 4130861, 8 pages
http://dx.doi.org/10.1155/2016/4130861

Research Article

Annotating the Function of the Human Genome with Gene Ontology and Disease Ontology

Yang Hu,[1] Wenyang Zhou,[1] Jun Ren,[1] Lixiang Dong,[2] Yadong Wang,[3] Shuilin Jin,[4] and Liang Cheng[5]

[1] *School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, China*
[2] *School of Software, Harbin Institute of Technology, Harbin 150001, China*
[3] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*
[4] *Department of Mathematics, Harbin Institute of Technology, Harbin 150001, China*
[5] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China*

Correspondence should be addressed to Yadong Wang; ydwang@hit.edu.cn, Shuilin Jin; jinsl@hit.edu.cn, and Liang Cheng; liangcheng@hrbmu.edu.cn

Received 2 June 2016; Revised 24 July 2016; Accepted 27 July 2016

Academic Editor: Xing Chen

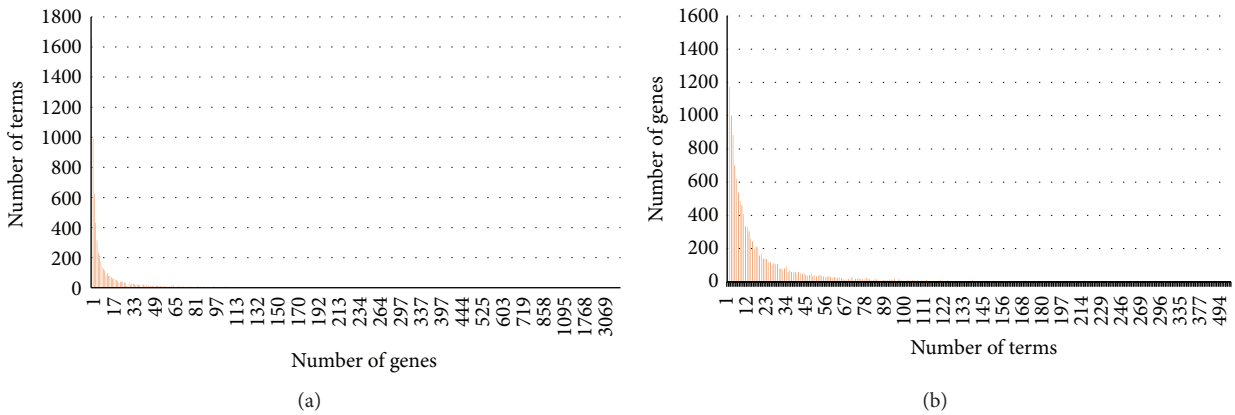
Increasing evidences indicated that function annotation of human genome in molecular level and phenotype level is very important for systematic analysis of genes. In this study, we presented a framework named Gene2Function to annotate Gene Reference into Functions (GeneRIFs), in which each functional description of GeneRIFs could be annotated by a text mining tool Open Biomedical Annotator (OBA), and each Entrez gene could be mapped to Human Genome Organisation Gene Nomenclature Committee (HGNC) gene symbol. After annotating all the records about human genes of GeneRIFs, 288,869 associations between 13,148 mRNAs and 7,182 terms, 9,496 associations between 948 microRNAs and 533 terms, and 901 associations between 139 long noncoding RNAs (lncRNAs) and 297 terms were obtained as a comprehensive annotation resource of human genome. High consistency of term frequency of individual gene (Pearson correlation = 0.6401, $p = 2.2e-16$) and gene frequency of individual term (Pearson correlation = 0.1298, $p = 3.686e-14$) in GeneRIFs and GOA shows our annotation resource is very reliable.

1. Introduction

The human genome is the complete set of nucleic acid sequence for human beings [1]. Researches on sequence of the human genome aim at exploring the functions of genes [2–5]. Human genes consisting of sequences could play diverse roles based on their functions in molecular level in balancing the body. Once the balance is lost by lack or enhancement of the functions of genes, diseases could be induced [6–9].

Previous studies focused on identifying the functions of the protein-coding genes in molecular level based on their encoded proteins. For example, through investigating p53 protein, Brain and Jenkins [10] exposed that TP53 gene is potentially capable of inhibiting mammalian replicative DNA synthesis by blocking the DNA strand separation step during replication origin recruitment. Based on a case control study, Benzon Larsen et al. [11] determined that ADH polymorphisms, which modify the rate of ethanol oxidation to acetaldehyde, were associated with breast cancer risk.

As a growing number of protein-coding genes identified, lots of functional terms emerged. For ease of comparing the functions of genes, these terms needed to be normalized. To this end, ontology was introduced to standardize the functional terms of genes. Among existing ontologies, Gene Ontology (GO) [12] is one of the earliest and most frequently used vocabularies, which focuses on describing biological process (BP), molecular function (MF), and cell component (CC) of genes. Since appearing in 2000, a large number of databases recording the functions of genes were annotated to the GO. The functional annotation of human protein-coding genes was provided at GO Annotation (GOA) databases [13], which involves a nonredundant set of annotations to



(a) (b)

Figure 1: Distribution of functional terms and genes in the annotation results. (a) Histogram of the number of genes associated with individual functional term. (b) Histogram of the number of functional terms associated with individual gene.

the human proteome. In comparison with the GO, Disease Ontology (DO) [14] focuses on standardizing the functional terms of genes at phenotype level. And disease terms in Gene Reference into Function (GeneRIF) [15] were annotated to the DO [16–18].

Recently, large-scale sequence analysis at genomic and transcriptomic level has shown that more than 98% of genome sequence cannot encode protein [19, 20], and microRNA genes and long noncoding RNA (lncRNA) genes constitute a large portion of them [21]. In comparison with protein-coding genes, the functions of microRNA genes and lncRNA genes are difficult to be identified [22]. However, these noncoding genes play an important role at molecular level and phenotype level [23–27]. For example, at molecular level, qPCR and in silico hybridization revealed that miR-124 and miR-155 can be directly involved in the transcriptional regulation of Runt-related transcription factor 2 (RUNX2) and receptor activator of nuclear factor kappa-B ligand (RANKL) genes [28]. At phenotype level, Huang et al. identified that underexpression of miR-345 is associated with prostate cancer [29]. At present, microRNA- and lncRNA-related diseases in HMDD [30] and LncRNADisease [31] have been manually annotated by Medical Subject Headings (MeSH) [32]. And several recent works proved more relationship between miRNA and diseases would be detected yet [33–35].

Although a few of databases have been annotated to gene functional vocabularies, a comprehensive annotation resource recording the functions of human genes had not yet appeared. For example, in our knowledge, no databases of noncoding genes were annotated to functional vocabularies at molecular level. This may be caused by the lack of resources that record the functions of protein-coding genes and noncoding genes simultaneously. Fortunately, GeneRIFs [15] provides a brief (up to 255 character) functional description of each gene in the NCBI database, and these functional descriptions could be annotated to vocabularies, such as DO and GO.

In this paper, we presented a framework, Gene2Function, to annotate the function of human genome with GO and DO.

Table 1: The statistical information of associations between genes and terms.

The number of genes	The number of terms	The number of associations between genes and terms
mRNA		
13,148	7,182	288,869
MicroRNA		
948	533	9,496
lncRNA		
139	297	901

After annotating GeneRIF, a comprehensive resource involving protein-coding genes, microRNA genes, and lncRNA genes could be obtained. The resource could be accessed from http://www.bio-annotation.cn/gene2function/.

2. Results

2.1. Mapping Genes to Gene Ontology and Disease Ontology. After annotating GeneRIFs by GO and DO (see Section 3), 288,869 associations between 13,148 mRNAs and 7,182 terms, 9,496 associations between 948 microRNAs and 533 terms, and 901 associations between 139 lncRNAs and 297 terms were obtained. The statistical information is shown in Table 1.

Figure 1(a) demonstrates the histogram of the number of genes associated with terms of GO and DO in the annotation results. 1,657 functional terms (23.0%) are associated with only one gene, while 3,924 functional terms (54.5%) are associated with more than three genes. The histogram of the number of terms associated with individual gene is shown in Figure 1(b). 1,375 genes (9.9%) are associated with only one functional terms, while 10,273 genes (74.3%) are associated with more than three genes.

The top ten terms ordered by the number of gene annotations and the top ten genes ordered by the number of term annotations are shown in Tables 2 and 3, respectively. Not surprisingly, several general terms in the top layer of the

Table 2: The top ten terms ordered by the number of gene annotations.

Term ID	Term name	Number of genes
GO:0005623	Cell	7,524
GO:0005488	Binding	5,011
GO:0065007	Biological regulation	4,846
GO:0023052	Signaling	4,466
GO:0032502	Developmental process	3,521
GO:0009058	Biosynthetic process	3,346
DOID:162	Cancer	3,139
GO:0006351	Transcription, DNA-templated	3,121
DOID:305	Carcinoma	3,069
GO:0040007	Growth	3,011

Table 3: The top ten genes ordered by the number of term annotations.

HGNC gene ID	Gene symbol	Number of functional terms
HGNC:11998	TP53	828
HGNC:11892	TNF	792
HGNC:6018	IL6	683
HGNC:12680	VEGFA	669
HGNC:11766	TGFB1	664
HGNC:3236	EGFR	560
HGNC:7176	MMP9	521
HGNC:391	AKT1	517
HGNC:7794	NFKB1	494
HGNC:6025	CXCL8	473

DAG have a larger number of genes associated with them, such as cell, binding, and developmental process (Table 2). The most prevalent disease terms appearing in the annotation result is cancer, which is associated with 3,139 genes (22.7% of all the terms). When we look at the genes associated with many terms, TP53 is the most prevalent genes appearing in the annotation result, which is associated with 828 terms (11.5% of all the genes).

2.2. Comparing with Existing Ontology Annotation Resources. To validate the performance of our annotation result, we compared the result with the previous prevalent annotation resources GOA [13], in which human gene is manually annotated to GO. To ensure the exact evaluation, DO annotations of GeneRIFs were discarded, and annotations Inferred from Electronic Annotations (IEA) of GOA were removed.

In total, we obtained 196,423 associations between 4,613 GO terms and 13,107 genes in GeneRIFs and 168,246 associations between 13,920 GO terms and 16,724 genes in GOA. Only 10,658 associations and 3,375 GO terms appeared in both annotation resources. In comparison, both of them have more common genes (11,816).

Figures 2(a) and 2(b) demonstrate the histogram of the number of genes per GO term, and the histogram of the number of GO terms per gene in annotations of GeneRIFs and GOA, respectively. Obviously, more GO terms (4,545) could be annotated to only one gene in GOA than that (1,114) in GeneRIFs. In contrast, more genes (1,671) could be annotated to only one term in GeneRIFs than that (1,499) in GOA.

Table 4: Data sources.

Data source	Web site (date of download)
GeneRIF	http://www.ncbi.nlm.nih.gov/gene/about-generif (Jun 2016)
HGNC	http://www.genenames.org/ (Jun 2016)
GO & GOA	http://geneontology.org/ (Jun 2016)
DO	http://disease-ontology.org/ (Jun 2016)

In order to evaluate the consistency, we compared the term frequency of individual gene and gene frequency of individual term in GeneRIFs and GOA. As a result, term frequency of individual gene in GeneRIF was significant positively correlated with it in GOA (Pearson correlation $\gamma^2 = 0.6401$, $p = 2.2e - 16$; Figure 2(c)), and gene frequency of individual term in GeneRIF was also significantly positively correlated with it in GOA (Pearson correlation $\gamma^2 = 0.1298$, $p = 3.686e - 14$; Figure 2(d)). Considering that GOA is most frequency used annotation resource, annotations of GeneRIFs should be also reliable.

2.3. A Network Visualization Based on the Functional Annotation of the Human Genome. Information in the annotation result can be used to describe the relationship among multiple genes or multiple terms. To this end, we create a bipartite network that describes the relationships between three genes (RNF2, RNF8, and RPS6) and 79 terms (Figure 3). Within this network, 33 terms are annotated to RNF2, 37 terms are annotated to RNF8, and 37 terms are annotated to RPS6. At the centre of the figure, 6 terms involving translation, execution phase of apoptosis, breast cancer, biological regulation, binding, and apoptotic process are related to all of these three genes. Using our annotation result, one can create this type of bipartite network as needed.

3. Materials and Methods

3.1. Data Collection

3.1.1. GeneRIF. GeneRIF was downloaded in June 2016 (Table 4). It involves five columns for describing tax identifier, NCBI gene ID, PubMed Unique Identifier (PMID), updated date, and function description. After extracting function descriptions of human genes, 650,079 descriptions remained.

3.1.2. Normalized Gene Symbol Vocabulary. The Human Genome Organisation Gene Nomenclature Committee (HGNC) [36] is responsible for approving unique symbols and names for human loci, including protein-coding genes and noncoding genes, to allow unambiguous scientific communication. In this paper, genes in GeneRIFs were normalized to HGNC gene symbols.

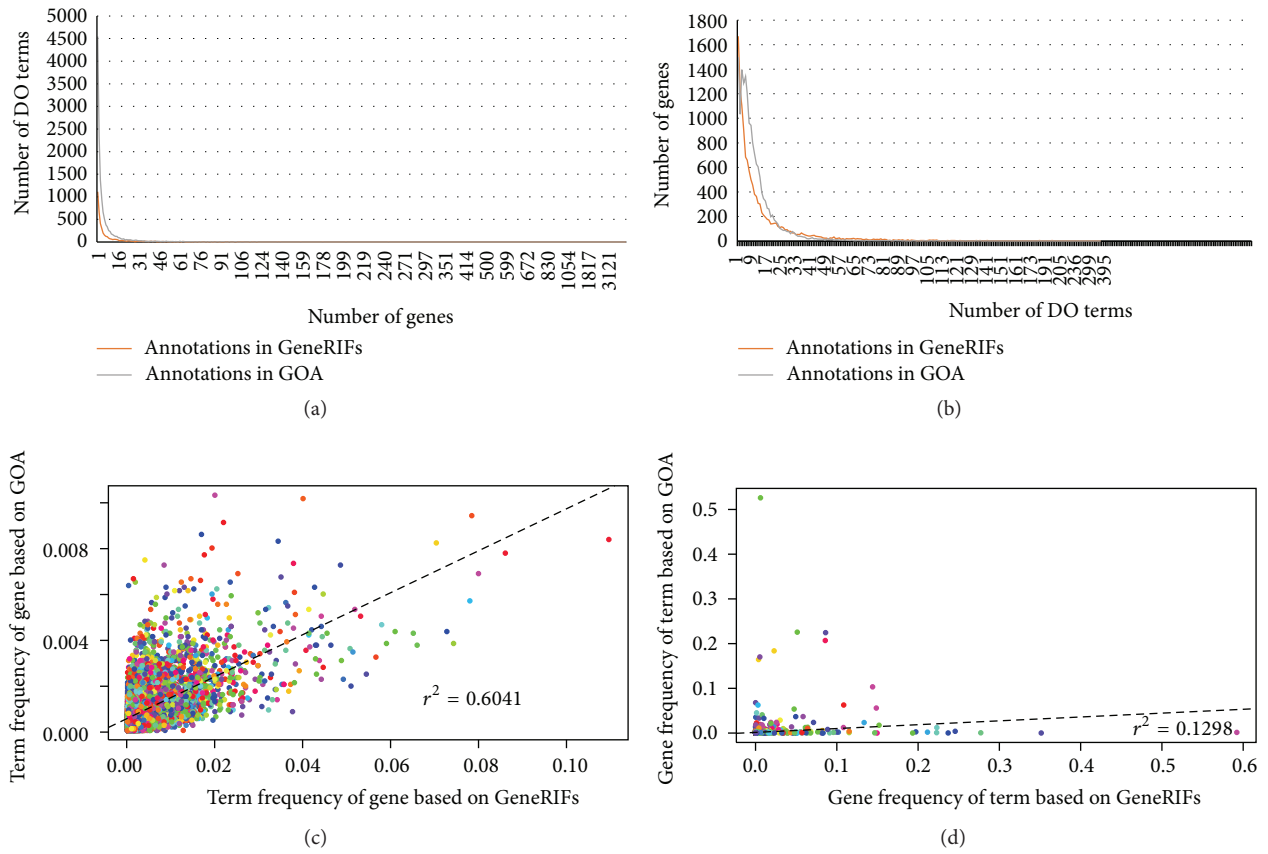


FIGURE 2: The comparison of annotations in GeneRIFs and with annotations in GOA. (a) Histogram of the number of genes associated with individual GO term. (b) Histogram of the number of DO terms associated with individual gene. (c) The correlation between term frequency of gene by GeneRIFs and GOA. (d) The correlation between gene frequency of term by GeneRIFs and GOA.

3.1.3. Ontologies and Annotations. As shown in Figure 4, GO organized BP terms in the Directed Acyclic Graph (DAG) by "IS_A" relationship. Currently, GO contains 55,565 "IS_A" relationships between 28,654 BP terms, 12,375 "IS_A" relationships between 10,159 MF terms, and 5,618 "IS_A" relationships between 3,907 CC terms. GOA was compared with our annotation result. After removing IEA and getting rid of duplicate records of GOA, 168,246 associations between 13,920 GO terms and 16,724 genes remained.

DO is a first ontology to organize terms around human disease, which describes each disease by a unique identifier, a disease name, and its synonymous. In the current version, it involved 7,124 "IS_A" relationships between 6,920 disease terms.

3.2. Method for Annotating Human Genome. As shown in Figure 5(a), we presented a framework, Gene2Function, to annotate the function of human genome. Firstly, a raw text of GeneRIF with functional description should be annotated by a text mining tool named Open Biomedical Annotator (OBA) [37], which provided an ontology-based web service that annotates public datasets with biomedical ontology concepts based on their textual metadata. As a result, the functional description will be mapped to the corresponding ontologies, such as GO and DO. Then, the Entrez gene identifier will be converted into a normalized gene symbol. Here, HGNC was exploited for normalizing and labelling the locus type of gene, such as protein-coding genes, microRNA genes, and lncRNA genes. Finally, each GeneRIF could be annotated to a triple involving gene symbol, locus type, and functional description.

All the GeneRIFs could be annotated based on the annotation framework. Figure 5(b) gives an example of annotating a GeneRIF with GO. "Enzyme activity" is a synonym of "catalytic activity (GO:0003824)," which was identified by OBA. And Entrez gene identifier "9" was converted into "NAT1 (HGNC:7645)" based on HGNC. Through the annotation framework, the annotation triple "mRNA, NAT1, catalytic activity" could be obtained.

