REVIEW ARTICLE

A Brief Survey for MicroRNA Precursor Identification Using Machine Learning Methods

Zheng-Xing Guan[1], Shi-Hao Li[1], Zi-Mei Zhang[1], Dan Zhang[1], Hui Yang[1] and Hui Ding[1,*]

[1]Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

ARTICLE HISTORY

Received: October 24, 2019
Revised: January 24, 2020
Accepted: January 30, 2020

DOI:
10.2174/1389202921666200214125102

Abstract: MicroRNAs, a group of short non-coding RNA molecules, could regulate gene expression. Many diseases are associated with abnormal expression of miRNAs. Therefore, accurate identification of miRNA precursors is necessary. In the past 10 years, experimental methods, comparative genomics methods, and artificial intelligence methods have been used to identify pre-miRNAs. However, experimental methods and comparative genomics methods have their disadvantages, such as time-consuming. In contrast, machine learning-based method is a better choice. Therefore, the review summarizes the current advances in pre-miRNA recognition based on computational methods, including the construction of benchmark datasets, feature extraction methods, prediction algorithms, and the results of the models. And we also provide valid information about the predictors currently available. Finally, we give the future perspectives on the identification of pre-miRNAs. The review provides scholars with a whole background of pre-miRNA identification by using machine learning methods, which can help researchers have a clear understanding of progress of the research in this field.

Keywords: microRNA, precursor, identification, machine learning methods, benchmark dataset, feature extraction, prediction algorithm.

1. INTRODUCTION

MicroRNA (miRNA) is an endogenous small non-coding RNA that can regulate the expression of other genes [1]. The earliest discovered miRNA gene was lin-4 from *C. elegant* [2], but it did not attract the attention of the scientific community at that time. It was not until the discovery of the second miRNA (named let-7) in 2000 that miRNAs came into everyone's sight [3]. After that, with the publication of a large number of papers about miRNAs, the biogenesis process of miRNAs was elaborated. In animals, the transcription of most miRNAs is mediated by RNA polymerase II (Pol II) [4]. Pri-miRNA is generally thousands of nucleotides (nt) long sequences with a stem-loop structure inside. The stem-loop structure in pri-miRNA will be cut by the endonuclease Drosha in the nucleus, resulting in a length of about 70nt pre-miRNA [5]. The pre-miRNA is then transported into the cytoplasm by Exportin V and Ran-GTP cofactor [6, 7]. The pre-miRNA is further cleaved by another endonuclease Dicer to form a double-stranded RNA (miRNA / miRNA*) [8]. Subsequently, one strand of the duplex, denoted with an asterisk (*), is normally degraded. The other strand is the mature miRNA, which will form an RNA-induced silencing complex (RISC) with other proteins and perform its regulatory functions by interacting with their target mRNAs [9]. The biogenesis of animal miRNA is shown in Fig. (**1**). Therefore, in animals, mature miRNAs are single-stranded RNAs with a length of about 22nt, which will play the role of translational repression and mRNA cleavage. In plants, the biological process of miRNA is very different from that of animals. The specific process can be seen in [10], and will not be introduced here.

MiRNAs play key roles in many biological processes, such as growth and development [11-13], cell proliferation [14], cell apoptosis [15], cell differentiation [16] and fat metabolism [17]. The abnormal expression of miRNAs has been found in many human diseases, particularly in cancer [18, 19]. In recent years, miRNA-based cancer treatment and drug development have attracted researchers' attention [20-25]. Therefore, how to accurately identify miRNAs has become a rapidly developing research field. At the beginning, the identification of novel miRNA genes was almost always achieved by direct cloning of endogenous small RNAs and high-throughput sequencing [26, 27]. Those low expression miRNAs or highly tissue specific, time-specific miRNAs are challenging to identify by experimental means. In this case, many methods based on comparative genomics have been proposed, like MiRscan [28], miRseeker [29] and MiRAlign [30]. The rationale of comparative methods is based on the conservation of pre-miRNA-like hairpin secondary structures in closely related genomes [31]. They could find conserved pre-miRNAs in closely related species, but many novel pre-miRNAs are missed [32]. Besides, the method is still time-consuming.

*Address correspondence to this author at the Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China; Tel/Fax: 028-83208232; E-mail: hding@uestc.edu.cn

1389-2029/20 $65.00+.00

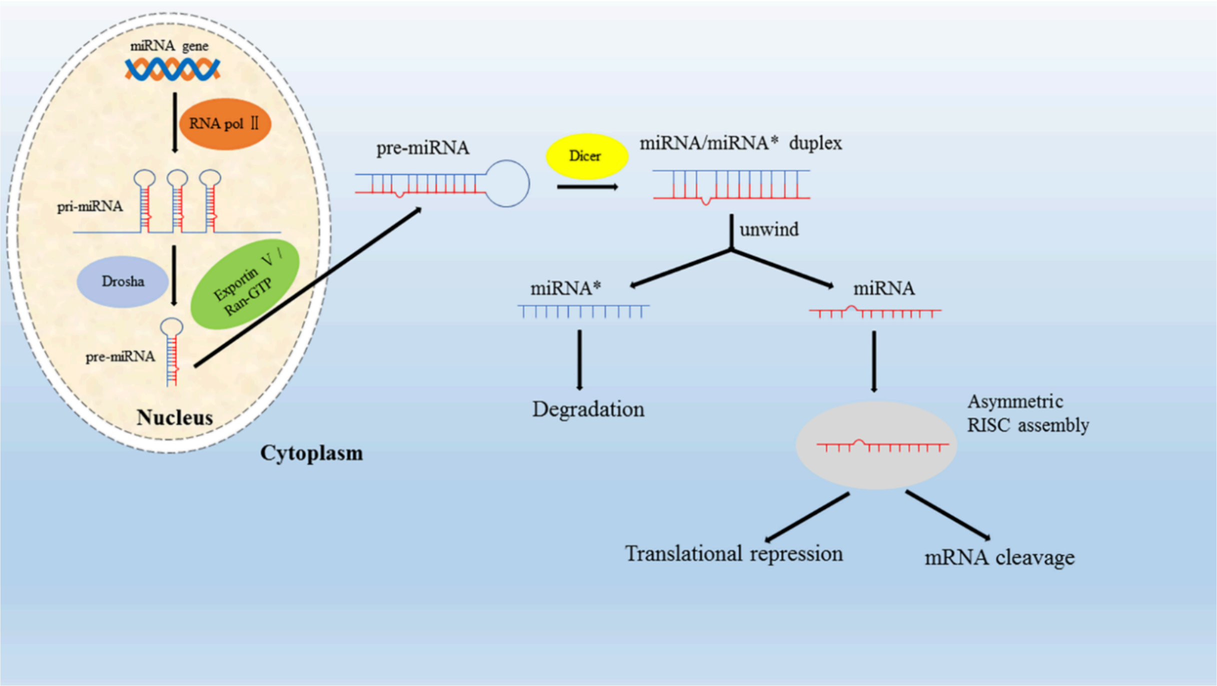

Fig. (1). The schematic diagram of miRNA biogenesis. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Although some miRNAs have been found, more miRNAs are still not discovered. It is obvious that wet-experiments are expensive and time-consuming for detecting miRNA genes. In the post-genomic era, more and more genomic sequence data were available, which provide an opportunity to computational identification of pre-miRNA. However, the capability to identify miRNAs is also very limited for comparative genomics. Consequently, it is urgent to design a powerful predictive computational method to discover new pre-miRNAs. Thus, artificial intelligence methods have been applied in the field. These artificial intelligence-based methods firstly transfer pre-miRNAs into a vector. Then the positive and negative samples are used to train the prediction model. Typically, the inputs are features of candidate pre-miRNA. The outputs would be 1 or 0 indicating pre-miRNA or not pre-miRNA. Through machine learning, a sequence can be easily judged as a pre-miRNA or not. Based on the ratio between positive and negative data, these predictors can be divided into two types: one is the balanced data (or low unbalanced data) based predictors [32-43] and others are the highly unbalanced data based predictors [31, 44-46]. Moreover, these predictors also can be roughly divided into human pre-miRNA predictors [31-35, 37, 38, 40, 42, 43], plant pre-miRNA predictors [36, 39] and multi-species pre-miRNA predictors [41, 44-47] according to the species. As early as 2005, Sewer *et al.* extracted 40 distinctive 'markers' from the hairpin structure and successfully predicted new viral miRNAs by using a support vector machine (SVM) [48]. In the same year, Xue *et al.* proposed a descriptor that formulates local contiguous structure-sequence characteristics from pre-miRNAs, and then combined with SVM to construct the triplet-SVM classifier [33]. In 2007, Kwang *et al.* obtained hairpin features and built the miPred classifier in conjunction with SVM [32]. At the same time, Peng *et al.* proposed a random forest (RF)-based prediction model called MiPred [34]. Both classifiers can identify pre-miRNAs. In 2009, based on 29 features extracted by Kwang *et al.* [32], Rukshan *et al.* used 48 features combined with SVM to build a classifier called microPred [31]. Later, in 2011, Ping *et al.* used SVM to construct the classifier PlantMiRNAPred [36]. Subsequently, in 2013, to solve the sample imbalance, increase the practicality of cross-species sequences and reduce computation time, Adamd *et al.* developed the HuntMi software [44]. In recent years, great progress has been made in predicting pre-miRNAs based on machine learning algorithms. For large-scale prediction of plant pre-miRNAs, Meng *et al.* proposed miPlantPreMat [49]. In 2015, Van *et al.* proposed a new approach to deal with the imbalance of training data in the identification of miRNA precursors [50]. They combined a sequence of weakened SVM component classifiers with the boosting method to construct a prediction model (called miRBoost), which has a reliable classification performance and fast running speed. Meanwhile, Liu *et al.* considered the correlated information in their model, which has a good improvement in the identification of human pre-miRNAs [38]. Then, in 2016, they proposed another feature extraction method based on the previous work, called Pseudo distance structure status pair composition (PseDPC) [40]. Zou *et al.* applied BP neural network to achieve good pre-miRNAs identification on a variety of species [41]. Stegmayer *et al.* used a deepSOM-based method to achieve clustering, which solved the problem well [45]. Tav *et al.* built a web server based on an algorithm called miRNAFold which can help researchers quickly predict pre-miRNAs in the genome [47]. At the same time, Yao *et al.* predicted plant pre-miRNA by energy features [39]. In 2017, Khan *et al.* proposed the MicroR-Pred model for identifying pre-microRNAs in humans [51]. In 2018, Yones *et al.* designed miRNAss based on semi-supervised learning [46]. In 2019, Zheng *et al.* applied convolutional neural networks to the prediction of pre-miRNA [42]. Fu *et al.* achieved better results in human pre-miRNA prediction [43]. All of these studies have yielded exciting results in their respective concerns and will greatly assist in further accurately identifying miRNAs, which have important implications for miRNAs-related drug development and treatment.

To build a pre-miRNAs predictor, the overall process required is shown in Fig. (**2**). Therefore, to help scholars have a good understanding of the identification of pre-miRNAs based on machine learning, this article will elaborate from the following six aspects: (1) Benchmark data generation; (2) Sample description; (3) Prediction methods; (4) Performance evaluation; (5) Published results; (6) Conclusion and perspectives [52-55].

