Computational and Structural Biotechnology Journal 23 (2024) 813–833

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review Article

Databases and computational methods for the identification of piRNA-related molecules: A survey

Chang Guo [a], Xiaoli Wang [b], Han Ren [a,c,*]

[a] Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510420, China
[b] Institute of Reproductive Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
[c] Laboratory of Language and Artificial Intelligence, Guangdong University of Foreign Studies, Guangzhou 510420, China

ARTICLE INFO

Keywords:
piRNA
Machine learning
Deep learning
Computational methods
piRNA–disease association prediction

ABSTRACT

Piwi-interacting RNAs (piRNAs) are a class of small non-coding RNAs (ncRNAs) that plays important roles in many biological processes and major cancer diagnosis and treatment, thus becoming a hot research topic. This study aims to provide an in-depth review of computational piRNA-related research, including databases and computational models. Herein, we perform literature analysis and use comparative evaluation methods to summarize and analyze three aspects of computational piRNA-related research: (i) computational models for piRNA-related molecular identification tasks, (ii) computational models for piRNA–disease association prediction tasks, and (iii) computational resources and evaluation metrics for these tasks. This study shows that computational piRNA-related research has significantly progressed, exhibiting promising performance in recent years, whereas they also suffer from the emerging challenges of inconsistent naming systems and the lack of data. Different from other reviews on piRNA-related identification tasks that focus on the organization of datasets and computational methods, we pay more attention to the analysis of computational models, algorithms, and performances that aim to provide valuable references for computational piRNA-related identification tasks. This study will benefit the theoretical development and practical application of piRNAs by better understanding computational models and resources to investigate the biological functions and clinical implications of piRNA.

1. Introduction

Piwi-interacting RNAs (piRNAs) are a class of small non-coding RNA (ncRNAs) [1] that are 26–32 nucleotides in length [1,2], slightly longer than other small ncRNAs. piRNAs have attracted the attention of many researchers in the fields of molecular biology, genomics, and biomedicine since they were first isolated from the vas deferens of male mice in 2006 [3]. Initial studies have found that piRNAs are specifically expressed in animal germ cells and play an important role in germline integrity and stem cell development [4]. Within the past decade, growing evidence has continuously extended our knowledge of piRNAs, showing that they are also involved in many biological processes such as transposon silencing, histone modification, translational control, DNA methylation, and viral defense [5,6]. Such evidence reveals the mechanistic insights of piRNAs in the regulation of gene expression in both germ and somatic cells as well as diseases caused by piRNA dysregulation, which may help identify new biomarkers and therapeutic targets for many diseases [7]. For example, piR-hsa-25781, piRhsa-28467, piR-hsa-1177, piR-hsa-26593, and piR-hsa-29114 may be effective biomarkers for Alzheimer's disease [8], whereas piR-823 may be strongly associated with breast cancer [9], gastric cancer [10], kidney cancer [11], rectal cancer [12] and liver cancer [13]. Therefore, studying the mechanisms of piRNAs in biological processes and the potential association between piRNAs and diseases will undoubtedly help to better understand the function of piRNAs and the pathogenesis of related diseases. Recently, a review article [14] comprehensively summarized the latest advances in piRNA biology, including the biogenesis, function and mechanism of piRNAs, as well as their novel roles in Drosophila and mouse germline development and human infertility, cancer and neurological diseases. This article provides an important background and reference for this study.

Currently, research on piRNAs has mainly focused on their molecular mechanisms and functions in various diseases. Biological experiments contribute to accurately identifying piRNAs and finding piRNA–disease

* Correspondence to: Laboratory of Language Engineering and Computing, Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou 510420, China.
E-mail address: hanren@gdufs.edu.cn (H. Ren).

https://doi.org/10.1016/j.csbj.2024.01.011
Received 11 September 2023; Received in revised form 31 December 2023; Accepted 15 January 2024
Available online 22 January 2024

associations; however, such work always requires delicate and expensive experimental settings [15], such as gene knockout [16] and RNA interference [17]. In addition, because the number of piRNAs and diseases is continuously increasing, finding an association between them through biological experiments is time-consuming. In recent years, computational analysis approaches, such as classical machine learning algorithms and deep learning algorithms, have been applied in piRNA-related research. For example, Liu et al. [18] and Wang et al. [19] used support vector machine (SVM) [20] and convolutional neural network (CNN) [21] approaches to identify piRNAs from different ncRNAs. Wei et al. [22] and Zheng et al. [23] employed random forest (RF) [24] and Graph ATtention neural networks (GAT) [25] to predict potential associations between piRNAs and diseases. In addition, large-scale piRNA databases such as piRNABank [26] and piRBase [27–29] have been built for computational analysis. The hypothesis of this study was that computational methods can effectively assist in the identification of piRNAs and the prediction of piRNA–disease associations; therefore, a thorough review of piRNA-related computational models and resources can provide valuable information for the functional research and clinical application of piRNAs.

In this study, we reviewed the use of computational methods for piRNA molecular identification and piRNA–disease association prediction. The results of this study can help analyze and predict the features, clusters, targets, and functions of piRNAs, provide a theoretical basis and computational tools for revealing the biological mechanisms of piRNAs, help investigate and evaluate the association of piRNAs with human diseases, especially reproductive, cardiac, neurological, and cancer diseases, and provide new molecular biomarkers and drug targets for disease diagnosis, treatment, and prevention. In the piRNA molecular identification task, researchers construct different classifiers or clusters to identify piRNAs and their related molecules based on machine learning methods, using the sequence, structure, expression, and functional features of piRNAs, as well as the interaction features of piRNAs with other molecules. Using deep learning methods, different neural network layers were constructed to learn the embedding representation of piRNAs and their related molecules and then identify piRNAs and their related molecules. For example, Costa et al. [30] proposed a deep learning model, piRNet, to help doctors and researchers quickly discover transposon-derived piRNAs, thus improving the understanding of the role of piRNAs in transposon silencing and genome stability. Khan et al. [31] proposed a model named 2 L-piRNADNN to provide doctors and researchers with a fast and accurate method and tool for identifying functional piRNAs, thus improving the utilization efficiency and value of functional piRNAs.

In addition, in the piRNA–disease association prediction task, since the number and function of known piRNAs are still very limited, the experimental verification of the association between piRNAs and diseases is time-consuming and resource-intensive. Using computational methods to predict the association between piRNAs and diseases can help doctors and researchers quickly discover the role of piRNAs in different diseases, thus improving the understanding and identification of diseases. For example, iPiDA-sHN [32] predicted that piR-has-1849, piR-hsa-23209, piR-hsa-23210, piR-hsa-15023, and piR-hsa-1823 were associated with Alzheimer's disease, and these piRNAs showed different expression levels in patients with Alzheimer's disease than in normal people and may be related to neuronal damage and repair. iPiDi-PUL [22] predicted that piR-hsa-1849 was associated with breast cancer, and it was downregulated in breast cancer tissues and may be used as a potential biomarker for breast cancer. iPiDA-LTR [33] predicted that piR-hsa-23210 and piR-hsa-23209 were associated with multiple diseases, and their target genes played important roles in human spermatogenesis and nervous system development and may be key therapeutic targets. iPiDA-GCN [34] predicted that piR-hsa-31280 and piR-hsa-8245 were associated with cardiovascular diseases, and they were abnormally expressed in cardiovascular disease tissues and may be related to cardiac function and repair.

Efforts have been made to summarize the computational knowledge in piRNA-related research. Liu et al. [35] reviewed computational methods and resources for identification tasks related to piRNAs, where feature engineering approaches and data sources for piRNA-related identification were summarized taxonomically. Zhang et al. [36] investigated existing piRNA databases and outlined computational methods for piRNA functions. Briefly, these studies retrospectively analyzed piRNA-related computational resources and methods. However, they involved a few summative investigations into current computational models for piRNA-related tasks, including architectures, algorithms, and performance comparisons, which are essential as an informative reference for piRNA-related computational studies. In addition, not all up-to-date computational studies, especially state-of-the-art models for piRNA-related prediction, have been included in these overviews.

This study provides a comprehensive review of the computational resources and methods for piRNA-related prediction tasks. The aim of this study was to help researchers better understand the recent advances in computational piRNA studies, especially emerging deep learning technologies. To this end, this study investigates computational models for these tasks, including architectures, algorithms, and performances, and discusses research challenges and perspectives. This study offers practical guidance for the computational modeling of piRNA-related prediction tasks.

The remainder of this manuscript is organized as follows. Section 2 presents the relevant dataset resources. Section 3 reviews the performance evaluation metrics of the computational methods for piRNA-related tasks. Section 4 summarizes the computational methods used for piRNA-related identification. Section 5 focuses on computational models for piRNA–disease association prediction. Finally, Section 6 discusses the limitations and challenges of the current computational methods for piRNA-related tasks and provides perspectives for future research.

2. Databases

With the development of high-throughput sequencing technologies, an increasing number of piRNAs has been identified. Two public databases, NONCODE [37] and NCBI Genetic Expression Omnibus [38], were the primary data sources for the construction of initial piRNA datasets. With the accumulation of piRNA-related knowledge, datasets for various piRNA-related computational tasks have recently been established. Such datasets can be taxonomically divided into six types: piRNA comprehensive annotation databases, piRNA cluster annotation databases, piRNA target databases, piRNA isoform annotation databases, PIWI-bound piRNA databases, and piRNA–disease association databases; the last four of which are mainly built for specific tasks.

2.1. piRNA comprehensive annotation databases

These databases are piRNAdb [39], piRNABank [26] and piRBase [27–29], which provide comprehensive information on RNA sequences, alignments, clusters, genomic elements, and targets. piRBase is the largest comprehensive piRNA annotation database available. It was first released in 2014 and updated to its third version in 2021. Currently, it contains more than 181,000,000 unique piRNA sequences from 44 species.

2.2. piRNA cluster databases

piRNA clusters are piRNA-producing loci that are transcribed from one or both DNA strands into large precursor piRNAs, which are then converted into mature piRNAs [40]. piRNA cluster databases, including piRNAQuest [2] and piRNAclusterDB [41,42], focus on the annotation of piRNA cluster information. piRNAQuest annotates the piRNA clusters of humans, mice, and rats and provides piRNA feature motifs and

expression profiles; it is a database that uses the piRNA cluster as a central resource, which was first released in 2014 and updated to the third version in 2021. Currently, the piRNAclusterDB contains piRNA sequences that have expanded from 12 to 51 species.

2.3. piRNA target databases

To facilitate the development of piRNA target prediction, piRNA-target [43] and piRTarBase [44] are built. piRNA targets contain information about human piRNAs, such as annotations, sequences, parental genes, targets, expression, mutations, and methylation profiles. However, they are not publicly accessible. piRTarBase uses pirScan [45], a piRNA target prediction tool, to predict piRNA target sites in published mRNA and small RNA sequencing data. These databases can help users to identify functional piRNA targets.

2.4. piRNA isoform databases

piRNA isoforms arise from their maturation process, especially the trimming step, which can modulate the ends of piRNAs, thereby resulting in variations in piRNA length [46]. IsopiRBank [47] was the first integrative database to focus on piRNA isoforms. Through the analysis of 2154 small RNA sequencing datasets from four species, including *Homo sapiens*, *Mus musculus*, lion, and *Drosophila melanogaster*, IsopiRBank collected 874,913,9 piRNA isoforms using the CPSS algorithm [48,49].

2.5. PIWI-bound piRNA databases

piRNA-IPdb [50] is a database based on PIWI protein-bound piRNA sequences for analyzing the data and functions of small ncRNA. This database provides a more accurate and reliable source of data on piRNAs. It includes 18,821,815 piRNA sequences that are only bound to PIWI proteins screened from piRBase; features of PIWI protein-bound piRNAs, such as length distribution, first uridine preference, and reverse complementary overlap, to verify the characteristics and classification of piRNAs; and relationships between piRNAs and miRNAs, such as overlap, complementarity, and co-expression, to discover the possible duality and regulatory mechanisms between piRNAs and miRNAs.

2.6. piRNA–disease association databases

piRDisease [51] was the first database to study the piRNA–disease association; however, the website of the database is currently inaccessible. piRPheno [52] is a database that provides a reference for the association between piRNA and diseases. It contains 9057 associations between 474 piRNAs and 204 diseases. Each association is assigned a clinical correlation label with a confidence score according to the evidence supported by the experiments.

The databases mentioned above provide many features, including comprehensive piRNA-related data, experimental piRNA–disease associations, piRNA sequence and location information, disease semantics, and functional information. These databases effectively facilitate computational studies on piRNAs. These databases are summarized in Table 1.

Existing piRNA databases were further classified into three categories according to the data collection methods: methods based on small RNA sequencing, methods based on experimental verification, and methods based on literature collection. Among them, the data collection method based on sequencing can yield a large number of piRNA sequences and cluster information, reflecting the expression level and distribution characteristics of piRNAs, making it suitable for discovering new piRNAs and studying their biological functions. However, this method requires high-throughput sequencing technology and professional data analysis software and has a high cost, large data volume, high analysis difficulty, and a certain sequencing error rate and bias. The data collection method based on experimental verification can directly observe and verify the targets and interactions of piRNAs, reflecting the regulatory mechanism and effect of piRNAs, and is suitable for studying the functional genomics and proteomics of piRNAs. However, experimental verification methods require specific experimental conditions and materials, are affected by experimental design and operation, have low data reproducibility and comparability, and have difficulty covering the whole genome range. Data collection methods based on the literature can use existing research results and data, save time and resources, and discover new knowledge and rules suitable for data mining, knowledge graphs, and other scenarios. However, this method requires the screening and evaluation of the quality and credibility of the literature. Data integrity and consistency are poor, and there may be omissions and biases; thus, it is difficult to reflect on the latest research progress.

In Table 1, it can be seen that some databases use multiple information collection methods at the same time, which increases their information volume and reliability. However, different databases have different definitions and annotations for piRNAs, which may affect their accuracy and consistency. Researchers must compare and verify the information when using these databases. Different piRNA databases have different characteristics and applicabilities, and appropriate databases can be selected for queries and analyses according to different research purposes. Therefore, the selection of appropriate databases should be based on the specific needs and objectives of the research and should comprehensively consider factors such as data source, volume, quality, and annotation information of the database. In addition, these databases provide many features, including comprehensive piRNA-related data, experimental piRNA–disease association, piRNA sequence and location information, disease semantics, and functional information. These databases have effectively promoted the computational research on piRNAs.