4. Discussion

The importance of the functional annotations of genes had been reflected in the previous annotation resource, such as

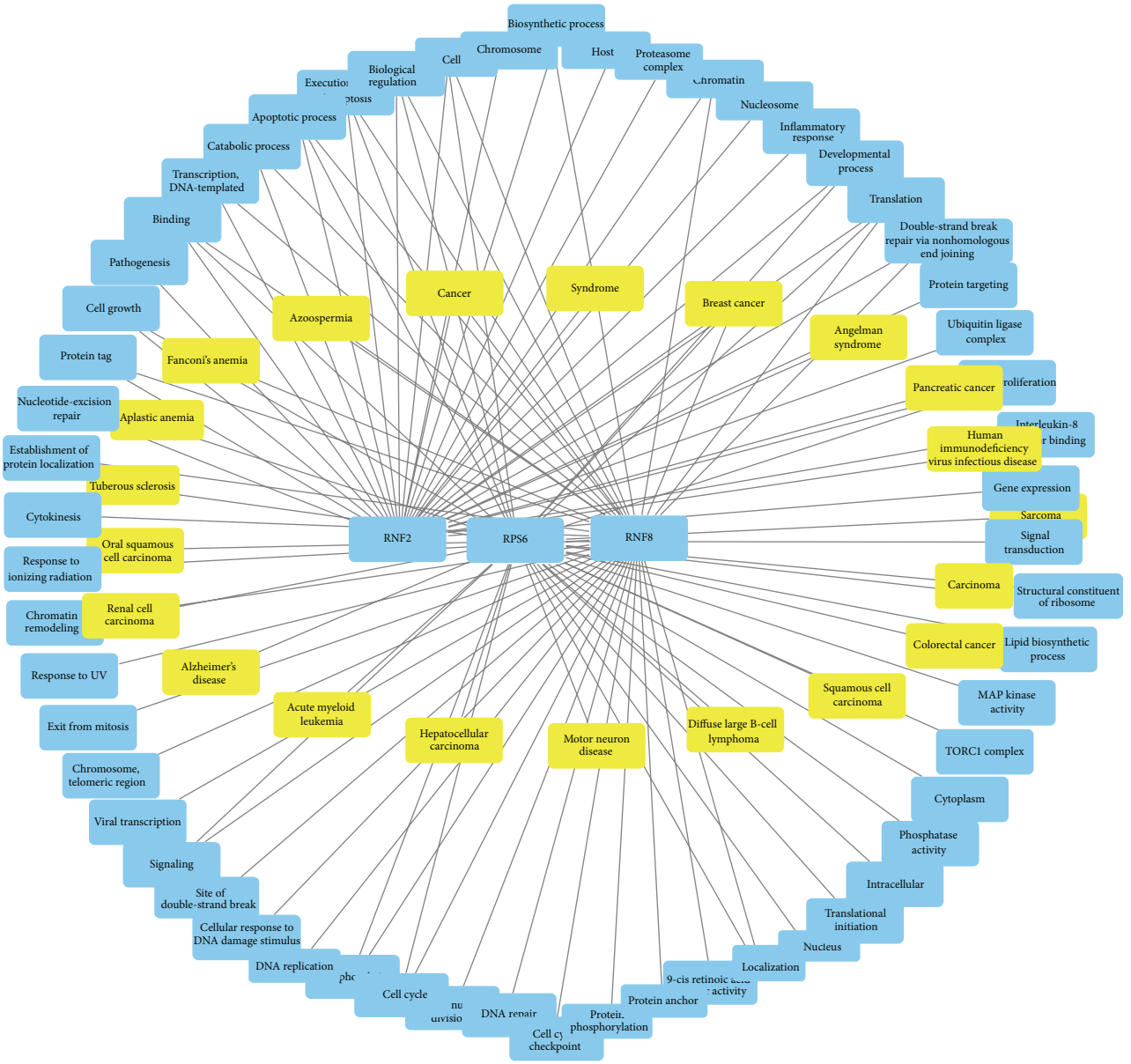


Figure 3: A bipartite network demonstrating the relationship between genes and terms. Rectangles with yellow represent DO terms, three rectangles with blue in the center of the figure indicate DO terms, and other rectangles are GO terms. An edge is placed between a gene and a term of GO and DO if the gene relates with the term.

GOA. Unfortunately, functional annotation resources of noncoding RNA are very few, which lead to the lack of a comprehensive annotation resource involving protein-coding genes, microRNA genes, and lncRNA genes. With the largest number of noncoding genes in the human genome, it is urgent to provide functional annotation of these genes. In this study, we presented a framework, Gene2Function, for annotating GeneRIFs. As a result, a comprehensive functional annotation resource of human genome was obtained based on the framework, which could be accessed at http://www.bio-annotation.cn/gene2function/. To evaluate the reliability, our annotation result was compared with a prevalent resource GOA. Subsequently, a network visualization of connectivity of genes by their functional terms shows the usability of the annotation result.

The annotation framework is based on a text mining tool OBA [37]. Under the framework, the functional terms of descriptions of GeneRIFs were annotated to GO and DO terms. And gene symbols were mapped to a normalized vocabulary of human gene HGNC [36], which makes it easy to distinguish the locus type of gene, such as protein-coding RNA, microRNA, and lncRNA.

The consistency test of the GeneRIFs and GOA (Figures 2(c) and 2(d)) shows the reliability of our annotation result. Because of a small amount of common associations between

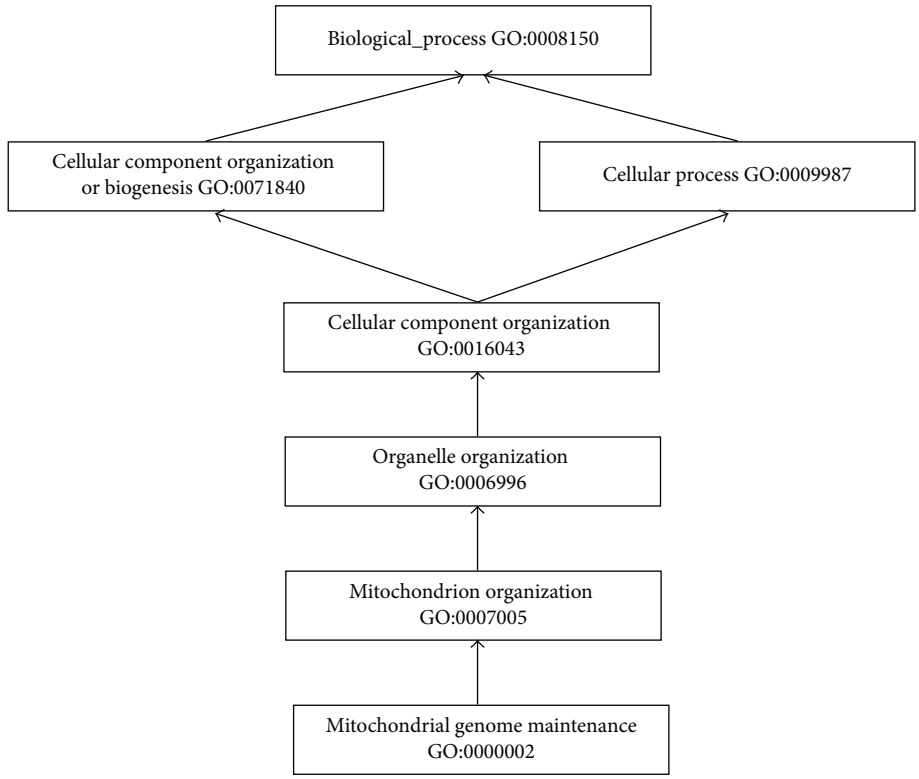


Figure 4: A subgraph of the DAG for BP term "Mitochondrial genome maintenance (GO:0000002)." The arrow symbol represents an "IS_A" link of GO. For example, "Mitochondrial genome maintenance (GO:0000002)" is linked to "Mitochondrion organization (GO:0007005)" by an "IS_A" relationship.

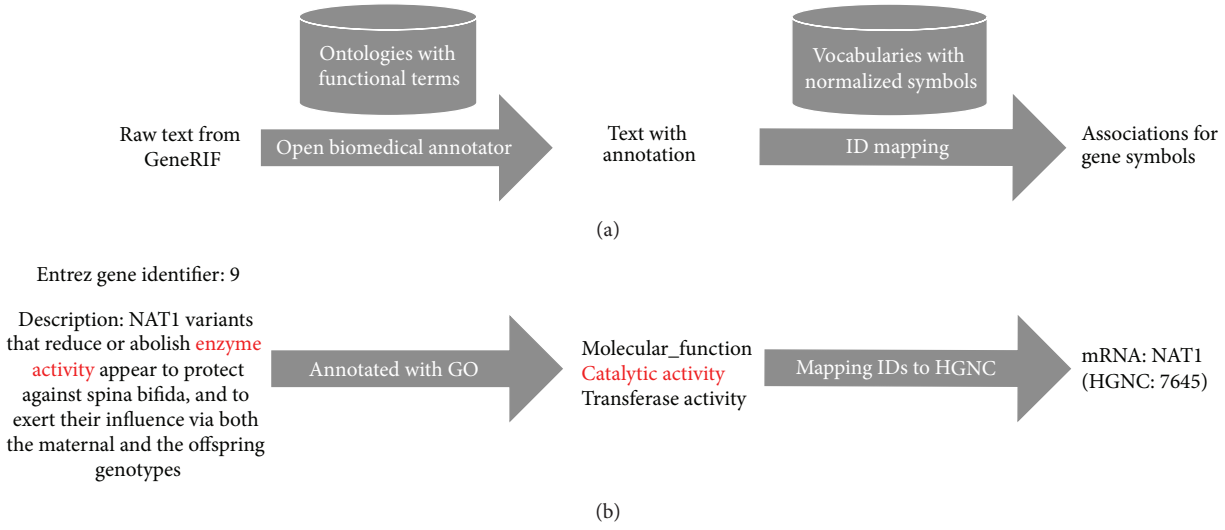


Figure 5: Diagram of functional annotation of human genome. (a) A framework to annotate functional description of human genome to ontologies. (b) An example of annotating a GeneRIF.

genes and GO terms in both annotation resources, they could be complementary in the usage of protein-coding RNA annotation. More GO terms were annotated in GOA (see Section 2) suggesting it is more deep and serious than our annotation results. In comparison, advantage of GeneRIFs is that not only protein-coding genes but also microRNA genes and lncRNA genes could be annotated with GO and other function terms (Table 1).

Competing Interests

The authors declare that they have no competing interests.

References

[1] J. C. Venter, M. D. Adams, E. W. Myers et al., "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

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

[3] S. M. Hollenberg, C. Weinberger, E. S. Ong et al., "Primary structure and expression of a functional human glucocorticoid receptor cDNA," *Nature*, vol. 318, no. 6047, pp. 635–641, 1985.

[4] W. E. Holmes, J. Lee, W.-J. Kuang, G. C. Rice, and W. I. Wood, "Structure and functional expression of a human interleukin-8 receptor," *Science*, vol. 253, no. 5025, pp. 1278–1280, 1991.

[5] R. R. Schumann, S. R. Leong, G. W. Flaggs et al., "Structure and function of lipopolysaccharide binding protein," *Science*, vol. 249, no. 4975, pp. 1429–1431, 1990.

[6] H. Shi, G. Zhang, M. Zhou et al., "Integration of multiple genomic and phenotype data to infer novel miRNA-disease associations," *PLoS ONE*, vol. 11, no. 2, Article ID e0148521, 2016.

[7] M. Zhou, Y. Sun, Y. Sun et al., "Comprehensive analysis of lncRNA expression profiles reveals a novel lncRNA signature to discriminate nonequivalent outcomes in patients with ovarian cancer," *Oncotarget*, vol. 7, no. 22, pp. 32433–32448, 2016.

[8] M. Zhou, W. Xu, X. Yue et al., "Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma," *Oncotarget*, vol. 7, no. 20, pp. 29720–29738, 2016.

[9] H. Shi, J. Xu, G. Zhang et al., "Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes," *BMC Systems Biology*, vol. 7, article 101, 2013.

[10] R. Brain and J. R. Jenkins, "Human p53 directs DNA strand reassociation and is photolabelled by 8-azido ATP," *Oncogene*, vol. 9, no. 6, pp. 1775–1780, 1994.

[11] S. Benzon Larsen, U. Vogel, J. Christensen et al., "Interaction between *ADH1C* $Arg^{272}Gln$ and alcohol intake in relation to breast cancer risk suggests that ethanol is the causal factor in alcohol related breast cancer," *Cancer Letters*, vol. 295, no. 2, pp. 191–197, 2010.

[12] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[13] E. Camon, M. Magrane, D. Barrell et al., "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic Acids Research*, vol. 32, pp. D262–D266, 2004.

[14] W. A. Kibbe, C. Arze, V. Felix et al., "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic Acids Research*, vol. 43, no. 1, pp. D1071–D1078, 2015.

[15] Z. Lu, K. B. Cohen, and L. Hunter, "Generif quality assurance as summary revision," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '07)*, pp. 269–280, January 2007.

[16] J. D. Osborne, J. Flatow, M. Holko et al., "Annotating the human genome with Disease Ontology," *BMC Genomics*, vol. 10, supplement 1, article S6, 2009.

[17] W. Xu, H. Wang, W. Cheng et al., "A framework for annotating human genome in disease context," *PLoS ONE*, vol. 7, no. 12, Article ID e49686, 2012.

[18] K. Peng, W. Xu, J. Zheng et al., "The Disease and Gene Annotations (DGA): an annotation resource for human disease," *Nucleic Acids Research*, vol. 41, no. 1, pp. D553–D560, 2013.

[19] P. Kapranov, A. T. Willingham, and T. R. Gingeras, "Genome-wide transcription and the implications for genomic organization," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 413–423, 2007.

[20] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[21] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[22] Y. Huang, N. Liu, J. P. Wang et al., "Regulatory long non-coding RNA and its functions," *Journal of Physiology and Biochemistry*, vol. 68, no. 4, pp. 611–618, 2012.

[23] J. Sun, H. Shi, Z. Wang et al., "Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network," *Molecular BioSystems*, vol. 10, no. 8, pp. 2074–2081, 2014.

[24] M. Zhou, X. Wang, H. Shi et al., "Characterization of long non-coding RNA-associated ceRNA network to reveal potential prognostic lncRNA biomarkers in human ovarian cancer," *Oncotarget*, vol. 7, no. 11, pp. 12598–12611, 2016.

[25] J. Sun, M. Zhou, H. Yang, J. Deng, L. Wang, and Q. Wang, "Inferring potential microRNA-microRNA associations based on targeting propensity and connectivity in the context of protein interaction network," *PLoS ONE*, vol. 8, no. 7, Article ID e69719, 2013.

[26] Y. A. Huang, X. Chen, Z. H. You, D. S. Huang, and K. C. Chan, "ILNCSIM: improved lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 18, pp. 25902–25914, 2016.

[27] X. Chen, Y. A. Huang, X. S. Wang, Z. H. You, and K. C. Chan, "FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model," *Oncotarget*, 2016.

[28] A. P. Daiwile, S. Sivanesan, A. Izzotti et al., "Noncoding RNAs: possible players in the development of fluorosis," *BioMed Research International*, vol. 2015, Article ID 274852, 10 pages, 2015.

[29] Y.-A. Huang, Z.-H. You, X. Chen, K. Chan, and X. Luo, "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC Bioinformatics*, vol. 17, no. 1, article 148, 2016.

[30] 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, no. 1, pp. D1070–D1074, 2014.

[31] G. Chen, Z. Wang, D. Wang et al., "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic Acids Research*, vol. 41, no. 1, pp. D983–D986, 2013.

[32] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, pp. 265–266, 2000.

[33] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 193–203, 2016.

[34] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.

[35] Q. Zou, J. Li, Q. Hong et al., "Prediction of microRNA-disease associations based on social network analysis methods," *BioMed*

Research International, vol. 2015, Article ID 810514, 9 pages, 2015.

[36] E. A. Bruford, M. J. Lush, M. W. Wright, T. P. Sneddon, S. Povey, and E. Birney, "The HGNC database in 2008: a resource for the human genome," *Nucleic Acids Research*, vol. 36, no. 1, pp. D445–D448, 2008.

[37] C. Jonquet, N. H. Shah, and M. A. Musen, "The open biomedical annotator," *AMIA Summits on Translational Science Proceedings*, vol. 2009, pp. 56–60, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 7562085, 14 pages
http://dx.doi.org/10.1155/2016/7562085

Research Article

Human Ribosomal RNA-Derived Resident MicroRNAs as the Transmitter of Information upon the Cytoplasmic Cancer Stress

Masaru Yoshikawa[1] and Yoichi Robertus Fujii[1,2]

[1]*Pharmaco-MicroRNA Genomics, Graduate School of Pharmaceutical Sciences, Advanced Pharmaceutical Science Center, Nagoya City University, Nagoya 467-8603, Japan*
[2]*Retroviral Genetics Group, Graduate School of Pharmaceutical Sciences, Advanced Pharmaceutical Science Center, Nagoya City University, Nagoya 467-8603, Japan*

Correspondence should be addressed to Masaru Yoshikawa; c162806@ed.nagoya-cu.ac.jp
and Yoichi Robertus Fujii; fatfuji@hotmail.co.jp

Received 18 April 2016; Accepted 19 June 2016

Academic Editor: Xing Chen

Dysfunction of ribosome biogenesis induces divergent ribosome-related diseases including ribosomopathy and occasionally results in carcinogenesis. Although many defects in ribosome-related genes have been investigated, little is known about contribution of ribosomal RNA (rRNA) in ribosome-related disorders. Meanwhile, microRNA (miRNA), an important regulator of gene expression, is derived from both coding and noncoding region of the genome and is implicated in various diseases. Therefore, we performed *in silico* analyses using M-fold, TargetScan, GeneCoDia3, and so forth to investigate RNA relationships between rRNA and miRNA against cellular stresses. We have previously shown that miRNA synergism is significantly correlated with disease and the miRNA package is implicated in memory for diseases; therefore, quantum Dynamic Nexus Score (DNS) was also calculated using MESer program. As a result, seventeen RNA sequences identical with known miRNAs were detected in the human rRNA and termed as rRNA-hosted miRNA analogs (rmiRNAs). Eleven of them were predicted to form stem-loop structures as pre-miRNAs, and especially one stem-loop was completely identical with *hsa-pre-miR-3678* located in the non-rDNA region. Thus, these rmiRNAs showed significantly high DNS values, participation in regulation of cancer-related pathways, and interaction with nucleolar RNAs, suggesting that rmiRNAs may be stress-responsible resident miRNAs which transmit stress-tuning information in multiple levels.

1. Introduction

It has recently been revealed that dysregulation of ribosome biogenesis is implicated in various diseases termed ribosomopathy such as Diamond-Blackfan anemia (DBA), Shwachman-Diamond syndrome (SDS), X-linked dyskeratosis congenita (DKC), Treacher Collins syndrome (TCS), and cartilage hair hypoplasia (CHH) [1–3]. The most studied ribosomopathy, DBA, is a rare congenital hypoplastic anemia and its pathogenesis is associated with defects in various ribosomal protein (RP) genes such as RPS19, RPS24, RPL5, and RPL11. Mutation in RPS and RPL genes results in significant reduction in the amount of mature 40S and 60S subunit, respectively [4]. Other ribosomopathies, SDS, DKC, TCS, and CHH, are caused by gene defects on SBDS, DKC1, TCOF1, and RMRP, respectively, which encode proteins involved in ribosome biogenesis [2]. However, what mechanism is linked to these proteins in the pathogenesis of ribosomopathies? Whether cancer is related to them? These are still unsolved.

Ribosomal RNA (rRNA) is the most abundant noncoding RNA gene in cells and is essential for the structure and function of ribosomes. All four eukaryotic rRNAs, such as 18S, 5.8S, 28S, and 5S, are highly conserved across human and related species, and their biogenesis is strictly regulated by several mechanisms [5–8]. RNA45S, also called RN45S,

the 45S gene, or rDNA, is an operon containing 18S, 5.8S, and 28S RNA genes [8–11]. On the other hand, the 5S RNA gene is coded alone. Among eukaryotes, the RNA45S and 5S RNA genes are transcribed by Pol I and Pol III, respectively [12, 13]. The first step in rRNA gene transcription in humans is the formation of the preinitiation complex (PIC) on the core promoter and the upstream control element of rDNA. PIC attracts Pol I, and a full-length rRNA precursor called 47S rRNA is transcribed. 47S rRNA is processed into 45S rRNA by cleaving fixed positions on the $3'$ and $5'$ external transcribed spacers (ETS) in the nucleus and is then divided into 21S and 32S rRNA by either of the two processes [8]. Finally, 18S-E, 6S, and 28S rRNAs are generated through various mechanisms and transported into the cytoplasm to construct the mature ribosomal complex.