2. BENCHMARK DATASETS

2.1. Published MiRNAs-Related Databases

A database is a collection of data stored together in a way that can be shared with multiple users. The database maintainer can perform operations such as adding, querying, updating, and deleting data in the database. Users can query regarding data and perform data download operations in the database. With the massive accumulation of biological data, it is very troublesome to find all data in a particular field from literature. Therefore, more and more databases have been developed to facilitate researchers to query data [56-60]. Some of these databases are specifically developed to store miRNA sequence information [61-63]. Here, we will give a brief introduction to these miRNA sequence databases.

The microRNA Registry database [61] is a database for storing miRNA information. In the early days, some data on miRNA-related work are derived from this database [49, 50].

The miRbase [62] sequence database is a comprehensive database that provides information on published miRNA sequence data, annotations, predicted gene targets, *etc.* It provides a convenient online query service that allows users to search for known miRNA and target information online using keywords or sequences. It is the most widely used database today.

The plant microRNA database (PMRD) [63] was a database of plant microRNAs. The database attempts to integrate large amounts of data on plant microRNAs. Although many of its functions are not available today, we can still find some plant miRNA and its target gene sequence data.

Besides, some miRNA-related databases, such as: miRTarBase [64] and starBase [65], have also been developed. These databases can be obtained by the URLs in Table **1**.

2.2. Published Benchmark Datasets

A benchmark dataset is a collection of data after processing constructed by researchers for different purposes. The data of the benchmark data set may come directly from the relevant database or from the latest published papers where the data have not been included in the databases. Even some researchers perform experiments by themselves to obtain the corresponding data for constructing an objective and strict benchmark dataset. Constructing a useful benchmark dataset is very important because it will significantly affect the performance of the predictor. In most of the existing pre-miRNAs identification work, the researchers obtained experimentally validated pre-miRNAs from a published database and then constructed a positive sample after slight processing. However, so far there is no golden rule for constructing negative samples. A classical method for construct-

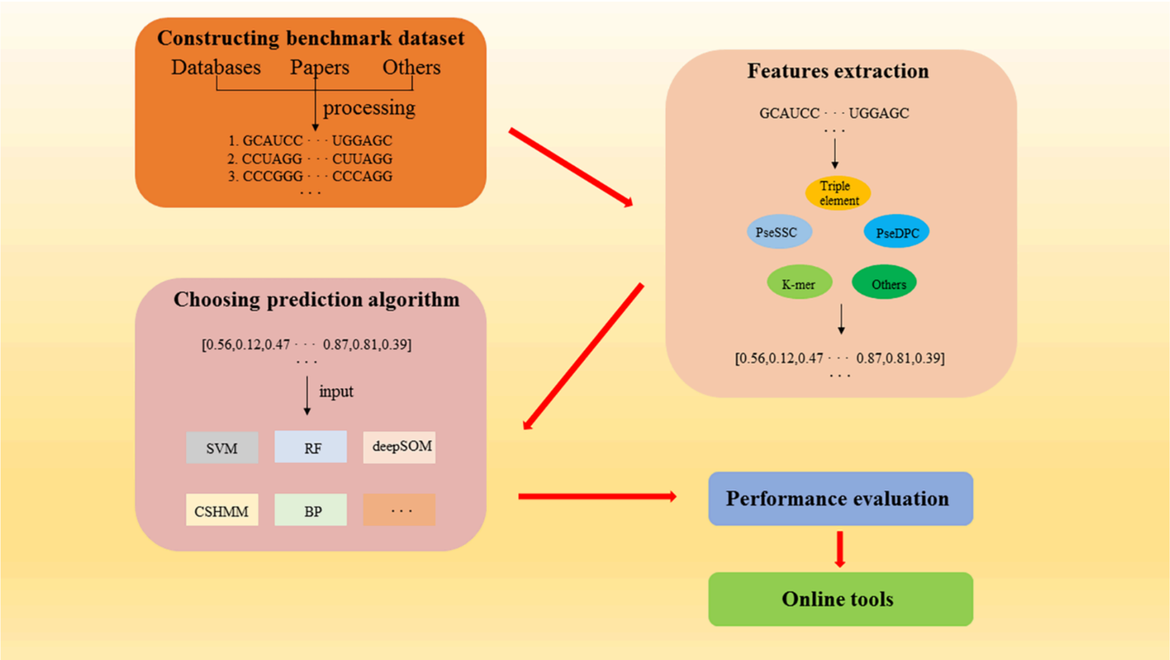

Fig. (2). The flow diagram for the pre-miRNA identification. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 1. URLs for miRNA-related databases.

Database	URL	References
miRBase	http://www.mirbase.org/	[62]
PMRD	http://bioinformatics.cau.edu.cn/PMRD/	[63]
The microRNA Registry	http://www.sanger.ac.uk/Software/Rfam/mirna/	[61]
miRTarBase	http://mirtarbase.mbc.nctu.edu.tw/php/index.php	[64]
starBase	http://starbase.sysu.edu.cn/	[65]

ing pre-miRNA and pseudo pre-miRNA benchmark datasets was proposed by Xue *et al.* [33]. They downloaded 207 pre-miRNAs in *Homo sapiens* from the miRNA registry database, and then removed those sequences with multiple loops in the secondary structure, resulting in 193 positive samples. The negative samples were extracted from the protein coding sequences (CDSs). Finally, they got 8494 pseudo pre-miRNAs as negative samples. The negative samples constructed by Xue *et al.* were used in subsequent works [31, 32, 34, 38, 40, 51], although the positive samples were different. Moreover, Rukshan *et al.* believe that a pre-miRNA classifier must not only be able to distinguish between real pre-miRNAs and pseudo-pre-miRNAs but also have the capability to identify pre-miRNAs from other ncRNAs. Therefore, their negative samples contain Xue's negative samples and some human other ncRNAs [31]. In machine learning-based pre-miRNA identification works, a good negative sample dataset is very similar to the positive sample dataset, because the predictor based on such data will have a useful generalization capability. Just like the criteria proposed by Xue *et al.*, the fundamental purpose is to make the obtained pseudo pre-miRNAs more similar to real pre-miRNAs. But Zou *et al.* believed that negative samples obtained by means of the criterion filtering are not sufficiently similar to positive samples, therefore, they designed a workflow to obtain high quality negative samples [37]. James *et al.* also presented a framework to improve miRNA prediction in non-human genomes [66]. In order to avoid the problem derived from constructing negative samples, Stegmayer *et al.* used a cluster-based approach to achieve pre-miRNA identification [45]. Unlike traditional binary classifiers, which need to construct positive and negative samples, they only have to construct a positive dataset and get some unlabeled sequences. In addition to some of the above datasets, there are some other datasets that were constructed for different purposes. Some datasets constructed by previous works are given in Table **2**.

The benchmark dataset is the fundamental of a prediction model. If a benchmark dataset is of poor quality, the predictors built on such a dataset will not have good results when tested, even lead to bias results. Therefore, researchers are very concerned about the construction of benchmark datasets. Generally, the construction of positive samples is very simple, but constructing negative samples is difficult because there are few or even no experimental-confirmed negative data. In pre-miRNA prediction, some methods have been proposed for constructing negative samples. Other works use labeled and unlabeled data instead of positive and negative samples, which is a remarkable improvement because it avoids problems false positives in negative sample dataset. It is important to construct a more objective and strict negative dataset.

3. FEATURE EXTRACTION

Formulating samples with a vector is a key step for classification [67-83]. It is well known that a pre-miRNA sequence is a string of characters consisting of A, U, C, and G, which cannot be directly calculated *via* computer. Therefore, we need to use the feature extraction method to transform the string into a vector that can be computed *via* computer. The obtained feature vector should reflect the original nature of the sequence as more as possible. In the pre-miRNA identification works, the extracted features can be classified into the following three aspects: (1) sequence-structure-based features; (2) primary-sequence-based features; (3) physics structure-based features. Here we will introduce these features and the methods.

3.1. Sequence-Structure Based Features

Sequence-structure based features consider their secondary structure information (*i.e.* the pairing of each nucleobase). The secondary structure of RNA is usually available through off-the-shelf prediction tools, such as RNAfold [84, 85].

3.1.1. The Triplet Structure-Sequence Elements

Since the distribution of local contiguous subsequences of real pre-miRNAs was observed to be different from pseudo pre-miRNAs, this property is proposed to recognize real pre-miRNAs [33]. In the secondary structure predicted by RNAfold, only two types of nucleotides, namely paired or unpaired occur. The paired nucleotides are replaced by brackets ("(" or ")"), and the unpaired nucleotides are replaced by points ("."). The difference between the left and right brackets is not distinguished. Therefore, when considering the pairing of any 3 consecutive nucleotides, there are $2\times2\times2 = 8$ combinations. In addition, when considering the middle of the three nucleotides, a total of $4 \times 8 = 32$ possible structure-sequence combinations will be obtained. Therefore, an RNA sequence can be converted to a 32-dimensional vector by calculating the frequencies of the 32 components.

Table 2. Published benchmark datasets.

Dataset	Positive	Negative	Species
D1 [33]	193	8494	*H. sapiens*
D2 [34]	426	8494	*H. sapiens*
D3 [32]	323	8494	*H. sapiens*
D4 [31]	691	9248	*H. sapiens*
D5 [36]	1906	2122	Plants
D6 [44]	1406	81228	*H. sapiens*
	231	28359	*A. thaliana*
	7053	218154	Animals
	2172	114929	Plants
	237	839	Virus
	691	9248	From D4
D7 [37]	16520	14661	Not detailed
D8 [38]	1612	8489	*H. sapiens*
D9 [45]	1406	81228	*H. sapiens*
	231	28359	*A. thaliana*
	7053	218154	Animals
	2172	114929	Plants
D10 [39]	3044	5186	Plants

3.1.2. Pseudo Structure Status Composition (PseSSC)

Stimulated by the PseAAC [86-89] in computational proteomics [90, 91] and PseKNC [92-94] in computational genomics, Liu *et al.* proposed the pseudo structure status composition (PseSSC) [38] to increase the prediction performance. The core idea of PseSSC can be explained in two parts. First, the frequency of *n* adjacent status is calculated to reflect the short-range correlation information of the sequence. Then the global structure-order information generated by the interaction between status pairs is captured through embedding a series of correlation factors. The details can be found from [38].

3.1.3. Pseudo Distance Structure Status Pair Composition (PseDPC)

In order to capture the distance-related structure status information for the RNA sequence, Liu *et al.* introduced a new concept called "distance structure status pair" or just "distance-pair" [40]. The main idea of PseDPC is similar to PseSSC. The difference between the two is that when calculating the frequency of status, PseDPC takes the frequency of two non-adjacent state sites into account, while PseSSC considers the frequency of occurrence of the n adjacent status site. Details about the method can be obtained from [40].

3.2. Primary-Sequence Based Features

Primary-sequence based features are those obtained directly from the primary sequence without regard to the pairing of each nucleobase.