3. piRNA-related molecule identification

Over the past few decades, computational efforts have been made to identify unexplored piRNA characteristics and the association between piRNAs and diseases using traditional machine learning algorithms, most of which are highly dependent on feature engineering methods. To this end, the features of piRNA-related tasks have been investigated, such as piRNA sequence features (k-mer-based features) [53], piRNA cluster features [54,55], physicochemical features [56], thermodynamic features, and their combinations [35]. These features contribute to the performance of machine learning algorithms, but they require handcrafted labeling. To alleviate the manual work, the research community has turned to deep learning technologies to automatically learn feature representations based on neural network models [57]. For example, deep feedforward neural networks (DFNN) [30] and CNN [19] have been employed to learn feature representations of piRNA sequences, whereas GAT [23] has been built to learn the structural topology of piRNA–disease association networks for piRNA identification. Meanwhile, some studies also retained handcrafted features, which helped alleviate the problem of data sparsity [56,58,59]. Fig. 1 shows the general computational framework for piRNA-related identification.

This section summarizes the computational methods for piRNA-related identification tasks, which are further divided into five categories: piRNA identification, transposon-derived piRNA identification, functional piRNA identification, piRNA target prediction, and piRNA cluster identification. All tasks have different goals; however, some of them may share the same features and algorithms. Table 2 lists the computational methods used for these tasks.

3.1. piRNA identification

Computational piRNA identification can be treated as a classification task, and it always includes two steps: 1) constructing experimental

Table 1
piRNA database resources.

Type	Name	Collection methods	Species	Description	Advantages	Disadvantages	URL
piRNA comprehensive annotation database	piRNAdb [39]	sequencing and literature	*Homo sapiens, Mus musculus, Rattus norvegicus, Cricetulus griseus, C. elegans, and Drosophila melanogaster*	This database contains piRNA database resources of 16 data sets of 6 species.	User-friendly, provides powerful storage and search functions, contains information on various species, and updates frequently.	There is no strict definition and annotation of piRNA, which may lead to false positives or false negatives. The quality and consistency of the data need to be improved, and the information on piRNA function and disease relevance is lacking.	https://www.pirnadb.org/
	piRNABank [26]	sequencing and literature	Humans, mice, rats	piRNABank contains 23,439 human-, 39,986 mouse-, and 38,549 rat-specific piRNA sequences.	Contains the most piRNA sequences and datasets and provides various search and download options.	There is no strict definition and annotation of piRNA.	http://pirnabank.ibab.ac.in/
	piRBase [27–29]	sequencing and literature	44 species including humans	PiRBase is currently the largest piRNA sequence data repository.	Contains the most comprehensive piRNA information, provides rich annotations and visualization functions, contains piRNA-related epigenetic data and disease information, and updates frequently.	The web interface and operation are somewhat complex, and the information on piRNA function and disease relevance is lacking.	http://bigdata.ibp.ac.cn/piRBase/
piRNA cluster database	piRNAclusterDB [41]	sequencing and literature	12 species including humans	The first database resource regards the piRNA genome cluster as a biological source of piRNA.	Provides detailed information and classification of piRNA clusters, contains information on various species, and updates frequently.	The database has a small amount of data and a narrow coverage.	http://www.smallrnagroup-mainz.de/piRNAclusterDB.html
	piRNAQuest [2]	sequencing and literature	Humans, mice, rats	The database contains piRNA sequences from 41,749 humans, 890,078 mice, and 66,758 rats.	Provides prediction and analysis tools for piRNA sequences and clusters.	The data quality is uneven and the annotation information is lacking.	http://bicresources.jcbose.ac.in/zhumur/pirnaquest
piRNA target database	piRNAtarget [43]	sequencing and literature	Humans	This database is mainly used to study the functions of human piRNA and its targets.	Provides prediction and analysis tools for piRNA targets.	The database has a small amount of data and a narrow coverage.	http://120.108.102.11/sophia/piRNAtarget
	piRTarBase [44]	experiment and literature	*C. elegans*	The database will use the pirScan tool for prediction and experimental methods to discover new piRNA targets.	Provides experimental verification information on piRNA targets and contains information on various species.	The database has a small amount of data and a narrow coverage.	http://cosbi6.ee.ncku.edu.tw/piRTarBase
piRNA isoform database	IsopiRBank [47]	sequencing and literature	*Homo sapiens, Mus musculus, Danio rerio, and Drosophila melanogaster*	This database is the first data resource to focus on the piRNA isoforms.	Provides annotation and analysis functions for isomiR piRNA, data is professional and focused.	The database has a small amount of data and a narrow coverage.	http://mcg.ustc.edu.cn/bsc/isopir/index.html
PIWI-bound piRNA database	piRNA-IPdb [50]	experiment and sequencing	23 species including mice and rats	piRNA-IPdb database contains 23 datasets screened from piRBase. It collects about 18.9 million unique sequences of piRNA that are bound to PIWI proteins, covering different developmental stages, biological samples, and PIWI protein types.	Provides interaction information on piRNA and PIWI proteins, contains information on various species, and updates frequently.	There is no strict definition and annotation of piRNA, which may lead to false positives or false negatives. The quality and consistency of the data need to be improved, and the information on piRNA clusters, transposons, targets, function, and disease relevance is lacking.	https://ipdb2.shinyapps.io/ipdb2/
piRNA–disease association database	piRDisease [51]	literature	Humans, mice, rats	The number of piRNAs:4796, diseases: 28, associations:7939	Provides association data of piRNA and disease, data is reliable and accurate.	The data volume is small, and the current data website is inaccessible.	http://www.piwirna2disease.org/index.php
	piRPheno [52]	Literature	Humans	The number of piRNAs:474, diseases: 204, associations:9057	Provides association data of piRNA and disease, data is reliable and accurate.	The data volume is small but updated in time.	http://www.biomedical-web.com/pirpheno

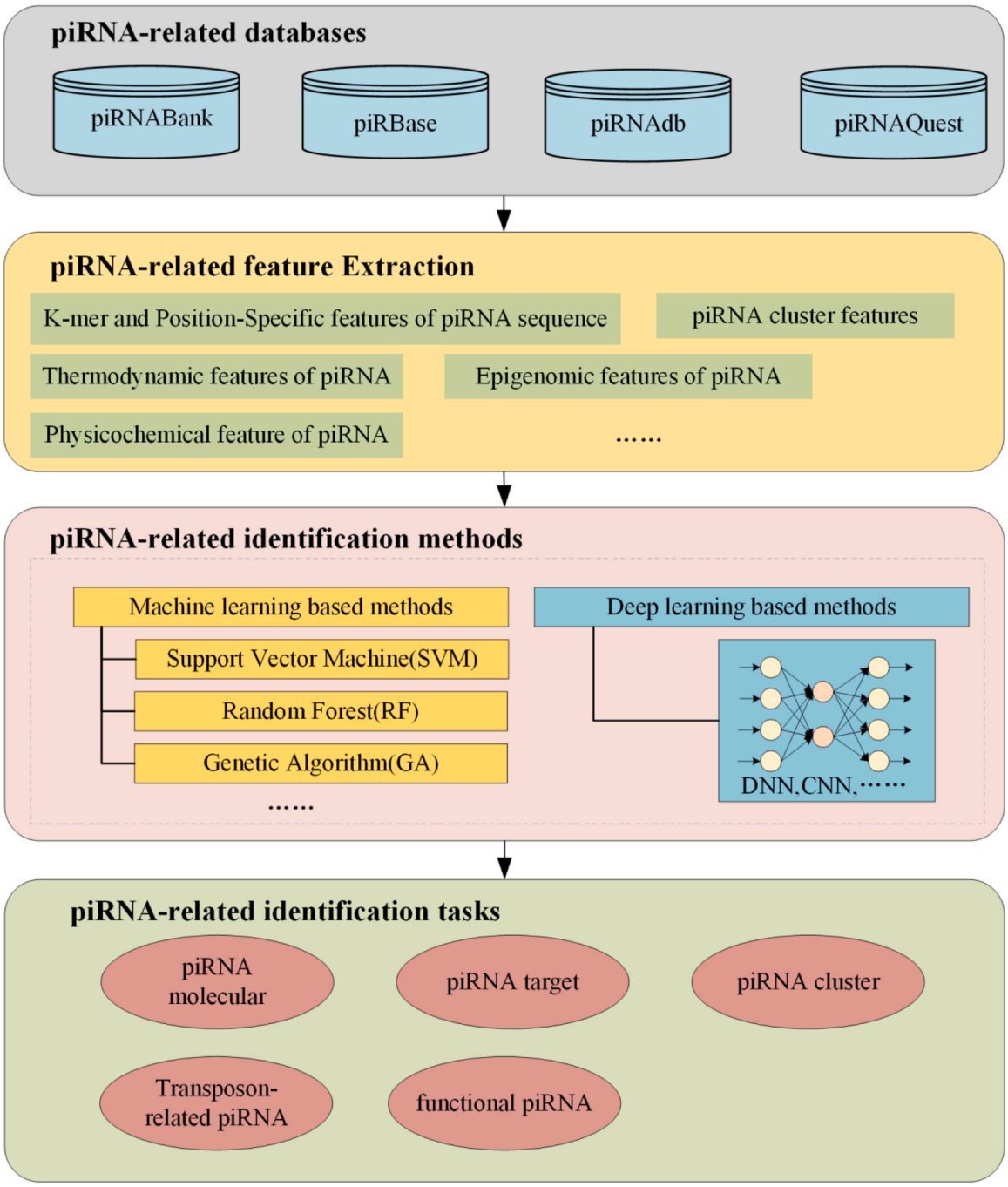

Fig. 1. The general calculation flow of Piwi-interacting RNA-related identification tasks.

datasets in which positive samples are from piRNA database resources and negative ones are from homologous databases of non-piRNAs [19]; 2) employing computational models to identify piRNAs based on the experimental datasets.

The initial work on traditional machine learning was conducted by Betel et al.[3], who proposed a position-specific scoring matrix (PSSM) algorithm to identify mouse piRNAs. The experimental results of PSSM demonstrated the effectiveness of the SVM-based model for piRNA identification, and further studies were conducted to improve its performance. Zhang et al. [53] extracted piRNA sequence features using the k-mer algorithm [60] and employed an SVM classifier to identify piRNAs in the NONCODE database 2.0 [37]. Brayet et al. [54] built a multiple-kernel-based SVM algorithm called piRPred, in which the kernels were built to represent heterogeneous features. Pian et al. [61] proposed a two-stage approach by first clustering piRNAs of homologous families via n-gram models and then identifying potential piRNAs using SVM-based classifiers. In addition, other statistical methods were used to identify piRNAs. For example, Liu et al. [62] applied the computational biology tool Teiresias [63] to identify motifs of variable lengths in mouse piRNA and non-piRNA sequences and built features for classification based on them. Menor et al. [64] utilized an empirical Bayesian kernel method to predict mature miRNAs and piRNAs, thereby avoiding the need for a direct genome reference. Rahiman et al. [65] developed a method based on feature calculations to identify secondary and tertiary piRNA structures.

Traditional machine learning methods rely on effective feature engineering [35]. To this end, researchers have continuously focused on building appropriate features for task-specific modeling. For example,

Table 2
Methods of piRNA-related identification.

Type	Category	Method	Algorithm	Description	Advantage	Disadvantage	Date
piRNA molecule identification	Classical machine learning methods	Betel et al. [3]	SVM	Identifying piRNA uses specific score matrix algorithm (PSSM) based on specific position only mice datdset	Based on the position-specific scoring matrix, it can capture the conservation and variation of piRNA sequences	Only applicable to mouse piRNA, without considering piRNA of other species	2007
		piRNAPredictor [53]	Fisher	Identifying piRNA based on k-MER sequence features	Based on the k-mer algorithm, it can extract the global features of piRNA sequences	Ignore the local features and structural information of piRNA sequences	2011
		piRPred [54]	Multi-kernels SVM	Combining the four features of piRNA to identify piRNA	Based on the multi-kernel SVM algorithm, it can use different types of heterogeneous features to improve the classification performance	Need to adjust the parameters of multiple kernel functions, which increases the computational complexity	2014
		Pibomd [62]	SVM	Identifying piRNA based on motif feature and SVM algorithm	Based on the Teiresias tool, it can identify variable-length motifs in piRNA and non-piRNA sequences and construct classification features	Need to set the minimum and maximum lengths of the motifs, which may cause feature redundancy or missing	2014
		McRUM [64]	Bayesian method	Relying on the nucleotide composition of the read to classify miRNA and piRNA	Based on the empirical Bayesian kernel method, it avoids the need for direct genome reference	Need to rely on known data, which may not be able to discover new piRNA	2015
		Rahiman et al. [65]	SVM based on L1 Gaussian kernel	Identifying piRNA by parameter-based method only on human datdaset	Developed a feature calculation-based method to improve the diversity of features	Need to predict the structure of piRNA, which increases the computation time	2015
		Liu et al. [18]	SVM	Identifying piRNA based on sequence feature and SVM algorithm	Based on weighted k-mer and wildcard-weighted k-mer, and considering the relative importance of nucleotides	It does not consider the structural features of piRNA, which may lose some information	2016
		Seyeddokht et al. [66]	SVM	A method for identifying human piRNA based on SVM algorithm	Using 48 heterogeneous features to encode piRNA, improving the comprehensiveness of features	Need to normalize and select features, which may introduce noise	2016
		piRNAdetect [116]	SVM	N-gram model was used to extract the features of prediction sequence	Based on the n-gram model, it can cluster piRNA molecules of the same homologous family, reducing the classification difficulty	Need to determine the length of the n-gram first, which may affect the feature selection	2017
		IpiRId [55]	multi-kernel fusion SVM	This method uses a large number of heterogeneity features	Using 12 different types of features to improve the accuracy and robustness of classification	Need to adjust the parameters of the twelve kernel functions, which increases the computational complexity	2017
		piRNAPred [67]	SVM	The algorithm the secondary structure, thermodynamic and physicochemical properties of RNA and other hybridization features	Based on hybrid features, it can consider the k-mer nucleotide composition, secondary structure, thermodynamics and physicochemical properties of piRNA	Need to normalize and select features, which may introduce noise	2019
	Deep learning methods	V-ELMpiRNAPred [61]	V-ELM	This method integrates short sequence motif features and K-mer features	Based on the extreme learning machine (ELM), it can provide generalization performance and fast learning speed	Need to merge multiple independent ELMs and make decisions through voting methods, which increases the computational complexity	2017
		piRNN [19]	CNN	The first used deep learning to identify piRNA	Based on CNN and FCNN, it can obtain the feature representation of piRNA, improving the classification accuracy	Needs a lot of parameters, resulting in high computational cost	2018
		Khan et al. [71]	parallel DNN	The model is highly extensible, ault tolerance and scalability	Based on parallel multi-layer DNN model, it can identify piRNA sequences, improving the computational efficiency	Need to use the Spark framework to parallelize the computation of large-scale DNN nodes, which increases the system complexity	2020