RNA45S genes in humans are located on chromosomes (Chr) 13, 14, 15, 21, and 22 [14]. These acrocentric chromosomes have multiple copies of the 45S RNA gene on the p12 region in their short arms. This tandemly repeated rRNA gene copy is commonly called an rDNA repeat or rRNA gene cluster, and each repeating unit consists of a nontranscribed spacer (NTS) and the RNA45S gene. RNA45S also contains a $5'$ ETS, an internal transcribed spacer (ITS), and a $3'$ ETS in addition to the 18S, 5.8S, and 28S rRNA genes. On the other hand, the 5S rRNA gene is only located at the q42 region of chromosome 1. The copy number of rDNA is important for normal cell functions although the majority of rDNA copies are transcriptionally silent; therefore, reduced rDNA copy number after cell stress is repaired by a specific amplification system. It has also been reported that perturbation in the copy number and stability of rRNA gene caused by mutations in rRNA-related enzymes or cell senescence are linked to various cellular dysfunctions and insufficiency of genome integrity [15–18].

MicroRNA (miRNA) is an essential regulator of gene expression and a member of the small noncoding RNA family, which are RNAs approximately 22 nucleotides long [19]. Sequence complementarity-based interactions between miRNA and its target mRNA suppress and occasionally augment the translation of mRNAs into proteins [20–23]. One miRNA regulates multiple mRNAs; thus, one mRNA is targeted by multiple miRNAs [24–26]. Almost all functional genes in humans are under the control of miRNAs [27]. Therefore, alterations in the miRNA profile after injury, infection, or chemical treatment can alter various functions, such as immunoreactivity, cell proliferation and differentiation, apoptosis, and carcinogenesis [28–32]. The expression profile of miRNA genes is deeply associated with a considerable number of human diseases including cancer [28, 29]. There are several reports about miRNA dynamics after cell stresses, for instance, participation of poly(ADP-ribose) in controlling miRNA activity in the cytoplasm [33]. In the deep insight of miRNA-disease relationship, it needs huge efforts to make complete data for clinical validation of miRNA-mRNA associations in diseases. Therefore, it has been shown that computational analysis is required for miRNA research and increasing number of disease-related miRNA databases and computational analyses have recently been established [34, 35]. The miRNA genes are scattered throughout the genome, and miRNAs are created through many complexed processing pathways [36–38]. Most miRNA genes are transcribed by RNA polymerase II (Pol II) as hairpin-shaped primary miRNA (pri-miRNA), and the pri-miRNA is processed into pre-miRNA after cleavage of the $5'$-cap and $3'$-polyA tail by the microprocessor complex, which is composed of Drosha and DiGeorge syndrome chromosomal region 8 (DCGR8). These are the RNase III proteins and double-stranded RNA binding proteins, respectively. Subsequently, pre-miRNA is exported to the cytoplasm by exportin-5 and further processed into the miRNA: miRNA* duplex by cleavage of the $5'$- and $3'$-termini and loop domain by Dicer, which is an RNase III-like protein. This duplex is finally loaded into the RNA-induced silencing complex, and a duplex chain is selected thermodynamically to function as mature miRNA [39]. However, some noncanonical pathways are used to mature miRNA [36, 40]. For example, dme-mir-1003 is the first discovered mirtron, which is a pri-miRNA that exists as an intron of pre-mRNA and is processed into pre-miRNA without the Drosha canonical processor [41]. This means that all protein-coding, noncoding, intergenic, and intragenic regions can become miRNA hosts.

According to the RNA wave 2000 model advocated by Fujii, miRNA genes are the RNA information genes with four critical characteristics: (1) the miRNA gene is a mobile genetic element that induces transcriptional and posttranscriptional silencing via networking processes; (2) the RNA information supplied by miRNA genes expands to intracellular, intercellular, intraorgan, interorgan, intraspecies, and interspecies under a lifecycle in the global environment; (3) mobile miRNAs self-proliferate; and (4) cells contain resident and genomic miRNAs [42, 43]. miRNAs can be classified into genomic and resident miRNAs. The former are miRNAs preserved in DNA as miRNA genes, and the latter are miRNAs stored in a non-DNA form. The greatest difference between genomic and resident miRNAs is the expression regulatory mechanism. Most known miRNAs are genomic because their expression levels are controlled by a specific transcriptional factor, RNA polymerase, and so forth [44–46]. However, some miRNAs, such as mmu-miR-712, dme-miR-10404, and hsa-miR-663, are typical resident miRNAs because they do not require specific transcriptional factors or nucleases to exert their functions [47, 48].

In particular, it is anticipated that resident miRNAs and other cytoplasmic RNAs play more important roles in cells with unique cytoplasmic or genomic characteristics, such as erythrocytes, spermatozoa, and oocytes, than those of other cells. Erythrocytes contain diverse and abundant RNA species, including cytoplasmic miRNAs that contribute to regulating erythropoiesis and malarial resistance, although erythrocytes have been thought to contain no RNA because they are anucleated [49, 50]. Given that erythrocytes are the most abundant cell in blood, a large number of erythrocyte-contained miRNAs may be circulating. Spermatozoa are characterized by minimal cytoplasm and extremely condensed DNA. However, various RNAs are abundant in the cytoplasm of spermatozoa, such as rRNA, transfer RNA (tRNA), piwi-interacting RNA, and miRNA, and have important roles before and after fertilization [51–53]. Oocytes

are transcriptionally silent cells; therefore, the many pooled mRNA and noncoding RNAs in the cytoplasm, such as miRNAs, are essential to complete late oogenesis and early embryogenesis without de novo transcription [53, 54]. Only the resident RNAs in these cells are considered information transmitters or memorizing devices, rather than DNA. Furthermore, as miRNA is self-reproducible, an identical miRNA could become both genomic and resident miRNA [42]. The quantities of tRNA and rRNA decrease under stress, suggesting that resident miRNAs help with biological regulation under stress [55]. Thus, we hypothesized that cytoplasmic tRNA and/or rRNA is a pool of self-reproducible resident miRNAs.

tRNA is another functional noncoding RNA that is most abundant (approximately 10% of RNAs) in cells next to rRNA (approximately 80% of RNAs). Recent studies have discovered that transfer RNA-derived RNA fragments (tRFs) are generated from tRNAs as terminal functional products. In the case of murine gammaherpesvirus 68, viral miRNAs were generated by Pol III [56]. Further, a number of endoribonucleases including Dicer and Angiogenin are implicated in the production of tRFs from tRNA transcripts [57, 58]. tRFs exist in various species, such as humans, cows, flies, and plants, and work as gene expression regulators, similar to miRNA [57, 59–63]. Some tRFs were listed as miRNAs in the miRBase (now dead entries). Other common characteristics between miRNAs and tRFs are their interactions with Argonaute (AGO) proteins, significant changes in expression levels during disease and aging, and circulation in a steady form [61–63]. Several reports have shown that tRFs are occasionally more abundant than miRNAs [64].

We considered the possibility that RNA fragments may be derived from rRNA in a manner similar to how tRFs are derived from tRNA because several tRF-related endoribonucleases have common activity of nuclease [58]. Till date, to the best of our knowledge, only a few studies have reported biogenesis and functions of rRNA-derived miRNAs or miRNA-like fragments, although many rRNA-annotated fragments of miRNA-like size have been detected in deep sequencing data from RNA studies [65]. Chak et al. revealed that the novel miRNA hairpin named mir-10404/mir-ITS1 exists in the ITS1 region of *Drosophila* rDNA [47]. Son et al. also discovered that mmu-miR-712 is coded in ITS2 of mouse 45S precursor RNA (Rn45s) and hsa-miR-663 is coded in the ITS1 region of human RNA45S [48]. Furthermore, Drosha-related proteins are included in rRNA processing pathways [66]. These ITS-derived miRNAs are supposed to be generated upon degradation of the ITS region, similar to the generation of mirtrons in the nucleoplasm or cytoplasm.

The effects of tRNA or rRNA degradation and processing on cell activities in response to stress are important. The small RNA molecules derived through this process play an important role in the transition from fine-tuning to stress-tuning functions. Other ncRNA species such as SINE, especially human Alu elements, have also been revealed to be contained in nucleolus and control the size of nucleoli adopting to cell circumstances [67].

Therefore, we examined whether rRNAs contain functional small RNAs and confirmed the relationship between ribosome and disease shown in previous studies. Moreover, how rRNA-hosted microRNA analogs (rmiRNAs) contribute to the stress response as nongenomic memory in the nucleolus and cytoplasm was also investigated using multiple computer-based tools and databases to find stress-tuning RNA interaction in transcriptional and posttranscriptional level. The quantum relationships among miRNAs were also calculated as Dynamic Nexus Score (DNS) by MESer program that we have previously developed and its significance in stress response was discussed.

2. Method

2.1. Sequence Data Collection. All miRNA sequence data used in this study were downloaded as miRNA.dat, hairpin.fa, and mature.fm from miRBase (http://www.mirbase.org) in release 21 (June 2014) [68]. This includes 2,588 and 1,915 mature miRNA sequences of human and mouse, respectively. Sequences of rRNAs were obtained from European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) release 127 (April 6, 2016) and National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) in FASTA format [69]. After comparing and merging latest rRNA sequence data, two rRNA coding sequences, RNA45S and human rDNA complete repeating unit, were selected as the source of rRNA sequence (Supplemental Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/7562085). The latter source involves the sequence of the former but their sequences have some differences even in common regions, for instance, slightly polymorphisms in 18S and moderate ones in 28S, ITS, and ETS region. Sequences of tRNA and tRF were also obtained from Genomic tRNA database (GtRNAdb, http://gtrnadb.ucsc.edu) and tRFdb (http://genome.bioch.virginia.edu/trfdb), respectively [70, 71].

2.2. Definition of Passenger Strand. Passenger strands of miRNAs whose guide strands were found in the rRNA sequences were researched referring to stem-loop structure in miRBase. If a passenger strand is not recorded in miRBase, a sequence which is complement to the guide strand was defined as the passenger strand in this study.

2.3. Secondary Structure Prediction. To determine the secondary structures of found miRNA-like sequences, M-fold was used in a condition of 37 Celsius degrees and 1 M NaCl. Any other options which influence prediction results were set in default (RNA sequence is linear, percent suboptimality number is 5, upper bound on the number of computed foldings is 50, the window parameter is default, the maximum interior/bulge loop size is 30, the maximum asymmetry of an interior/bulge loop is 30, and the maximum distance between paired bases is no limit).

2.4. Chromosome Confirmation. For browsing miRNA locations on each chromosome visually, UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38) was used.

2.5. Calculations of DNS. Dynamic Nexus Score (DNS) was prepared as a quantum-based score for evaluating quantum

Table 1: Detected mature miRNAs from rRNA gene and adjacent region.

miR name	Mature sequence	Region	Location
miR-663a	AGGCGGGGCGCCGCGGGACCGC	5′ ETS	2049–2071
miR-663b	GGUGGCCCGGCCGUGCCUGAGG	5′ ETS	2113–2135
miR-1268a	CGGGCGUGGUGGUGGGGG	3′ ETS	13102–13119
miR-1268b	CGGGCGUGGUGGUGGGGGUG	3′ ETS	13102–13121
miR-1275	GUGGGGGAGAGGCUGUC	(NTS)	42294–42310
miR-3648	AGCCGCGGGGAUCGCCGAGGG	5′ ETS	2513–2533
miR-3656	GGCGGGUGCGGGGGUGG	28S	8524–8540
miR-3687	CCCGGACAGGCGUUCGUGCGACGU	(5′ ETS)	2888–2911
miR-4417	GGUGGGCUUCCCGGAGGG	5′ ETS	2412–2429
miR-4466	GGGUGCGGGCCGGCGGGG	(5′ ETS)	631–648
miR-4488	AGGGGGCGGGCUCCGGCG	28S	8510–8527
miR-4492	GGGGCUGGGCGCGCGCC	28S	10851–10867
miR-4508	GCGGGGCUGGGCGCGCG	28S	10849–10865
miR-4516	GGGAGAAGGGUCGGGGC	28S	11049–11065
miR-4532	CCCCGGGGAGCCCGGCG	28S	11227–11243
miR-6087	UGAGGCGGGGGGGCGAGC	28S	12007–12024
miR-6724	CUGGGCCCGCGGCGGGCGUGGGG	(NTS)	42320–42342

Seventeen sequences homologous to mature human miRNAs were detected from rRNA gene coding region. Note that miR-1268a and miR-1268b were found in only RNA45S and miR-1275, miR-3687, miR-4466, and miR-6724 were found in only rDNA-repeating unit. This data might be caused by differences in base alignment between two rRNA sequence data.

interactions between or among miRNAs [72]. DNS calculation of rRNA-derived miRNA and tRF was performed by using the original program, MESer (http://meser.mirna-academy.org). Computational results were statistically analyzed with Microsoft Office Excel 2013 (Microsoft Japan Co., Ltd., Tokyo, Japan).

2.6. Target Prediction and Ontology Analysis. Putative targets of rmiRNAs were predicted under the seed theory by using TargetScan (http://www.targetscan.org/) [73]. Validated targets of rmiRNAs were confirmed in miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/) [74]. Selection of top 10 targets in miRTarBase was conducted by referring to the number of validation methods (primary) and the number of reports (secondary). Categorization of putative target genes in Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways was accomplished by using GeneCoDis3 web service (http://genecodis.cnb.csic.es/) [75].

2.7. Alu Sequence and Target Site Prediction. Sequence data of human Alu family were downloaded from SINE Base (http://sines.eimb.ru), last update May 28, 2015 [76]. Target sites of rmiRNAs in Alu sequences were predicted using RNA22 version 2.0 (https://cm.jefferson.edu/rna22/) with default settings (sensitivity of 63%, specificity of 61%, seed size of 7, allow maximum of 1 UN-paired bases in seed, minimum number of paired-up bases in heteroduplex being 12, maximum folding energy for heteroduplex being −12 Kcal/mol, and maximum number of G : U wobbles allowed in seed region being no limit) [77].

3. Results

3.1. Pre-miRNA Sequence in Human rRNA. To investigate whether miRNAs also exist in human rRNA, we firstly collected base sequences of pre-miRNAs, mature miRNAs, and rDNA. Then the sequence of rRNA and its adjacent regions, RNA45S, and rDNA-repeating unit, respectively, were searched for 2,588 human pre-miRNA sequences by using a simple C++ based detection program we developed for this study. As a result, an identical sequence to pre-miR-3687 was detected from rDNA-repeating unit although the known location of the miR-3687 gene was distinct from rRNA coding region. However, other 2,587 pre-miRNA sequences were not found in any rRNA-related sequences. For further similar sequencing research, a detection of mature miRNA sequences instead of pre-miRNA from rRNA gene was also performed. Subsequently, seventeen RNA alignments identical to human mature miRNAs, namely, miR-663a, miR-663b-3p, miR-1268a, miR-1268b, miR-1275, miR-3648, miR-3656-3p, miR-3687-3p, miR-4417, miR-4466, miR-4488, miR-4492-3p, miR-4508, miR-4516, miR-4532, miR-6087, and miR-6724, were detected from the human rRNA sequences (Table 1). In detail, miR-1268a and miR-1268b were detected from only RNA45S, and miR-1275, miR-3687, and miR-6724 were detected from only rDNA-repeating unit.

Among these detected miRNAs, miR-1268a, miR-3648, miR-3687, miR-4508, and miR-6724 were originated in rDNA containing chromosomes, Chr 15, Chr 21, Chr 21, Chr 15, and Chr 21, respectively. However, their locations were different from rRNA coding regions (data not shown). This suggests that the detected miRNAs might also have been transcribed

from rRNA-related regions as well as above Chr loci or that the miRNAs may be generated through further processing of transcribed rRNA gene.

To further examine whether the detected miRNAs could form miRNA/miRNA* duplex and/or stem-loop structure like miR-3687, their passenger strand sequences were also searched in the rDNA repeat region. Passenger strand sequences were determined referring to their putative stem-loop structure in miRBase. As a result, passenger strands of miR-663a, miR-3648, miR-3687, miR-6087, and miR-6724 were found nearby their guide strands (Figure 1(a)). This result indicated that these rRNA-hosted miRNA-like RNAs could form stem-loop or at least miRNA/miRNA* duplex. Intriguingly, a passenger strand of miR-3687 was detected from RNA45S although its guide strand was not detected. Since these rRNA-hosted RNA pieces, especially miR-3687, possess high concordance rate to each known pre-miRNA sequence, therefore we termed them rRNA-hosted miRNA analog (rmiRNA).

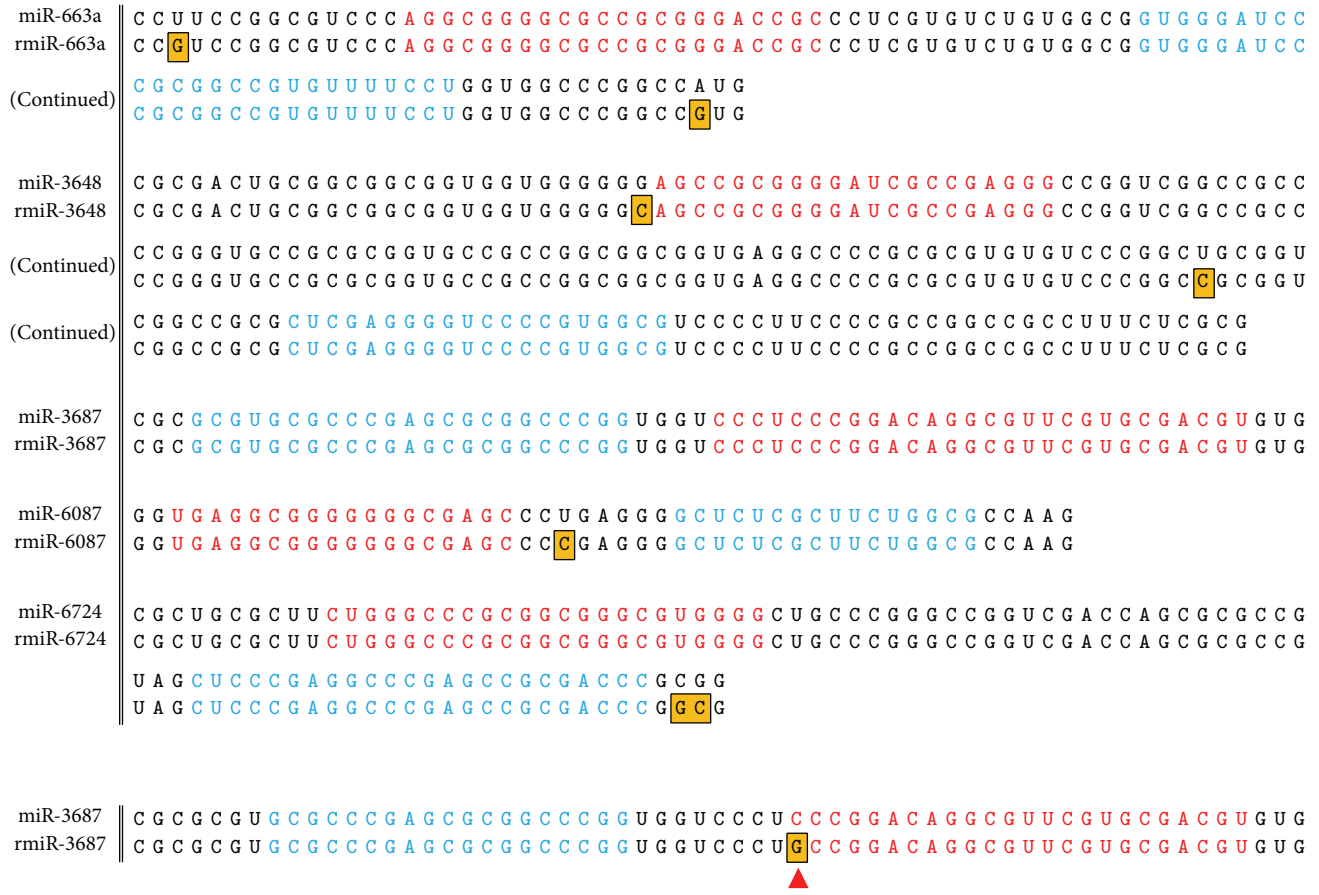
For more rigorous verification, putative precursor sequences of rmiRNAs were predicted by referring to the sequences and structures of known human pre-miRNAs identical to detected rmiRNAs (Figure 1(a)). Subsequently, secondary structures of pre-rmiRNA sequences were predicted by using M-fold software. The same prediction for canonical pre-miRNA sequences were also performed and used as positive controls for comparison, and it was proven that all of five pre-rmiRNA candidates could form hairpin-loop structures which have high similarities to that original pre-miRNAs form (Figure 1(b)).