3.2.1. K-mer Sequence Component

The K-mer sequence component has achieved great success in representing RNA sequences, and it has been applied to lots of work of pre-miRNA identification [32, 37, 41, 51, 95] and other bioinformatics prediction. For a given sequence R, nucleotide frequencies % $x_1x_2\cdots x_k$ are computed, where $x_1x_2\cdots x_k$ represents any adjacent k nucleobases in R, and $x_i \in \{\mathrm{A,U,C,G}\}(i = 1{,}2,\cdots,k)$. Therefore, we will get 4^k possible combinations in total. By calculating the frequency of each combination in R, we can get a 4^k-dimensional vector. The most important point of using the K-mer method is to determine the value of *k*. If the value of *k* is too large, then the resulting feature vector dimension will be too high, which will cause a high dimensional disaster and thus reduce the predictive performance of the model [95]. Therefore, the maximum value of *k* is set 6 in most of cases.

3.2.2. One-Hot Encoding

One-hot encoding is a feature extraction method widely used in the field of deep learning. An RNA sequence can consist of four different bases: A, U, C, and G. Through one-hot encoding, each base in the RNA sequence can be represented by a 4-dimensional vector, such as A: [1,0,0,0], U: [0,1,0,0], C [0,0,1,0], G [0,0,0,1]. In this way, an RNA of length L can be represented by a 4×L-dimensional matrix.

3.2.3. Primary Sequence Features Based on Mutual Information (PSFMI)

PSFMI is a method proposed by Fu *et al.* for extracting sequence information between bases [43]. In pre-miRNAs, the continuous subsequence of length *n* can be represented by *n*-grams. When *n* = 2, four bases can produce ten different combinations without considering the order of the two bases. Therefore, by calculating the interdependence between two bases on a given pre-miRNA sequence, 10-dimensional features about mutual information (MI) can be obtained.

The pre-miRNA structural information predicted by RNAfold software from sequences can be used as features by machine-learning algorithms. Therefore, Fu *et al.* proposed SSFMI (secondary structure features based on mutual information), which only had a slight modification on the basis of PSFMI, to extract secondary structure features that are based on mutual information [43]. An RNA sequence consisting of 4 bases (A, U, C, G) can be predicted by RNAfold software to obtain a sequence that reflects its structure. This structure-related sequence contains three symbols (left bracket "(", right bracket "(" and point "."). The calculation of PSFMI is based on n-grams composed of 4 types of bases. The slight difference between SSFMI and PSFMI is that n-grams are composed of the above three symbols when calculating SSFMI. Therefore, it is easy to know that there are 6 different combinations for 2-grams and 10 different combinations for 3-grams when calculating SSFMI. The detailed calculation formula for SSFMI is the same as PSFMI. The above-mentioned method is discussed in [43].

3.3. Physics Structure-Based Features

Physics structure-based features are those features that describe some properties of pre-miRNA structures, such as the minimum free energy (MFE) and different base pairing, *etc.*

Since pre-miRNAs have a lower folding free energy than random sequences [96], unlike tRNAs and rRNAs, the performance of the final predictive model can be improved by embedding the MFE feature of the secondary structure of the pre-miRNA. Zu *et al.* used the Vienna RNA software package to predict the MFE of the pre-miRNA secondary structure and used it as a one-dimensional feature vector [34]. Kwang *et al.* also used adjusted MFE, MFEI1 and MFEI2 as part of the feature vector [32]. In addition, MFE is also used by researchers in a variety of different forms [31, 36].

In a pre-miRNA sequence, the stem-loop structure is a very important characteristic. Pre-miRNA base pairing includes not only traditional Watson-Crick pairing but also G-U pairing. Therefore, pairing information is often used as a feature vector to describe pre-miRNA.

Zou *et al.* included G-U pairing information in their extracted feature vectors [41]. To better represent pre-miRNAs, Rukshan *et al.* also added various types of pairing information, such as the number of base pairs in the secondary structure and average base pairs per stem, *etc.*, to their newly introduced features [31].

In addition to the above two types of features, in order to achieve sound predictive effects, some other properties of pre-miRNA, such as base pair distance, degree of compactness [32], structure entropy, structure enthalpy, melting energy [31], the p-value of randomization test [34], *etc.*, are also used by researchers.

As mentioned above, the ultimate goal of feature extraction method is to generate features that can reflect the intrinsic properties of the pre-miRNA. For feature extraction methods based on primary sequences, they are simple and intuitive. However, it is difficult to reflect the structural properties of pre-miRNA, which are different from pseudo pre-miRNA. Therefore, sequence structure-based features are often used to describe pre-miRNA samples. However, the sequence structure-based features are calculated by using prediction software. If the performance of a prediction software is not very good, it is not reliable to use such information to perform prediction. Consequently, new feature extraction methods that could reflect the inherent nature of pre-miRNA sequences without noise are urgently needed.

4. PREDICTION ALGORITHMS

In the identification of pre-miRNA using machine learning methods, the third step is to choose a suitable prediction algorithm. Here, we will briefly introduce these algorithms in pre-miRNA prediction.

4.1. Support Vector Machine (SVM)

SVM is a supervised learning algorithm introduced by Vapnik for the first time [97]. The core idea of SVM is to map low-dimensional indivisible data to high-dimensional Hilbert space by kernel function and maximize the interval between two categories to achieve the classification effect. SVM can only handle a fixed dimension of the input vector. The samples can be converted into fixed-dimensional feature vectors by the above feature extraction methods. The feature vectors used for training will be input into SVM to determine the classification hyperplane in high-dimensional space. Later new samples will be directed into that space and their categories will be determined by their position relative to the hyperplane. SVM will have good results in small sample datasets and has been well applied in many bioinformatics research [98-104]. For ease of use, the LIBSVM was developed [105]. The most commonly used kernel function in SVM is Radial Basis Function (RBF). The grid tool provided by LIBSVM was used to determine the values of kernel function parameter *g* and regularization penalty parameter *C*.

4.2. Random Forest (RF)

Random forest (RF) is a classifier consisting of an ensemble of tree-structured classifiers [106]. The essence of RF is integrated learning, and its basic unit is the decision tree. Based on the decision tree as the base learner to build the bagging [107] integration, RF introduces the random attribute selection in the training process of the decision tree. For each tree, the training set they use is sampled from the total training set using the sampling with replacement method. When training the nodes of each tree, the features used are randomly selected from all features in a certain proportion without replacement. For RF, its each decision tree will produce a classification result and the final output of RF is generated by voting all the results.

RF is simple and easy to implement, but surprisingly, it exhibits powerful performance in many real-world tasks [108-112]. It performs well on large data sets and is able to assess the importance of various features in the classification work. When the classified dataset is an unbalanced dataset, the RF can balance the error. And for estimating missing data and maintaining accuracy when a large proportion of the data was missed, the algorithm is still effective [113].

4.3. Back Propagation (BP) Neural Network

BP neural network is a multi-layer feedforward neural network. Its main features are: the signal is forward propagating, and the error is back propagating. The basic structure of a neural network is an input layer-intermediate layer (hidden layer)-output layer, and its basic constituent unit is the neuron. The number of neurons in input layer is same as the dimension of the input data, and the number of neurons in the output layer is same as the number of the data to be fitted. The number of neurons in the hidden layer and the number of layers needs to be set by the designer according to some rules and objectives. The goal of the BP algorithm is to minimize the cumulative error on the training set.

In the forward propagation process, the input mode in the input layer reaches the output layer after passing through multiple hidden layers. The neuron state of each layer is only affected by the neurons connected to it in the previous layer. If the desired output cannot be obtained at the output layer, then the backward propagation is performed, and the error signal is returned along the original connection path. By modifying the weights of the neurons, the error signal decreases along the direction of the gradient descent and eventually reaches a certain threshold.

4.4. Self-Organizing Maps (SOM)

SOM is an unsupervised clustering method proposed by Kohonen in 1982 [114]. The basic idea of SOM is that each neuron in the network output layer gains a chance to respond to the input layer through competition, and finally only one neuron wins. The effect of the winning neuron on its neighboring neurons is from near to far, from excitement to inhibition, and the connection weights associated with the winning neurons are transformed in a direction that favors its competition. The goal of SOM is to represent complex high-dimensional input patterns into a simpler low-dimensional discrete map, with prototype vectors that can be visualized in a two-dimensional lattice structure, while preserving the proximity relationships of the original data as much as possible [45]. Since SOMs have the capability of identifying similar input patterns in the feature space, by assigning them to the same neuron or a group of adjacent neurons on the map [115], it can be used in the prediction work of miRNA precursors. Based on SOM, Georgina proposes a hierarchy of SOM in deep levels (deepSOM) to deal with high class-imbalance problems [45]. The deepSOM was trained with some labeled positive and unlabeled data. In deepSOM, those neurons with well-known labeled data are defined as miRNA neurons. During training, only sequences clustered in miRNA neurons remain for further training the next level of deepSOM. After training several nested SOM, the best pre-miRNAs can be identified as the ones that remain close to the prototypes of the miRNA neurons in the last deep level. More details about deepSOM are mentioned in [45].

4.5. Context-Sensitive Hidden Markov Model (CSHMM)

The hidden Markov model (HMM) is efficient in modeling short-term dependencies between adjacent samples. However, when symbols are distant from each other, HMM does not work well. Because some RNAs form a hairpin structure, the regular HMM is not very effective for them. Therefore, Byung *et al.* introduced the concept of CSHMM that is capable of modeling strong pairwise correlations between distant symbols [116]. The CSHMM model consists of three kinds of states (namely S_n, P_n and C_n). S_n are single-emission states, which are exactly the same as the ones used in traditional HMMs. P_n are pairwise-emission states and C_n are context-sensitive states. These two states always exist in pairs. On the basis of the concept of CSHMM, Ashwin made a slight adjustment and applied it to the prediction of pre-miRNA [35]. A complete CSHMM contains the structure and probability of moving from one state to another (emission and transition probabilities). The CSHMM structure proposed by them has two context sensitive states which are linked to the same pairwise-emission state through a stack. The overall structure of their CSHMM and computing method of transition probabilities are described in detail in [35].

4.6. Convolutional Neural Network (CNN)

Convolutional neural networks (CNNs), originally invented for computer vision, can automatically extract features by filters/kernels. CNNs have already proven to be successful for image classification and many natural language processing (NLP) tasks and achieved good results. Therefore, more and more researchers apply CNN to the field of bioinformatics [42, 117]. Convolutional neural networks usually consist of an input layer, convolutional layers, activation functions, pooling layers, and fully connected layers. The input layer is where the data is input. In the prediction work of pre-miRNA, the features extracted from the RNA sequence will be input in the form of the matrix. Convolutional layers are the core cornerstone of convolutional neural networks. Each convolutional layer consists of several convolutional units, and the parameters of each convolutional unit are optimized through back-propagation algorithms. Local connections and weight sharing are two major characteristics of convolutional layers. The purpose of the convolution operation is to further extract the input features to obtain more informative features. The result calculated by the convolution layer needs to be transformed by the activation function. The activation function is very important. It can bring non-linear results, and non-linear can make us fit various functions well. The pooling layer is usually after the convolution layer, which is mainly used for feature dimension reduction, compressing the number of data and parameters, reducing overfitting, and improving the fault tolerance of the model. The pooling layer contains two types: maximum pooling and average pooling. The fully connected layer is after the last pooling layer. The purpose of the fully connected layer is to calculate the data obtained by the previous convolution and pooling operations to obtain the result. The neurons between two adjacent layers of the fully connected layer are usually fully connected, but in order to avoid overfitting, the dropout operation is usually performed. In the last layer of the fully connected layer, we can get the output. The framework of a convolutional neural network model (the number of convolutional layers, pooling layers, and fully connected layers) is set according to different problems.