(*continued on next page*)

Table 2 (*continued*)

Type	Category	Method	Algorithm	Description	Advantage	Disadvantage	Date
		LSTM4piRNA [74]	LSTM	A piRNA identification method suitable for the analysis of large-scale databases.	It does not require manual feature selection and can automatically learn sequence features.	It cannot fully predict all piRNAs and needs to integrate biological prior information to overcome this limitation.	2023
piRNA cluster identification		proTRAC [117]	probability analysis	The first tool for piRNA cluster identification, visualization and analysis.	Based on statistical learning, it comprehensively considers the features of cluster candidates	Need to assume that non-piRNA follows a uniform distribution, which may not conform to the actual situation	2012
		piClust [118]	DBSCAN	The tool prevides web server, and inputs a small RNA-seq data, and outputs piRNA clusters.	Based on clustering algorithm, it does not need to assume the uniform distribution of non-piRNA	Need to determine the threshold, which may affect the clustering results	2014
		PILFER [119]	probability analysis	The tool is state-of-the-art (SOTA) on piRNA cluster identification task.	Based on sliding window, it has higher accuracy, less memory and time consumption	Need to determine the size of the sliding window	2017
transposon-related piRNA identification	Classical machine learning methods	Piano [77]	SVM	The tool is a piRNA annotation program using piRNA-transposon interaction information	Based on the structural information of piRNA-transposon sequences, discriminant features are constructed	Expert feature selection is required	2014
		piPipes [120]	MACS2	A pipeline analysis method is provided for piRNA and transposon analysis	Based on the ensemble learning method, multiple features of transposon-derived piRNA are integrated to improve the classification performance	The parameters of the ensemble learning need to be adjusted, which increases the computational complexity	2015
		Li et al. [87]	GA-WE	Six sequence derived features were extracted to represent piRNA sequence	Based on the genetic algorithm weighted ensemble method, 23 features are used to identify transposon-derived piRNA, improving the diversity of features	The fitness function of the genetic algorithm needs to be determined, which may affect the weight of the features	2016
	Deep learning methods	piRNet [30]	DFNN	A deep learning model for human piRNA classification	Based on the deep learning model, the feature expression ability is improved	A large number of parameters are required, resulting in high computational cost	2021
piRNA target prediction	Classical machine learning methods	Chan et al. [107]	frequency distribution	A kind of frequency distribution method is proposed to identify piRNA molecules of multiple species	Usually refer to miRNA target site prediction tools to find potential piRNA target sites	A large number of experimentally validated datasets and task-specific feature selection are required	2016
		pirnaPre [104]	SVM	A SVM classifier was used to identify piRNA targets on mRNA at the genome-wide level	Using CLIP-Seq features and position-derived features, the accuracy of the model is improved	Insufficient generalization ability for other species, and contains unverified negative samples, which may lead to an increase in false positive rate	2016
	Deep learning methods	Yang et al. [115]	CNN&Muti-head attention&MLP	A deep learning model based on multi-head attention and MLP by using piRNA sequences one-hot encoded to predict piRNA target sites.	Based on the multi-head attention network to extract the binding rules of piRNA-mRNA	A large number of parameters are required, resulting in high computational cost	2021
		Singh and Mallick [121]	-	In this article, the authors used a combination of miRanda algorithm and MFOLD program to analyze the sequence and structural features of piRNA target sites based on CLIP-Seq data.	Using the CLIP-Seq-based dataset, the sequence and structural features of piRNA target sites were analyzed, and some features that can improve the accuracy of piRNA target prediction were proposed	No prediction model was built, only feature analysis was performed	2021
Functional piRNA identification	Classical machine learning methods	2L-piRNA [91]	two-layer SVM	Further identifying whether the piRNA sequence has the role of instructing target mRNA deadenylation	Using a two-layer ensemble classification model, the accuracy and robustness of the classification are improved	The feature extraction method is relatively simple, without considering the structural information of the sequence	2017
		2L-piRNAPred [93]	two-layer SVM	The optimized feature vectors are used for prediction	Based on 2 L-piRNA, more features are added, further improving the accuracy of the classification	The dimension of the feature space increases, resulting in high computational cost	2018

(*continued on next page*)

Table 2 (*continued*)

Type	Category	Method	Algorithm	Description	Advantage	Disadvantage	Date
		2lpiRNApred [58]	SRC&SVMMDRBF	Feature selection algorithm based on LFE-GM is used to optimize features	A feature selection algorithm is proposed, which can reduce the dimension of the feature space and reduce the computational cost	The effect of feature selection may be affected by the dataset, and it does not have universality	2020
	Deep learning methods	2L-piRNADNN [31]	Multilayer Perceptron	Using two layers of deep neural networks model	Using deep neural network to build a classification model, improving the accuracy	Using dinucleotide autocovariance (DAC) to represent the sequence, ignoring the global sequence information, and deep neural network is prone to overfitting	2020
		2S-piRCNN [98]	CNN	Using two layers of CNN model	Using convolutional neural network (CNN) to build a classification model, which can effectively capture the local features of the sequence and avoid overfitting	CNN has more parameters and requires a larger dataset to train	2020
		piRNA (2L)-PseKNC [56]	Multilayer Perceptron	Using two layers of deep neural networks model	Based on the neural network model, using structural information and global sequence information to represent the sequence, improving the sequence expression ability	Using deep neural network as the classification model, there is a risk of overfitting	2020
		Deep-piRNA [95]	Multilayer Perceptron	improving the prediction accuracy of piRNA molecules and their functions using a multi-layer deep neural network model.	Using four feature extraction methods to represent the sequence, improving the diversity and richness of the sequence	Using deep neural networ as the classification model, there is a risk of overfitting	2022
		piRNA-CNN [99]	CNN	Identifying piRNA based on Word2vec and CNN is the best method at present	Based on Word2Vec to obtain the interpretable representation of piRNA, which can capture the semantic information of the sequence	Word2Vec training requires a lot of data, and CNN has more parameters, increasing the computational complexity	2021
		Liu et al. [101]	GloVe	Using a deep learning model based on natural language processing techniques for sequence embedding with attention mechanism to predict exosomal piRNAs.	Based on the GloVe algorithm to pre-train subsequence vectors, and using self-attention mechanism to build a sequence embedding model, improving the accuracy and interpretability	GloVe algorithm training requires a lot of data, and self-attention mechanism has high computational complexity	2022

Liu et al. [18] proposed a piRNA identification method based on sequence features, including weighted k-mer, weighted k-mer with wildcards, location-specific value, and piRNA length feature. Seyed-dokht et al. [66] used 48 heterogeneous features to encode piRNAs, including sequence and structural features. Boucheham et al. [55] proposed a machine learning model with 12 kernels, each of which employed a different feature type. Monga et al. [67] introduced a classification model, piRNAPred, which considered the hybrid features of piRNAs, including k-mer nucleotide composition, secondary structure, and thermodynamic and physicochemical properties.

With the development of deep learning technologies [68], the research community has turned to deep neural network (DNN) models to explore more effective feature learning methods for piRNA identification. For example, Cao et al. [69] proposed a classification model called V-ELM to identify piRNAs. This model is based on an extreme learning machine (ELM) [70], which is a single-hidden layer feedforward network that offers generalization performance and fast learning speed. The V-ELM model improves it by merging multiple independent ELMs and making decisions using a voting method. Pian et al. [61] proposed an optimized version of V-ELM by employing multiple short sequence motif features. These features indicate the typical characteristics of piRNA sequences that are helpful for identification. Wang et al. [19] proposed a method called piRNN using a CNN to obtain feature representations of piRNAs and a fully connected neural network for classification. This method achieved $>$ 90 % accuracy in identifying piRNAs in *Caenorhabditis elegans*, *Drosophila melanogaster*, rats, and humans. However, deep learning-based models always have a large number of parameters, which may lead to high computational costs. To address this problem, Khan et al. [71] proposed a parallel multilayer DNN model [72] to identify piRNA sequences, where the Spark framework [73] was used to calculate the number of nodes of a large-scale DNN in parallel. To suit the analysis of large-scale databases, Chen et al. [74] proposed a novel deep learning-based method for piRNA identification, named LSTM4piRNA. The method utilizes compact LSTM networks to effectively analyze RNA sequences from massive datasets to identify piRNA. The method achieved excellent performance on the piRBase database.

3.2. Transposon-derived piRNA identification

piRNAs play key roles in transposon silencing in the germline by protecting the germline genome from transposon expression [75]. Transposon-derived piRNAs are produced during transposon silencing, and identifying transposon-derived piRNAs helps to better understand their functions [76].

Transposon-derived piRNAs contain structural information about the piRNA transposon sequence, which can be leveraged to build discriminative features. Based on this, Wang et al. [77] proposed a classification algorithm called Piano to identify transposon-derived piRNAs based on RNAplex [78], a tool used to find RNA-binding sites. However, it always

requires elaborate feature selection. In contrast, ensemble learning helps mitigate the effect of inappropriate modeling and feature combinations [79–81]. Luo et al. [82] utilized an ensemble learning method to identify piRNAs by integrating six features of transposon-derived piRNAs: the spectrum profile [83,84], mismatch profile [85], subsequence profile [85], PSSM [86], pseudo-dinucleotide composition [83,84] and local structure sequence triplet elements [77]. Luo et al. [87] proposed a genetic algorithm-based weighted ensemble method to identify transposon-derived piRNAs with 23 features. Table 3 shows the performance of current machine learning methods for transposon-derived piRNA identification.

Deep learning-based methods help mitigate human effort in feature selection for transposon-derived piRNA identification. For example, Costa et al. [30] proposed a deep learning model called piRNet, which included eight DFNNs, each of which had different hyperparameter configurations. The best configuration was investigated during the training stage and utilized in test one. The experimental results showed that piRNet outperformed two traditional machine learning models, SVM and RF, and the neural network model, piRNN.

3.3. Functional piRNA identification

Functional piRNAs are a class of small ncRNAs that form RNA–protein complexes by interacting with the piwi-subfamily of Argonaute proteins. They are mainly involved in the epigenetic and post-transcriptional silencing of transposable elements and other spurious or repeat-derived transcripts in germ cells, but they can also regulate other genetic elements [88]. Unlike nonfunctional piRNAs, functional piRNAs target mRNA deadenylation, which promotes mRNA stability and translation efficiency [89,90]. Predicting functional piRNAs can facilitate an in-depth understanding and research of piRNA systems. Other piRNAs may have unknown roles or associations with diseases. Therefore, predicting them can also help reveal their true functions and potential applications in living organisms.

Computational efforts have been made to identify functional piRNAs. Formally, determining whether an RNA molecule is functional piRNA or other piRNA, non-piRNA is a tri-classification task. However, it is practically treated as a two-stage process, that is, to first determine whether an RNA molecule is a piRNA and then judge whether it is a functional molecule. For example, Liu et al. [91] proposed a two-layer ensemble classification model, 2L-piRNA [92] which achieved accuracies of 86.1 % and 77.6 % at the first and second stages, respectively. Li et al. [93] built an improved version named 2 L-piRNAPred with more features and achieved 89.0 % and 84.0 % accuracy in the first and second stages, respectively. However, the abundance of features results in increased computational costs. To address this, Zuo et al. [58] proposed a feature selection algorithm, Luca Fuzzy Entropy and Gaussian Membership function, to reduce the dimensions of the feature space.

Recent studies on functional piRNA identification have turned to deep learning technologies, although most still follow a two-stage strategy. For example, Khan et al. [31] proposed a model named 2 L-piRNADNN, where a DNN was first employed to classify each RNA sequence into two classes, piRNA and non-piRNA, and the other was then employed to determine whether it was functional or non-functional if it was a piRNA. The experimental results indicated that the two models achieved accuracies of 91.81 % and 84.52 %, respectively. However, this model adopts a simple double nucleotide autocovariance for sequence representation [84], ignoring global sequence information [94]. Khan et al. [56] proposed an improvement by leveraging structure information and global sequence-order information for sequence representation; the experimental results showed that the two models achieved accuracies of 94.73 % and 85.21 %, respectively. They also built Deep-piRNA [95] using four feature extraction methods, namely normalized Moreau-Broto autocorrelation, Z-curve-12-bit, single nucleotide composition, and dinucleotide composition, to represent RNA sequences and achieved accuracies of 96.13 % and 85.54 % in each stage, respectively. However, DNNs are prone to overfitting during the training process [96,97]; hence, researchers tend to utilize more reliable neural network models. For example, Ali et al. [98] proposed a two-layer model named 2S-piRCNN to identify functional piRNAs based on a CNN and achieved accuracies of 93.60 % and 90.10 % in the first and second layers, respectively. Tahir et al. [99] employed Word2Vec [100], a distributed feature representation method, to obtain an interpretable representation of piRNA, and proposed a two-layer model piRNA-CNN that was also based on CNN.

Some studies have leveraged the advantages of natural language processing technologies to model functional piRNA sequences. For example, Liu et al. [101] developed an end-to-end model in which a piRNA sequence is treated as a sentence and each k-mer subsequence is treated as a word in the sentence. They pre-trained subsequence vectors using the GloVe algorithm and built a sequence-embedding model with a self-attention mechanism to extract different aspects of multiple vector representations from piRNA sequences. The model achieved an accuracy of 82.0 % in identifying exosomal piRNAs. They also revealed the key subsequences of exosomal piRNAs via an attention mechanism.

3.4. piRNA target identification

piRNAs can identify and silence their targets through complementary pairing with mRNAs, thereby participating in important biological processes, such as transposon suppression, germline development, and genome stability [89,102,103]. Therefore, the accurate identification of piRNA target sites is of great value for revealing the functional mechanisms and biological significance of piRNAs. However, owing to the complexity and diversity of piRNA targeting rules as well as the lack of large-scale experimental validation datasets, the prediction of piRNA target sites still faces huge challenges [44,45,104].

Computational methods for piRNA target site prediction are mainly divided into two categories: rule-based and machine learning-based methods [105]. Rule-based methods follow the principle of sequence complementarity, namely the base pairing between piRNAs and mRNAs [44,45,106], to screen potential target sites, and usually draw on miRNA target site prediction tools [89,107–111]. However, these methods ignore factors that may affect the efficiency and specificity of piRNA targeting, such as the sequence and structural features around the target site, location of the target site on the mRNA, and interaction between the target site and the PIWI protein [112,113]. Computational methods utilize the biological features of piRNAs to train models to identify potential piRNA target sites [89,107–111]. For example, Yuan et al. [104] built an SVM-based prediction model using CLIP-seq and

Table 3
The performance of current machine learning methods on transposon-derived piRNA identification.