3.2. Exploration for Noncanonical Passenger Strands and Precursors. Since one pre-rmiRNA sequence is identical with pre-miR-3687 and four pre-rmiRNAs which have high similarity to known pre-miRNAs were detected in rDNA, we thought that other twelve rmiRNAs also could form stem-loop structure with different style. To examine this hypothesis, we have carefully investigated adjacent regions of detected guide strands. Primarily, some RNA sequences were clipped out as putative pre-rmiRNA. Each of them contained guide strand sequence and had the same length to its canonical pre-miRNA. Next, these putative pre-rmiRNAs were compared with its canonical pre-miRNA in base sequences and then secondary structures. Of twelve putative pre-rmiRs, pre-rmiR-663a showed the highest similarity to pre-miR-663b in both base sequence and precursor structure (Supplemental Figure 1A). Other three pre-rmiRNAs, pre-rmiR-3656, pre-rmiR-4417, pre-rmiR-4466, and pre-rmiR-4508 also showed high similarity in secondary structures to pre-miR-3656, pre-miR-4417, pre-miR-4466, and pre-miR-4508, respectively, although their precursor sequences showed low similarities to canonical ones (Supplemental Figure 2A). These results implied that these rmiRNAs have obtained new passenger strand to maintain their function as mature miRNAs. On the other hand, putative pre-rmiR-1268b sequence generated in accordance with the rules above did not form stem-loop structure according to M-fold prediction. However, we found that pre-rmiRNA-1268b could construct stem-loop structure with a slight modification such as lengthening of the 3′ terminal region (Supplemental Figure 1A).

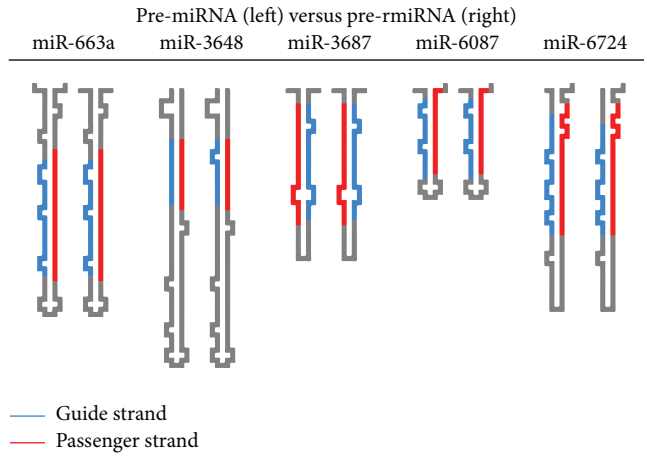
Furthermore, it was ascertained that the left six rmiRNAs, namely, rmiR-1268a, rmiR-1275, rmiR-4488, rmiR-4492, rmiR-4516, and miR-4532, also could form stem-loop structure by further modification. We conceived an idea of "reversed pattern" of primary structure; for instance, miR-1275 usually exists as 5p sequence in pre-miR-1275 but might exist as 3p sequence in pre-rmiR-1275. To examine this idea, broader region analysis was performed and some new candidate rmiRNA sequences were predicted (Supplemental Figure 1B). As a result, it was confirmed that all of new rmiRNAs can form well-ordered stem-loop structure (Table 2 and Supplemental Figure 2B).

3.3. DNS Computation and Comparison. Because concordance of so many sequences must not be detected accidentally, it is natural to consider hidden mechanisms on the background. We previously developed a quantum-based score, Dynamic Nexus Score (DNS), to evaluate miRNA/miRNA interactions and demonstrated that biological activity of the miRNA synergy is positively correlated with DNS value. The average DNS value among mature rmiRNAs was calculated through MESer computer program. DNSs of 1,032 human tRFs and all of 2,588 human miRNAs were also calculated as controls. Surprisingly, the average DNS of rmiRNAs marked 130.23; it was much higher than that of tRFs (40.76) and all human miRNAs (38.31) (Supplemental Figures 3A and 3B). Additionally, DNS values between tRFs, rmiRNAs, and all miRNAs were also calculated. As a result, it was confirmed that the miRNA pairs including rmiRNAs had relatively high DNS values (Supplemental Figure 3C), and this meant that rmiRNAs might induce miRNA-miRNA synergy to accelerate their biofunctions.

3.4. Target Prediction and Ontology Analysis. To investigate targets of rmiRNAs, we used TargetScan and miRTarBase. TargetScan was used for collecting putative target genes predicted by the seed theory-based algorithm; in contrast, miRTarBase was used for collecting experimentally validated target genes. In this experiment, we focused on top 5 high DNS of rmiRNAs, namely, rmiR-1268, rmiR-3656, rmiR-4466, rmiR-6087, and rmiR-6724, and these rmiRNAs were located at separated regions of the rDNA-repeating unit, such as 5′ ETS, 28S rRNA, 3′ ETS, and NTS. Top 10 targets of top 5 DNS rmiRNAs (total 50 targets) were extracted from both TargetScan and miRTarBase (Supplemental Table 2); subsequently, their classification in GO biological process (BP), molecular function (MF), and cellular component (CC) were performed and their results were listed through GeneCodis3 web tool (Figures 2(a) and 2(b)). Intriguingly, three gene ontology (GO) terms, namely, nucleus (CC), protein binding (MF), and nucleotide binding (MF), were commonly ranked on top 3 place between TargetScan and miRTarBase. Moreover, almost all their biofunctions were commonly related to gene regulation such as transcription and nucleotide binding although the greater parts of the GO analysis results were different in detail. Contributions of total 50 targets in biological pathway were also analyzed



(a)



(b)

FIGURE 1: An overview of pre-rmiRNAs with high similarity to canonical pre-miRNAs. (a) Sequence comparison between detected precursor-rmiRNAs (pre-rmiRNAs) and their canonical pre-miRNAs. Mature (guide) miRNA strand is colored with red and passenger is colored with blue. Differences in base sequences between pre-rmiRNAs and pre-miRNAs are highlighted in yellow and box lines. (b) Comparison of secondary structure of pre-rmiRNAs with that of their canonical pre-miRNAs. These rmiRNAs contain the identical sequences to canonical ones in both guide and passenger. This group contains only a few polymorphisms in loop and terminal region.

Table 2: A list of all determined pre-rmiRNA sequence and its location.

rmiR name	Sequence of pre-rmiRNAs	Region	Location
rmiR-663a	CCGUCCGGCGUCCCAGGCGGGGCGCCGCGGGACCGCCCUCGUGUCUGUGGCGGUGGGAUCCCGCGGCCGUGUUUUCCUGGUGGCCCGGCCGUG	ETS	2028–2119
rmiR-663b	GGGGCCGAGGGCCGUCCGGCGUCCCAGGCGGGGCGCCGCGGGACCGCCCUCGUGUCUGUGGCGGUGGGAUCCCGCGGCCGUGUUUUCCUGGUG-GCCCGGCCGUGCCUGAGGUUUC	ETS	2025–2140
rmiR-1268a	CUUCCUCCCUCCCGGCCUCUCCCGCCGACCGCGGGCGUGGUGGUGGGGGU	3′ ETS	13071–13123
rmiR-1268b	CCGCGGGCGUGGUGGUGGGGGUGUGGGGGGGAGGGCGCGCGACCCCGGUCGGCGCGCCCCGCUUC	3′ ETS	13099–13163
rmiR-1275	AGCCCGGCUGGCCCGGUGGCGCCAGAGCUGUGGCCGGUCGCUUGUGAGUCACAGCUCUGGCGUGCAGGUUUAUGUGGGGGAGAGGCUGUCGCU	(NTS)	42221–42313
rmiR-3648	CGCGACUGCGGCGGCGGUGGUGGGGGCAGCCGCGGGGAUCGCCGAGGGCCGGUCGGCCGCCCCGGGUGCCGCGCGGUGCCGCCGGCGGCGGUG-AGGCCCCGCGCGUGUGUCCCGGCCGCGGUCGGCCGCGCUCGAGGGGUCCCCGUGGCGUCCCCUUCCCCGCCGGCCGCCUUUCUCGCG	ETS	2486–2665
rmiR-3656	CUCCCUUCCCCCGCCGCCCCUCCUCCUCCUCCCCGGAGGGGGCGGGCUCCGGCGGGUGCGGGGGUGGGC	28S	8474–8542
rmiR-3687	CGCGCGUGCGCCCGAGCGCGGCCCGGUGGUCCCUCCCGGACAGGCGUUCGUGCGACGUGUG	(ETS)	2854–2914
rmiR-3687*	CGCGCGUGCGCCCGAGCGCGGCCCGGUGGUCCCUGCCGGACAGGCGUUCGUGCGACGUGUG	ETS	2857–2917
rmiR-4417	GCGUGGGGCCCGGUGGGCUUCCCGGAGGGUUCCGGGGGUCGGCCUGCGGCGCGU	ETS	2400–2454
rmiR-4466	UCGCGGGUGCGGGCCGGCGGGGUCCUCUGACGCGGCAGACAGCCCUGCCUGUCG	(ETS)	627–680
rmiR-4488	CCGCCCUCCCUUCCCCCGCCGCCCCUCCUCCUCCUCCCCGGAGGGGGCGGGCUCCGGCGGGUGCGGGGGUGGGCGG	28S	8468–8544
rmiR-4492	GGGGCGCGAAGCGGGGCUGGGCGCGCGCCGCGGCUGGACGAGGCGCCGCCGCCCCCCCCACGCCCGGGGCACCCCCCUCGCGGCCC	28S	10838–10924
rmiR-4508	GGCGCGAAGCGGGGCUGGGCGCGCGCCGCGGCUGGACGAGGCGCCGCCGCCCCCCCCACGCCCGGGGCAC	28S	10841–10910
rmiR-4516	CCGUCCUCCCCCCUCCCCGGGGGAGCGCCGCGUGGGGGCGGCGGCGGGGGGAGAAGGGUCGGGGCGG	28S	11001–11067
rmiR-4532	GACGCGAGCCGGGCCCUUCCCGUGGAUCGCCCCAGCUGCGGCGGGCGUCGCGGCCGCCCCCGGGGAGCCCGGCGGGCGCCGGCGC	28S	11169–11254
rmiR-6087	GGUGAGGCGGGGGGGCGAGCCCCGAGGGGCUCUCGCUUCUGGCGCCAAG	28S	12005–12052
rmiR-6724	CGCUGCGCUUCUGGGCCCGCGGCGGGCGUGGGGCUGCCCGGGCCGGUCGACCAGCGCGCCGUAGCUCCCGAGGCCCGAGCCGCGACCCGGCG	(NTS)	42310–42401

Note that some of them are overlapping each other. *rmiR-3687 indicates the sequence identical with pre-miR-3687 except for a point mutation in the guide sequence (see the lower part of Figure 1(a)).

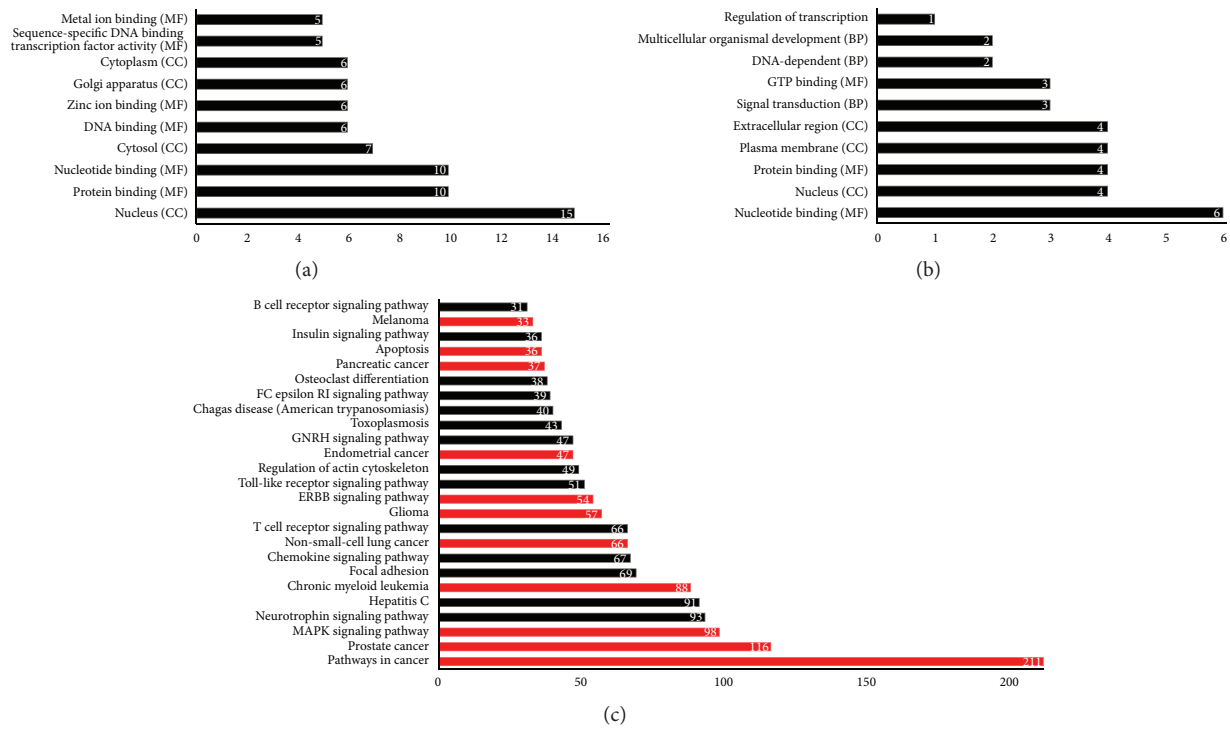


FIGURE 2: GO and KEGG analysis for predicted targets of rmiRNAs. (a) GO characterization of top 10 targets of the top 5 high DNS rmiRNAs in miRTarBase. (b) GO characterization of top 10 putative targets of the top 5 high DNS rmiRNAs in TargetScan. (c) KEGG pathway annotations of putative target genes having less than −0.1 cumulative weighted context++ score in TargetScan. Cancer and cancer-related pathways were colored with red.

using GeneCodis3 with Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway option. However, no pathway was presented in both cases of predicted target (TargetScan) and validated targets (miRTarBase). Therefore, to investigate in larger scale, we extracted all predicted targets in TargetScan having more than 0.1 cumulative weighted context++ score. KEGG pathway analysis of these targets was conducted and various biological pathways were successfully indicated. The results showed that the majority of putative targets were related to cancer or cancer-related pathways such as MAPK signaling and ERBB signaling (Figure 2(c)). This suggests that rmiRNAs have an inclination to target cancer-related genes and might have some important roles in anticancer or antistress pathways.

3.5. Prediction of rmiRNA Targets in Alu. Numerous Alu element-containing RNAs exist in the nucleolus and participate in the synthesis of rRNAs. Therefore, to seek RNA-RNA relationship between nucleolar function and rmiRNA, target sites of rmiRNAs inside Alu sequence were searched. Six branched members of Alu family, namely, Alu consensus, AluJo, AluSz, AluSc, AluSp, and AluY, were processed using RNA22 tool and then 5, 8, 7, 5, 4, and 3 putative rmiRNA target sites were detected, respectively (Supplemental Table 3). Several rmiRNA target sites were conserved among these Alu sequences. Nine of 17 rmiRNAs have potent target sites on Alu sequences and 3 of 5 high DNS rmiRNAs such as miR-4466, miR-6087, and miR-6724-5p were included.

4. Discussion

Ribosomes, large ribonucleoprotein complexes composed of various RPs and rRNAs, are a molecular machine that translate mRNAs into proteins and exists in all living cells [78, 79]. Proper function of the ribosome is essential for normal cell activities; therefore, modification and assembling of RPs and rRNAs are strictly and intricately regulated in the ribosome construction process [78–80]. Ribosomal dysfunction is associated with various diseases represented by ribosomopathies such as DBA, TCS, and SDS and is caused by mutation in ribosome-related genes, ribosomal haploinsufficiency, cellular stresses from chemical or infection, and so forth [1–3]. Several ribosomopathies increase the risk of carcinogenesis and cancer cells often have abnormality in the ribosome function due to mutation in RP and ribosome-related processor genes that cause ribosomopathies [4].

The synthesis of the ribosome itself largely contributes to malignant cell proliferation [81–83]. Ribosomal biogenesis is generally upregulated in the G1 phase dividing cells because enhanced protein synthesis is required to produce viable daughter cells [84]. Thus, inhibiting ribosomal synthesis causes G1 phase arrest and hinders cell proliferation reversibly in normal cells [85]. The upregulation of ribosomal synthesis is also observed in various cancer cells [82, 83]; therefore, inhibiting ribosomal synthesis has been recognized as a potent and novel anticancer strategy [81, 86–88]. This method particularly showed apoptosis-inducing effects in

various malignant cells that synthesized ribosomes at a high rate and exhibited sufficient efficacy with only 3 h transient treatments. This inhibition was accomplished by deleting ribosomal protein genes or by cell treatments with actinomycin D, doxorubicin, 5-fluorouracil, and CX-5461, which have significant tumor suppressing effects [85, 88, 89]. However, this antitumor effect of inhibiting ribosomal biogenesis is not dependent on suppressing protein synthesis but on ribosomal biogenesis itself [81]. This anticancer mechanism presumably depends on stabilizing p53 by inhibiting a p53 degrading protein, called Murine Double Minute 2, by competitively combining with a ribosomal protein, which becomes free because it no longer participates in the construction of ribosomes [81, 88]. Additional evidences declaring associations between ribosome and cellular dysfunction have been reported. 28S rRNA has been identified as a novel fusion partner of carcinoma-related genes such as BCL6, BCL11B, IGKV3-20, and COG1 in gastric lymphoma or hematopoietic tumors [90, 91]. Mutation or inhibition of specific genes associated with ribosome construction pathway such as DKC1, AROS, and several snoRNAs have been identified to impair normal cell functions and sometimes cause carcinogenesis in various cell types [92–95]. Angiogenin-mediated rRNA transcription has been revealed to be related to squamous cell carcinoma [96]. Almost all these ribosome-related genes are considered as potent therapeutic targets for ribosomopathies.