This section introduces several prediction algorithms used in pre-miRNA prediction work. The SVM and RF are the two most commonly used methods because their performance is always better on smaller sample data. Most of pre-miRNA prediction works are based on small datasets. Both the CSHMM and BP neural networks have achieved good results in the prediction of pre-miRNA. The deepSOM based on SOM is an excellent method on the basis of clustering ideas, because it not only solves the problem of data imbalance, but also avoids the problems that may be encountered in constructing negative samples. The CNN is a good attempt to apply deep learning to this field. As the amount of data increases, this method should be taken advantage of.

5. PERFORMANCE EVALUATION

The 5-fold or 10-fold cross-validation and a jackknife test [118-123] are always used to evaluate the performance of predictors. The performance evaluation metrics commonly used in pre-miRNA prediction work are sensitivity (*Sn*); specificity (*Sp*); accuracy (*Acc*); Matthew's correlation coefficient (*MCC*); Geometric mean (*Gm*). In addition to the above metrics, precision (*Pr*), recall (*Re*) (*i.e.* sensitivity), *F1* score and predictor calculation time (*T*) are also used to evaluate the performance of the predictor.

The receiver operating characteristic (ROC) curve [124-126] is also often used to evaluate the performance of the constructed model, which can intuitively reflect the trend of the performance when selecting different thresholds.

6. PUBLISHED RESULTS

Based on benchmark dataset D1, Xue *et al.* proposed the triplet structure-sequence elements for feature extraction [33]. SVM-based model could achieve an overall accuracy of 90.7% on three independent human test sets. When the model (called triplet-SVM) was used for a cross-species test set containing 11 species, it achieved an overall accuracy of 90.9% [33]. In order to improve the prediction accuracy of human pre-miRNAs, Peng *et al.* combined the triplet structure-sequence elements, MFE, and P-value to represent RNA sequence [34]. After comparing SVM with RF, they chose RF as the prediction algorithm. The RF-model (called MiPred) obtained accuracy of 96.68% with MCC of 0.94 based on benchmark dataset D2. It can be seen from the results that the improvement of accuracy is obvious, which implies that the performance of the model can be improved by embedding the physics structure features and selecting the appropriate prediction algorithm.

Based on benchmark dataset D3, Kwang *et al.* encoded the RNA sequence with a 29-dimensional vector, which could reflect the RNA global and intrinsic folding attributes [32]. They developed a classifier called miPred using SVM algorithm [32]. The miPred achieved the accuracy of 93.5% on an independent human test set. In order to verify the generalization capability of the miPred, a non-human independent test set, functional ncRNAs and mRNAs were used to evaluate the performance of the miPred [32]. The miPred reported an overall accuracy of 95.64% for the test set, a mean Sp of 76.15% for functional ncRNAs and 87.1% for mRNAs.

Since the sequences in benchmark dataset D4 contained both hairpin secondary structures and structures having multi-branched loops, Rukshan used 48 features (including 29 features from [32] and 19 new physics structure features) to construct the feature vector [31]. They used Jeffries–Matusita distance (J-M) [127] for feature selection and SVM as prediction algorithm. After using SMOTE technique [128] to handle data imbalanced problem, the classifier (called microPred) obtained sensitivity of 90.02%, specificity of 97.28% and Gm of 93.58% in the 5-fold cross-validation (CV) on the basis of benchmark dataset D4 [31]. When tested with 6095 non-human animal pre-miRNAs and 139 virus pre-miRNAs, microPred could produce 92.71% (5651/6095) and 94.24% (131/139) recognition rates, respectively [31].

In order to obtain efficient prediction on plant pre-miRNAs, Ping *et al.* constructed the benchmark dataset D5 [36]. They designed a training sample selection method (miSampleSelection) that selected the positive/negative training samples according to the sample distribution in the positive families/negative groups. By using the strategy, they eventually got 960 positive samples and 960 negative samples from benchmark dataset D5. In order to encode these samples, 115 features (including 48 features from [31], 32 features from [33], 32 structured triplet composition features from stems, and 3-dimensional new physics structure features) were extracted initially. After considering the effects of information gain and feature similarity (IG-FS), they selected 68 features from the 115 features to represent RNA. By combining these 68 features with the SVM algorithm, they built a classifier called PlantMiRNAPred. The classifier achieved >90% accuracy on an independent plant test set from eight plant species [36].

On the basis of benchmark dataset D6, Adma developed a novel method for dealing with the class imbalance problem called ROC-select [44], which was based on a threshold score function produced by traditional classifiers. They used 21 features selected from [31] and another 7 sequence-related and structure-related features to encode RNA. By comparison with multiple algorithms, RF was chosen as the classification algorithm. The RF-model (called HuntMi) had good performance for predicting new microRNA in human, *Arabidopsis*, animals, plants and viruses [44].

Based on dataset D7, Zou selected 1155 human pre-miRNAs and 1155 pseudo human pre-miRNAs [37]. They used 98-dimensional feature vector (64-dimensional 3-mer sequence composition, 32-dimensional triplet structure-sequence elements, 2-dimensional physics structure features) to represent RNA. Finally, an online system called miRNApre was specifically developed for human pre-miRNA identification using SVM. In a 10-fold CV, the miRNApre got the Gm of 98% and accuracy of 98.1%. Based on miRNApre, they developed a miRNA mining tool called mirnaDetect, which can be applied to find potential miRNAs in genome-scale data [37].

Based on the benchmark dataset D8, a total of three work was carried out. First of all, in 2015, Liu *et al.* used 1612 human pre-miRNAs and 1612 pseudo pre-miRNAs from D8 as the training set [38]. None of the sequences included ≥80% pairwise sequence identity with any other. They used PseSSC and the augmented PseSSC (ExPseSSC) as feature extraction methods to construct two SVM-classifiers: iMcRNA-PseSSC and iMcRNA-ExPseSSC. The jackknife test was used to evaluate the performance of the two classifiers. The accuracy achieved by iMcRNA-PseSSC was 85.76% with the MCC equal to 0.72. The corresponding results achieved by iMcRNA-ExPseSSC were 89.86% and 0.80 for accuracy and MCC, respectively [38]. Then, in 2016, using the same training set as [38], Liu *et al*. combined PseDPC with SVM to construct the classifier iMiRNA-PseDPC [40]. The iMiRNA-PseDPC obtained the accuracy of 87.69% with MCC of 0.75 in the jackknife test. Finally, in 2017, in order to better reflect the original nature of RNA sequences, Khan *et al.* used a set of hybrid features (including the triplet structure-sequence elements features, PseDPC features and 2,3,4,5,6-mer features) to encode sequences [51]. In order to avoid high dimensional disasters, the partial least squares (PLS) technique [129] was used to select features. Based on SVM and RF algorithm, they proposed the prediction model called MicroR-Pred, which achieved the accuracy of 88.4% for RF-model and 93.9% for SVM-model in the jackknife test.

Stegmayer constructed dataset D9 for dealing with high class-imbalance problem in pre-miRNA prediction [45]. Most of the samples in the benchmark dataset D9 are unlabeled sequences, while the labeled sequences (*i.e.*, positive samples)

Table 3. Pre-miRNA predictors methods.

Predictor	Feature			Model Type	Years
	Type	Number	Selection		
triplet-SVM [33]	s-s	32	None	SVM	2005
miPred [32]	se, st	29	None	SVM	2007
MiPred [34]	st, s-s	34	None	RF	2007
microPred [31]	se, st	21	J-M	SVM	2009
Ashwin *et al.* [35]	Not detailed	Not detailed	None	CSHMM	2010
PlantMiRNAPred [36]	se, st, s-s	68	IG-FS	SVM	2011
HuntMi [44]	se, st, s-s	28	None	RF	2013
miRNApre [37]	se, st, s-s	98	None	SVM	2014
iMcRNA-PseSSC / iMcRNA-ExPseSSC [38]	se, st, s-s	113 / 93	None	SVM	2015
plantMirP [39]	se, s-s	Not detailed	None	SVM	2016
iMiRNA-PseDPC [40]	s-s	725	None	SVM	2016
miRNAFold [47]	se, st, s-s	55	None	Features-based search	2016
deepSOM [45]	se, st	29	None	deepSOM	2016
Zou *et al.* [41]	se, st, s-s	98	None	BP Neural Network	2016
MicroR-Pred [51]	se, s-s	40	PLS	SVM	2017
miRNAss [46]	se, st, s-s	Not detailed	None	Semi-supervised	2018
CNN-filter6-128 [42]	se	656	None	CNN	2019
Fu *et al.* [43]	se, st, s-s	55	None	SVM	2019

Note: Feature type includes sequence-structure features (s-s), primary sequence features (se), physics structure features (st); Feature number refers to the number of features that were ultimately used to build the model.

account for only a small part. They proposed the deepSOM to overcome this problem. They used features from [44] to encode RNA sequences and 10-fold CV for evaluating model performance. The deepSOM had an excellent performance in predicting novel pre-miRNAs in many species of animals and plants [45].

Based on dataset D10, Yao *et al.* used knowledge-based energy features to formulate the RNA sequence [39]. They inputted these features into the SVM and built a predictor called plantMirp. Compared to miPlantPreMat [49] and PlantMiRNAPred [36], the plantMirP has had better results in plant pre-miRNA prediction, which achieved a promising sensitivity of 92.61% and a specificity of 98.88%.

In addition to the above pre-miRNA identification work, Ashwin used CSHMM for predicting miRNA sequences and their classifier showed a sensitivity of about 85% with a specificity of about 97-98% on human miRNA sequences [35]. Zou employed BP neural network together with 98-dimensional features for human microRNA precursor identification and got a precision of 95.53% and recall of 96.67% [41]. The BP method had also achieved good results in multiple species. In 2016, Fariza *et al.* presented a web server dedicated to miRNA precursors identification at a large scale in genomes [47], which was based on an algorithm called miRNAFold. The miRNAFold algorithm was based on some of the criteria (12 criteria for the longest exact stem, 17 criteria for the longest non-exact stem and 26 criteria for the hairpin) that observed from the pre-miRNA structure. Only the sequence in the sliding window satisfied a certain percentage of the criteria can be predicted as a possible pre-miRNA [130]. To enable efficient and speedy genome-wide predictions of novel miRNAs, a semi-supervised learning method was proposed [46]. MiRNAss was tested with the genome-wide data of *A. thaliana*, *Caenorhabditis elegans* and *Anopheles gambiae*, and it reported Gm of 84.82%, 87.61% and 93.34% respectively [46]. In 2019, Zheng *et al.* combined the one-hot encoding method with CNN to build a prediction model for human pre-miRNA, and it achieved an accuracy of 0.92 [42]. Fu *et al.* extracted sequence and structural features that are based on mutual information from RNA sequences to distinguish between pre-miRNA and pseudo pre-miRNA [43]. These features were applied to train a SVM model, which produced an exhilarating effect on human pre-miRNA predictions [43].