Dataset	Method	AUC (H/M/D)			ACC (H/M/D)		
		Human	Mouse	Drosophila	Human	Mouse	Drosophila
Balanced	Piano [77]	0.592	0.445	0.741	0.560	0.537	0.692
	Ensemble Learning [82]	0.920	0.924	0.994	0.807	0.810	0.958
	GA-WE [87]	0.932	0.937	0.995	0.839	0.838	0.959
Imbalanced	Piano [77]	0.449	0.441	0.804	0.747	0.744	0.712
	Ensemble Learning [82]	0.922	0.928	0.995	0.836	0.849	0.965
	GA-WE [87]	0.935	0.939	0.996	0.869	0.889	0.964

position-derived features. However, such models have insufficient generalization capabilities for other species owing to their task-specific biological features. It is trained on experiment datasets [114] with unverified negative samples, which may lead to an increased false alarm rate. To solve this problem, Yang et al. [115] proposed a deep learning method.

The method automatically learned the motif features of piRNAs using convolutional filtering and squeezing-and-excitation blocks and extracted piRNA-mRNA binding rules via multihead attention networks. Additionally, they proposed acquiring a validated negative set by leveraging experimentally validated positive sets. The experimental results showed that the method achieved state-of-the-art performance on a task with an AUC of 95.7 %. Singh and Mallick [121] analyzed the sequence and structural features of piRNA target sites using CLIP-Seq-based datasets. They investigated the distribution, folding free energy, miRanda score, and nucleotide composition of piRNA target sites in the 30 UTR, CDS, and 50 UTR regions and found significant differences between the IP+ set (true targets) and the IP− set (hypothetical targets). Based on this, they proposed features that could improve the accuracy of piRNA target prediction.

3.5. piRNA cluster identification

The majority of piRNA sequences are located in a small number of genomic regions known as clusters [122,123], which was first defined by Aravin et al. [102] in 2006. The length of a piRNA cluster generally ranges from 1 to 100 kilobase (kb), encoding 10–4500 piRNAs [124]. Identifying piRNA clusters helps to better understand the biological production mechanisms and functions of piRNAs [123,125].

Current methods and tools for piRNA cluster identification are primarily based on statistical learning. Rosenkranz et al. [117] developed the proTRAC software to identify piRNA clusters, where the characteristics of cluster candidates, including the number of normalized hits to total hit ratio and extent of strand bias, were considered synthetically. However, this method assumes that the non-piRNAs follow a uniform distribution. Jung et al. [118] proposed a piClust model by leveraging the clustering algorithm DBSCAN [126] to identify piRNA clusters via k-dist analysis without assuming a uniform distribution for non-piRNAs. PILFER [119] is a state-of-the-art piRNA cluster identification tool that uses sliding windows to observe sequences, such as proTRAC. It identifies piRNA clusters by integrating the expression of reads with spatial information. Compared with proTRAC and piClust, PILFER runs with higher accuracy and consumes less memory and time.

3.6. Limitation

However, computational methods based on machine and deep learning can discover, identify, analyze, and predict piRNAs and their related molecules in a fast, accurate, flexible, and innovative manner, thus promoting the functional research and application of piRNAs. However, current machine learning and deep learning algorithms have the following limitations:

The main limitations of machine learning methods are as follows: (1) the need to manually design and extract effective features, which requires significant domain knowledge and annotation work, and the selection and combination of features may affect the generalization ability and interpretability of the model; (2) the need to choose appropriate classifiers and parameters, which requires an adjustment and optimization process, and different classifiers may have different performances on different datasets and tasks; and (3) the need to deal with the data imbalance and sparsity problem, which may lead to model bias and variance, and reduce the accuracy and robustness of the model. For example, in a piRNA identification task, Betel et al. [3] used the PSSM algorithm, which requires manual selection of the piRNA length and position specificity, and the performance of PSSM was affected by the noise and variation of the sequence. In the functional piRNA identification task, Liu et al. [18] used weighted k-mers and other features that required manual determination of k values and weights, and these features may not capture the structural information of piRNAs. In the piRNA target identification task, Yuan et al. [104] used an SVM classifier, which required manual selection of the kernel function and regularization parameters, and the performance of the SVM was limited by the data scale and dimension.

Although deep learning methods have significantly improved the prediction and identification performance, their limitations are as follows: (1) the need for a large amount of annotated data and computing resources, which may lead to the difficulty and cost of data acquisition and processing, as well as the time and space consumption of model training and testing; (2) the need to design a reasonable network structure and loss function, which requires a deep understanding of the model principle and mechanism, and different network structures and loss functions may have different adaptabilities to different datasets and tasks; and (3) the need to solve model overfitting and underfitting problems, which may lead to a decrease in model generalization ability and robustness, as well as the lack of model interpretability and credibility. For example, in the piRNA identification task, Wang et al. [19] used a CNN model that required a large number of piRNA and non-piRNA sequences as training data and the structure and parameters of the CNN needed to be adjusted according to the dataset and task. In a transposon-derived piRNA identification task, Costa et al. [30] used the piRNet model, which requires a large amount of computing resources to train and test, and the performance of piRNet was affected by data quality and distribution. In a functional piRNA identification task, Liu et al. [101] used the self-attention mechanism, which is required to design a reasonable loss function and attention weight. However, the principle and mechanism of the self-attention mechanism may be difficult to explain.

4. piRNA–disease association prediction

Identifying disease-related piRNAs and their relevance to pathogenesis are important issues in biomedical and clinical research. The academic community has continuously focused on the automatic prediction of piRNA–disease associations, and some studies have summarized related literature from biological and computational perspectives [35,36]. However, a few studies have focused on reviewing the architecture and algorithms of these computational models for piRNA–disease association prediction. This section provides a summary of the computational models for piRNA–disease association prediction, which could provide a reference for model construction and performance evaluation in future studies.

Non-coding RNAs with similar biological functions are often associated with similar diseases [22,127]. In other words, the association between an unexplored non-coding RNA and diseases can be legitimately speculated based on the association between diseases and RNAs that are similar to it. This issue can be decomposed into two computational tasks: finding relations between RNA and disease and modeling such known relations to predict unknown ones. The former can be viewed as the construction of association networks, where nodes are RNA molecules and diseases, while edges represent relations between them; the latter is a machine learning procedure leveraging knowledge from association networks. Fig. 2 shows a pipeline of the piRNA–disease association prediction models, where a piRNA–disease association network via diverse similarity measures based on piRNA and disease datasets is built first, then a feature learning procedure is conducted to obtain the semantic representation of piRNA and disease, and finally, a classifier is built to predict the association between them.

Recently, some studies have employed multi-omics data integration methods to explore the role of piRNAs in cancer subtype identification [128,129]. These studies have revealed the differential expression patterns of piRNAs between tumor and normal tissues, as well as among different subtypes of tumors. Therefore, the expression profiles of

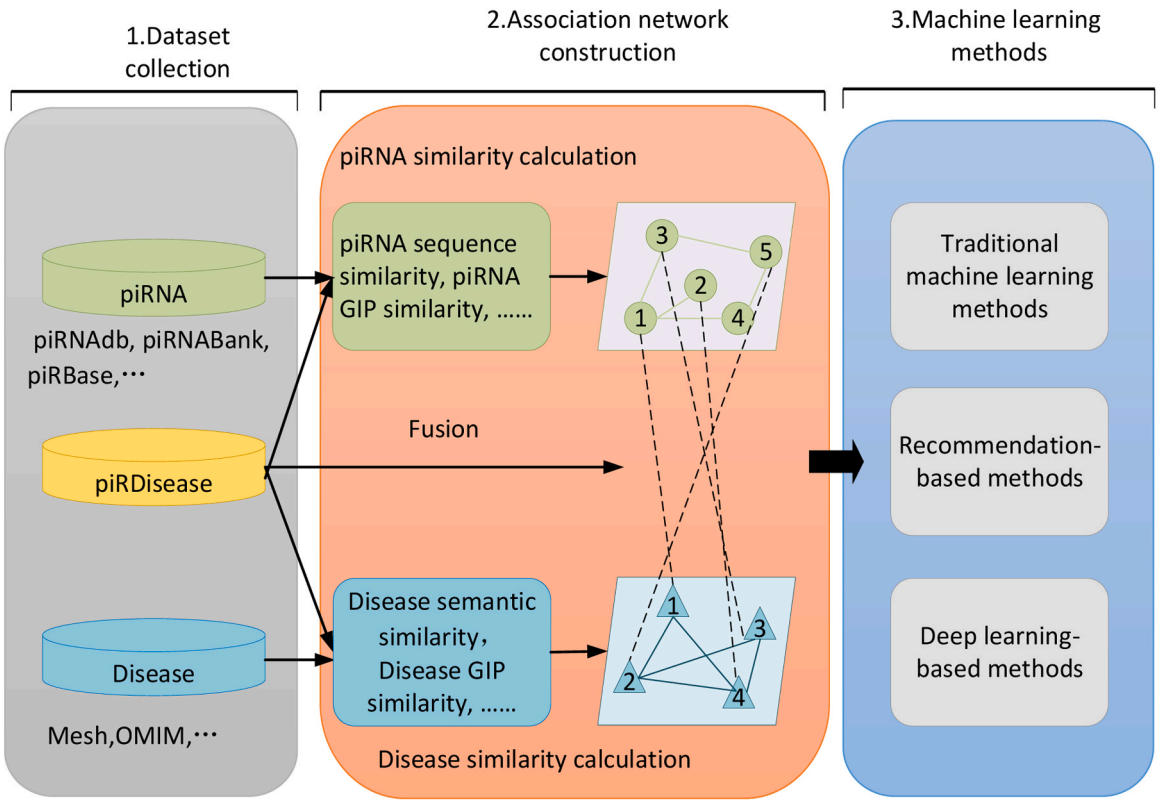

Fig. 2. The computational flow of the Piwi-interacting RNA (piRNA) and disease association prediction. Step 1: Obtaining experimental data from public databases; Step 2: Calculating the similarity and constructing association networks from experimental data; Step 3: Adopting different methods to predict piRNA–disease association.

piRNAs can be used to identify cancer subtypes. Moreover, some studies have identified SNPs that may affect the biogenesis or binding of piRNAs or their targets, and have shown their association with cancer risk or prognosis [130]. These studies have demonstrated the potential of piRNAs as novel biomarkers and therapeutic targets for cancer. However, the role of piRNAs in other diseases, such as neurodegenerative diseases, metabolic diseases, cardiovascular diseases, etc., remains largely unknown. Therefore, it is necessary to develop computational models for piRNA–disease association prediction, which can help to uncover the underlying mechanisms of piRNAs in disease pathogenesis and provide new insights for diagnosis and treatment.

4.1. Association network construction

Association network construction by extracting effective similarity features of RNA and disease is an important step in building a reliable

Table 4
Similarity calculation of piRNA and disease.

Similarity	Formula	Notation	Description
Disease semantic Similarity [156]	$DS_{sem}(d_i,d_j) = \frac{\sum_{t\in T(d_i)\cap T(d_j)}\left(D_{d_i}(t)+D_{d_j}(t)\right)}{D_{sv}(d_i)+D_{sv}(d_j)}$ Where : $D_{sv}(d) = \sum_{t\in T(d)} D_d(t)$ $D_d(t) = \begin{cases} 1, & if t = d \\ \max\{\Delta * D(t'), t' \in children\ \ of\ \ t & if t \neq d \end{cases}$	Disease semantics can be expressed as $DAG(d) = (d, T(d), E(d))$ by DAG. Where T(d) is the set of disease, disease $t \in T(d)$, Δ is the semantic contribution factor, and the value of this factor is usually 0.5. $D_d(t)$ is the contribution node of t to d. if t = d, the value of $D_d(t)$ is 1, otherwise, the value is $\Delta * D(t')$.	The formula could be used to calculate the semantic values of these two diseases and their semantic similarity.
Disease GIP Similarity [132]	$DS_{GIP} = \exp\left(-\lambda_d \| A(d_i) - A(d_j) \|^2\right)$ Where: $\lambda_d = \frac{1}{N_d}\sum_{n=1}^{N_d} \| T(d_n) \|^2$	Where $A(d_i)$ denotes the association between disease d_i and all piRNA in the sample. Where N_d denotes the number of diseases in the sample.	The disease GIP similarity is calculated by the known association matrix between piRNA and diseases.
piRNA sequence similarity [59]	$PS_{seq} = \frac{\sum_{i=1}^{N}(S_{p_i} - \overline{K_{p_i}})\sum_{j=1}^{N}(S_{p_j} - \overline{K_{p_j}})}{\sqrt{\sum_{i=1}^{N}(S_{p_i} - \overline{K_{p_i}})^2 \sum_{j=1}^{N}(S_{p_j} - \overline{K_{p_j}})^2}}$ Where: $S_{p_i} = \frac{(k-mer\ \ count\ \ in\ \ P_{seq})}{length(P_{seq}) - k + 1}$	Where P_{seq} denotes the sequence of piRNA, and S_{p_i} denotes the frequency score for each k-mer in i-th P_{seq}. Where $\overline{K_p}$ is the mean of S_{p_i} of each piRNA sequence.	The piRNA of related function often have related k-mer content [157]. Through the sequence similarity of piRNA, it can be inferred that piRNA has similar functions.
piRNA GIP similarity [132]	$PS_{GIP} = \exp\left(-\lambda_p \| A(p_i) - A\left(p_j\right) \|^2\right)$ Where: $\lambda_p = \frac{1}{N_p}\sum_{m=1}^{N_p} \| T(p_m) \|^2$	Where $A(p_i)$ denotes the association between piRNA p_i and all diseases in the sample. Where N_p denotes the number of piRNA in the sample.	The piRNA GIP similarity is calculated by the known association matrix between piRNA and diseases.

RNA-disease association predictor [131]. Such feature-building methods depend on piRNA and disease association information, which usually comes from RNA similarity, disease similarity, and RNA-disease association data sources. In this section, methods used for association network construction and fusion are reviewed.

4.1.1. Network construction

The main step in constructing an association network of piRNAs and diseases is to measure the similarity degrees between them, which are generally calculated by an assortment of similarity calculation methods for piRNAs and diseases. These methods are summarized in Table 4.

Two types of similarity knowledge are employed to build a piRNA–disease association network: piRNA similarity and disease similarity. The first employs piRNA sequence similarity [59], which works by calculating the similarity of two piRNA sequences via the k-mer algorithm and piRNA Gaussian Interaction Profile (GIP) kernel similarity [132]. GIP similarity measures the similarity of interaction patterns of biological entities [133] by extracting their associated features via Gaussian kernel functions and is widely utilized for similarity measurements of miRNA [133–135], circular RNA (circRNA) [136] and long ncRNA (lncRNA) [137,138]. The second method employs disease semantic similarity, which is measured by estimating the ratio of the DAG [139] that a disease pair shares and disease GIP similarity, which can be used to calculate GIP similarity from a known association matrix between piRNA and disease.