Similar to rRNA, tRNA is the second most abundant noncoding RNA and contains many miRNA-like fragments as tRFs. Various tRFs are generated from mature or pre-tRNAs by ordered cleavage processing and they have similar characteristics to miRNAs, such as evolutionary conservation, target RNA recognition, translational regulation, circulation, and interaction with AGO proteins [57, 59–63, 97, 98]. In addition, miRNAs derived from pre-tRNA or tRNA–miRNA encoded by tRNA genes have been reported [59, 99]. Therefore, it is appropriate that miRNAs or miRNA-like functional fragments are also generated from rRNA gene-related regions.

As a sequel to computational analysis, we found that 17 pre-miRNA-like arrays, which contained identical sequences to known human mature miRNAs, were located in the rRNA gene coding region; therefore, we called them rDNA-hosted pre-miRNA analogs (rmiRNA). We also performed supplemental examination with mouse miRNAs and rDNA and found several sequences which are identical to miR-696, miR-712-5p, miR-712-3p, miR-714, miR-466i-5p, miR-5099, and miR-6538. Of these miRNAs, miR-696, miR-712, and miR-714 have already been reported as rRNA-locating ones [47]. According to Son et al., mmu-miR-712 is located at ITS2 of mouse rRNA gene (Rn45s) and hsa-miR-663 is located at ITS1 region of human rRNA gene (RNA45S) [48]. In our research, miR-712 was likewise found in ITS2 of Rn45s; however, hsa-miR-663 was not found in ITS1 but $5'$ ETS of RNA45S. This discrepancy may be caused by the differences in detection tool or the version of RNA45S sequence data. For instance, they used MirEval web tool for sequence analysis and their study probably may be based on the older version of rRNA sequence data.

We also discovered that almost all of these rmiRNAs formed stem-loop structures and were located in RNA45S rather than the NTS (Figure 1, Supplemental Figure 2, and Table 2). Putative pre-rmiRNAs which have very similar sequences with their original miRNAs also showed similarity to their original miRNAs in the thermodynamic stability of stem-loop structures (Supplemental Figure 4A). Moreover, putative pre-miRNAs which have moderate substitutions in their sequences except the guide strands likewise showed similar stabilities and some showed more stable stem-loop structures than their original pre-miRNAs. This indicates that rmiRNAs have adequate potential stability to form stem-loop structures regarded as pre-miRNA. These results suggest that a large number of rmiRNAs are continually transcribed because rRNA is the most abundant noncoding RNA in eukaryotic cells. rDNA is transcribed so frequently that the rDNA region on the genome forms multicopies in the nucleolus [100]. Furthermore, rmiRNA may be noncanonically generated from rRNA in the cytoplasm because cytoplasmic RNA is a miRNA source [55].

All pooled miRNAs, such as rmiRNAs and/or tRFs, are important for immunoreactivity, transcriptional regulation, gene mobility, and cytoplasmic memory [17]. Sharma et al. indicated that paternal diet alters RNA information, such as population and composition of tRFs, in spermatozoa and influences progeny phenotype [101]. Likewise, although it was not confirmed in this study whether rDNA-hosted miRNAs would work *in vivo* with known identical miRNAs, rmiRNAs may function as resident miRNAs and participate in determining genotype and memorization. The number of stress-sensitive miRNAs is insufficient to exert an immediate response to cell damage if these miRNAs are generated only by transcription from DNA. Therefore, rapid generation of miRNAs from ready-made RNAs, such as rRNA and tRNA, should be considered. In addition, rmiRNAs and their identical miRNAs may work together because homologous miRNAs at different loci function together [102].

According to previous reports, rRNA-contained miRNAs such as miR-663, miR-1275, miR-3648, miR-3656, miR-3687, miR-4417, and miR-4516 are associated with tumor suppression, carcinomas, neuronal differentiation, breast cancer, breast cancer/neuronal differentiation, breast cancer, and regulation of signal transducer and activator of transcription 3, respectively [103–108]. However, the functions of residual rRNA-hosted miRNAs remain unclear. This motivated us to predict the targets of rmiRNAs and a number of putative and validated target genes which were associated with cancer-related pathways were found (Figure 2(c)). Therefore, RNA-RNA and/or RNA-protein interactions may participate in cancer-related functions. This indicates that rmiRNAs, in addition to ribosome-associated proteins and snoRNAs, might also be implicated in cell dysregulation and dysfunctions linked to ribosomopathies in multiple steps such as transcription, posttranscription, and biofunction.

In our previous study, the DNS was positively correlated with the strength of miRNA/miRNA synergies [72, 109]. As these rmiRNAs commonly have high DNS values, rmiRNA-derived miRNAs may also function as an activity booster of other miRNAs (Supplemental Figure 2). This characteristic

would be effective for quick responses to cell emergencies that are not severe enough to cause changes in intracellular RNA composition [10, 16, 110, 111]. In this study, most of the top 10 rmiRNA targets were predicted to play roles in gene regulation and participate in cancer-related pathways (Figure 2(c)). This finding indicates that the rRNA copy number and expression level may be directly associated with rmiRNA generation and regulation of the biological reactions to cell stress leading to carcinogenesis.

Given the known mechanisms of pre-rRNA processing and that of miRNA generation from rRNA, maturation of rmiRNA occurs as follows: (1) rRNA genes containing rmiRNAs are transcribed as RNA45S by Pol I [7, 9]; (2a) the RNA45S ITS and ETS are degraded by XRN1 or other nucleases after the rRNA matures, and pre-rmiRNAs located in these regions are generated simultaneously [7, 9, 48]; (2b) pre-rmiRNAs located on 18S rRNA are biologically generated upon the degradation of mature rRNAs in the ribosome or degradation of pre-rRNA in response to stress [110–112]. (3) Drosha, Dicer, or its related proteins and enzymes process pre-rmiRNAs into mature rmiRNAs. The last step in which Drosha and Dicer participate in rRNA processing has been observed in several studies. RNase III enzymes including Drosha and Dicer have a miRNA-independent role in RNA processing, because the depletion of Dicer or Drosha impairs rRNA processing but does not affect the exonuclease activities required for rRNA processing [113, 114]. Fukuda et al. revealed that the DEAD-box RNA helicase p68 (Ddx5) and p72 (Ddx17), which are subunits of the Drosha complex, are required for pre-rRNA and pri-miRNA processing. Woolnough et al. reported that the human Ago2 protein binds rRNA and interferes with the transcription of nascent human rRNA via binding with Pol III and the transcription factor III complex on the gene [66, 115]. These data indicate that rRNA processing is closely related to the miRNA processing enzymes and its related proteins. In contrast, Chak et al. reported that the generation of miR-10404 and endo-siRNA from the rRNA gene is unaffected by mutations in Drosha, Pasha, Ago2, or Dcr-2 but by Dcr-1 in *Drosophila* [47]. Son et al. demonstrated that the generation of pre-miR-712 is dependent on XRN1 but independent of Drosha and DGCR8 in mice [48]. Pre-miR-712 processing is a mirtron-like, but it remains unknown whether Drosha and Dicer contribute to generating rRNA-derived miRNA because the details of the roles of Drosha, Dicer, and related proteins in rRNA processing are unknown. However, these findings suggest that rRNA-derived miRNAs could be generated in both Drosha-dependent and Drosha-independent pathways.

The number of repeated rDNA arrays is strongly associated with cell senescence, gene integrity, and ribosomal function, although the majority of rDNA is inactive [15–17]. Moreover, rDNA cluster size differs among species and individuals and even in individual cells when the cells are responding to DNA damage or when the rRNA repeat number is being amplified [15–17, 116]. As these differences are inherited, it is certain that rRNA and rRNA-hosted miRNAs participate in cell identity [117]. The ETS and ITS regions are not highly conserved as compared to 18S, 5.8S, and 28S RNA. All three previously reported ITS- or ETS-derived miRNAs, such as miR-663, miR-712, and miR-10404, are human-, mouse-, and fly-specific miRNAs, respectively, and they are well conserved intraspecifically [47, 48], suggesting that variations in rmiRNAs and rDNA copy number contribute to evolution, particularly the inheritance of acquired characteristics.

Nucleolus, where rRNA is transcribed and processed, is the largest structure in the nucleus formed at rRNA coding regions on chromosomes and composed of diverse specific proteins and RNAs [118]. It has been revealed that some miRNAs exist and function in nucleolus. For instance, miR-206, a highly expressed miRNA in skeletal muscle, and several other miRNAs are detected in the nucleolus as well as in the cytoplasm with *in situ* hybridization [119]. Subsequently, it has been shown by deep sequencing that a set of miRNAs present in the nucleus rather than in the cytoplasm and some of them tend to accumulate at the nucleolus [120]. RNA interference (RNAi) factors such as AGO protein, Dicer, and TRBP are also found in the cell nuclei, suggesting that miRNA machinery is active even in the nucleolus [121]. Moreover, it has recently been reported that Alu element-containing Pol II transcripts (aluRNA) are abundant in nucleolus [67]. Alu element is the most abundant SINE family that comprises about 10% of the genome and exists in both noncoding and coding region including introns and 3′ UTR of mRNA transcripts [122, 123]. There are growing evidences that a portion of mRNAs have Alu-derived sequence in their 3′ UTR which can be targeted by a set of miRNAs [124–126]. Since it has been reported that the transcriptional rate of Alu is upregulated upon cellular stress and strongly influences the nucleolar size and pre-rRNA transcript rate [67, 127, 128], we supposed that aluRNAs might also be regulated by miRNAs. In our investigation, several miRNAs were detected from rRNA sequence as rmiRNA, and half of these rmiRNAs have potential target sites in Alu family sequences (Supplemental Table 3). Although it was not confirmed in this study whether rmiRNAs really regulate aluRNAs, at least, the possibility that rmiRNAs might interact with aluRNA in the nucleolus and contribute to the regulation of ribosomal function and composition upon cellular stress as a ribosomal feedback machinery was implied.

It was technically difficult to distinguish the origins of the sources using ready-made technologies, because the mature rmiRNA sequences, such as rmiR-663a/b and rmiR-1268a/b, were identical between the rDNA and non-rDNA genes. Therefore, in this study, we performed *in silico* analyses to by-pass this problem. No rmiRNA was detected from the 5S rRNA gene but the AGO2 protein binds to 5S rRNA [115], and AGO2 has Slicer activity [129], suggesting that various rRNA-derived specific miRNAs with different mature sequences to annotated miRNAs, that is, novel miRNAs, may be generated from the rRNA coding region. Furthermore, mature rmiRNA may have been generated in another form, such as loop miRNAs [130]. Numerous undefined RNA fragments derived from well-known RNAs or other noncoding RNAs might unveil the RNA wave enigma and implication of tumorigenesis. Therefore, additional laboratory and clinical investigations are required for discovery of the nascent

human miRNAs and for decipherment of precise interaction among miRNAs, noncoding RNAs, and human cancer.

5. Conclusion

Seventeen rDNA-hosted miRNA analogs (rmiRNAs) were found in rRNA coding region by *in silico* analyses. These rmiRNAs might be generated from rRNA upon construction or degradation of ribosomes. The majority of predicted targets of rmiRNAs were stress- or cancer-related genes and it was indicated that rmiRNAs could also target AluRNA in nucleolus, suggesting that rmiRNAs may regulate ribosomal function at multiple levels adopting to cellular stress. While rmiRNAs showed significantly high DNS values compared to those of normal miRNAs and tRFs, rmiRNAs may efficiently boost bioactivities of other miRNAs to attenuate cell stress and tumorigenesis as a quantum memory device and a member of the resident miRNA genes. Altogether, rmiRNAs would be implicated in human ribosomopathy. In future, rmiRNA mimics or anti-rmiRNA agents may be developed to cancer therapy and there is some possibility that rmiRNAs in serum could be applied for prognosis and/or diagnosis of ribosomopathy.

Competing Interests

The authors have declared that no competing interests exist.

References

[1] F. Luft, "The rise of a ribosomopathy and increased cancer risk," *Journal of Molecular Medicine*, vol. 88, no. 1, pp. 1–3, 2010.

[2] A. Narla and B. L. Ebert, "Ribosomopathies: human disorders of ribosome dysfunction," *Blood*, vol. 115, no. 16, pp. 3196–3205, 2010.

[3] P. C. Yelick and P. A. Trainor, "Ribosomopathies: global process, tissue specific defects," *Rare Diseases*, vol. 3, no. 1, Article ID e1025185, p. 10, 2015.

[4] K. M. Goudarzi and M. S. Lindström, "Role of ribosomal protein mutations in tumor development (Review)," *International Journal of Oncology*, vol. 48, no. 4, pp. 1313–1324, 2016.

[5] R. N. Nazar, T. O. Sitz, and H. Busch, "Sequence homologies in mammalian 5.8S ribosomal RNA," *Biochemistry*, vol. 15, no. 3, pp. 505–508, 1976.

[6] G. E. Zentner, A. Saiakhova, P. Manaenkov, M. D. Adams, and P. C. Scacheri, "Integrative genomic analysis of human ribosomal DNA," *Nucleic Acids Research*, vol. 39, no. 12, pp. 4949–4960, 2011.

[7] S. Frenk, D. Oxley, and J. Houseley, "The nuclear exosome is active and important during budding yeast meiosis," *PLoS ONE*, vol. 9, no. 9, Article ID e107648, 2014.

[8] A. K. Henras, C. Plisson-Chastang, M.-F. O'Donohue, A. Chakraborty, and P.-E. Gleizes, "An overview of pre-ribosomal RNA processing in eukaryotes," *Wiley Interdisciplinary Reviews: RNA*, vol. 6, no. 2, pp. 225–242, 2015.

[9] W. A. Decatur and M. J. Fournier, "RNA-guided nucleotide modification of ribosomal and other RNAs," *The Journal of Biological Chemistry*, vol. 278, no. 2, pp. 695–698, 2003.

[10] B. E. H. Maden, C. L. Dent, T. E. Farrell, J. Garde, F. S. McCallum, and J. A. Wakeman, "Clones of human ribosomal DNA containing the complete 18S-rRNA and 28S-rRNA genes. Characterization, a detailed map of the human ribosomal transcription unit and diversity among clones," *Biochemical Journal*, vol. 246, no. 2, pp. 519–527, 1987.

[11] K.-D. Chang, S.-A. Fang, F.-C. Chang, and M.-C. Chung, "Chromosomal conservation and sequence diversity of ribosomal RNA genes of two distant *Oryza* species," *Genomics*, vol. 96, no. 3, pp. 181–190, 2010.

[12] S. J. Goodfellow and J. C. B. M. Zomerdijk, "Basic mechanisms in RNA polymerase I transcription of the ribosomal RNA genes," in *Epigenetics: Development and Disease*, T. K. Kundu, Ed., vol. 61 of *Subcellular Biochemistry*, pp. 211–236, Springer, Dordrecht, The Netherlands, 2013.

[13] M. Ciganda and N. Williams, "Eukaryotic 5S rRNA biogenesis," *Wiley Interdisciplinary Reviews: RNA*, vol. 2, no. 4, pp. 523–533, 2011.

[14] A. S. Henderson, D. Warburton, and K. C. Atwood, "Ribosomal DNA connectives between human acrocentric chromosomes," *Nature*, vol. 245, no. 5420, pp. 95–97, 1973.

[15] S. Ide, T. Miyazaki, H. Maki, and T. Kobayashi, "Abundance of ribosomal RNA gene copies maintains genome integrity," *Science*, vol. 327, no. 5966, pp. 693–696, 2010.

[16] T. Kobayashi, "Ribosomal RNA gene repeats, their stability and cellular senescence," *Proceedings of the Japan Academy Series B: Physical and Biological Sciences*, vol. 90, no. 4, pp. 119–129, 2014.

[17] J. H. Malone, "Balancing copy number in ribosomal DNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 9, pp. 2635–2636, 2015.

[18] J. Diesch, R. D. Hannan, and E. Sanij, "Perturbations at the ribosomal genes loci are at the centre of cellular dysfunction and human disease," *Cell & Bioscience*, vol. 4, no. 1, article 43, 2014.

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

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

[21] Q. Zhang, J. Yao, G. W. Smith, and C. Dong, "Identification of a novel microRNA important for melanogenesis in alpaca (*Vicugna pacos*)," *Journal of Animal Science*, vol. 93, no. 4, pp. 1622–1631, 2015.

[22] K. J. Rayner, "MicroRNA-155 in the heart: the right time at the right place in the right cell," *Circulation*, vol. 131, no. 18, pp. 1533–1535, 2015.

[23] S. Srikantan, B. S. Marasa, K. G. Becker, M. Gorospe, and K. Abdelmohsen, "Paradoxical microRNAs: individual gene repressors, global translation enhancers," *Cell Cycle*, vol. 10, no. 5, pp. 751–759, 2011.

[24] F. Grey, R. Tirabassi, H. Meyers et al., "A viral microRNA down-regulates multiple cell cycle genes through mRNA 5′UTRs," *PLoS Pathogens*, vol. 6, no. 6, Article ID e1000967, 2010.

[25] Y. Hashimoto, Y. Akiyama, and Y. Yuasa, "Multiple-to-multiple relationships between MicroRNAs and target genes in gastric cancer," *PLoS ONE*, vol. 8, no. 5, Article ID e62589, 2013.

[26] L. Xu, W.-Q. Dai, X.-F. Xu, F. Wang, L. He, and C.-Y. Guo, "Effects of multiple-target anti-microRNA antisense oligodeoxyribonucleotides on proliferation and migration of gastric cancer cells," *Asian Pacific Journal of Cancer Prevention*, vol. 13, no. 7, pp. 3203–3207, 2012.

[27] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[28] Y. R. Fujii, "Oncoviruses and pathogenic microRNAs in humans," *The Open Virology Journal*, vol. 3, no. 1, pp. 37–51, 2009.

[29] A. M. Ardekani and M. M. Naeini, "The role of microRNAs in human diseases," *Avicenna Journal of Medical Biotechnology*, vol. 2, no. 4, pp. 161–179, 2010.

[30] M. Gama-Carvalho, J. Andrade, and L. Brás-Rosário, "Regulation of cardiac cell fate by microRNAs: implications for heart regeneration," *Cells*, vol. 3, no. 4, pp. 996–1026, 2014.

[31] A. Shenoy and R. H. Blelloch, "Regulation of microRNA function in somatic stem cell proliferation and differentiation," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 9, pp. 565–576, 2014.

[32] W.-F. Lai and P. M. Siu, "MicroRNAs as regulators of cutaneous wound healing," *Journal of Biosciences*, vol. 39, no. 3, pp. 519–524, 2014.

[33] A. K. L. Leung, S. Vyas, J. E. Rood, A. Bhutkar, P. A. Sharp, and P. Chang, "Poly(ADP-ribose) regulates stress responses and microRNA activity in the cytoplasm," *Molecular Cell*, vol. 42, no. 4, pp. 489–499, 2011.

[34] 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.

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

[36] V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small RNAs in animals," *Nature Reviews Molecular Cell Biology*, vol. 10, no. 2, pp. 126–139, 2009.

[37] P. W. C. Hsu, H.-D. Huang, S.-D. Hsu et al., "miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes," *Nucleic Acids Research*, vol. 34, pp. D135–D139, 2006.