The methods used in the above model are shown in Table **3**. It is necessary to build an online tool based on the model because it can avoid complicated mathematical calculations when people use the model. For easy access, Table **4** gives the

Table 4. Availability of pre-miRNA predictors.

Predictor	Type	URL	Species	Pre-miRNA *vs* ncRNA (Yes/No)	Prediction from Genome (Yes/No)
Triplet-SVM	Package	http://bioinfo.au.tsinghua.edu.cn/mirnasvm/	*H. sapiens*	No	No
miPred	Script	https://web.bii.a-star.edu.sg/archive/stanley/ Publications/Supp_materials/06-002-supp.html	*H. sapiens*	Yes	No
PlantMiRNAPred	Web server	http://nclab.hit.edu.cn/PlantMiRNAPred/	Plants	No	No
PlantMirP	Script	https://github.com/yygen89/plantMirP	**Plants**	No	No
HuntMi	Package	http://adaa.polsl.pl/agudys/huntmi/huntmi.htm	*H. sapiens*, *A. thaliana*, animals, plants, virus	Yes	No
iMcRNA-PseSSC / iMcRNA-ExPseSSC	Web server	http://bioinformatics.hitsz.edu.cn/iMcRNA/	*H. sapiens*	No	No
iMiRNA-PseDPC	Web server	http://bioinformatics.hitsz.edu.cn/iMiRNA-PseDPC/	***H. sapiens***	No	No
miRNAFold	Web server	https://evryrna.ibisc.univ-evry.fr/miRNAFold	**Any**	No	**Yes**
deepSOM	Web server	http://fich.unl.edu.ar/sinc/web-demo/deepsom/	***H. sapiens*,** ***A. thaliana*, animals, plants**	Yes	No
miRNAss	Package	https://sourceforge.net/projects/sourcesinc/files/mirnass/	**Any**	No	No
CNN-filter6-128	Script	https://github.com/zhengxueming/cnnMirtronPred	*H. sapiens*	No	No

Note: Predictors in bold are the predictors recommended in this review and their main applicable species.

tools that are currently available and gives a brief introduction to their usefulness.

As mentioned above, there are still many pre-miRNA prediction tools available, and each of these tools has its own focus, which makes it difficult to say which one is better than another. So here, we give some suggestions for using tools to predict pre-miRNA as follows. For human pre-miRNA prediction, the method proposed by Fu *et al.* has achieved good results, but it does not provide a published tool, which makes it very inconvenient to use [43]. The iMiRNA-PseDPC [40] is recommended because it not only has good results but also has a user-friendly web server. For plant pre-miRNA prediction, the plantMirP has achieved good results so far [39]. For multi-species pre-miRNA prediction, both the deepSOM [45] and miRNAss [46] are recommended. The deepSOM is more suitable for new scholars because it has a web server. However, the miRNAss has a better performance on genome-wide pre-miRNA prediction., The potential of pre-miRNAs can be predicted directly from the whole genome, the miRNAFold is a right choice [47]. Users can choose the appropriate predictor according to their purposes. The relevant information can be obtained from Table **4**.

CONCLUSION AND PERSPECTIVES

Identification of miRNAs is the first step toward understanding their biological characteristics. In this article, we reviewed the work on pre-miRNA identification using machine learning and summarized the benchmark dataset, feature extraction method, prediction algorithm, and the results of models.

A good benchmark dataset is essential. We found that in the current pre-miRNA identification work, positive samples are usually from real pre-miRNAs in the database, while negative samples are constructed in many ways. But they all have a common core idea that is to make the constructed negative samples more similar to the positive samples, which will make the constructed models have better generalization capability. In other methods, namely non-binary classifiers, to identify pre-miRNAs, they usually use a set of unlabeled data instead of negative samples to avoid the problems that may arise from constructing negative samples.

It is very important to extract features from RNA sequences that reflect their original natures [131]. In this paper, we divide these features into three aspects: (1) sequence-structure based features; (2) primary-sequence based features; (3) physics structure-based features. In the sequence-structure based features, the triplet structure-sequence elements were first proposed, which can well reflect the local continuous sequence-structural properties of RNA. The PseSSC and PseDPC can reflect global or long-range sequence-structure information. The above three features can be extracted from RNA through a web server tool called repRNA [132]. In the primary-sequence based features, we introduced the K-mer composition, which reflects short-range information of RNA primary sequences and is widely used in many bioinformatics works. Some of the properties extracted from RNA were also generated into physics structure-based features which may be used to distinguish between pre-miRNAs and pseudo pre-miRNAs. In addition, hybrid features were also used by researchers. When features were extracted from RNA sequences, these features may be of high dimensionality, resulting in high dimensional disasters, or high redundancy between features that can affect predictor performance. Therefore, after extracting features, it is necessary to make a feature selection. Currently, features selection methods, such as Jeffries–Matusita distance, information gain and feature similarity, the partial least squares, have been used in pre-miRNA identification. Other methods, such as minimal-redundancy-maximal-relevance (mRMR) [133], F-score [134], analysis of variance (ANOVA) [70], should be considered in the future.

Choosing a suitable prediction algorithm will ensure that the model has a good prediction accuracy. SVM and RF are the most commonly used prediction algorithms in the current pre-miRNA identification work. The SVM produces a better effect on small sample data. In addition, BP neural networks, CSHMM, deepSOM and CNN are also used to identify pre-miRNAs.

Many predictors have achieved good results. However, most of the predictor's training sets are balanced or low unbalanced samples from the benchmark dataset. This is very inconsistent with the actual situation. The number of real pre-miRNAs is very small compared with the number of pseudo pre-miRNAs in realistic data. If we used the particularly unbalanced data as a training set, the proposed model will tend the category with large samples when making predictions, which will seriously affect the performance of the model. Some strategies have been proposed to deal with this problem. Rukshan used the SMOTE method [31]. Adam developed a method called ROC-select [44]. Yones *et al.* built a depth model to automate the problem [45]. In the future, we hope that more methods and models can be developed to solve this problem.

Compared with previous reviews, this paper has the following differences [135]. In this review, we have described the benchmark dataset, feature extraction method, prediction algorithm, and model performance for pre-miRNA prediction. Through the review, we believe that even a new student can have a preliminary understanding of the work in this field. New progress in the field was included. The review will help researchers understand the progress of work in this field. The datasets given in this article will be helpful to researchers. Besides, the available URLs and recommended usages given in Table **4** will be convenient for researchers who want to use the pre-miRNA predictor.

Nowadays, researchers have increasingly focused on the identification of pre-miRNAs on more species. Some predictors have been able to directly predict possible pre-miRNAs from genomes or could act on wide-genome data. With the accumulation of data, more and more predictors will be constructed for the identification of pre-miRNA in different species. At the same time, we hope that more feature extraction methods will be developed in the future. Besides, we can also consider applying deep learning [136-140] to this field.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Ambros, V. The functions of animal microRNAs. *Nature,* **2004**, *431*(7006), 350-355.
http://dx.doi.org/10.1038/nature02871 PMID: 15372042

[2] Ruvkun, G.; Giusto, J. The *Caenorhabditis elegans* heterochronic gene lin-14 encodes a nuclear protein that forms a temporal developmental switch. *Nature,* **1989**, *338*(6213), 313-319.
http://dx.doi.org/10.1038/338313a0 PMID: 2922060

[3] Reinhart, B.J.; Slack, F.J.; Basson, M.; Pasquinelli, A.E.; Bettinger, J.C.; Rougvie, A.E.; Horvitz, H.R.; Ruvkun, G. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans. Nature,* **2000**, *403*(6772), 901-906.
http://dx.doi.org/10.1038/35002607 PMID: 10706289

[4] Lee, Y.; Kim, M.; Han, J.; Yeom, K.H.; Lee, S.; Baek, S.H.; Kim, V.N. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.,* **2004**, *23*(20), 4051-4060.
http://dx.doi.org/10.1038/sj.emboj.7600385 PMID: 15372072

[5] Lee, Y.; Ahn, C.; Han, J.; Choi, H.; Kim, J.; Yim, J.; Lee, J.; Provost, P.; Rådmark, O.; Kim, S.; Kim, V.N. The nuclear RNase III Drosha initiates microRNA processing. *Nature,* **2003**, *425*(6956), 415-419.
http://dx.doi.org/10.1038/nature01957 PMID: 14508493

[6] Kim, V.N. MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends Cell Biol.,* **2004**, *14*(4), 156-159.
http://dx.doi.org/10.1016/j.tcb.2004.02.006 PMID: 15134074

[7] Bohnsack, M.T.; Czaplinski, K.; Gorlich, D. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA,* **2004**, *10*(2), 185-191.
http://dx.doi.org/10.1261/rna.5167604 PMID: 14730017

[8] Knight, S.W.; Bass, B.L. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans. Science,* **2001**, *293*(5538), 2269-2271.
http://dx.doi.org/10.1126/science.1062039 PMID: 11486053

[9] Gregory, R.I.; Chendrimada, T.P.; Cooch, N.; Shiekhattar, R. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell,* **2005**, *123*(4), 631-640.
http://dx.doi.org/10.1016/j.cell.2005.10.022 PMID: 16271387

[10] Millar, A.A.; Waterhouse, P.M. Plant and animal microRNAs: similarities and differences. *Funct. Integr. Genomics,* **2005**, *5*(3), 129-135.
http://dx.doi.org/10.1007/s10142-005-0145-2 PMID: 15875226

[11] Kittelmann, S.; McGregor, A.P. Modulation and evolution of animal development through microRNA regulation of gene expression. *Genes (Basel),* **2019**, *10*(4), 10.
http://dx.doi.org/10.3390/genes10040321 PMID: 31027314

[12] López-Ruiz, B.A.; Juárez-González, V.T.; Sandoval-Zapotitla, E.; Dinkova, T.D. Development-related miRNA expression and target regulation during staggered *in vitro* plant regeneration of Tuxpeño VS-535 maize cultivar. *Int. J. Mol. Sci.,* **2019**, *20*(9), 20.
http://dx.doi.org/10.3390/ijms20092079 PMID: 31035580

[13] Sun, Y.; Gao, Y.; Song, T.; Yu, C.; Nie, Z.; Wang, X. MicroRNA-15b participates in the development of peripheral arterial disease by modulating the growth of vascular smooth muscle cells. *Exp. Ther. Med.,* **2019**, *18*(1), 77-84.
http://dx.doi.org/10.3892/etm.2019.7552 PMID: 31258640

[14] Xia, M.M.; Shen, X.Y.; Niu, C.M.; Xia, J.; Sun, H.Y.; Zheng, Y. [MicroRNA regulates Sertoli cell proliferation and adhesion]. *Yi*

Chuan, **2018**, *40*(9), 724-732.
PMID: 30369476

[15] Zhang, J.; Xu, Y.; Liu, H.; Pan, Z. MicroRNAs in ovarian follicular atresia and granulosa cell apoptosis. *Reprod. Biol. Endocrinol.,* **2019**, *17*(1), 9.
http://dx.doi.org/10.1186/s12958-018-0450-y PMID: 30630485