4.1.2. Network fusion

The goal of network fusion is to identify additional connections between piRNAs and diseases. As the types of nodes in piRNA and disease similarity networks differ, some studies have constructed heterogeneous networks to fuse them. Additional knowledge, such as piRDisease [51], can be used to bridge piRNAs and diseases in fused networks. Experimental results have shown that such fused networks provide more associated information regarding piRNAs and diseases, making the predictor more accurate [140].

4.2. Prediction methods

Biological experiments for identifying disease-associated piRNAs are expensive and time-consuming. Accordingly, the research community has turned to computational methods for such tasks in recent years, such as automatic miRNA-disease association prediction [133–135], lncRNA–disease association prediction [137,138] and circRNA–disease association prediction [136].

Although most of the methods used for the above tasks can be applied to the piRNA–disease association prediction task, the current piRNA–disease association prediction methods mostly focus on the correlation between piRNA expression changes and diseases without delving into the specific roles and mechanisms of piRNAs in the occurrence and development of diseases. This is because functional research on piRNAs is still in its infancy, and there is insufficient experimental data and theoretical support to explain the causal relationship between piRNAs and diseases. Therefore, we believe that computational methods can serve as an auxiliary means to provide valuable candidate piRNA–disease associations for experimental research but cannot completely replace experimental verification.

Not until the first piRDisease database that annotated piRNA–disease associations was released, which provided a benchmark dataset for computational methods in the piRNA–disease association prediction task, an increasing number of researchers began to use computational methods to explore piRNA–disease associations. piRDisease is a manually curated database that collects data on 7939 experimentally supported associations between 4796 piRNAs and 28 diseases. However, the database also has some limitations, such as data sparsity, lack of negative samples, inconsistency in piRNA identifiers, and unclear disease classification[141]. Therefore, the current computational methods still have room for improvement in understanding the role and mechanism of piRNAs in diseases.

In this section, we review the proposed computational models for piRNA–disease association prediction, which can be categorized into three techniques: traditional machine learning methods, recommendation-based methods, and deep learning-based methods. The models are summarized in Table 5.

4.2.1. Traditional machine learning methods

Zheng et al. [140] proposed a computational model, APDA, to predict the association between piRNAs and diseases using the benchmark dataset piRDisease V1.0. They investigated the effect of features on prediction performance by employing two groups of features: the former obtained feature representation via collaborative filtering, and the latter utilized several feature representation methods: a correlation matrix between piRNA and diseases, GIP kernel similarity matrix of piRNA and disease, piRNA sequence similarity matrix, and disease semantic similarity matrix. Experiments showed that the model using the features of the latter group achieved better performance, indicating that the integration of multiple features improves the prediction ability of the task.

However, the task of piRNA–disease association prediction suffers from a lack of high-quality negative samples during training. To address this problem, Wei et al. [22] proposed a model called iPiDi-PUL to predict the association between piRNAs and diseases based on positive unlabeled learning (PUL) [142,143]. Such a method is always employed to learn a model from positive and unlabeled samples and is widely used in non-coding RNA-disease association prediction [144–146]. In iPiDA-PUL, the experimental dataset contains 4350 piRNAs, 21 diseases, and 5002 association pairs, and features are extracted by principal component analysis. An ensemble learning strategy was adopted to train multiple RFs with different depths for prediction. Finally, the model uses the average score of all the classifiers to make the final decision. The ensemble learning strategy is illustrated in Fig. 3.

4.2.2. Recommendation-based methods

Essentially, piRNA–disease association prediction can be viewed as a recommendation task, where the associations between piRNAs and diseases are regarded as recommendation behaviors between items and users, and thus can be solved using principles based on recommendation. For example, Zheng et al. [147] proposed a structural perturbation method based on heterogeneous networks to predict associations between piRNAs and diseases. The heterogeneous network consisted of three subnetworks: piRNA similarity, disease similarity, and piRNA–disease association networks. A piRNA similarity network was constructed using sequence information and GIP kernel similarity. The disease similarity network was derived from functional information and GIP kernel similarity based on gene–disease association information. The structural perturbation method evaluates the predictability of unknown associations in a heterogeneous network by randomly selecting edges and calculating the eigenvalues and eigenvectors of the perturbed network. This model does not require negative examples and can effectively utilize multisource information. The model exhibits high performance and robustness on a benchmark dataset and a new dataset. Zhang et al. [33] proposed an approach to identify piRNA–disease association, following the idea of learning to rank (LTR) [148,149] proposed by Wei et al. [150]. Specifically, such an identification or classification problem is formalized as a search task, where the target piRNA and diseases are viewed as the query and the document candidates, respectively, and the association between the piRNA and diseases is positively correlated to the ranking position of the disease. The LTR strategy for the piRNA–disease association prediction task is shown in Fig. 4. In other words, the higher the ranking position of a disease, the more relevant the piRNA and disease. To implement the LTR method, they utilized LambdMART, a gradient-boosted decision tree, to build the learning model, and employed two types of methods, namely machine learning-based and collaborative filtering-based methods, to calculate

Table 5
piRNA and disease association prediction models.

Algorithm	Methods	Similarity computation	piRNA	disease	association	AUC	Advantage	Disadvantage	Case study	# of New associations predicted
iPiDA-sHN [32]	CNN&SVM	Disease semantic similarity, piRNA sequence similarity and GIP similarity of piRNA and diseases	4350	21	5002	0.8576	Using a two-step PUL strategy based on SVM, high-quality negative samples are screened from random negative samples, and CNN is used to capture the nonlinear relationship between piRNA and disease.	The parameters of CNN are not optimized, which may affect the generalization ability and robustness of the model.	'Alzheimer's disease' and 'Head and neck cancer'	12
iPiDA-PUL [22]	Random Forest	Disease semantic similarity, piRNA sequence similarity and GIP similarity of piRNA and diseases	4350	21	5002	0.8830	Using PUL learning method to avoid the problem of lacking high-quality negative samples, and using ensemble learning strategy to improve the prediction ability.	The feature extraction method is simple, and does not consider the attribute information of piRNA and disease.	'Head and neck squamous cell carcinoma','Breast cancer ','Alzheimer's disease 'and 'Gastric cancer '	15
DFL-PiDA [158]	CNN&ELM	Disease semantic similarity, piRNA sequence similarity and GIP similarity of piRNA and diseases	4350	21	5002	09042	Using convolutional denoising autoencoder to extract features, eliminate noise and redundant information, and using ELM for fast prediction.	The training of convolutional denoising autoencoder may require a lot of computational resources and time, and the performance of ELM also depends on the selection and initialization of hidden layer nodes.	None	0
piRDA [6]	CNN&SVM	One-hot features of piRNA and diseases	4350	21	5002	0.9510	Using one-hot encoding method to capture the hidden information of piRNA and disease.	The known and similar association pairs of piRNA and disease are not considered. In addition, the one-hot encoding method may lead to high-dimensional sparse feature representation, which increases the computational complexity.	Cardiovascular disease,Renal cell carcinoma and Alzheimer's disease	13
iPiDA-LTR [33]	SVM&LR&RF&CF	Disease semantic similarity and piRNA sequence similarity	4350	21	5002	0.9543	Using learning ranking method, transforming the prediction problem into a search task, and modeling the piRNA-disease association from	The performance of learning ranking method also depends on the selection and evaluation of ranking metrics.	piR-hsa-23210 and piR-hsa-15023	9

(*continued on next page*)

Table 5 (*continued*)

Algorithm	Methods	Similarity computation	piRNA	disease	association	AUC	Advantage	Disadvantage	Case study	# of New associations predicted
							a global perspective.			
ETGPDA	Embedding transformation GCN	Disease semantic similarity, piRNA sequence similarity and GIP similarity of piRNA and diseases	4350	21	5002	0.9603	Using heterogeneous graph convolutional network and multi-source attention mechanism to extract the low-dimensional embeddings of piRNA and disease, and using embedding transformation module to solve the problem of embedding space inconsistency.	The performance of embedding transformation module also depends on the selection and optimization of transformation function, and it may introduce additional parameters and complexity.	'Alzheimer's disease' and 'Head and neck cancer'	13
APDA [140]	Autoencoder &Random Forest	Disease semantic similarity, piRNA sequence similarity and GIP similarity of piRNA and diseases	4503	27	5214	0.9088	Using multiple feature representation methods, integrating the attribute and association information of piRNA and disease.	Simply integrating different feature representation methods may cause inconsistency and conflict, which affects the stability of the model.	None	0
SPRDA [59]	Matrix Completion	Disease semantic similarity, Disease functional similarity, piRNA sequence similarity and GIP similarity of piRNA and diseases	501	22	1212	0.9529	Using structural perturbation method and multi-source information for prediction, effectively avoiding the problem of lacking high-quality negative samples.	Because the structural perturbation method is based on randomly selected edges to evaluate the predictability of the network, it may cause some important edges to be ignored or some irrelevant edges to be considered.	None	0
GAPDA [23]	GAT	Disease semantic similarity, piRNA sequence similarity and GIP similarity of piRNA and diseases	501	22	1212	0.9038	Using GAT to learn the hidden representation of nodes in the network, which can capture the complex relationship between nodes.	There may be data sparsity or incompleteness problems. And GAT has high computational complexity.	None	0
MSRDA [155]	Stacked autoencoders	Disease semantic similarity, piRNA sequence similarity and GIP similarity of piRNA and diseases	501	22	1212	0.9184	Using stacked autoencoder (SAE) to perform deep abstract representation of multi-source data, which can eliminate noise and redundant information.	The network structure information is not considered, which may ignore some potential associations.	None	0
iPiDA-GBNN [153]	GrownNet	Disease semantic similarity, piRNA sequence similarity	5184	33	8002	-	Using GrowNet to predict piRNA-disease association, which can effectively	The training time of GrowNet is long, which may affect the efficiency of the model.	Alzheimer's disease and Parkinson's disease	

(*continued on next page*)

Table 5 (*continued*)

Algorithm	Methods	Similarity computation	piRNA	disease	association	AUC	Advantage	Disadvantage	Case study	# of New associations predicted
		and GIP similarity of piRNA and diseases					handle nonlinear and high-dimensional data.			
iPiDA-GCN [34]	GCN	Disease semantic similarity and piRNA sequence similarity	10,149	19	11,981	0.7149	Designing two GCN modules (Asso-GCN and Sim-GCN) to extract information from piRNA-disease association network and two similarity networks respectively, which can enhance the diversity and robustness of the representation.	Three different networks need to be built, which may cause data sparsity or incompleteness problems, and also increase the computational complexity of the model.	Cardiovascular disease,Renal cell carcinoma, Alzheimer's disease and Parkinson's disease	14
iPiDA-SWGCN	Supplementarily Weighted GCN	Disease semantic similarity and piRNA sequence similarity	10,149	19	11,981	0.8178	Proposing a complementary weighting strategy, which integrates various basic predictors to supplement the potential associations in the sparse piRNA-disease network, and enriches the network structure information.	The parameters of the complementary weighting matrix need to be adjusted, and the attribute information of piRNA and disease is not considered.	Cardiovascular disease,Renal cell carcinoma and Parkinson's disease	19 15
PDA-PRGCN	GCN	Disease semantic similarity, piRNA sequence similarity and GIP similarity of piRNA and diseases	4350	21	4993	0.9630	Proposing a subgraph projection strategy, which extracts more topological information; designing a residual-based node feature enhancement algorithm, which obtains high-quality initial representation; introducing a dual-loss mechanism, which optimizes the performance of the model by cross-entropy loss and sensitivity-specificity loss.	There may be data sparsity or incompleteness problems; the parameters of the dual-loss mechanism need to be adjusted.	Breast neoplasm, Renal cell carcinoma,Head and neck neoplasms and Alzheimer's disease	0

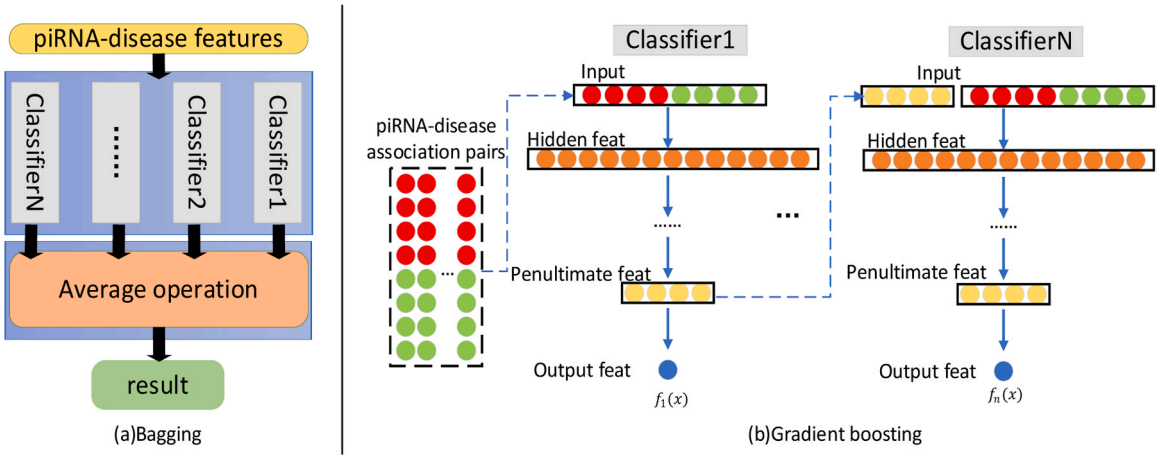

Fig. 3. Ensemble learning strategy. (a) iPiDA-PUL adopts the Bootstrap AGGregating method to train various random forest classifiers, and the unknown Piwi-interacting RNA–disease associations were predicted based on the average of the scores of all classifiers. (b) iPiDA-GBNN utilized Gradient boosting methods to build a complex model GrowNet incrementally with a multilayer network.

