EasyChair Preprint

№ 201

A new method for identification of pre-microRNAs based on hybrid features

Yuanlin Ma, Zuguo Yu, Guosheng Han and Vo Anh

May 31, 2018

RESEARCH

A new method for identification of pre-microRNAs based on hybrid features

Yuanlin Ma[1], Zuguo Yu[1,2*], Guosheng Han[1] and Vo Anh[1,3]

*Correspondence: yuzuguo@aliyun.com
[1]Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Hunan 411105 P.R. China
Full list of author information is available at the end of the article

Abstract

Background: The identification of pre-microRNAs (precursor microRNAs) helps us to understand the regulatory mechanism of biological processes. Currently, machine learning is the most popular method for pre-microRNA identification. However, most methods mainly focus on secondary structure information of pre-microRNA, while ignoring sequence-order information and sequence evolution information.

Results: In this work, we use three different methods to extract features of the pre-microRNAs at different levels. We first extract features from PSI-BLAST profiles and Hilbert-Huang transform, which contain rich sequence evolution information and sequence-order information respectively. We then get properties of small molecular networks of pre-microRNAs, which contain refined secondary structure information. We extract 591 features in total. After extraction, we use support vector machine (SVM) as our classifier, and use the maximum relevance and minimum redundancy (mRMR) method for feature selection. Finally, we construct a new predictor *MicroRNA−NHPred* by using the optimal feature set. The performance of *MicroRNA−NHPred* is quite promising compared to other popular miRNA predictors. It achieves an accuracy of up to 94.83%.

Conclusions: The higher prediction accuracy achieved by our proposed method is attributed to the design of a comprehensive feature set on the sequence and secondary structure, which are capable of characterizing the sequence evolution information and sequence-order information, and global and local information of pre-microRNAs secondary structure. Therefore, it is a valuable method to pre-microRNAs identification.

Keywords: Pre-microRNA; PSI-BLAST profiles; Hibert-Huang transform; Network; mRMR; SVM

Background

MicroRNAs (miRNAs) are small single-strand, non-coding RNAs (about 22 nucleotides in length), which play significant regulatory roles in various biological processes of animals, plants and viruses [1, 2]. There are many forms of miRNAs, including primary miRNAs (pri-miRNAs), mature miRNAs and precursor microRNAs (pre-microRNAs). Mature miRNAs are usually cleaved from ~ 90nt pre-microRNAs which are derived from processing of a long pri-miRNA by a ribonucluease [3]. In fact, pre-microRNAs is the earliest and most widely studied, and many commercialized miRNA libraries take this form. With the advent of the post genome era and the development of sequencing technology, how to find miRNAs from millions of reads has been one of the hot topics in bioinformatics. Detecting miRNAs by

experimental techniques in biology is expensive and time-consuming. What's more, it is difficult to identify directly the lowly expressed miRNAs or the miRNAs that are expressed in the specific tissues or in the developmental stage. Computational methods have provided potential pre-microRNAs candidates for biologists. Because miRNAs are too short, the traditional feature engineering approaches [4] are usually failed to extract features based on their sequences and structures. Therefore, computational methods is usually to identify the pre-microRNAs instead of miRNAs.

At present, there are many methods to identify pre-microRNAs, which are mainly divided into four categories. The first category contains the earliest methods which are based on searching homologous genes [5]. The search process is a typical alignment problem of sequences and structures, the main alignment algorithms including the Smith-Waterman algorithm [5], the FASTA algorithm, the BLAST algorithm [6, 7, 8, 9], etc. However, these methods can only find highly homologous miRNAs with known miRNA sequences and require a large amount of computational resource for whole genome. The second category contains comparative genome methods which predict miRNAs in the study of species early stages. In the process of prediction, these methods mainly utilize the conserved characteristics of miRNAs and their precursor sequences in multiple species to search for the conserved sequences in the intergenic region. These sequences have a better secondary structure of stem ring. Based on comparative genomics, the limitation of predicting miRNAs is that the predicted miRNA candidates are highly conserved in multiple species, and these methods cannot be used to predict miRNAs which are not conserved [10, 11, 12, 13]. At the same time, these methods are also subject to challenges of both time complexity and space complexity. The third category is based on conservation of binding sites of miRNA which are the short sequences of miRNA that bind the target mRNA. These short sequences have conserved properties among multiple species [14, 15, 16]. The miRNAs and the target mRNAs usually have perfect complementary features in plants, while it does not match well in animals. Therefore, this kind of methods is usually used in plants. The fourth category is based on machine learning methods [17, 18, 19, 20, 21]. Machine learning uses the information on sequences, structural and thermodynamic energy of pre-microRNAs. These methods can discover new, non-homologous pre-microRNAs. So, machine learning is the main method for miRNA prediction and identification at present. The difficulty of the method is how to select the positive/ negative samples which are able to describe sufficiently the whole sample space and how to find a better distinction between true/ false pre-microRNAs. In addition, high false positive rates and computational complexity likely occur in the prediction of whole genome data. Thus, further improvement in sensitivity and specificity of the pre-microRNA classification is necessary. It is also a desirable task to explore a solution based on machine learning prediction.