[16] Chen, P.; Zhang, H.; Sun, X.; Hu, Y.; Jiang, W.; Liu, Z.; Liu, S.; Zhang, X. microRNA-449a modulates medullary thymic epithelial cell differentiation. *Sci. Rep.,* **2017**, *7*(1), 15915.
http://dx.doi.org/10.1038/s41598-017-16162-2 PMID: 29162901

[17] Chen, Z.; Chu, S.; Wang, X.; Fan, Y.; Zhan, T.; Arbab, A.A.I.; Li, M.; Zhang, H.; Mao, Y.; Loor, J.J.; Yang, Z. MicroRNA-106b regulates milk fat metabolism *via* ATP binding cassette subfamily A member 1 (ABCA1) in bovine mammary epithelial cells. *J. Agric. Food Chem.,* **2019**, *67*(14), 3981-3990.
http://dx.doi.org/10.1021/acs.jafc.9b00622 PMID: 30892026

[18] Liao, Z.; Li, D.; Wang, X. Cancer diagnosis from isomiR expression with machine learning method. *Curr. Bioinform.,* **2018**, *13*, 57-63.
http://dx.doi.org/10.2174/1574893611666160609081155

[19] Tang, W.; Wan, S.; Yang, Z.; Teschendorff, A.E.; Zou, Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics,* **2018**, *34*(3), 398-406.
http://dx.doi.org/10.1093/bioinformatics/btx622 PMID: 29028927

[20] Rupaimoole, R.; Slack, F.J. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.,* **2017**, *16*(3), 203-222.
http://dx.doi.org/10.1038/nrd.2016.246 PMID: 28209991

[21] Xue, J.; Yang, J.; Luo, M.; Cho, W.C.; Liu, X. MicroRNA-targeted therapeutics for lung cancer treatment. *Expert Opin. Drug Discov.,* **2017**, *12*(2), 141-157.
http://dx.doi.org/10.1080/17460441.2017.1263298 PMID: 27866431

[22] Zambrano, T.; Salazar, L.A. microRNAs and response to statins in patients with hypercholesterolemia: from basic research to precision medicine. *Pharmacogenomics,* **2018**, *19*(9), 748-751.
http://dx.doi.org/10.2217/pgs-2018-0051 PMID: 29785870

[23] Cheng, L.; Hu, Y.; Sun, J.; Zhou, M.; Jiang, Q. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics,* **2018**, *34*(11), 1953-1956.
http://dx.doi.org/10.1093/bioinformatics/bty002 PMID: 29365045

[24] Cheng, L.; Sun, J.; Xu, W.; Dong, L.; Hu, Y.; Zhou, M. OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.,* **2016**, *6*, 34820.
http://dx.doi.org/10.1038/srep34820 PMID: 27703231

[25] Zhang, X.; Zou, Q.; Rodriguez-Paton, A.; Zeng, X. Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans. Comput. Biol. Bioinformatics,* **2019**, *16*(1), 283-291.
http://dx.doi.org/10.1109/TCBB.2017.2776280 PMID: 29990255

[26] Lagos-Quintana, M.; Rauhut, R.; Lendeckel, W.; Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science,* **2001**, *294*(5543), 853-858.
http://dx.doi.org/10.1126/science.1064921 PMID: 11679670

[27] Lau, N.C.; Lim, L.P.; Weinstein, E.G.; Bartel, D.P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science,* **2001**, *294*(5543), 858-862.
http://dx.doi.org/10.1126/science.1065062 PMID: 11679671

[28] Ruby, J.G.; Stark, A.; Johnston, W.K.; Kellis, M.; Bartel, D.P.; Lai, E.C. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res.,* **2007**, *17*(12), 1850-1864.
http://dx.doi.org/10.1101/gr.6597907 PMID: 17989254

[29] Lai, E.C.; Tomancak, P.; Williams, R.W.; Rubin, G.M. Computational identification of Drosophila microRNA genes. *Genome Biol.,* **2003**, *4*(7), R42.
http://dx.doi.org/10.1186/gb-2003-4-7-r42 PMID: 12844358

[30] Wang, X.; Zhang, J.; Li, F.; Gu, J.; He, T.; Zhang, X.; Li, Y. MicroRNA identification based on sequence and structure alignment. *Bioinformatics,* **2005**, *21*(18), 3610-3614.
http://dx.doi.org/10.1093/bioinformatics/bti562 PMID: 15994192

[31] Batuwita, R.; Palade, V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics,* **2009**, *25*(8), 989-995.
http://dx.doi.org/10.1093/bioinformatics/btp107 PMID: 19233894

[32] Ng, K.L.; Mishra, S.K. *De novo* SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics,* **2007**, *23*(11), 1321-1330.
http://dx.doi.org/10.1093/bioinformatics/btm026 PMID: 17267435

[33] Xue, C.; Li, F.; He, T.; Liu, G.P.; Li, Y.; Zhang, X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics,* **2005**, *6*, 310.
http://dx.doi.org/10.1186/1471-2105-6-310 PMID: 16381612

[34] Jiang, P.; Wu, H.; Wang, W.; Ma, W.; Sun, X.; Lu, Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.,* **2007**, *35*(Web Server issue), W339-W344.
http://dx.doi.org/10.1093/nar/gkm368 PMID: 17553836

[35] Agarwal, S.; Vaz, C.; Bhattacharya, A.; Srinivasan, A. Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). *BMC Bioinformatics,* **2010**, *11*(Suppl. 1), S29.
http://dx.doi.org/10.1186/1471-2105-11-S1-S29 PMID: 20122201

[36] Xuan, P.; Guo, M.; Liu, X.; Huang, Y.; Li, W.; Huang, Y. PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. *Bioinformatics,* **2011**, *27*(10), 1368-1376.
http://dx.doi.org/10.1093/bioinformatics/btr153 PMID: 21441575

[37] Wei, L; Liao, M; Gao, Y. Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set. *BMC Bioinformatics,* **2014**, *11*(Suppl. 1), S29.
http://dx.doi.org/10.1109/TCBB.2013.146

[38] Liu, B.; Fang, L.; Liu, F.; Wang, X.; Chen, J.; Chou, K.C. Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS One,* **2015**, *10*(3), e0121501.
http://dx.doi.org/10.1371/journal.pone.0121501 PMID: 25821974

[39] Yao, Y.; Ma, C.; Deng, H.; Liu, Q.; Zhang, J.; Yi, M. plantMirP: an efficient computational program for the prediction of plant pre-miRNA by incorporating knowledge-based energy features. *Mol. Biosyst.,* **2016**, *12*(10), 3124-3131.
http://dx.doi.org/10.1039/C6MB00295A PMID: 27472470

[40] Liu, B.; Fang, L.; Liu, F.; Wang, X.; Chou, K.C. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. *J. Biomol. Struct. Dyn.,* **2016**, *34*(1), 223-235.
http://dx.doi.org/10.1080/07391102.2015.1014422 PMID: 25645238

[41] Jiang, L.; Zhang, J.; Xuan, P.; Zou, Q. BP neural network could help improve pre-miRNA identification in various species. *BioMed Res. Int.,* **2016**, *2016,* 9565689.
http://dx.doi.org/10.1155/2016/9565689 PMID: 27635401

[42] Zheng, X.; Xu, S.; Zhang, Y.; Huang, X. Nucleotide-level convolutional neural networks for pre-miRNA classification. *Sci. Rep.,* **2019**, *9*(1), 628.
http://dx.doi.org/10.1038/s41598-018-36946-4 PMID: 30679648

[43] Fu, X.; Zhu, W.; Cai, L.; Liao, B.; Peng, L.; Chen, Y.; Yang, J. Improved pre-miRNAs identification through mutual information of pre-miRNA sequences and structures. *Front. Genet.,* **2019**, *10*, 119.
http://dx.doi.org/10.3389/fgene.2019.00119 PMID: 30858864

[44] Gudyś, A.; Szcześniak, M.W.; Sikora, M.; Makałowska, I. HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC Bioinformatics,* **2013**, *14*, 83.
http://dx.doi.org/10.1186/1471-2105-14-83 PMID: 23497112

[45] Stegmayer, G; Yones, C; Kamenetzky, L. High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM. *IEEE/ACM Trans. Comput. Biol. Bioinform.,* **2017**, *14*, 1316-26.
http://dx.doi.org/10.1109/TCBB.2016.2576459

[46] Yones, C.; Stegmayer, G.; Milone, D.H.; Sahinalp, C. Genome-wide pre-miRNA discovery from few labeled examples. *Bioinformatics,* **2018**, *34*(4), 541-549.
http://dx.doi.org/10.1093/bioinformatics/btx612 PMID: 29028911

[47] Tav, C.; Tempel, S.; Poligny, L.; Tahi, F. miRNAFold: a web server for fast miRNA precursor prediction in genomes. *Nucleic Acids Res.,* **2016**, *44*(W1), W181-W184.
http://dx.doi.org/10.1093/nar/gkw459 PMID: 27242364

[48] Pfeffer, S.; Sewer, A.; Lagos-Quintana, M.; Sheridan, R.; Sander, C.; Grässer, F.A.; van Dyk, L.F.; Ho, C.K.; Shuman, S.; Chien, M.; Russo, J.J.; Ju, J.; Randall, G.; Lindenbach, B.D.; Rice, C.M.; Simon, V.; Ho, D.D.; Zavolan, M.; Tuschl, T. Identification of microRNAs of the herpesvirus family. *Nat. Methods,* **2005**, *2*(4), 269-276.

http://dx.doi.org/10.1038/nmeth746 PMID: 15782219

[49] Meng, J.; Liu, D.; Sun, C.; Luan, Y. Prediction of plant pre-microRNAs and their microRNAs in genome-scale sequences using structure-sequence features and support vector machine. *BMC Bioinformatics,* **2014**, *15*, 423.
http://dx.doi.org/10.1186/s12859-014-0423-x PMID: 25547126

[50] Tran, Vdu.T.; Tempel, S.; Zerath, B.; Zehraoui, F.; Tahi, F. miRBoost: boosting support vector machines for microRNA precursor classification. *RNA,* **2015**, *21*(5), 775-785.
http://dx.doi.org/10.1261/rna.043612.113 PMID: 25795417

[51] Khan, A.; Shah, S.; Wahid, F.; Khan, F.G.; Jabeen, S. Identification of microRNA precursors using reduced and hybrid features. *Mol. Biosyst.,* **2017**, *13*(8), 1640-1645.
http://dx.doi.org/10.1039/C7MB00115K PMID: 28686281

[52] Yang, W.; Zhu, X.J.; Huang, J. A brief survey of machine learning methods in protein sub-Golgi localization. *Curr. Bioinform.,* **2019**, *14*, 234-240.
http://dx.doi.org/10.2174/1574893613666181113131415

[53] Lv, H.; Zhang, Z.M.; Li, S.H.; Tan, J.X.; Chen, W.; Lin, H. Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinform.,* **2019**, bbz048.
http://dx.doi.org/10.1093/bib/bbz048 PMID: 31157855

[54] Stephenson, N.; Shane, E.; Chase, J. Survey of machine learning techniques in drug discovery. *Curr. Drug Metab.,* **2019**, *20*(3), 185-193.
http://dx.doi.org/10.2174/1389200219666180820112457 PMID: 30124147