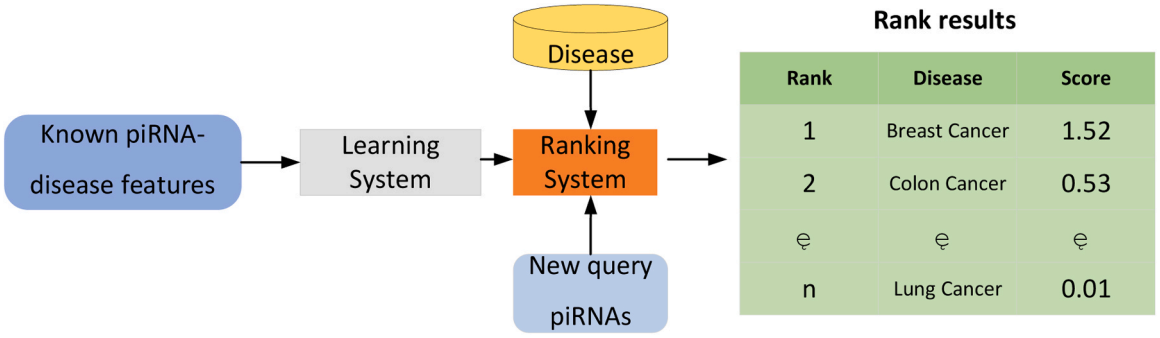

Fig. 4. Learning To Rank (LTR) strategy. In iPiDA-LTR, the known Piwi-interacting RNA (piRNA)–disease association feature sets were fed into the Lambda MART model for learning, and then the rank score of diseases associated with new query piRNAs was obtained using the ranking system.

association scores. The advantage of this model is that it models the piRNA–disease association from a global perspective, thereby reducing the probability of false-positive errors.

4.2.3. Deep learning-based methods

Deep learning-based methods focus on feature representation learning and achieve significant performance improvements in the prediction of disease-related RNAs. Wei et al. [32] adopted a two-step PUL method called iPiDA-sHN, in which a classifier was trained with positive and random negative samples in the first step, and then another classifier was trained with positive and high-quality negative samples acquired from the first step. The two-step PUL method is described in Fig. 5, and the following piRDA adopts it. A deep learning CNN model was utilized to extract high-level features of piRNA–disease associations. This study was based on Shrivastava et al. [151]. The unlabeled samples were ranked by their prediction scores, and one-third of the samples in the intermediate layer were regarded as reliable negative samples.

Ji et al. [150] proposed a deep feature learning model, DFL-PiDA, that utilizes a convolutional denoising autoencoder neural network to capture piRNA–disease association features by considering piRNA and disease similarity features as the input data. They utilized ELM, a feedforward neural network with a single hidden layer, to predict the association between piRNAs and diseases. According to comparative experiments, such a model is more efficient in predicting latent piRNA–disease associations.

Ali et al. [6] proposed a piRDA model based on the two-step strategy proposed by Wei et al. [32]. Compared with prior work, they used a one-hot encoding method [152] to encode raw piRNA sequences and disease semantics. By doing so, the concealed information on piRNAs and diseases can be captured by neural network models without losing the contextual information between them. To alleviate the class imbalance problem, they adopted a bootstrapping method to train the model using partitioned sample blocks, each of which contained approximately the same number of positive and negative samples. A comparative experiment revealed that the piRNA model outperformed the iPiDA-PUL and iPiDA-sHN models.

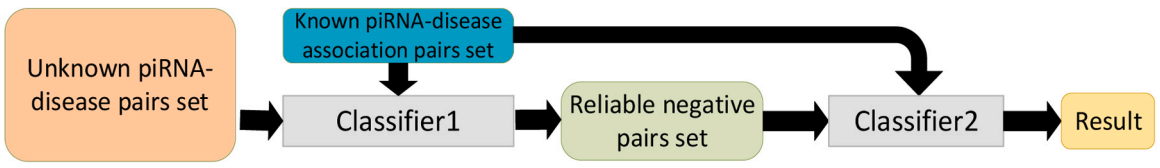

Fig. 5. Two-step positive unlabeled learning strategy. iPiDA-sHN and piRDA both use this method to obtain high-quality negative samples for Piwi-interacting RNA–disease association prediction.

Qian et al. [153] proposed a prediction model, iPiDA-GBNN, based on the algorithm of gradient-boosting neural networks [154]. They employed multiple similarity features, such as GIP kernel similarity, Jaccard similarity, and sequence similarity. They also utilized a stacked autoencoder and multilayer neural network to extract features and eliminate noise. For prediction, they utilized GrowNet, a gradient-boosting framework based on weak learners, which helped achieve a good performance using trivial classifiers.

Zheng et al. [155] proposed a decision support system based on multisource information and stacked autoencoders, called MSRDA, to predict potential piRNA–disease associations. MSRDA constructed feature descriptors using piRNA sequence information, disease semantic information, piRNA GIP kernel similarity, and disease GIP kernel similarity. Stacked autoencoders were then used to perform feature denoising and abstraction. Finally, an RF model was used for classification and prediction. In the 5-fold cross-validation, MSRDA achieved an average AUC value of 0.9184 ± 0.0015, demonstrating the effectiveness of introducing multisource information and stacked autoencoders for improving the performance of piRNA–disease association prediction.

Zheng et al.[23] proposed a model called GAPDA, which first employed GAT [25], a graph-based representation learning model, to learn piRNA and disease representations. The model aimed to build a graph in which the nodes in a neighborhood were automatically weighted via a self-attention mechanism. This method helped capture hidden association features between nodes and learned the structural features of the piRNA–disease association network at the node level. The attention mechanism was implemented using masked attention and a multihead attention method. The performance of GAPDA in the benchmark dataset showed convincing results in piRNA–disease association prediction. The authors also made a comparative experiment between GAPDA and the prior model APDA proposed by them, finding that the attention-based method achieved better performance than collaborative filtering-based and attribute-based methods.

Hou et al. [34] proposed a model called iPiDA-GCN based on a graph convolutional network (GCN), which is a graph-based representation learning model. Their idea is similar to that of Zheng et al.[23], who attempted to capture the underlying relationship patterns in graph-structured data for piRNA–disease association prediction. To this end, they built two GCN models: Asso-GCN, which learned the feature representation of association information from heterogeneous nodes of piRNAs and diseases, and Sim-GCN, which captured similarity features from homogeneous nodes among piRNAs or diseases. The two models were pipelined to obtain the feature representations of piRNAs and diseases.

* In this table, we list the AUC values of computational methods based on same benchmark dataset, and use "_" to represent AUC values of methods based on different benchmark dataset. "None" means that the computational method did not conduct case study, and could not predict new associations."None" means that the computational method did not conduct case studies, and could not predict new associations.

A comparative experiment showed that the model outperformed other state-of-the-art models, including iPiDA-PUL [22], iPiDA-sHN [32] and piRDA [6], verifying the viewpoint advocated by authors that the GCN model effectively captured nonlinear relation patterns from complex association networks between piRNAs and diseases.

Zhang et al. [159] proposed a computational method called PDA-PRGCN, which uses a GCN, subgraph projection, feature augmentation, and dual-loss mechanism strategies to transform piRNA–disease association prediction into a graph link prediction task. They constructed a heterogeneous graph consisting of piRNA–piRNA subgraphs, disease–disease subgraphs, known piRNA–disease subgraphs, and learned node embeddings using GCN layers. They conducted extensive experiments on the main and piRDisease datasets and compared them with existing methods, demonstrating superior performance and robustness.

Meng et al. [160] proposed the ETGPDA model based on an embedded transformation GCN. Compared with previous studies, they used an embedding transformation module that can map the embeddings of piRNAs and diseases to the same space, thus improving the accuracy and efficiency of prediction. To utilize multisource information, they constructed a heterogeneous network based on the similarity information of piRNAs and diseases and the known piRNA–disease associations and used a GCN with an attention mechanism to extract the low-dimensional embeddings of piRNAs and diseases. Finally, they obtained piRNA–disease association scores by calculating the cosine similarity of piRNA and disease embeddings. In the 5-fold cross-validation, ETGPDA achieved an AUC value of 0.9603, which was superior to those of the other five selected computational models. Case studies based on head and neck squamous cell carcinoma and Alzheimer's disease have also confirmed the superior performance of ETGPDA. The advantage of ETGPDA is that it can integrate multisource data information, use a GCN to extract embeddings, and use an embedding transformation module to solve the problem of an inconsistent embedding space. However, it also has some limitations, such as dependence on known piRNA–disease associations and sparsity of the original data. In the future, the authors will consider introducing more similar information on piRNAs and diseases, and further study the different association types and interactions of piRNAs and diseases to provide more powerful help for biological experiments.

Hou et al. [161] proposed an iPiDA-SWGCN model based on a GCN to predict potential piRNA–disease associations. Compared with previous studies, they used a supplementary weighting strategy to solve the problem of high sparsity and Boolean representation of the piRNA–disease network. They integrated various basic prediction factors, supplemented the potential piRNA–disease association in the sparse piRNA–disease network, and assigned different confidence levels to the original piRNA–disease association to perform feature learning and node representation in the GCN. The experimental results showed that iPiDA-SWGCN outperformed other state-of-the-art models, such as iPiDA-PUL [22], iPiDA-sHN [32], piRDA [6], GAPDA [23], and iPiDA-GCN [34] and could predict new piRNA–disease associations. The authors also discussed the application of the supplementary weighting strategy in other link prediction tasks, as well as the value and challenges of unverified piRNA–disease associations in biological research.

4.3. Limitation

In summary, computational methods have achieved remarkable results in piRNA–disease association prediction. However, these computational methods have limitations. The main limitations of traditional machine learning methods are as follows: (1) the need to manually design and select features that may not fully utilize the multisource information of piRNA and disease, such as sequence, function, and semantics, and (2) the need for a large number of positive and negative samples to train the model, but the negative samples of piRNA–disease associations are difficult to obtain, which may lead to class imbalance and overfitting problems. For example, APDA [140] used the features obtained by collaborative filtering, but this method may ignore some sparse or novel associations, and (3) the need to adjust multiple hyperparameters and thresholds, which may affect the generalization ability and stability of the model. For example, iPiDA-PUL[22] used the PUL method; however, in this method, determining the number and quality of unlabeled samples, as well as the integrated learning strategy, was required.

The recommended methods significantly improved the prediction and identification performance, but their limitations were as follows: (1) complex graph structures need to be constructed, which may increase the computational overhead and memory consumption. For example, SPRDA [147] uses a heterogeneous network, but the structural perturbation method of this network may introduce noise and error; (2) the need to consider the association strength between piRNAs and diseases,

rather than just binary association, which may require more fine-grained evaluation metrics and ranking methods. For example, PDA-LTR [33] used a ranking learning method, but determining the ranking and loss functions, as well as dealing with the problem of different lengths of queries and documents were required in this method.

Deep learning-based methods are currently the best computational methods for predicting piRNA–disease associations, but they have the following limitations: (1) the need for a large amount of data and computing resources to train DNNs, which may lead to overfitting and gradient vanishing problems. For example, iPiDA-sHN [32] used a CNN model, but determining the number of layers, size and number of convolutional kernels, activation function, and optimizer was required, and (2) the need to reasonably select and combine different deep learning models, which may cause model incompatibility and conflict problems. For example, iPiDA-GCN [34] uses two GCN models; however, determining the adjacency matrix of the graph, attention mechanism, aggregation function, and fusion method was required in this model.

In addition, the data quality and limitations of piRNA–disease association databases exert an important effect on computational models. First, the data in the piRNA–disease association databases were collected from different studies, and most of the associations in the literature were derived from high-throughput sequencing technology and clinical experimental verification. Although these associations have been verified relatively and reliably, they may also introduce false-positive or false-negative results, leading to noise and data inaccuracy. Second, the data in the piRNA–disease association databases can only reflect the correlation between piRNAs and disease, but cannot reveal the function and mechanism of piRNAs in disease. This requires combining other bioinformatics data, such as gene expression, protein interactions, and signaling pathways, to construct a more complete piRNA–disease association network and perform more in-depth analysis and interpretation. Finally, the data in the piRNA–disease association databases also have some incompleteness and imbalance; that is, some piRNA or disease association information is missing or sparse, whereas some piRNA or disease association information is excessive or redundant. This requires the use of data augmentation or dimensionality reduction methods to improve the coverage and diversity of the data while reducing its redundancy and complexity.

5. Conclusion and perspectives

5.1. Conclusion

In this review, we summarize the databases, computational methods, and evaluation metrics for piRNA-related tasks to provide useful information for future work. The main contributions and innovations of this study are: (1) We proposed a systematic framework, which divided the piRNA computational-related tasks into five identification tasks and one prediction task, which helps to clarify the research context and development trend of computational methods in piRNA-related tasks, (2) We introduced the databases for piRNA-related tasks in detail, including the data sources, data contents, and data collection methods, which help understand the data characteristics and data quality of piRNA-related tasks, (3) We introduced the computational methods for piRNA-related tasks in detail, which help understand the method principles and method performance of piRNA-related tasks, and (4) We paid special attention to the piRNA–disease association prediction task and reviewed the computational methods for the piRNA–disease association prediction task in detail, including the association network construction, feature representation, and prediction methods, as well as their performance comparison, which help promote the development and application of the piRNA–disease association prediction task. This study can help researchers understand the current situation and challenges of piRNA computational research, as well as future development directions.

However, our study has some limitations, mainly owing to insufficient data and methods for piRNA-related tasks. First, there is still no consistent piRNA-naming system in the databases; for example, piR-hsa-237 in the piRBase database is called hsa_piR_000001 in the piRNABank database and hsa-piR-1 in the piRNAdb database, which makes the integration of knowledge difficult. Second, most computational methods require manual work to perform feature engineering, which is expensive and time-consuming. Third, many databases have imbalanced positive and negative samples and lack high-quality negative samples that have been verified experimentally. Fourth, the performance of the computational methods in practical tasks is unsatisfactory. These limitations should be addressed in future studies.

5.2. Future development directions

Based on the above discussion, we propose prospects for future research. First, with the advancement in piRNA research, the construction of piRNA databases and tools has faced new challenges and opportunities. First, a unified piRNA naming system, such as an miRNA database, should be established to facilitate the integration of knowledge. Second, piRNA databases must be constantly updated and expanded to accommodate the growth of piRNA data from different species, tissues, and conditions. Simultaneously, piRNA databases also need to improve the quality and credibility of the data, as well as provide more functional and associative information, such as piRNA modification, structure, interaction, regulation, and biological function. Third, piRNA tools must improve the compatibility and reproducibility of datasets so that users can utilize them on different platforms and environments. In addition, piRNA tools also need to upgrade their algorithms to improve performance and accuracy, as well as provide more visualization and interactive analysis functions to facilitate users in exploring and understanding the complexity and diversity of piRNAs. Some existing piRNA tools, such as piRBase, piRNAQuest, and piRTarBase, have made efforts and contributions to these aspects but still need to be further improved and optimized.