[38] M. Fernandez-Mercado, L. Manterola, and C. H. Lawrie, "MicroRNAs in lymphoma: regulatory role and biomarker potential," *Current Genomics*, vol. 16, no. 5, pp. 349–358, 2015.

[39] A. Khvorova, A. Reynolds, and S. D. Jayasena, "Functional siRNAs and miRNAs exhibit strand bias," *Cell*, vol. 115, no. 2, pp. 209–216, 2003.

[40] C. A. Melo and S. A. Melo, "Biogenesis and physiology of MicroRNAs," in *Non-Coding RNAs and Cancer*, M. Fabbri, Ed., pp. 5–24, Springer, New York, NY, USA, 2014.

[41] J. G. Ruby, C. H. Jan, and D. P. Bartel, "Intronic microRNA precursors that bypass Drosha processing," *Nature*, vol. 448, no. 7149, pp. 83–86, 2007.

[42] Y. R. Fujii, "RNA genes: retroelements and virally retroposable microRNAs in human embryonic stem cells," *The Open Virology Journal*, vol. 4, pp. 63–75, 2010.

[43] Y. R. Fujii, "The RNA gene information: retroelement-MicroRNA entangling as the RNA quantum code," in *MicroRNA Protocols*, S.-Y. Ying, Ed., pp. 47–67, Humana Press, Totowa, NJ, USA, 2013.

[44] J. L. Wiesen and T. B. Tomasi, "Dicer is regulated by cellular stresses and interferons," *Molecular Immunology*, vol. 46, no. 6, pp. 1222–1228, 2009.

[45] Z. Yang and L. Wang, "Regulation of microRNA expression and function by nuclear receptor signaling," *Cell & Bioscience*, vol. 1, no. 1, article 31, 2011.

[46] S. Kuchen, W. Resch, A. Yamane et al., "Regulation of microRNA expression and abundance during lymphopoiesis," *Immunity*, vol. 32, no. 6, pp. 828–839, 2010.

[47] L.-L. Chak, J. Mohammed, E. C. Lai, G. Tucker-Kellogg, and K. Okamura, "A deeply conserved, noncanonical miRNA hosted by ribosomal DNA," *RNA*, vol. 21, no. 3, pp. 375–384, 2015.

[48] D. J. Son, S. Kumar, W. Takabe et al., "The atypical mechanosensitive microRNA-712 derived from pre-ribosomal RNA induces endothelial inflammation and atherosclerosis," *Nature Communications*, vol. 4, article 3000, 2013.

[49] J. F. Doss, D. L. Corcoran, D. D. Jima, M. J. Telen, S. S. Dave, and J. Chi, "A comprehensive joint analysis of the long and short RNA transcriptomes of human erythrocytes," *BMC Genomics*, vol. 16, article 952, 2015.

[50] G. LaMonte, N. Philip, J. Reardon et al., "Translocation of sickle cell erythrocyte microRNAs into *Plasmodium falciparum* inhibits parasite translation and contributes to malaria resistance," *Cell Host & Microbe*, vol. 12, no. 2, pp. 187–199, 2012.

[51] W. S. Ward and D. S. Coffey, "DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells," *Biology of Reproduction*, vol. 44, no. 4, pp. 569–574, 1991.

[52] M. Jodar, S. Selvaraju, E. Sendler, M. P. Diamond, and S. A. Krawetz, "The presence, role and clinical use of spermatozoal RNAs," *Human Reproduction Update*, vol. 19, no. 6, pp. 604–624, 2013.

[53] A. Govindaraju, A. Uzun, L. Robertson et al., "Dynamics of microRNAs in bull spermatozoa," *Reproductive Biology and Endocrinology*, vol. 10, article 82, 2012.

[54] B. Barckmann and M. Simonelig, "Control of maternal mRNA stability in germ cells and early embryos," *Biochimica et Biophysica Acta (BBA)—Gene Regulatory Mechanisms*, vol. 1829, no. 6-7, pp. 714–724, 2013.

[55] Q. Jin, Z. Xue, C. Dong, Y. Wang, L. Chu, and Y. Xu, "Identification and characterization of MicroRNAs from tree peony (*Paeonia ostii*) and their response to copper stress," *PLoS ONE*, vol. 10, no. 2, Article ID e0117584, 2015.

[56] T. A. Reese, J. Xia, L. S. Johnson, X. Zhou, W. Zhang, and H. W. Virgin, "Identification of novel microRNA-like molecules generated from herpesvirus and host tRNA transcripts," *Journal of Virology*, vol. 84, no. 19, pp. 10344–10353, 2010.

[57] Y. S. Lee, Y. Shibata, A. Malhotra, and A. Dutta, "A novel class of small RNAs: tRNA-derived RNA fragments (tRFs)," *Genes & Development*, vol. 23, no. 22, pp. 2639–2649, 2009.

[58] C. Megel, G. Morelle, S. Lalande, A.-M. Duchêne, I. Small, and L. Maréchal-Drouard, "Surveillance and cleavage of eukaryotic tRNAs," *International Journal of Molecular Sciences*, vol. 16, no. 1, pp. 1873–1893, 2015.

[59] D. Green, W. D. Fraser, and T. Dalmay, "Transfer RNA-derived small RNAs in the cancer transcriptome," *Pflügers Archiv-European Journal of Physiology*, vol. 468, no. 6, pp. 1041–1047, 2016.

[60] K. W. Diebel, K. Zhou, A. B. Clarke, and L. T. Bemis, "Beyond the ribosome: extra-translational functions of tRNA fragments," *Biomarker Insights*, vol. 11, supplement 1, pp. 1–8, 2016.

[61] E. Casas, G. Cai, and J. D. Neill, "Characterization of circulating transfer RNA-derived RNA fragments in cattle," *Frontiers in Genetics*, vol. 6, article 271, pp. 1–7, 2015.

[62] S. Karaiskos, A. S. Naqvi, K. E. Swanson, and A. Grigoriev, "Age-driven modulation of tRNA-derived fragments in Drosophila and their potential targets," *Biology Direct*, vol. 10, article 51, 2015.

[63] G. Loss-Morais, P. M. Waterhouse, and R. Margis, "Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets," *Biology Direct*, vol. 8, no. 1, article 6, 2013.

[64] S. R. Selitsky, J. Baran-Gale, M. Honda et al., "Small tRNA-derived RNAs are increased and more abundant than microRNAs in chronic hepatitis B and C," *Scientific Reports*, vol. 5, article no. 7675, 2015.

[65] H. Wei, B. Zhou, F. Zhang et al., "Profiling and identification of small rDNA-derived RNAs and their potential biological functions," *PLoS ONE*, vol. 8, no. 2, Article ID e56842, 2013.

[66] T. Fukuda, K. Yamagata, S. Fujiyama et al., "DEAD-box RNA helicase subunits of the Drosha complex are required for processing of rRNA and a subset of microRNAs," *Nature Cell Biology*, vol. 9, pp. 604–611, 2007.

[67] M. Caudron-Herger, T. Pankert, J. Seiler et al., "Alu element-containing RNAs maintain nucleolar structure and function," *The EMBO Journal*, vol. 34, no. 22, pp. 2758–2774, 2015.

[68] A. Kozomara and S. Griffiths-Jones, "MiRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic Acids Research*, vol. 42, no. 1, pp. D68–D73, 2014.

[69] R. Gibson, B. Alako, C. Amid et al., "Biocuration of functional annotation at the European nucleotide archive," *Nucleic Acids Research*, vol. 44, pp. D58–D66, 2016.

[70] P. P. Chan and T. M. Lowe, "GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes," *Nucleic Acids Research*, vol. 44, no. 1, pp. D184–D189, 2016.

[71] P. Kumar, S. B. Mudunuri, J. Anaya, and A. Dutta, "tRFdb: a database for transfer RNA fragments," *Nucleic Acids Research*, vol. 43, no. 1, pp. D141–D145, 2015.

[72] M. Yoshikawa, T. Osone, and Y. Fujii, "MicroRNA memory I: the positive correlation between synergistic effects of microRNAs in cancer and a novel quantum scoring system," *Journal of Advances in Medical and Pharmaceutical Sciences*, vol. 5, no. 4, pp. 1–16, 2016.

[73] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *eLife*, vol. 4, no. 2015, Article ID e05005, 2015.

[74] C.-H. Chou, N.-W. Chang, S. Shrestha et al., "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database," *Nucleic Acids Research*, vol. 44, no. 1, pp. D239–D247, 2016.

[75] D. Tabas-Madrid, R. Nogales-Cadenas, and A. Pascual-Montano, "GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics," *Nucleic Acids Research*, vol. 40, no. 1, pp. W478–W483, 2012.

[76] N. S. Vassetzky and D. A. Kramerov, "SINEBase: a database and tool for SINE analysis," *Nucleic Acids Research*, vol. 41, no. 1, pp. D83–D89, 2013.

[77] K. C. Miranda, T. Huynh, Y. Tay et al., "A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes," *Cell*, vol. 126, no. 6, pp. 1203–1217, 2006.

[78] D. L. J. Lafontaine and D. Tollervey, "The function and synthesis of ribosomes," *Nature Reviews Molecular Cell Biology*, vol. 2, no. 7, pp. 514–520, 2001.

[79] M. Gamalinda and J. L. Woolford, "Paradigms of ribosome synthesis: lessons learned from ribosomal proteins," *Translation*, vol. 3, no. 1, Article ID e975018, 2015.

[80] D. Kressler, E. Hurt, and J. Baβler, "Driving ribosome assembly," *Biochimica et Biophysica Acta (BBA)—Molecular Cell Research*, vol. 1803, no. 6, pp. 673–683, 2010.

[81] E. Brighenti, D. Treré, and M. Derenzini, "Targeted cancer therapy with ribosome biogenesis inhibitors: a real possibility?" *Oncotarget*, vol. 6, no. 36, pp. 38617–38627, 2015.

[82] L. Montanaro, D. Treré, and M. Derenzini, "Nucleolus, ribosomes, and cancer," *The American Journal of Pathology*, vol. 173, no. 2, pp. 301–310, 2008.

[83] L. B. Maggi Jr. and J. D. Weber, "Nucleolar adaptation in human cancer," *Cancer Investigation*, vol. 23, no. 7, pp. 599–608, 2005.

[84] G. Thomas, "An encore for ribosome biogenesis in the control of cell proliferation," *Nature Cell Biology*, vol. 2, no. 5, pp. E71–E72, 2000.

[85] S. Volarević, M. J. Stewart, B. Ledermann et al., "Proliferation, but not growth, blocked by conditional deletion of 40S ribosomal protein S6," *Science*, vol. 288, no. 5473, pp. 2045–2047, 2000.

[86] S. S. Negi and P. Brown, "Transient rRNA synthesis inhibition with CX-5461 is sufficient to elicit growth arrest and cell death in acute lymphoblastic leukemia cells," *Oncotarget*, vol. 6, no. 33, pp. 34846–34858, 2015.

[87] S. S. Negi and P. Brown, "rRNA synthesis inhibitor, CX-5461, activates ATM/ATR pathway in acute lymphoblastic leukemia, arrests cells in G2 phase and induces apoptosis," *Oncotarget*, vol. 6, no. 20, pp. 18094–18104, 2015.

[88] F. Scala, E. Brighenti, M. Govoni et al., "Direct relationship between the level of p53 stabilization induced by rRNA synthesis-inhibiting drugs and the cell ribosome biogenesis rate," *Oncogene*, vol. 35, no. 8, pp. 977–989, 2015.

[89] S. Sulic, L. Panic, M. Barkic, M. Mercep, M. Uzelac, and S. Volarevic, "Inactivation of S6 ribosomal protein gene in T lymphocytes activates a p53-dependent checkpoint response," *Genes & Development*, vol. 19, no. 24, pp. 3070–3082, 2005.

[90] Y.-W. Chen, X.-T. Hu, A. C. Liang et al., "High BCL6 expression predicts better prognosis, independent of BCL6 translocation status, translocation partner, or BCL6-deregulating mutations, in gastric lymphoma," *Blood*, vol. 108, no. 7, pp. 2373–2383, 2006.

[91] S. Kobayashi, T. Taki, H. Nagoshi et al., "Identification of novel fusion genes with 28S ribosomal DNA in hematologic malignancies," *International Journal of Oncology*, vol. 44, no. 4, pp. 1193–1198, 2014.

[92] C. Bellodi, M. McMahon, A. Contreras et al., "H/ACA small RNA dysfunctions in disease reveal key roles for noncoding RNA modifications in hematopoietic stem cell differentiation," *Cell Reports*, vol. 3, no. 5, pp. 1493–1502, 2013.

[93] J. R. P. Knight, A. E. Willis, and J. Milner, "Active regulator of SIRT1 is required for ribosome biogenesis and function," *Nucleic Acids Research*, vol. 41, no. 7, pp. 4185–4197, 2013.

[94] S. B. Sondalle and S. J. Baserga, "Human diseases of the SSU processome," *Biochimica et Biophysica Acta (BBA—Molecular Basis of Disease*, vol. 1842, no. 6, pp. 758–764, 2014.

[95] Y. Moon, "Ribosomal alteration-derived signals for cytokine induction in mucosal and systemic inflammation: noncanonical pathways by ribosomal inactivation," *Mediators of Inflammation*, vol. 2014, Article ID 708193, 10 pages, 2014.

[96] L. Chen and G.-F. Hu, "Angiogenin-mediated ribosomal RNA transcription as a molecular target for treatment of head and neck squamous cell carcinoma," *Oral Oncology*, vol. 46, no. 9, pp. 648–653, 2010.

[97] N. Guzman, K. Agarwal, D. Asthagiri et al., "Breast cancer-specific miR signature unique to extracellular vesicles includes 'microRNA-like' tRNA fragments," *Molecular Cancer Research*, vol. 13, no. 5, pp. 891–901, 2015.

[98] P. Kumar, J. Anaya, S. B. Mudunuri, and A. Dutta, "Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets," *BMC Biology*, vol. 12, article 78, 2014.

[99] K. W. Diebel, L. M. Oko, E. M. Medina et al., "Gammaherpesvirus small noncoding RNAs are bifunctional elements that regulate infection and contribute to virulence in vivo," *mBio*, vol. 6, no. 1, Article ID e01670-14, 2015.

[100] R. Y. L. Tsai and T. Pederson, "Connecting the nucleolus to the cell cycle and human disease," *The FASEB Journal*, vol. 28, no. 8, pp. 3290–3296, 2014.

[101] U. Sharma, C. C. Conine, J. M. Shea et al., "Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals," *Science*, vol. 351, no. 6271, pp. 391–396, 2016.

[102] F. Lovat, M. Fassan, P. Gasparini et al., "miR-15b/16-2 deletion promotes B-cell malignancies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 37, pp. 11636–11641, 2015.

[103] W. Zang, Y. Wang, T. Wang et al., "miR-663 attenuates tumor growth and invasiveness by targeting eEF1A2 in pancreatic cancer," *Molecular Cancer*, vol. 14, no. 1, article 37, 2015.

[104] I. O. Fawzy, M. T. Hamza, K. A. Hosny, G. Esmat, H. M. El Tayebi, and A. I. Abdelaziz, "MiR-1275: A single microRNA that targets the three IGF2-mRNA-binding proteins hindering tumor growth in hepatocellular carcinoma," *FEBS Letters*, vol. 589, no. 17, pp. 2257–2265, 2015.

[105] N. Matamala, M. T. Vargas, R. González-Cámpora et al., "Tumor MicroRNA expression profiling identifies circulating microRNAs for early breast cancer detection," *Clinical Chemistry*, vol. 61, no. 8, pp. 1098–1106, 2015.

[106] R. Shu, W. Wong, Q. H. Ma et al., "APP intracellular domain acts as a transcriptional regulator of miR-663 suppressing neuronal differentiation," *Cell Death and Disease*, vol. 6, no. 2, Article ID e1651, 2015.

[107] R. Murria, S. Palanca, I. de Juan et al., "Immunohistochemical, genetic and epigenetic profiles of hereditary and triple negative breast cancers. Relevance in personalized medicine," *American Journal of Cancer Research*, vol. 5, no. 7, pp. 2330–2343, 2015.

[108] S. Chowdhari and N. Saini, "hsa-miR-4516 mediated downregulation of STAT3/CDK6/UBE2N plays a role in PUVA induced apoptosis in keratinocytes," *Journal of Cellular Physiology*, vol. 229, no. 11, pp. 1630–1638, 2014.

[109] T. Osone, M. Yoshikawa, and Y. Fujii, "MicroRNA memory II: a novel scoring integration model for prediction of human disease by microRNA/microRNA quantum multi-interaction," *Journal of Advances in Medical and Pharmaceutical Sciences*, vol. 5, no. 3, pp. 1–18, 2016.

[110] C. Allmang, P. Mitchell, E. Petfalski, and D. Tollervey, "Degradation of ribosomal RNA precursors by the exosome," *Nucleic Acids Research*, vol. 28, no. 8, pp. 1684–1691, 2000.

[111] G. N. Basturea, M. A. Zundel, and M. P. Deutscher, "Degradation of ribosomal RNA during starvation: comparison to quality control during steady-state growth and a role for RNase PH," *RNA*, vol. 17, no. 2, pp. 338–345, 2011.

[112] Q. Wang, I. Lee, J. Ren, S. S. Ajay, Y. S. Lee, and X. Bao, "Identification and functional characterization of tRNA-derived RNA fragments (tRFs) in respiratory syncytial virus infection," *Molecular Therapy*, vol. 21, no. 2, pp. 368–379, 2013.

[113] T. M. Johanson, A. M. Lew, and M. M. W. Chong, "MicroRNA-independent roles of the RNase III enzymes Drosha and Dicer," *Open Biology*, vol. 3, no. 10, Article ID 130144, 2013.

[114] X.-H. Liang and S. T. Crooke, "Depletion of key protein components of the RISC pathway impairs pre-ribosomal RNA processing," *Nucleic Acids Research*, vol. 39, no. 11, pp. 4875–4889, 2011.

[115] J. L. Woolnough, B. L. Atwood, and K. E. Giles, "Argonaute 2 binds directly to tRNA genes and promotes gene repression in *cis*," *Molecular and Cellular Biology*, vol. 35, no. 13, pp. 2278–2294, 2015.

[116] D. M. Stults, M. W. Killen, H. H. Pierce, and A. J. Pierce, "Genomic architecture and inheritance of human ribosomal RNA gene clusters," *Genome Research*, vol. 18, no. 1, pp. 13–18, 2008.

[117] Y. Robertus Fujii, "The xenotropic microRNA gene information for stem cell researches and clinical applications," *Stem Cell Discovery*, vol. 3, no. 1, pp. 32–36, 2013.

[118] Y. W. Lam and L. Trinkle-Mulcahy, "New insights into nucleolar structure and function," *F1000Prime Reports*, vol. 7, article 48, 2015.

[119] J. C. Ritland Politz, E. M. Hogan, and T. Pederson, "MicroRNAs with a nucleolar location," *RNA*, vol. 15, no. 9, pp. 1705–1715, 2009.

[120] B. Bai, H. Liu, and M. Laiho, "Small RNA expression and deep sequencing analyses of the nucleolus reveal the presence of nucleolus-associated microRNAs," *FEBS Open Bio*, vol. 4, pp. 441–449, 2014.

[121] K. Gagnon, L. Li, Y. Chu, B. Janowski, and D. Corey, "RNAi factors are present and active in human cell nuclei," *Cell Reports*, vol. 6, no. 1, pp. 211–221, 2014.

[122] A. J. Mighell, A. F. Markham, and P. A. Robinson, "Alu sequences," *FEBS Letters*, vol. 417, no. 1, pp. 1–5, 1997.