[55] Lai, H.Y.; Feng, C.Q.; Zhang, Z.Y.; Tang, H.; Chen, W.; Lin, H. A brief survey of machine learning application in cancerlectin identification. *Curr. Gene Ther.,* **2018**, *18*(5), 257-267.
http://dx.doi.org/10.2174/1566523218666180913112751 PMID: 30209997

[56] Zhang, T.; Tan, P.; Wang, L.; Jin, N.; Li, Y.; Zhang, L.; Yang, H.; Hu, Z.; Zhang, L.; Hu, C.; Li, C.; Qian, K.; Zhang, C.; Huang, Y.; Li, K.; Lin, H.; Wang, D. RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.,* **2017**, *45*(D1), D135-D138. PMID: 27543076

[57] Liang, Z.Y.; Lai, H.Y.; Yang, H.; Zhang, C.J.; Yang, H.; Wei, H.H.; Chen, X.X.; Zhao, Y.W.; Su, Z.D.; Li, W.C.; Deng, E.Z.; Tang, H.; Chen, W.; Lin, H. Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics,* **2017**, *33*(3), 467-469.
PMID: 28171531

[58] Cheng, L.; Wang, P.; Tian, R.; Wang, S.; Guo, Q.; Luo, M.; Zhou, W.; Liu, G.; Jiang, H.; Jiang, Q. LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.,* **2019**, *47*(D1), D140-D144.
http://dx.doi.org/10.1093/nar/gky1051 PMID: 30380072

[59] Cheng, L.; Yang, H.; Zhao, H.; Pei, X.; Shi, H.; Sun, J.; Zhang, Y.; Wang, Z.; Zhou, M. MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief. Bioinform.,* **2019**, *20*(1), 203-209.
http://dx.doi.org/10.1093/bib/bbx103 PMID: 28968812

[60] Hu, B.; Zheng, L.; Long, C.; Song, M.; Li, T.; Yang, L.; Zuo, Y. EmExplorer: a database for exploring time activation of gene expression in mammalian embryos. *Open Biol.,* **2019**, *9*(6), 190054.
http://dx.doi.org/10.1098/rsob.190054 PMID: 31164042

[61] Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res.,* **2004**, *32*(Database issue), D109-D111.
http://dx.doi.org/10.1093/nar/gkh023 PMID: 14681370

[62] Kozomara, A.; Birgaoanu, M.; Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.,* **2019**, *47*(D1), D155-D162.
http://dx.doi.org/10.1093/nar/gky1141 PMID: 30423142

[63] Zhang, Z.; Yu, J.; Li, D.; Zhang, Z.; Liu, F.; Zhou, X.; Wang, T.; Ling, Y.; Su, Z. PMRD: plant microRNA database. *Nucleic Acids Res.,* **2010**, *38*(Database issue), D806-D813.
http://dx.doi.org/10.1093/nar/gkp818 PMID: 19808935

[64] Chou, C.H.; Shrestha, S.; Yang, C.D.; Chang, N.W.; Lin, Y.L.; Liao, K.W.; Huang, W.C.; Sun, T.H.; Tu, S.J.; Lee, W.H.; Chiew, M.Y.; Tai, C.S.; Wei, T.Y.; Tsai, T.R.; Huang, H.T.; Wang, C.Y.; Wu, H.Y.; Ho, S.Y.; Chen, P.R.; Chuang, C.H.; Hsieh, P.J.; Wu, Y.S.; Chen, W.L.; Li, M.J.; Wu, Y.C.; Huang, X.Y.; Ng, F.L.; Buddhakosai, W.; Huang, P.C.; Lan, K.C.; Huang, C.Y.; Weng, S.L.; Cheng, Y.N.; Liang, C.; Hsu, W.L.; Huang, H.D. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.,* **2018**, *46*(D1), D296-D302.
http://dx.doi.org/10.1093/nar/gkx1067 PMID: 29126174

[65] Li, J.H.; Liu, S.; Zhou, H.; Qu, L.H.; Yang, J.H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.,* **2014**, *42*(Database issue), D92-D97.
http://dx.doi.org/10.1093/nar/gkt1248 PMID: 24297251

[66] Peace, R.J.; Biggar, K.K.; Storey, K.B.; Green, J.R. A framework for improving microRNA prediction in non-human genomes. *Nucleic Acids Res.,* **2015**, *43*(20), e138.
http://dx.doi.org/10.1093/nar/gkv698 PMID: 26163062

[67] Xu, Z.C.; Feng, P.M.; Yang, H.; Qiu, W.R.; Chen, W.; Lin, H. iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics,* **2019**, *35*(23), 4922-4929.
http://dx.doi.org/10.1093/bioinformatics/btz358 PMID: 31077296

[68] Qu, K.Y.; Wei, L.Y.; Zou, Q. A review of DNA-binding proteins prediction methods. *Curr. Bioinform.,* **2019**, *14*, 246-254.
http://dx.doi.org/10.2174/1574893614666181212102030

[69] Lin, H; Liang, ZY; Tang, H. identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans Comput. Biol. Bioinform.,* **2019**, *16*, 1316-1321.
http://dx.doi.org/10.1109/TCBB.2017.2666141

[70] Tang, H.; Zhao, Y.W.; Zou, P.; Zhang, C.M.; Chen, R.; Huang, P.; Lin, H. HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.,* **2018**, *14*(8), 957-964.
http://dx.doi.org/10.7150/ijbs.24174 PMID: 29989085

[71] Song, J.; Wang, Y.; Li, F. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinform.,* **2019**, *20*(2), 638-658.
http://dx.doi.org/10.1093/bib/bby028 PMID: 29897410

[72] Loh, S.K.; Low, S.T.; Chai, L.E. A review of computational approaches to predict gene functions. *Curr. Bioinform.,* **2018**, *13*, 373-386.
http://dx.doi.org/10.2174/1574893612666171002113742

[73] Li, B.Q.; Zhang, Y.H.; Jin, M.L. Prediction of protein-peptide interactions with a nearest neighbor algorithm. *Curr. Bioinform.,* **2018**, *13*, 14-24.
http://dx.doi.org/10.2174/1574893611666160711162006

[74] Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T.T.; Wang, Y.; Webb, G.I.; Smith, A.I.; Daly, R.J.; Chou, K.C.; Song, J. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics,* **2018**, *34*(14), 2499-2502.
http://dx.doi.org/10.1093/bioinformatics/bty140 PMID: 29528364

[75] Zhao, W.; Feng, Y.E. Identify protein 8-class secondary structure with quadratic discriminant algorithm based on the feature combination. *Lett. Org. Chem.,* **2017**, *14*, 625-631.

[76] Yuan, L.Z.; Yong, E.F.; Wei, Z. Using quadratic discriminant analysis to predict protein secondary structure based on chemical Shifts. *Curr. Bioinform.,* **2017**, *12*, 52-56.
http://dx.doi.org/10.2174/1574893611666160628074537

[77] Cao, R.; Freitas, C.; Chan, L.; Sun, M.; Jiang, H.; Chen, Z. ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules,* **2017**, *22*(10), 22.
http://dx.doi.org/10.3390/molecules22101732 PMID: 29039790

[78] Ding, H.; Deng, E.Z.; Yuan, L.F.; Liu, L.; Lin, H.; Chen, W.; Chou, K.C. iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Res. Int.,* **2014**, *2014,* 286419.
http://dx.doi.org/10.1155/2014/286419 PMID: 24991545

[79] Feng, P.M.; Lin, H.; Chen, W. Identification of antioxidants from sequence information using naïve Bayes. *Comput. Math. Methods Med.,* **2013**, *2013,* 567529.
http://dx.doi.org/10.1155/2013/567529 PMID: 24062796

[80] Long, CS; Li, W; Liang, PF Transcriptome comparisons of multi-species identify differential genome activation of mammals embryogenesis. *IEEE Access,* **2018**, *7*, 7794-802.
http://dx.doi.org/10.1109/ACCESS.2018.2889809

[81] Basith, S.; Manavalan, B.; Shin, T.H.; Lee, G. SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids,* **2019**, *18*, 131-141.
http://dx.doi.org/10.1016/j.omtn.2019.08.011 PMID: 31542696

[82] Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC

site prediction using effective feature representation. *Mol. Ther. Nucleic Acids,* **2019**, *16*, 733-744.
http://dx.doi.org/10.1016/j.omtn.2019.04.019 PMID: 31146255

[83] Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics,* **2019**, *35*(16), 2757-2765.
http://dx.doi.org/10.1093/bioinformatics/bty1047 PMID: 30590410

[84] Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res.,* **2003**, *31*(13), 3429-3431.
http://dx.doi.org/10.1093/nar/gkg599 PMID: 12824340

[85] Hofacker, I.L.; Priwitzer, B.; Stadler, P.F. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics,* **2004**, *20*(2), 186-190.
http://dx.doi.org/10.1093/bioinformatics/btg388 PMID: 14734309

[86] Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins,* **2001**, *43*(3), 246-255.
http://dx.doi.org/10.1002/prot.1035 PMID: 11288174

[87] Yang, H.; Tang, H.; Chen, X.X.; Zhang, C.J.; Zhu, P.P.; Ding, H.; Chen, W.; Lin, H. Identification of secretory proteins in *Mycobacterium tuberculosis* using pseudo amino acid composition. *BioMed Res. Int.,* **2016**, *2016*, 5413903.
http://dx.doi.org/10.1155/2016/5413903 PMID: 27597968

[88] Tang, H.; Chen, W.; Lin, H. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. *Mol. Biosyst.,* **2016**, *12*(4), 1269-1275.
http://dx.doi.org/10.1039/C5MB00883B PMID: 26883492

[89] Chen, X.X.; Tang, H.; Li, W.C.; Wu, H.; Chen, W.; Ding, H.; Lin, H. Identification of bacterial cell wall lyases *via* pseudo amino acid composition. *BioMed Res. Int.,* **2016**, *2016*, 1654623.
http://dx.doi.org/10.1155/2016/1654623 PMID: 27437396

[90] Zuo, Y.; Li, Y.; Chen, Y.; Li, G.; Yan, Z.; Yang, L. PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics,* **2017**, *33*(1), 122-124.
http://dx.doi.org/10.1093/bioinformatics/btw564 PMID: 27565583

[91] Zuo, Y.; Lv, Y.; Wei, Z.; Yang, L.; Li, G.; Fan, G. iDPF-PseRAAAC: A web-server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition. *PLoS One,* **2015**, *10*(12), e0145541.
http://dx.doi.org/10.1371/journal.pone.0145541 PMID: 26713618

[92] Yang, H.; Lv, H.; Ding, H.; Chen, W.; Lin, H. iRNA-2OM: a sequence-based predictor for identifying 2′-O-methylation sites in *Homo sapiens*. *J. Comput. Biol.,* **2018**, *25*(11), 1266-1277.
http://dx.doi.org/10.1089/cmb.2018.0004 PMID: 30113871

[93] Chen, W.; Feng, P.M.; Lin, H.; Chou, K.C. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *BioMed Res. Int.,* **2014**, *2014*, 623149.
http://dx.doi.org/10.1155/2014/623149 PMID: 24967386

[94] Chen, W.; Zhang, X.; Brooker, J.; Lin, H.; Zhang, L.; Chou, K.C. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics,* **2015**, *31*(1), 119-120.
http://dx.doi.org/10.1093/bioinformatics/btu602 PMID: 25231908

[95] Chou, K.C. A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun.,* **1999**, *264*(1), 216-224.
http://dx.doi.org/10.1006/bbrc.1999.1325 PMID: 10527868

[96] Bonnet, E.; Wuyts, J.; Rouzé, P.; Van de Peer, Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics,* **2004**, *20*(17), 2911-2917.
http://dx.doi.org/10.1093/bioinformatics/bth374 PMID: 15217813

[97] *Statistical Learning Theory*; John Wiley and Sons Inc: New York, NY, USA, **1998**.