Although this study provided a relatively comprehensive summary of the computational methods for piRNA-related tasks, these methods still have some shortcomings and limitations that need to be addressed. In the future, more biological experimental results, sequencing information, and literature data will become available, and the rapid identification of potential piRNAs from massive amounts of data will become a focus of tool development. Therefore, we think that future research has the following directions: (1) Using parallel deep computing to process massive data, using distributed systems and high-performance hardware to accelerate the training and inference of models, and reducing the computational cost and time; (2) Solving or alleviating the data sparsity and imbalance problems in piRNA-related tasks, using some data augmentation or data dimensionality reduction methods, improving the coverage and diversity of the data, and reducing the redundancy and complexity of the data; (3) Improving the interpretability and credibility of computational methods, developing some visualization and analysis methods, and showing the relationship between the input and output of the model, as well as the key parameters and features of the model so as to better understand the biological functions and mechanisms of piRNA; (4) Using multisource heterogeneous data, including gene expression data, protein interaction data, and epigenetic data, to enrich the feature representation of piRNA, designing appropriate data fusion and representation learning methods, extracting and integrating the commonality and individuality of the data, and enhancing the generalization ability and robustness of the model; (5) Using multitask learning and transfer learning, improving the performance of piRNA identification and prediction, using the correlation and complementarity between different tasks, sharing and transferring knowledge, improving the efficiency and effectiveness of the model, using the similarity and difference between different domains or species, transferring and adapting knowledge, and improving the generalization ability and adaptability of the model.

In conclusion, this study provides a new perspective and idea for computational research on piRNA-related tasks but also has some limitations and challenges that need to be improved and optimized in future research. We hope that this study will stimulate more research interest and innovation and promote the development and progress of computational research on piRNA-related tasks.

Funding

This work was supported by the Major Project of Philosophy and Social Sciences of the Ministry of Education (No. 21JDA050), the Research Fund of the National Language Commission (No.YB145-2), the Guangdong Education Department Project Foundation (No. 2017KTSCX064), the Guangdong Philosophy and Social Sciences Foundation (No. GD20XZY01), Guangdong University of Foreign Studies Project Foundation (Nos. LAI202305, LEC2019ZBKT002, LEC2022ZBKT005), and Guangzhou Science and Technology Project Foundation (No. 202201010717).

CRediT authorship contribution statement

Conceptualization, Chang Guo and Han Ren; methodology, Chang Guo; investigation, Chang Guo, Xiaoli Wang and Han Ren; writing—original draft preparation, Chang Guo and Han Ren; writing—review and editing, Han Ren and Xiaoli Wang; funding acquisition, Han Ren. All authors have read and agreed to the published version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. Nat Rev Cancer 2018;18:5–18.
[2] Sarkar A, Maji RK, Saha S, Ghosh Z. piRNAQuest: searching the piRNAome for silencers. BMC Genom 2014;15:1–17.
[3] Betel D, Sheridan R, Marks DS, Sander C. Computational analysis of mouse piRNA sequence and biogenesis. PLoS Comput Biol 2007;3:e222.
[4] Rayford KJ, Cooley A, Rumph JT, Arun A, Rachakonda G, Villalta F, et al. piRNAs as modulators of disease pathogenesis. Int J Mol Sci 2021;22:2373.
[5] Wu W, Lu B-F, Jiang R-Q, Chen S. The function and regulation mechanism of piRNAs in human cancers. Histol Histopathol 2021;18323:807–16.
[6] Ali SD, Tayara H, Chong KT. Identification of piRNA disease associations using deep learning. Comput Struct Biotechnol J 2022;20:1208–17.
[7] Weng W, Li H, Goel A. Piwi-interacting RNAs (piRNAs) and cancer: emerging biological concepts and potential clinical implications. Biochim Et Biophys Acta (BBA)-Rev Cancer 2019;1871:160–9.
[8] Roy J, Sarkar A, Parida S, Ghosh Z, Mallick B. Small RNA sequencing revealed dysregulated piRNAs in Alzheimer's disease and their probable role in pathogenesis. Mol Biosyst 2017;13:565–76.
[9] Maleki Dana P, Mansournia MA, Mirhashemi SM. PIWI-interacting RNAs: new biomarkers for diagnosis and treatment of breast cancer. Cell Biosci 2020;10:1–8.
[10] Cui L, Lou Y, Zhang X, Zhou H, Deng H, Song H, et al. Detection of circulating tumor cells in peripheral blood from patients with gastric cancer using piRNAs as markers. Clin Biochem 2011;44:1050–7.
[11] Iliev R, Fedorko M, Machackova T, Mlcochova H, Svoboda M, Pacík D, et al. Expression levels of PIWI-interacting RNA, piR-823, are deregulated in tumor tissue, blood serum and urine of patients with renal cell carcinoma. Anticancer Res 2016;36:6419–23.
[12] Yin J, Jiang XY, Qi W, Ji CG, Xie XL, Zhang DX, et al. piR-823 contributes to colorectal tumorigenesis by enhancing the transcriptional activity of HSF 1. Cancer Sci 2017;108:1746–56.
[13] Tang X, Xie X, Wang X, Wang Y, Jiang X, Jiang H. The combination of piR-823 and eukaryotic initiation factor 3 B (EIF3B) activates hepatic stellate cells via upregulating TGF-β1 in liver fibrogenesis. Med Sci Monit: Int Med J Exp Clin Res 2018;24:9151.
[14] Wang X, Ramat A, Simonelig M, Liu M-F. Emerging roles and functional mechanisms of PIWI-interacting RNAs. Nat Rev Mol Cell Biol 2023;24:123–41.
[15] Chen X, Sun Y-Z, Guan N-N, Qu J, Huang Z-A, Zhu Z-X, et al. Computational models for lncRNA function prediction and functional similarity calculation. Brief Funct Genom 2019;18:58–82.
[16] Ernst C, Odom DT, Kutter C. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. Nat Commun 2017;8(1):10.
[17] Thakker DR, Natt F, Hüsken D, Maier R, Müller M, van der Putten H, et al. Neurochemical and behavioral consequences of widespread gene knockdown in the adult mouse brain by using nonviral RNA interference. Proc Natl Acad Sci USA 2004;101:17270–5.
[18] Liu Y, Zhang J, Li A, Liu Z, Zhang Y, Sun X. Detection of Piwi-interacting RNAs based on sequence features. Genet Mol Res 2016;15.
[19] Wang K, Hoeksema J, Liang C. piRNN: deep learning algorithm for piRNA prediction. PeerJ 2018;6:e5429.
[20] Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565–7.
[21] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. Pattern Recognit 2018;77:354–77.
[22] Wei H, Xu Y, Liu B. iPiDi-PUL: identifying Piwi-interacting RNA-disease associations based on positive unlabeled learning. Brief Bioinforma 2021;22: bbaa058.
[23] Zheng K, You Z-H, Wang L, Wong L, Chen Z-H. Inferring disease-associated Piwi-interacting RNAs via graph attention networks. In: Proceedings of international conference on intelligent computing, Springer; 2020, p. 239–50.
[24] Qi Y. Random forest for bioinformatics, Ensemble machine learning. Springer; 2012. p. 307–23.
[25] Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. stat 2017;1050:20.
[26] Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res 2008;36:D173–7.
[27] Zhang P, Si X, Skogerbø G, Wang J, Cui D, Li Y, et al. piRBase: a web resource assisting piRNA functional study. Database 2014;(2014).
[28] Wang J, Zhang P, Lu Y, Li Y, Zheng Y, Kan Y, et al. piRBase: a comprehensive database of piRNA sequences. Nucleic Acids Res 2019;47:D175–80.
[29] Wang J, Shi Y, Zhou H, Zhang P, Song T, Ying Z, et al. piRBase: integrating piRNA annotation in all aspects. Nucleic Acids Res 2022;50:D265–72.
[30] da Costa AH, Santos RACd, Cerri R. Investigating deep feedforward neural networks for classification of transposon-derived piRNAs. Complex Intell Syst 2022;8:477–87.
[31] Khan S, Khan M, Iqbal N, Hussain T, Khan SA, Chou K-C. A two-level computation model based on deep learning algorithm for identification of piRNA and their functions via Chou's 5-steps rule. Int J Pept Res Ther 2020;26:795–809.
[32] Wei H, Ding Y, Liu B. iPiDA-sHN: Identification of Piwi-interacting RNA-disease associations by selecting high quality negative samples. Comput Biol Chem 2020; 88:107361.
[33] Zhang W, Hou J, Liu B. iPiDA-LTR: Identifying piwi-interacting RNA-disease associations based on Learning to Rank. PLOS Comput Biol 2022;18:e1010404.
[34] Hou J, Wei H, Liu B. iPiDA-GCN: Identification of piRNA-disease associations based on Graph Convolutional Network. PLOS Comput Biol 2022;18:e1010671.
[35] Liu Y, Li A, Xie G, Liu G, Hei X. Computational methods and online resources for identification of piRNA-related molecules. Interdiscip Sci: Comput life Sci 2021; 13:176–91.
[36] Zhang T, Chen L, Li R, Liu N, Huang X, Wong G. PIWI-interacting RNAs in human diseases: databases and computational models, Briefings in Bioinformatics; 2022.
[37] He S, Liu C, Skogerbø G, Zhao H, Wang J, Liu T, et al. NONCODE v2. 0: decoding the non-coding. Nucleic Acids Res 2007;36:D170–2.
[38] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30: 207–10.
[39] Piuco R, Galante PA. piRNAdb: A piwi-interacting RNA database, bioRxiv; 2021.
[40] Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 2007;128:1089–103.
[41] Rosenkranz D. piRNA cluster database: a web resource for piRNA producing loci. Nucleic Acids Res 2016;44:D223–30.
[42] Rosenkranz D, Zischler H, Gebert D. piRNAclusterDB 2.0: update and expansion of the piRNA cluster database. Nucleic Acids Res 2022;50:D259–64.
[43] Jiang, B-R, Wu W-Y, Chien C-H, Tsai JJ, Chan W-L. piRNAtarget: The integrated database for mining functionality of piRNA and its targets. In: Proceedings of the 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE; 2016, p. 382–6.
[44] Wu W-S, Brown JS, Chen T-T, Chu Y-H, Huang W-C, Tu S, et al. piRTarBase: a database of piRNA targeting sites and their roles in gene regulation. Nucleic Acids Res 2019;47:D181–7.
[45] Wu W-S, Huang W-C, Brown JS, Zhang D, Song X, Chen H, et al. pirScan: a webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. Nucleic Acids Res 2018;46:W43–8.
[46] Gainetdinov I, Colpan C, Cecchini K, Arif A, Jouravleva K, Albosta P, et al. Terminal modification, sequence, length, and PIWI-protein identity determine piRNA stability. Mol Cell 2021;81:4826–42. e4828.
[47] Zhang H, Ali A, Gao J, Ban R, Jiang X, Zhang Y, et al. IsopiRBank: a research resource for tracking piRNA isoforms. Database 2018:2018.
[48] Zhang Y, Xu B, Yang Y, Ban R, Zhang H, Jiang X, et al. CPSS: a computational platform for the analysis of small RNA deep sequencing data. Bioinformatics 2012;28:1925–7.
[49] Wan C, Gao J, Zhang H, Jiang X, Zang Q, Ban R, et al. CPSS 2.0: a computational platform update for the analysis of small RNA sequencing data. Bioinformatics 2017;33:3289–91.
[50] Barreñada O, Larriba E, Brieño-Enriquez MA, Mazo Jd. piRNA-IPdb: a PIWI-bound piRNAs database to mining NGS sncRNA data and beyond. BMC Genom 2021;22:1–8.