[123] P. Deininger, "*Alu* Elements," in *Genomic Disorders*, pp. 21–34, Humana Press, Totowa, NJ, USA, 2006.

[124] E. Daskalova, V. Baev, V. Rusinov, and I. Minkov, "3′UTR-located ALU elements: donors of potential miRNA target sites and mediators of network miRNA-based regulatory interactions," *Evolutionary Bioinformatics Online*, vol. 2, pp. 103–120, 2006.

[125] Y. Hoffman, D. Dahary, D. R. Bublik, M. Oren, and Y. Pilpel, "The majority of endogenous microRNA targets within Alu elements avoid the microRNA machinery," *Bioinformatics*, vol. 29, no. 7, pp. 894–902, 2013.

[126] Y. Hoffman, Y. Pilpel, and M. Oren, "MicroRNAs and Alu elements in the p53-Mdm2-Mdm4 regulatory network," *Journal of Molecular Cell Biology*, vol. 6, no. 3, pp. 192–197, 2014.

[127] T. J. Gu, X. Yi, X. W. Zhao, Y. Zhao, and J. Q. Yin, "Alu-directed transcriptional regulation of some novel miRNAs," *BMC Genomics*, vol. 10, article 563, 2009.

[128] S. Lehnert, P. Van Loo, P. J. Thilakarathne, P. Marynen, G. Verbeke, and F. C. Schuit, "Evidence for co-evolution between human microRNAs and Alu-repeats," *PLoS ONE*, vol. 4, no. 2, Article ID e4456, 2009.

[129] K. Miyoshi, H. Tsukumo, T. Nagami, H. Siomi, and M. C. Siomi, "Slicer function of *Drosophila* Argonautes and its involvement in RISC formation," *Genes & Development*, vol. 19, no. 23, pp. 2837–2848, 2005.

[130] L.-L. Chak and K. Okamura, "Argonaute-dependent small RNAs derived from single-stranded, non-structured precursors," *Frontiers in Genetics*, vol. 5, article 172, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 4563524, 8 pages
http://dx.doi.org/10.1155/2016/4563524

Research Article

Ens-PPI: A Novel Ensemble Classifier for Predicting the Interactions of Proteins Using Autocovariance Transformation from PSSM

Zhen-Guo Gao,[1] Lei Wang,[1,2] Shi-Xiong Xia,[1] Zhu-Hong You,[1] Xin Yan,[3] and Yong Zhou[1]

[1]*School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China*
[2]*College of Information Science and Engineering, Zaozhuang University, Zaozhuang, Shandong 277100, China*
[3]*School of Foreign Languages, Zaozhuang University, Zaozhuang, Shandong 277100, China*

Correspondence should be addressed to Shi-Xiong Xia; xiasx@cumt.edu.cn and Zhu-Hong You; zhuhongyou@cumt.edu.cn

Received 9 April 2016; Accepted 8 May 2016

Academic Editor: Zheng Yin

Protein-Protein Interactions (PPIs) play vital roles in most biological activities. Although the development of high-throughput biological technologies has generated considerable PPI data for various organisms, many problems are still far from being solved. A number of computational methods based on machine learning have been developed to facilitate the identification of novel PPIs. In this study, a novel predictor was designed using the Rotation Forest (RF) algorithm combined with Autocovariance (AC) features extracted from the Position-Specific Scoring Matrix (PSSM). More specifically, the PSSMs are generated using the information of protein amino acids sequence. Then, an effective sequence-based features representation, Autocovariance, is employed to extract features from PSSMs. Finally, the RF model is used as a classifier to distinguish between the interacting and noninteracting protein pairs. The proposed method achieves promising prediction performance when performed on the PPIs of *Yeast*, *H. pylori*, and *independent datasets*. The good results show that the proposed model is suitable for PPIs prediction and could also provide a useful supplementary tool for solving other bioinformatics problems.

1. Introduction

Proteins are the most versatile and important macromolecules in life. They are vital for nearly all of the activity in the cell, including signaling cascades, metabolic cycles, and DNA transcription and replication [1]. Researchers found out that proteins rarely act as isolated agents to achieve their function. As expected, proteins are mutually matched with each other, forming a huge and complex network of Protein-Protein Interactions (PPIs) [2]. Therefore, research on PPIs has become the core issue of systems biology [3, 4].

So far, a variety of experimental techniques have been developed and designed for the detection of PPIs. The high-throughput techniques including Yeast Two-Hybrid (Y2H) screen [5–7], Tandem Affinity Purification (TAP) [2], and Mass Spectrometric Protein Complex Identification (MS-PCI) [6] spend considerable amounts of time, money, and manpower for detecting PPIs. In addition, PPIs obtained by biological experiments at present can only cover a small part of the whole PPIs network [8]. Therefore, the development of reliable computational methods which can improve the recognition efficiency has important significance [9–11].

A large number of *in silico* methods for predicting PPI have emerged [12–14]. These methods are usually based on the information of gene neighboring [15], gene coexpression [15], phylogenetic relationship [16], gene fusion events [17], three-dimensional structural information [18], and so on [19]. However, the application of these methods is limited [20, 21], because they need to rely on preknowledge of the protein. Recently, the methods based on the sequence information of protein amino acids for detecting PPI have been proposed [22–24]. For example, You et al. [25] used only protein sequence information to predict PPI, in which a kind of method called PCA-EELM (Principal Component Analysis-Ensemble Extreme Learning Machine) is designed. When performed on the PPIs data of *Saccharomyces cerevisiae*, this

model yields 87.00% prediction accuracy, 86.15% sensitivity, and 87.59% precision. Martin et al. [26] designed a model to detect PPIs by using the extended signature descriptor, which was extended to protein pairs. In order to verify the predictive ability of this method, when using 10-fold cross-validation applied on the *H. pylori* and *Yeast* datasets, the accuracy of this method is from 70% to 80%. Shen et al. [11] considered the residues local environments and designed the conjoint triad method. When performed on *human* PPIs dataset, this method has yielded 83.9% accuracy. Guo et al. [9] combined Support Vector Machine classifier with Automatic Covariance features extracted from the protein sequences to predict PPIs in *Saccharomyces cerevisiae*. The average prediction accuracy of the method reached 86.55%.

In this study, we presented a sequence-based method which combines the RF classifier and Autocovariance (AC) algorithm to predict the interacting protein pairs [9, 27, 28]. A novel protein feature representation is derived from Position-Specific Scoring Matrix (PSSM) [29], which gives the log-odds score of specific residue replacement based on specific location of evolutionary information. Then, an effective sequence-based protein representation, Autocovariance, is employed to extract features from PSSMs. The interaction among a certain number of amino acid sequences was calculated by AC algorithm. Thus, this model took into account the proximity effect and made it possible to find patterns throughout the sequence. Finally, the ensemble RF classifier is established, which is using the PSSM-derived features as input. In the experiments, the proposed model was evaluated on *Yeast* and *H. pylori* PPI datasets. The experiment results show that our model achieved 97.77% and 84.84% prediction accuracy with 95.57% and 82.77% sensitivity on these two datasets. In addition, we evaluate the proposed model on independent datasets of the *C. elegans*, *E. coli*, *H. sapiens*, and *M. musculus* PPIs and achieved 96.01%, 97.73%, 98.30%, and 96.81% prediction accuracy, respectively.

2. Materials and Methodology

2.1. Data Sources. In the experiments, we used nonredundant *Yeast* data, which was gathered in *Saccharomyces cerevisiae* core subset of the Database of Interacting Proteins (DIP) [30], and the version is DIP *20070219* by Guo et al. [9]. Two methods, Paralogous Verification Method (PVM) and Expression Profile Reliability (EPR) [31], have proven the reliability of the core subset. There are 5966 interaction pairs contained in the core subset. Sequences with less than 50 amino acid residues were removed because they might just be fragments. The final positive dataset was comprised of the remaining 5943 protein pairs. The CD-Hit [32, 33] algorithm was further used with less than forty percent identity to decrease pairwise sequence redundancy. By doing this, the rest of the 5594 protein pairs constructed the positive dataset. We chose 5594 additional protein pairs in different subcellular localization to construct the negative dataset. Finally, the complete dataset was constructed; it was composed of 11188 protein pairs, half of which were positive and the other half were negative.

We also tested our method using two-hybrid measurements of *H. pylori* introduced by Rain et al. [34]. The *H. pylori* dataset (available at http://www.cs.sandia.gov/~smartin/software.html) contains 2916 protein pairs. There are interacting pairs and noninteracting pairs, each accounting for fifty percent. This dataset provides a platform for comparing our approach and other approaches [25, 26, 35–38].

2.2. Position-Specific Scoring Matrix (PSSM). Position-Specific Scoring Matrix is first used in the detection of distantly related protein, which is proposed by Gribskov et al. [29]. Its feasibility has been verified in protein secondary structure prediction [39], prediction of disordered regions [40], and protein binding site prediction [41]. Structure of a PSSM is L rows and 20 columns. Suppose that $\text{PSSM} = \{\theta_{i,j} : i = 1, \ldots, L,\ j = 1, \ldots, 20\}$. Rows of the matrix represent the protein residues and columns represent the naive amino acids. Each matrix can be represented by the following formula:

$$\text{PSSM} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,20} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{L,1} & \theta_{L,2} & \cdots & \theta_{L,20} \end{bmatrix}, \quad (1)$$

where L is the length of the corresponding protein sequence and $\theta_{i,j}$ in the i row of PSSM meant the probability of the ith residue being mutated into type j of 20 native amino acids during the procession of evolutionary information in the protein from multiple sequence alignments.

In this experiment, we introduced the Position-Specific Iterated BLAST (PSI-BLAST) program [42] and *SwissProt* dataset on a local machine to produce PSSMs. PSI-BLAST is more sensitive compared to BLAST, particularly in the discovery of new members of a protein family. To generate the PSSM, PSI-BLAST needs sequence contrast with very high sensitivity between the input proteins and the proteins in the database, and all sequence entries in the *SwissProt* database have been carefully verified by computer tools and access to relevant literature through the experience of molecular biologists and protein chemists, so we put *SwissProt* database as the optimal comparison database in the experiment. And to get broad and high homologous sequences, we held the other parameters constant, where the e-value is set to 0.001 and the number of iterations is set to 3, respectively. Applications of PSI-BLAST and *SwissProt* database can be downloaded from http://blast.ncbi.nlm.nih.gov/Blast.cgi.

2.3. Autocovariance (AC). As one of the most effective analyzing sequences of vectors statistical tools, the AC has been widely used in protein family classification by researchers [43, 44], prediction of secondary structure content [45, 46], and protein interaction prediction [9]. AC is a variable expressed in a given protein sequence of two residues' average

correlation, which can be calculated by

$$\mathrm{AC}(\lambda, lg) = \frac{1}{L-lg}\sum_{\lambda=1}^{L-lg}\left(M_{\lambda,\theta} - \frac{1}{L}\sum_{\lambda=1}^{L}M_{\lambda,\theta}\right) \cdot \left(M_{(\lambda+lg),\theta} - \frac{1}{L}\sum_{\lambda=1}^{L}M_{\lambda,\theta}\right), \quad (2)$$

where lg is the distance between residues, λ represents the λth amino acid, L denotes the length of the protein sequence, and $M_{\lambda,\theta}$ indicates the matrix score of amino acid λ at position θ.

Using the above expression, the value of AC variable M can be figured out: $M = lg \times N$, where N is the number of descriptors. When all the data in the database complete the operation, each protein sequence was represented as a vector of AC variables; a protein pair was characterized by concatenating the vectors of two proteins in this protein pair.

2.4. Rotation Forest Classifier. Rotation Forest (RF) is a popular ensemble classifier and this idea originated from Random Forests classifier. Each decision tree in Rotation Forest is trained on the dataset in a rotated feature space. As a decision tree learning algorithm establishes the classification regions using hyperplanes parallel to the feature axes and a small rotation of axes may build an entirely different tree, the diversity of RF can be guaranteed by the transformation. Thus, RF model can enhance the accuracy for individual classifier and the diversity in the ensemble at the same time. It is more robust compared to the previously proposed ensemble systems, such as Random Forest [32, 47], Bagging [33, 48], and Boosting [49]. The RF algorithm is described as follows.

Assuming $\{x_i, y_i\}$ contains N training samples, wherein $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ is a D-dimensional feature vector. Suppose that X is the training sample set ($n \times D$ matrix), which is composed of n observation feature vector composition; S denote the feature set, and Y denote the corresponding labels, and then $X = (x_1, x_2, \ldots, x_n)^T$, $Y = (y_1, y_2, \ldots, y_n)^T$. Assume a feature set with an appropriate factor randomly divided into K subsets of the same size; in this case, the decision trees L in the forest can be expressed as $T_1, T_2, \ldots, T_L$, respectively. The execution steps of the training set for a single classifier T_i are shown below:

(1) Select the appropriate parameter K which is a factor of n; let S be randomly divided into K parts of the disjoint subsets; each subset contains a number of features, $C = n/k$.

(2) From the training dataset X, select the corresponding column of the feature in the subset $T_{i,j}$ and form a new matrix $X_{i,j}$, followed by a bootstrap subset of objects extracting 75 percent of X constituting a new training set $X'_{i,j}$.

(3) Matrix $X'_{i,j}$ is used as the feature transform for producing the coefficients in a matrix $M_{i,j}$, with jth column coefficient as the characteristic jth component.

(4) The coefficients obtained in the matrix $M_{i,j}$ are constructed as a sparse rotation matrix R_i, which is expressed as follows:

$$R_i = \begin{bmatrix} \lambda_{i,1}^{(1)}, \ldots, \lambda_{i,1}^{(C_1)} & 0 & \cdots & 0 \\ 0 & \lambda_{i,2}^{(1)}, \ldots, \lambda_{i,2}^{(C_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{i,k}^{(1)}, \ldots, \lambda_{i,k}^{(C_k)} \end{bmatrix}. \quad (3)$$

In the prediction period, the test sample x, generated by the classifier T_i of $d_{i,j}(XR_i^{\lambda})$ to determine x, belongs to class y_i. Next, the class of confidence is calculated by means of the average combination, and the formula is as follows:

$$\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}\left(XR_i^{\lambda}\right). \quad (4)$$

Then, assign the category with the largest $\mu_j(x)$ value to x.

3. Results and Discussions

3.1. Evaluation Measures. In this section, 5-fold cross-validation is used to evaluate the performance of the proposed method, in which all samples are split into five subsets. Therefore, one subset is the test set and the remaining four subsets are the training set. Evaluation criteria used in our study include overall prediction accuracy (Accu.), sensitivity (Sen.), precision (Prec.), and Matthews correlation coefficient (MCC). The calculation formulas are listed below:

$$\begin{aligned} \text{Accu.} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \\ \text{Sen.} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \\ \text{Prec.} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \\ \text{MCC} &= \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}, \end{aligned} \quad (5)$$

where True Positive (TP) represents the number of samples that are correctly detected as positive, True Negative (TN) represents the number of samples that are correctly detected as negative, False Positive (FP) represents the number of samples that are incorrectly detected as positive, and False Negative (FN) represents the number of samples that are incorrectly detected as negative. We also produce Receiver Operating Characteristic (ROC) [50] curves to assess the capability of the classifier. Typically, the threshold value of the classifier is 0.5 by default. When a new set of prediction results is accepted, the threshold value will be changed with the True

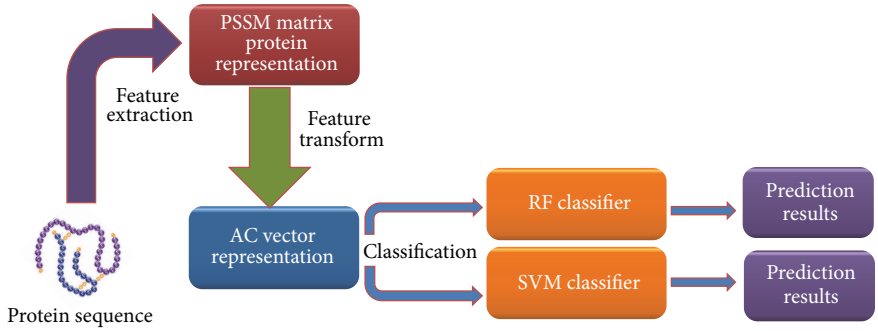


Figure 1: The workflow of our method.

Positive Rate versus the False Positive Rate; this change can be drawn out with graphics. In addition, the Area Under a Curve (AUC), with score ranges from 0 to 1, can also be expressed by the ROC curve. When a predictor of the AUC value is greater than another predictor, this predictor is regarded as a better one. The workflow of our method is shown in Figure 1.

3.2. Assessment of Prediction Ability. In order to achieve better results in the experiment, we used the grid search method to explore the parameters of the proposed model; concrete has parameter lg for AC and parameters K and L value for RF. Firstly, we discuss the parameters of AC; the maximal possible lg is the shortest sequence length (50 amino acids) on the *Yeast* dataset. In this experiment, several lgs ($lg = 5, 10, 15, 20, 25, 30, 35, 40, 45$) were evaluated in order to achieve the best performance of the protein sequences. The prediction results were shown in Figure 2. As seen from the curve in the graph, the prediction accuracy gradually increases when the parameters lg of the AC algorithm change from 5 to 40, and it decreases when the lg value changes from 40 to 45. There is a peak point with an average accuracy of 95.86% when the value of lg was 40. We can draw a conclusion; when the parameters lg of the AC algorithm are less than 40 or the number of amino acids is less than 40, protein sequences will lose some useful information, but larger lg may introduce noise rather than improvnig the performance of the model. So we set the value of lg as 40.

Secondly, we discuss the parameters of the RF. Based on previous studies, we chose PCA as Rotation Forest conversion method. Additionally, the J48 decision tree was selected as the base classifier from the WEKA database. In this experiment, two parameters (the number of feature subsets K and the number of decision trees L) were tested by the grid search method in the range of values to achieve better performance. Figure 3 shows the prediction results of different parameters. We can see that accuracy fluctuates at the beginning and then is slowly enhanced with the increase of L, but it seems to be not closely related to the increase of K. Considering the accuracy rate and the time cost of the algorithm, as a result, we obtained optimal parameters of $K = 20$ and $L = 3$. For the *H. pylori* dataset, we use the AC to extract features and RF validation with the same parameters with the *Yeast* dataset.

The 5-fold cross-validation method was introduced to reduce the dependence of the data on the prediction model

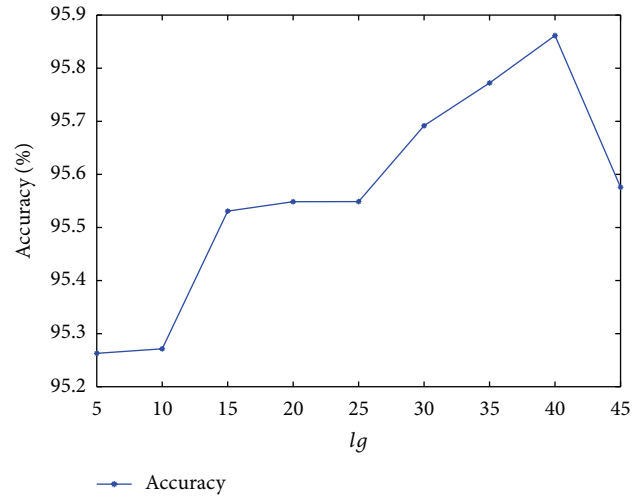


Figure 2: The average prediction accuracy corresponding to different lg of the AC algorithm in the proposed model.