[98] Dao, F.Y.; Lv, H.; Wang, F.; Feng, C.Q.; Ding, H.; Chen, W.; Lin, H. Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics,* **2019**, *35*(12), 2075-2083.
http://dx.doi.org/10.1093/bioinformatics/bty943 PMID: 30428009

[99] Feng, C.Q.; Zhang, Z.Y.; Zhu, X.J.; Lin, Y.; Chen, W.; Tang, H.; Lin, H. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics,* **2019**, *35*(9), 1469-1477.
http://dx.doi.org/10.1093/bioinformatics/bty827 PMID: 30247625

[100] Lai, H.Y.; Zhang, Z.Y.; Su, Z.D.; Su, W.; Ding, H.; Chen, W.; Lin, H. iProEP: a computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids,* **2019**, *17*, 337-346.
http://dx.doi.org/10.1016/j.omtn.2019.05.028 PMID: 31299595

[101] Zhu, X.J.; Feng, C.Q.; Lai, H.Y. Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Base. Syst.,* **2019**, *163*, 787-793.
http://dx.doi.org/10.1016/j.knosys.2018.10.007

[102] Manavalan, B.; Shin, T.H.; Lee, G. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget,* **2017**, *9*(2), 1944-1956.
PMID: 29416743

[103] Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.,* **2018**, *9*, 476.
http://dx.doi.org/10.3389/fmicb.2018.00476 PMID: 29616000

[104] Tang, H.; Cao, R.Z.; Wang, W. A two-step discriminated method to identify thermophilic proteins. *Int. J. Biomath.,* **2017**, *10*(4), 10.
http://dx.doi.org/10.1142/S1793524517500504

[105] Lin. C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.,* **2011**, *2*, 27.
https://doi.org/10.1145/1961189.1961199

[106] Breiman, L. Random forests. *Mach. Learn.,* **2001**, *45*, 5-32.
http://dx.doi.org/10.1023/A:1010933404324

[107] Breiman, L. Bagging predictors. *Mach. Learn.,* **1996**, *24*, 123-140.
http://dx.doi.org/10.1007/BF00058655

[108] Manavalan, B.; Lee, J.; Lee, J. Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS One,* **2014**, *9*(9), e106542.
http://dx.doi.org/10.1371/journal.pone.0106542 PMID: 25222008

[109] Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. PIP-EL: A new ensemble learning method for improved proinflammatory peptide predictions. *Front. Immunol.,* **2018**, *9*, 1783.
http://dx.doi.org/10.3389/fimmu.2018.01783 PMID: 30108593

[110] Manavalan, B.; Shin, T.H.; Kim, M.O.; Lee, G. AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. *Front. Pharmacol.,* **2018**, *9*, 276.
http://dx.doi.org/10.3389/fphar.2018.00276 PMID: 29636690

[111] Hasan, M.M.; Manavalan, B.; Khatun, M.S.; Kurata, H. i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. *Int. J. Biol. Macromol.,* **2019**, S0141-8130(19)38547-2.
http://dx.doi.org/10.1016/j.ijbiomac.2019.12.009 PMID: 31805335

[112] Hasan, M.M.; Manavalan, B.; Khatun, M.S.; Kurata, H. Prediction of S-nitrosylation sites by integrating support vector machines and random forest. *Mol Omics,* **2019**, *15*(6), 451-458.
http://dx.doi.org/10.1039/C9MO00098D PMID: 31710075

[113] Dao, F.Y.; Lv, H.; Wang, F.; Ding, H. Recent advances on the machine learning methods in identifying DNA replication origins in eukaryotic genomics. *Front. Genet.,* **2018**, *9*, 613.
http://dx.doi.org/10.3389/fgene.2018.00613 PMID: 30619452

[114] Kohonen, T. *Self-organized formation of topologically correct feature maps.,* **1988**.

[115] Milone, D.H.; Stegmayer, G.S.; Kamenetzky, L.; López, M.; Lee, J.M.; Giovannoni, J.J.; Carrari, F. *omeSOM: a software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. *BMC Bioinformatics,* **2010**, *11*, 438.
http://dx.doi.org/10.1186/1471-2105-11-438 PMID: 20796314

[116] Yoon, B.J.; Vaidyanathan, P.P. Context-sensitive hidden Markov models for modeling long-range dependencies in symbol sequences. *IEEE Trans. Signal Process.,* **2006**, *54*(11), 4166-4184.
http://dx.doi.org/10.1109/TSP.2006.880252

[117] Xue, L.; Tang, B.; Chen, W.; Luo, J. Prediction of CRISPR sgRNA activity using a deep convolutional neural network. *J. Chem. Inf. Model.,* **2019**, *59*(1), 615-624.
http://dx.doi.org/10.1021/acs.jcim.8b00368 PMID: 30485088

[118] Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.,* **2011**, *273*(1), 236-247.
http://dx.doi.org/10.1016/j.jtbi.2010.12.024 PMID: 21168420

[119] Tan, J.X.; Li, S.H.; Zhang, Z.M.; Chen, C.X.; Chen, W.; Tang, H.; Lin, H. Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.,* **2019**, *16*(4), 2466-2480.
http://dx.doi.org/10.3934/mbe.2019123 PMID: 31137222

[120] Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chen, W.; Chou, K.C. iDNA6mA-PseKNC: identifying DNA N^6-methyladenosine sites by

incorporating nucleotide physicochemical properties into PseKNC. *Genomics,* **2019**, *111*(1), 96-102.
http://dx.doi.org/10.1016/j.ygeno.2018.01.005 PMID: 29360500

[121] Chen, W.; Lv, H.; Nie, F.; Lin, H. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics,* **2019**, *35*(16), 2796-2800.
http://dx.doi.org/10.1093/bioinformatics/btz015 PMID: 30624619

[122] Manavalan, B.; Basith, S.; Shin, T.H.; Lee, D.Y.; Wei, L.; Lee, G. 4mCpred-EL: an ensemble learning framework for identification of DNA N^4-methylcytosine sites in the mouse genome. *Cells,* **2019**, *8*(11), 8.
http://dx.doi.org/10.3390/cells8111332 PMID: 31661923

[123] Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput. Struct. Biotechnol. J.,* **2019**, *17*, 972-981.
http://dx.doi.org/10.1016/j.csbj.2019.06.024 PMID: 31372196

[124] Metz, C.E. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest. Radiol.,* **1989**, *24*(3), 234-245.
http://dx.doi.org/10.1097/00004424-198903000-00012 PMID: 2753640

[125] Cheng, L.; Jiang, Y.; Ju, H.; Sun, J.; Peng, J.; Zhou, M.; Hu, Y. InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics,* **2018**, *19*(Suppl. 1), 919.
http://dx.doi.org/10.1186/s12864-017-4338-6 PMID: 29363423

[126] Cheng, L.; Zhuang, H.; Yang, S.; Jiang, H.; Wang, S.; Zhang, J. Exposing the causal effect of C-reactive protein on the risk of Type 2 diabetes mellitus: a mendelian randomization study. *Front. Genet.,* **2018**, *9*, 657.
http://dx.doi.org/10.3389/fgene.2018.00657 PMID: 30619477

[127] Kavzoglu, T.; Mather, P.M. The role of feature selection in artificial neural network applications. *Int. J. Remote Sens.,* **2002**, *23*, 2919-2937.
http://dx.doi.org/10.1080/01431160110107743

[128] Chawla, N.V.; Bowyer, K.W.; Hall, L.O. smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.,* **2002**, *16*, 321-357.
http://dx.doi.org/10.1613/jair.953

[129] Boulesteix, A.L.; Strimmer, K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.,* **2007**, *8*(1), 32-44.
http://dx.doi.org/10.1093/bib/bbl016 PMID: 16772269

[130] Tempel, S.; Tahi, F. A fast *ab-initio* method for predicting miRNA precursors in genomes. *Nucleic Acids Res.,* **2012**, *40*(11), e80.
http://dx.doi.org/10.1093/nar/gks146 PMID: 22362754

[131] Liu, D.; Li, G.; Zuo, Y. Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Brief. Bioinform.,* **2019**, *20*(5), 1826-1835.
http://dx.doi.org/10.1093/bib/bby053 PMID: 29947743

[132] Liu, B.; Liu, F.; Fang, L.; Wang, X.; Chou, K.C. repRNA: a web server for generating various feature vectors of RNA sequences. *Mol. Genet. Genomics,* **2016**, *291*(1), 473-481.
http://dx.doi.org/10.1007/s00438-015-1078-7 PMID: 26085220

[133] Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.,* **2005**, *27*(8), 1226-1238.
http://dx.doi.org/10.1109/TPAMI.2005.159 PMID: 16119262

[134] Lin, H.; Deng, E.Z.; Ding, H.; Chen, W.; Chou, K.C. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.,* **2014**, *42*(21), 12961-12972.
http://dx.doi.org/10.1093/nar/gku1019 PMID: 25361964

[135] Saçar, M.D.; Allmer, J. Machine learning methods for microRNA gene prediction. *Methods Mol. Biol.,* **2014**, *1107*, 177-187.
http://dx.doi.org/10.1007/978-1-62703-748-8_10 PMID: 24272437

[136] Hou, J.; Wu, T.; Cao, R.; Cheng, J. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins,* **2019**, *87*(12), 1165-1178.
http://dx.doi.org/10.1002/prot.25697 PMID: 30985027

[137] Peng, L.; Peng, M.M.; Liao, B. The advances and challenges of deep learning application in biological big data processing. *Curr. Bioinform.,* **2018**, *13*, 352-359.
http://dx.doi.org/10.2174/1574893612666170707095707

[138] Patel, S.; Tripathi, R.; Kumari, V. DeepInteract: deep neural network based protein-protein interaction prediction tool. *Curr. Bioinform.,* **2017**, *12*, 551-557.
http://dx.doi.org/10.2174/1574893611666160815150746

[139] Long, H.X.; Wang, M.; Fu, H.Y. Deep convolutional neural networks for predicting hydroxyproline in proteins. *Curr. Bioinform.,* **2017**, *12*, 233-238.
http://dx.doi.org/10.2174/1574893612666170221152848

[140] Cao, R.; Bhattacharya, D.; Hou, J.; Cheng, J. DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics,* **2016**, *17*(1), 495.
http://dx.doi.org/10.1186/s12859-016-1405-y PMID: 27919220