[51] Muhammad A, Waheed R, Khan NA, Jiang H, Song X. piRDisease v1. 0: a manually curated database for piRNA associated diseases. Database 2019;2019.
[52] Zhang W, Wu S, Zhang H, Guan W, Zeng B, Wei Y, Chan GC-F, Li W. piRPheno: A manually curated database to prioritize and analyze human disease related piRNAs, bioRxiv; 2020.
[53] Zhang Y, Wang X, Kang L. A k-mer scheme to predict piRNAs and characterize locust piRNAs. Bioinformatics 2011;27:771–6.
[54] Brayet J, Zehraoui F, Jeanson-Leh L, Israeli D, Tahi F. Towards a piRNA prediction using multiple kernel fusion and support vector machine. Bioinformatics 2014;30:i364–70.
[55] Boucheham A, Sommard V, Zehraoui F, Boualem A, Batouche M, Bendahmane A, et al. IpiRId: Integrative approach for piRNA prediction using genomic and epigenomic data. PLoS One 2017;12:e0179787.
[56] Khan S, Khan M, Iqbal N, Khan SA, Chou K-C. Prediction of piRNAs and their function based on discriminative intelligent model using hybrid features into Chou's PseKNC. Chemom Intell Lab Syst 2020;203:104056.
[57] Zheng J, Wang K. Emerging deep learning methods for single-cell RNA-seq data analysis. Quant Biol 2019;7:247–54.
[58] Zuo Y, Zou Q, Lin J, Jiang M, Liu X. 2lpiRNApred: A two-layered integrated algorithm for identifying piRNAs and their functions based on LFE-GM feature selection. RNA Biol 2020;17:892–902.
[59] Zheng K, You Z-H, Wang L, Wong L, Zhan Z-h. SPRDA: a matrix completion approach based on the structural perturbation to infer disease-associated Piwi-Interacting RNAs. bioRxiv 2020.
[60] Kurtz S, Narechania A, Stein JC, Ware D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genom 2008;9:1–18.
[61] Pian C, Chen Y-Y, Zhang J, Chen Z, Zhang G-L, Li Q, et al. V-ELMpiRNAPred: Identification of human piRNAs by the voting-based extreme learning machine (V-ELM) with a new hybrid feature. J Bioinforma Comput Biol 2017;15:1650046.
[62] Liu X, Ding J, Gong F. piRNA identification based on motif discovery. Mol Biosyst 2014;10:3075–80.
[63] Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinforma (Oxf, Engl) 1998;14:55–67.
[64] Menor MS, Baek K, Poisson G. Prediction of mature microRNA and piwi-interacting RNA without a genome reference or precursors. Int J Mol Sci 2015;16:1466–81.
[65] Rahiman AA, Ajitha, J, Chandra V. An integrated computational schema for analysis, prediction and visualization of piRNA sequences, International Conference on Intelligent Computing, Springer; 2015, pp. 744–50.
[66] Seyeddokht A, Aslaminejad AA, Masoudi-Nejad A, Nassiri M, Zahiri J, Sadeghi B. Computational detection of piRNA in human using support vector machine. Avicenna J Med Biotechnol 2016;8:36.
[67] Monga I, Banerjee I. Computational identification of piRNAs using features based on rna sequence, structure, thermodynamic and physicochemical properties. Curr Genom 2019;20:508–18.
[68] LeCun Y, Bengio Y, Hinton G. Deep learning. nature 2015;521:436–44.
[69] Cao J, Lin Z, Huang G-B, Liu N. Voting based extreme learning machine. Inf Sci 2012;185:66–77.
[70] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: a new learning scheme of feedforward neural networks. In: Proceedings of the 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541), IEEE; 2004, p. 985–90.
[71] Khan S, Khan M, Iqbal N, Li M, Khan DM. Spark-based parallel deep neural network model for classification of large scale RNAs into piRNAs and non-piRNAs. IEEE Access 2020;8:136978–91.
[72] Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-R, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 2012;29:82–97.
[73] Kim H, Park J, Jang J, Yoon S. Deepspark: a spark-based distributed deep learning framework for commodity clusters. arXiv Prepr arXiv 2016;1602:08191.
[74] Chen C-C, Chan Y-M, Jeong H. LSTM4piRNA: efficient piRNA detection in large-scale genome databases using a deep learning-based LSTM network. Int J Mol Sci 2023;24:15681.
[75] Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. PIWI-interacting RNAs: small RNAs with big functions. Nat Rev Genet 2019;20:89–108.
[76] Luo L, Li D, Zhang W, Tu S, Zhu X, Tian G. Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features. PLoS One 2016;11:e0153268.
[77] Wang K, Liang C, Liu J, Xiao H, Huang S, Xu J, et al. Prediction of piRNAs using transposon interaction and a support vector machine. BMC Bioinforma 2014;15:1–8.
[78] Tafer H, Hofacker IL. RNAplex: a fast tool for RNA–RNA interaction search. Bioinformatics 2008;24:2657–63.
[79] Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 2010;26:392–8.
[80] Zhang W, Liu J, Xiong Y, Ke M, Zhang K. Predicting immunogenic T-cell epitopes by combining various sequence-derived features. In: Proceedings of the 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE; 2013, p. 4–9.
[81] Zou Q, Guo J, Ju Y, Wu M, Zeng X, Hong Z. Improving tRNAscan-SE annotation results via ensemble classifiers. Mol Inform 2015;34:761–70.
[82] Dietterich TG. Ensemble learning, The handbook of brain theory and neural networks. 2002. p. 110–25.
[83] Liu B, Liu F, Wang X, Chen J, Fang L, Chou K-C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res 2015;43:W65–71.
[84] Liu B, Liu F, Fang L, Wang X, Chou K-C. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics 2015;31:1307–9.
[85] El-Manzalawy Y, Dobbs D, Honavar V. Predicting flexible length linear B-cell epitopes, Computational Systems Bioinformatics: (Volume 7), World Scientific; 2008, p. 121–32.
[86] Xia X. Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. Scientifica 2012;2012.
[87] Li D, Luo L, Zhang W, Liu F, Luo F. A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. BMC Bioinforma 2016;17:1–11.
[88] Burgess DJ. Defining piRNA expression. Nat Rev Genet 2013;14. 301-301.
[89] Gou L-T, Dai P, Yang J-H, Xue Y, Hu Y-P, Zhou Y, et al. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res 2014;24:680–700.
[90] Zhang L-M, Gao Q-X, Chen J, Li B, Li M-M, Zheng L, et al. A universal catalytic hairpin assembly system for direct plasma biopsy of exosomal PIWI-interacting RNAs and microRNAs. Anal Chim Acta 2022;1192:339382.
[91] Liu B, Yang F, Chou K-C. 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. Mol Ther-Nucleic Acids 2017;7:267–77.
[92] Liu B, Fang L, Long R, Lan X, Chou K-C. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 2016;32:362–9.
[93] Li T, Gao M, Song R, Yin Q, Chen Y. Support vector machine classifier for accurate identification of piRNA. Appl Sci 2018;8:2204.
[94] Ben-Bassat I, Chor B, Orenstein Y. A deep neural network approach for learning intrinsic protein-RNA binding preferences. Bioinformatics 2018;34:i638–46.
[95] Khan S, Khan M, Iqbal N, Rahman M, Karim MKA. Deep-piRNA: Bi-Layered Prediction Model for PIWI-Interacting RNA Using Discriminative Features. Comput, Mater Contin 2022;72:2243–58.
[96] M.A. Nielsen, Neural networks and deep learning, Determination press San Francisco, CA, USA2015.
[97] Yager RR, Kreinovich V. Universal approximation theorem for uninorm-based fuzzy systems modeling. Fuzzy Sets Syst 2003;140:331–9.
[98] Ali SD, Alam W, Tayara H, Chong K. Identification of functional piRNAs using a convolutional neural network. IEEE/ACM Trans Comput Biol Bioinforma 2020.
[99] Tahir M, Hayat M, Khan S, Chong K to. Prediction of Piwi-Interacting RNAs and Their Functions via Convolutional Neural Network. IEEE Access 2021;9:54233–40.
[100] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv Prepr arXiv 2013;1301:3781.
[101] Liu Y, Ding Y, Li A, Fei R, Guo X, Wu F. Prediction of exosomal piRNAs based on deep learning for sequence embedding with attention mechanism. In: Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE; 2022, pp. 158–61.
[102] Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature 2006;442:203–7.
[103] Klattenhoff, C, Theurkauf W. Biogenesis and germline functions of piRNAs; 2008.
[104] Yuan J, Zhang P, Cui Y, Wang J, Skogerbø G, Huang D-W, et al. Computational identification of piRNA targets on mouse mRNAs. Bioinformatics 2016;32:1170–7.
[105] Singh G, Roy J, Rout P, Mallick B. Genome-wide profiling of the PIWI-interacting RNA-mRNA regulatory networks in epithelial ovarian cancers. PLoS One 2018;13:e0190485.
[106] Zhang D, Tu S, Stubna M, Wu W-S, Huang W-C, Weng Z, et al. The piRNA targeting rules and the resistance to piRNA silencing in endogenous genes. Science 2018;359:587–92.
[107] Chan W-L, Yeh M-C, Wang J-D, Chang J-G, Tsai JJ. Genome-wide functional identification of maximal consensus patterns derived from multiple species pirnas. In: Proceedings of the 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE; 2016, p. 377–81.
[108] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[109] Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. elife 2015;4:e05005.
[110] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS, et al. Human microRNA targets. PLoS Biol 2004;2:e363.
[111] Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. Rna 2004;10:1507–17.
[112] Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y. Potent effect of target structure on microRNA function. Nat Struct Mol Biol 2007;14:287–94.
[113] Wang J-H, Chen W-X, Mei S-Q, Yang Y-D, Yang J-H, Qu L-H, et al. tsRFun: a comprehensive platform for decoding human tsRNA expression, functions and prognostic value by high-throughput small RNA-Seq and CLIP-Seq data. Nucleic Acids Res 2022;50. D421-D431.
[114] Zhang P, Kang J-Y, Gou L-T, Wang J, Xue Y, Skogerboe G, et al. MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. Cell Res 2015;25:193–207.

[115] Yang T-H, Shiue S-C, Chen K-Y, Tseng Y-Y, Wu W-S. Identifying piRNA targets on mRNAs in C. elegans using a deep multi-head attention network. BMC Bioinforma 2021;22:1–23.
[116] Chen C-C, Qian X, Yoon B-J. Effective computational detection of piRNAs using n-gram models and support vector machine. BMC Bioinforma 2017;18:103–9.
[117] Rosenkranz D, Zischler H. proTRAC-a software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinforma 2012;13:1–10.
[118] Jung I, Park JC, Kim S. piClust: a density based piRNA clustering algorithm. Comput Biol Chem 2014;50:60–7.
[119] Ray R, Pandey P. piRNA analysis framework from small RNA-Seq data by a novel cluster prediction tool-PILFER. Genomics 2018;110:355–65.
[120] Han BW, Wang W, Zamore PD, Weng Z. piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome-and CAGE-seq, ChIP-seq and genomic DNA sequencing. Bioinformatics 2015;31:593–5.
[121] Singh G, Mallick B. Predicting sequence and structural features of effective piRNA target binding sites. J Mol Recognit 2022;35:e2949.
[122] Kim VN. Small RNAs just got bigger: piwi-interacting RNAs (piRNAs) in mammalian testes. Genes Dev 2006;20:1993–7.
[123] Assis R, Kondrashov AS. Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. Proc Natl Acad Sci USA 2009;106:7079–82.
[124] Choudhuri S. Lesser known relatives of miRNA. Biochem Biophys Res Commun 2009;388:177–80.
[125] Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, et al. Characterization of the piRNA complex from rat testes. Science 2006;313:363–7.
[126] Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. kdd 1996:226–31.
[127] You Z-H, Huang Z-A, Zhu Z, Yan G-Y, Li Z-W, Wen Z, et al. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput Biol 2017;13:e1005455.
[128] Zhou J, Xie H, Liu J, Huang R, Xiang Y, Tian D, et al. PIWI-interacting RNAs: critical roles and therapeutic targets in cancer. Cancer Lett 2023:216189.
[129] Liu Y, Xie G, Li A, He Z, Hei X. Prediction of cancer-related piRNAs based on network-based stratification analysis. Int J Pattern Recognit Artif Intell 2022;36:2259002.
[130] Liu Y, Li A, Zhu Y, Pang X, Hei X, Xie G, et al. piRSNP: a database of piRNA-related SNPs and their effects on cancerrelated piRNA functions. Curr Bioinforma 2023;18:509–16.
[131] Li G, Luo J, Wang D, Liang C, Xiao Q, Ding P, et al. Potential circRNA-disease association prediction using DeepWalk and network consistency projection. J Biomed Inform 2020;112:103624.
[132] Van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 2011;27:3036–43.
[133] Li J, Zhang S, Liu T, Ning C, Zhang Z, Zhou W. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. Bioinformatics 2020;36:2538–46.
[134] Ding Y, Tian L-P, Lei X, Liao B, Wu F-X. Variational graph auto-encoders for miRNA-disease association prediction. Methods 2021;192:25–34.
[135] Yao Y, Ji B, Shi S, Xu J, Xiao X, Yu E, et al. IMDAILM: inferring miRNA-disease association by integrating lncRNA and miRNA data. IEEE Access 2019;8:16517–27.
[136] Wang L, You Z-H, Li Y-M, Zheng K, Huang Y-A. GCNCDA: a new method for predicting circRNA-disease associations based on graph convolutional network algorithm. PLoS Comput Biol 2020;16:e1007568.
[137] Wu X, Lan W, Chen Q, Dong Y, Liu J, Peng W. Inferring LncRNA-disease associations based on graph autoencoder matrix completion. Comput Biol Chem 2020;87:107282.
[138] Xuan P, Pan S, Zhang T, Liu Y, Sun H. Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. Cells 2019;8:1012.
[139] Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. Bioinformatics 2019;35:4730–8.
[140] Zheng K, You Z-H, Wang L, Li H-Y, Ji B-Y. Predicting human disease-associated piRNAs based on multi-source information and random forest. International conference on intelligent computing. Springer; 2020. p. 227–38.
[141] Ou-Yang L, Huang J, Zhang X-F, Li Y-R, Sun Y, He S, et al. LncRNA-disease association prediction using two-side sparse self-representation. Front Genet 2019;10:476.
[142] Mordelet F, Vert J-P. A bagging SVM to learn from positive and unlabeled examples. Pattern Recognit Lett 2014;37:201–9.
[143] Claesen M, De Smet F, Suykens JA, De Moor B. A robust ensemble approach to learn from positive and unlabeled data using SVM base models. Neurocomputing 2015;160:73–84.
[144] Chen X, Zhu C-C, Yin J. Ensemble of decision tree reveals potential miRNA-disease associations. PLoS Comput Biol 2019;15:e1007209.
[145] Zhang X, Zou Q, Rodriguez-Paton A, Zeng X. Meta-path methods for prioritizing candidate disease miRNAs. IEEE/ACM Trans Comput Biol Bioinforma 2017;16:283–91.
[146] Chen X, Liu X. A weighted bagging LightGBM model for potential lncRNA-disease association identification. In: Proceedings of international conference on bio-inspired computing: theories and applications, Springer; 2018, p. 307–14.
[147] Zheng K, Zhang X-L, Wang L, You Z-H, Ji B-Y, Liang X, et al. SPRDA: a link prediction approach based on the structural perturbation to infer disease-associated Piwi-interacting RNAs. Brief Bioinforma 2023;24:bbac498.
[148] He S, Guo F, Zou Q. MRMD2. 0: a python tool for machine learning with feature ranking and reduction. Curr Bioinforma 2020;15:1213–21.
[149] Figueroa A, Neumann G. Learning to rank effective paraphrases from query logs for community question answering. Twenty-Seven– AAAI Conf Artif Intell 2013.
[150] Wei H, Xu Y, Liu B. iCircDA-LTR: identification of circRNA–disease associations based on learning to rank. Bioinformatics 2021;37:3302–10.
[151] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. Proc IEEE Conf Comput Vis Pattern Recognit 2016:761–9.
[152] Buckman J, Roy A, Raffel C, Goodfellow I. Thermometer encoding: One hot way to resist adversarial examples. Int Conf Learn Represent 2018.
[153] Qian Y, He Q, Deng L. iPiDA-GBNN: Identification of Piwi-interacting RNA-disease associations based on gradient boosting neural network. In: Proceedings of the 2021 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE; 2021, p. 1045–50.
[154] Badirli S, Liu X, Xing Z, Bhowmik A, Doan K, Keerthi SS. Gradient boosting neural networks: Grownet. arXiv Prepr arXiv 2020;2002:07971.
[155] Zheng K, Liang Y, Liu Y-Y, Yasir M, Wang P. A decision support system based on multi-sources information to predict piRNA–disease associations using stacked autoencoder. Soft Comput 2022;26:11007–16.
[156] Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics 2010;26:1644–50.
[157] Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics 2014;30:31–7.
[158] Ji B, Luo J, Pan L, Xie X, Peng S. DFL-PiDA: Prediction of Piwi-interacting RNA-Disease Associations based on Deep Feature Learning. In: Proceedings of the 2021 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE; 2021, pp. 406–11.
[159] Zhang P, Sun W, Wei D, Li G, Xu J, You Z, et al. PDA-PRGCN: identification of Piwi-interacting RNA-disease associations through subgraph projection and residual scaling-based feature augmentation. BMC Bioinforma 2023;24:1–18.
[160] Meng X, Shang J, Ge D, Yang Y, Zhang T, Liu J-X. ETGPDA: identification of piRNA-disease associations based on embedding transformation graph convolutional network. BMC Genom 2023;24:279.
[161] Hou J, Wei H, Liu B. iPiDA-SWGCN: identification of piRNA-disease associations based on supplementarily weighted graph convolutional network. PLOS Comput Biol 2023;19:e1011242.