Table 1: 5-fold cross-validation results obtained by using the proposed method on *Yeast* dataset.

Testing set	Accu. (%)	Prec. (%)	Sen. (%)	MCC (%)
1	97.59	100.00	95.14	95.28
2	97.54	100.00	95.03	95.19
3	98.17	100.00	96.40	96.40
4	97.59	100.00	95.01	95.27
5	97.99	99.82	96.27	96.06
Average	*97.77 ± 0.29*	*99.96 ± 0.08*	*95.57 ± 0.70*	*95.64 ± 0.55*

[51–55]. Table 1 lists all of the prediction results; the prediction accuracies were greater than 97.54%, the precisions were greater than 99.82%, and the sensitivities were greater than 95.01%. Our proposed method can yield an average prediction accuracy of 97.77 ± 0.29%. The ROC curves performed on *Yeast* dataset were shown in Figure 4. In this figure, x-ray depicts False Positive Rate (FPR) while y-ray depicts True Positive Rate (TPR).

3.3. Comparison with the Proposed Method on H. pylori Dataset. For analyzing the ability of the proposed method

Table 2: 5-fold cross-validation results obtained by using the proposed method on *H. pylori* dataset.

Testing set	Accu. (%)	Prec. (%)	Sen. (%)	MCC (%)
1	85.76	87.45	82.87	75.52
2	83.53	82.65	84.38	72.49
3	86.11	87.55	83.57	76.02
4	81.99	83.27	79.51	70.42
5	86.82	90.88	83.55	77.06
Average	*84.84 ± 2.01*	*86.36 ± 3.40*	*82.77 ± 1.90*	*74.30 ± 2.76*

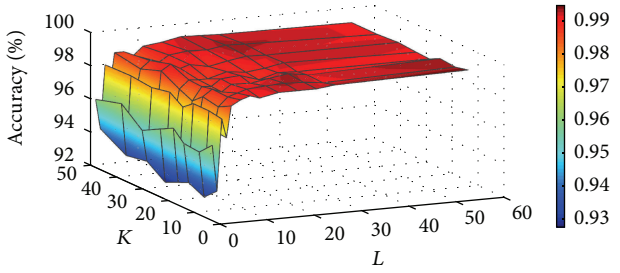


Figure 3: Accuracy surface obtained from Rotation Forest for optimizing regularization parameters K and L.

to predict PPIs, we tested its ability in different dataset. We used the proposed method to predict interactions on the *H. pylori* dataset. A total of 2916 proteins were included in this database, half of which were interacting pairs and the other half were noninteracting pairs. Our prediction results were shown in Table 2. We can see an accuracy, precision, sensitivity, and MCC of 84.84%, 86.36%, 82.77%, and 74.30%, respectively. The ROC curves performed on *H. pylori* dataset were shown in Figure 5.

3.4. Comparison with Previous Method. In order to more clearly assess the proposed method, we compared its results with the previous models on the *Yeast* dataset. As a classic classification algorithm, Support Vector Machine has a very superior performance in identifying interacting and noninteracting protein pairs. For example, Guo et al. [9] proposed a new method with Support Vector Machine combined with Autocovariance to predict Protein-Protein Interactions in *Yeast* dataset, and the results have proven its ability. Specifically, we use the same feature extraction method (AC) combined with PSSMs to compare the classification performance between Rotation Forest and SVM in the same dataset. We use grid search method to optimize the parameters of Support Vector Machine and set $c = 0.5$ and $g = 0.6$, respectively. The LIBSVM tools we adopted can be downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. As can be seen from Table 3, when using SVM to predict PPIs of *Yeast* dataset, we obtained excellent results with the accuracy, precision, sensitivity, and MCC of 95.86%, 96.46%, 95.21%, and 92.06%, respectively. Most of the SVM based methods produce average standard values that were lower than our method on *Yeast* dataset.

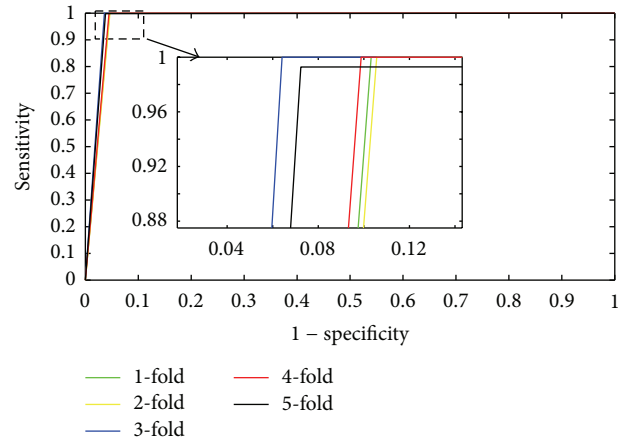


Figure 4: ROC curves performed by the proposed method on *Yeast* PPIs dataset.

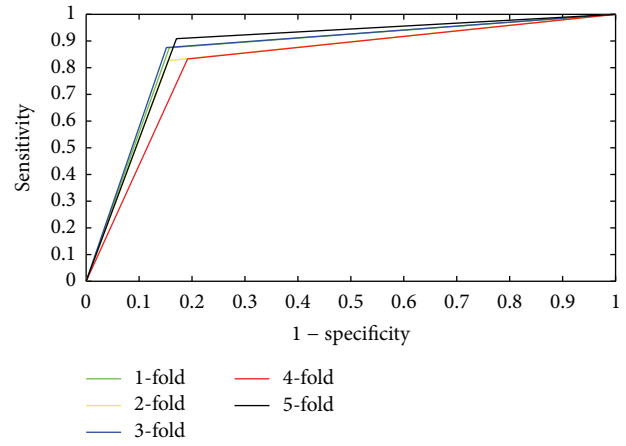


Figure 5: ROC curves performed by the proposed method on *H. pylori* dataset.

In addition, we also compared the other existing methods on the *Yeast* and *H. pylori* datasets. Table 3 shows the average results of the other six methods in the *Yeast* dataset; we can see that the accuracy results obtained by these methods are between 75.08% and 89.33%. The average accuracy, precision, sensitivity, and MCC values of these methods are lower than those of our method, which are 97.77%, 99.96%, 95.57%, and 95.64%, respectively. Table 4 shows the average predictive values of the six kinds of methods on the *H. pylori* dataset. We can see that the accuracy values obtained by these methods are between 75.80% and 87.50%, and the accuracy value of our proposed method is 84.84%, which also performs well in it.

3.5. Performance on Independent Dataset. Having achieved reasonably good results on the *Yeast* dataset and the *H. pylori* dataset, we decided to test the proposed method's performance on *independent datasets*. We built our final prediction model using all 11188 pairs of *Yeast* dataset as the training set with the parameters obtained by the grid search method; the value of lg is 40 in AC, the value of K is 20, and L is 3 in RF. The feature vector uses the feature

Table 3: Different methods on *Yeast* dataset performance comparison.

Model	Test set	Accu. (%)	Prec. (%)	Sen. (%)	MCC (%)
Guo et al.'s work [9]	ACC	89.33 ± 2.67	88.87 ± 6.16	89.93 ± 3.68	N/A
	AC	87.36 ± 1.38	87.82 ± 4.33	87.30 ± 4.68	N/A
You et al.'s work [25]	PCA-EELM	87.00 ± 0.29	87.59 ± 0.32	86.15 ± 0.43	77.36 ± 0.44
Yang et al.'s work [56]	Cod1	75.08 ± 1.13	74.75 ± 1.23	75.81 ± 1.20	N/A
	Cod2	80.04 ± 1.06	82.17 ± 1.35	76.77 ± 0.69	N/A
	Cod3	80.41 ± 0.47	81.86 ± 0.99	78.14 ± 0.90	N/A
	Cod4	86.15 ± 1.17	90.24 ± 0.45	81.03 ± 1.74	N/A
Zhou et al.'s work [57]	SVM + LD	88.56 ± 0.33	89.50 ± 0.60	87.37 ± 0.22	77.15 ± 0.68
Our method	*SVM + PSSM*	*95.86 ± 0.34*	*96.46 ± 0.50*	*95.21 ± 0.70*	*92.06 ± 0.62*
	RF + PSSM	*97.77 ± 0.29*	*99.96 ± 0.08*	*95.57 ± 0.70*	*95.64 ± 0.55*

Table 4: Different methods on *H. pylori* dataset performance comparison.

Model	Accu. (%)	Prec. (%)	Sen. (%)	MCC (%)
Phylogenetic bootstrap [35]	75.80	80.20	69.80	N/A
HKNN [36]	84.00	84.00	86.00	N/A
Ensemble of HKNN [37]	86.60	85.00	86.70	N/A
Signature products [26]	83.40	85.70	79.90	N/A
Boosting [38]	79.52	81.69	80.37	70.64
Ensemble ELM [25]	87.50	86.15	88.95	78.13
Our method	*84.84*	*86.36*	*82.77*	*74.30*

Table 5: Prediction results in *independent datasets*.

Species	Test pairs	Accu. (%)
C. elegans	4013	96.01
E. coli	6954	97.73
H. sapiens	1412	98.30
M. musculus	313	96.81

extraction method (AC) based on the PSSMs to extract from the four datasets as RF test input. Independent test dataset is composed of the four databases (*C. elegans*, *E. coli*, *H. sapiens*, and *M. musculus*) collected in DIP database. The results of our model are listed in Table 5; the prediction accuracies on *C. elegans*, *E. coli*, *H. sapiens*, and *M. musculus* are 96.01%, 97.73%, 98.30%, and 96.81%, respectively. Those results show the excellent performance of our approach in predicting the accuracy of the interactions of other species.

4. Conclusions

In this study, a stable and robust computational method based on the features extracted from PSSM has been proposed to predict PPIs. It is known that the main computational challenge for sequence-based methods for predicting PPIs is to find a suitable feature representation to fully describe the important information of protein interactions. To solve this problem, we here firstly extracted the features from the Position-Specific Scoring Matrices (PSSMs) using Autocovariance (AC) method. Then, Rotation Forest (RF) model is employed as a novel and accurate classifier for PPIs prediction with better performance than state-of-the-art SVM classifier. In order to evaluate the performance of the proposed method, five PPIs datasets, that is, *C. elegans*, *E. coli*, *H. pylori*, *H. sapiens*, and *M. musculus*, have been used to perform the comparisons. As expected, the experiments results showed that the proposed method performs better than the other methods. Consequently, the proposed approach can be considered as a powerful tool for predicting PPI.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Authors' Contributions

Zhen-Guo Gao and Lei Wang contributed equally to this work.

Acknowledgments

This work is supported by the National Science Foundation of China, under Grants 61373086, 61572506, and 61401385, in part by the Guangdong Natural Science Foundation under Grant 2014A030313555, and in part by the Shenzhen Scientific Research and Development Funding Program under Grant JCYJ20140418095735569.

References

[1] Z. Yin, T. Deng, L. E. Peterson et al., "Transcriptome analysis of human adipocytes implicates the NOD-like receptor pathway in obesity-induced adipose inflammation," *Molecular and Cellular Endocrinology*, vol. 394, no. 1-2, pp. 80–87, 2014.

[2] A.-C. 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.

[3] K. A. Theofilatos, C. M. Dimitrakopoulos, A. K. Tsakalidis, S. D. Likothanassis, S. T. Papadimitriou, and S. P. Mavroudi, "Computational approaches for the prediction of protein-protein interactions: a survey," *Current Bioinformatics*, vol. 6, no. 4, pp. 398–414, 2011.

[4] N. Tuncbag, G. Kar, O. Keskin, A. Gursoy, and R. Nussinov, "A survey of available tools and web servers for analysis of protein-protein interactions and interfaces," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 217–232, 2009.

[5] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[6] 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.

[7] 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.

[8] J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal, "Effect of sampling on topology predictions of protein-protein interaction networks," *Nature Biotechnology*, vol. 23, no. 7, pp. 839–844, 2005.

[9] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[10] Z.-H. You, Z. Yin, K. Han, D.-S. Huang, and X. Zhou, "A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network," *BMC Bioinformatics*, vol. 11, article 343, 2010.

[11] J. Shen, J. Zhang, X. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4337–4341, 2007.

[12] Z. Ji, B. Wang, S. P. Deng, and Z. You, "Predicting dynamic deformation of retaining structure by LSSVR-based time series method," *Neurocomputing*, vol. 137, pp. 165–172, 2014.

[13] 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.

[14] L. Zhu, Z.-H. You, and D.-S. Huang, "Increasing the reliability of protein-protein interaction networks via non-convex semantic embedding," *Neurocomputing*, vol. 121, pp. 99–107, 2013.

[15] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, no. 1, pp. S233–S240, 2002.

[16] R. Jothi, M. G. Kann, and T. M. Przytycka, "Predicting protein-protein interaction by searching evolutionary tree automorphism space," *Bioinformatics*, vol. 21, no. 1, pp. I241–I250, 2005.

[17] A. J. Enright, I. Illopoulos, N. C. Kyrpides, and C. A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events," *Nature*, vol. 402, no. 6757, pp. 86–90, 1999.

[18] Q. C. Zhang, D. Petrey, L. Deng et al., "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[19] Z. Yin, A. Sadok, H. Sailem et al., "A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes," *Nature Cell Biology*, vol. 15, no. 7, pp. 860–871, 2013.

[20] Z. Yin, X. Zhou, C. Bakal et al., "Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput RNAi screens," *BMC Bioinformatics*, vol. 9, no. 1, article 264, 2008.

[21] Y. Mao, Z. Xia, Z. Yin, Y. Sun, and Z. Wan, "Fault diagnosis based on fuzzy support vector machine with parameter tuning and feature selection," *Chinese Journal of Chemical Engineering*, vol. 15, no. 2, pp. 233–239, 2007.

[22] Z.-H. You, Y.-K. Lei, J. Gui, D.-S. Huang, and X. Zhou, "Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data," *Bioinformatics*, vol. 26, no. 21, pp. 2744–2751, 2010.

[23] Y.-K. Lei, Z.-H. You, Z. Ji, L. Zhu, and D.-S. Huang, "Assessing and predicting protein interactions by combining manifold embedding with multiple information integration," *BMC Bioinformatics*, vol. 13, supplement 7, article S3, 2012.

[24] Y.-K. Lei, Z.-H. You, T. Dong, Y.-X. Jiang, and J.-A. Yang, "Increasing reliability of protein interactome by fast manifold embedding," *Pattern Recognition Letters*, vol. 34, no. 4, pp. 372–379, 2013.

[25] Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC Bioinformatics*, vol. 14, supplement 18, article S10, 2013.

[26] S. Martin, D. Roe, and J.-L. Faulon, "Predicting protein-protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.

[27] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: a new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006.

[28] L. Nanni and A. Lumini, "Ensemble generation and feature selection for the identification of students with learning disabilities," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3896–3900, 2009.

[29] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Profile analysis: detection of distantly related proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 13, pp. 4355–4358, 1987.

[30] 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.

[31] C. M. Deane, Ł. Salwiński, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Molecular & Cellular Proteomics*, vol. 1, no. 5, pp. 349–356, 2002.

[32] D. R. Cutler, T. C. Edwards Jr., K. H. Beard et al., "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[33] P. L. Braga, A. L. I. Oliveira, G. H. T. Ribeiro, and S. R. L. Meira, "Bagging predictors for estimation of software project effort," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '07)*, pp. 1595–1600, Orlando, Fla, USA, August 2007.

[34] J. C. Rain, L. Selig, H. De Reuse et al., "The protein-protein interaction map of *Helicobacter pylori*," *Nature*, vol. 409, no. 6817, pp. 211–215, 2001, Erratum to *Nature*, vol. 409, no. 6821, article 743, 2001.

[35] J. R. Bock and D. A. Gough, "Whole-proteome interaction mining," *Bioinformatics*, vol. 19, no. 1, pp. 125–135, 2003.

[36] L. Nanni, "Hyperplanes for predicting protein-protein interactions," *Neurocomputing*, vol. 69, no. 1–3, pp. 257–263, 2005.

[37] L. Nanni and A. Lumini, "An ensemble of K-local hyperplanes for predicting protein-protein interactions," *Bioinformatics*, vol. 22, no. 10, pp. 1207–1210, 2006.

[38] B. Liu, J. Yi, A. Sv et al., "QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions," *BMC Genomics*, vol. 14, no. 8, article S3, 2013.

[39] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology*, vol. 292, no. 2, pp. 195–202, 1999.

[40] D. T. Jones and J. J. Ward, "Prediction of disordered regions in proteins from position specific score matrices," *Proteins: Structure, Function and Genetics*, vol. 53, no. 6, pp. 573–578, 2003.

[41] X.-W. Chen and J. C. Jeong, "Sequence-based prediction of protein interaction sites with an integrative method," *Bioinformatics*, vol. 25, no. 5, pp. 585–591, 2009.

[42] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[43] Y. Guo, M. Li, M. Lu, Z. Wen, and Z. Huang, "Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 1, pp. 55–60, 2006.

[44] M. Lapinsh, A. Gutcaits, P. Prusis, C. Post, T. Lundstedt, and J. E. S. Wikberg, "Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences," *Protein Science*, vol. 11, no. 4, pp. 795–805, 2002.

[45] Z. Lin and X.-M. Pan, "Accurate prediction of protein secondary structural content," *Journal of Protein Chemistry*, vol. 20, no. 3, pp. 217–220, 2001.

[46] C.-T. Zhang, Z.-S. Lin, Z. Zhang, and M. Yan, "Prediction of the helix/strand content of globular proteins based on their primary sequences," *Protein Engineering*, vol. 11, no. 11, pp. 971–979, 1998.

[47] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[48] T.-H. Lee and Y. Yang, "Bagging binary and quantile predictors for time series," *Journal of Econometrics*, vol. 135, no. 1-2, pp. 465–497, 2006.

[49] G. Rätsch, S. Mika, B. Schölkopf, and K.-R. Müller, "Constructing boosting algorithms from SVMs: an application to one-class classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1184–1199, 2002.

[50] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577, 1993.

[51] Z.-H. You, J. Li, X. Gao et al., "Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines," *BioMed Research International*, vol. 2015, Article ID 867516, 9 pages, 2015.

[52] Z.-H. You, K. C. C. Chan, and P. Hu, "Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest," *PLoS ONE*, vol. 10, no. 5, article e0125811, 2015.

[53] Y. Huang, Z. You, X. Gao, L. Wong, and L. Wang, "Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence," *BioMed Research International*, vol. 2015, Article ID 902198, 10 pages, 2015.

[54] Z.-H. You, J.-Z. Yu, L. Zhu, S. Li, and Z.-K. Wen, "A MapReduce based parallel SVM for large-scale predicting protein-protein interactions," *Neurocomputing*, vol. 145, pp. 37–43, 2014.

[55] Z.-H. You, S. Li, X. Gao, X. Luo, and Z. Ji, "Large-scale protein-protein interactions detection by integrating big biosensing data with computational model," *BioMed Research International*, vol. 2014, Article ID 598129, 9 pages, 2014.

[56] L. Yang, J.-F. Xia, and J. Gui, "Prediction of protein-protein interactions from protein sequence using local descriptors," *Protein and Peptide Letters*, vol. 17, no. 9, pp. 1085–1090, 2010.

[57] Y. Z. Zhou, Y. Gao, and Y. Y. Zheng, "Prediction of protein-protein interactions using local description of amino acid sequence," in *Advances in Computer Science and Education Applications, Part II*, vol. 02, pp. 254–262, Springer, Berlin, Germany, 2011.