Generally speaking, the problem of pre-microRNA identification can be viewed as a classification problem so that it can be tackled by machine learning methods. To implement a classification task, two major procedures are generally required: feature extraction and a machine learning classifier. In the past few decades, extracted features of pre-microRNAs are mainly divided into three categories: primary sequences, secondary structures and thermodynamical properties. Among them, the

k-mer sequence composition (based on the primary sequence) is the most successful technique for representation of pre-microRNAs [22]. Many studies have shown that most of pre-microRNAs have the properties of stem loop hair-pin structures [19]. Therefore, secondary structures can be predicted, and features derived from these structures, e.g, 32 local structure features in triplet-SVM, are used to predict human pre-microRNAs [19]. Energy characteristics are another kind of important features of pre-microRNAs [23]. The pre-microRNAs in a folded state have lower free energies than random sequences [24]. It is well studied that good features and positive/ negative (real/ pseudo pre-microRNA) datasets are the basis of constructing efficient classification models.

In this study, we use three different methods to extract features of the pre-microRNAs at different levels. To describe the local or short-range sequence order information and evolution information of pre-microRNAs, we introduce the PSI-BLAST profile into the analysis of pre-microRNAs for the first time. And then we introduce the Hilbert-Huang transform [25], which is a time-frequency analysis method, into pre-microRNA identification. We use it to describe the local and long-range relationship between sequence bases. We obtain the topological parameters of small molecular networks constructed from the secondary structures of pre-microRNAs, which contain refined secondary structure information. These features are carefully selected so that they can depict both global and local characteristics of pre-microRNAs. After the feature extraction, we use support vector machine (SVM) as our classifier, and use the maximum relevance and minimum redundancy (mRMR) [26] method for feature selection. Finally, a new predictor *MicroRNA−NHPred* is constructed by using the optimal feature set, which achieves an accuracy of up to 94.83%. This demonstrate that the new constructed predictor improves the sensitivity and specificity of precursor microRNA prediction.

Materials and methods

Datasets

In order to compare with previous works, we use the benchmark dataset in the works of Liu et al. [27, 28, 29, 30] and Khan et al. [31], which consists of positive samples (true pre-microRNAs) and negative samples (pseudo pre-microRNAs). As in the above works, we derived the positive samples from the miRBase (released on 20 June, 2013) [32], which is composed of 1872 experimentally confirmed pre-microRNA sequences of homo sapiens. These sequences were filtered by the CD-HIT software [33], and the redundant sequences were filtered out with a threshold of 80% sequence identity. Finally, we obtained 1612 true pre-microRNA sequences as positive samples. As in previous works [17, 18, 19, 24], we used 8494 human pseudo pre-microRNAs. The dataset of negative samples collected from human protein coding regions was downloaded from Xue et al. [19]. These sequences are very similar to the real pre-microRNAs in the sequence length, the minimum base pair of their stem of hairpin structure, and the maximum energy of secondary structure. In the same way as positive samples, we also used the CD-HIT software to filter the sequences, so that sequence similarity of the negative samples is kept below 80%. In order to solve the sample imbalance problem [27, 28], 1612 sequences are selected randomly as negative samples from the filtered sequences.

In order to further verify the performance of our method, we use an independent test set to verify it. This test set comes from the latest released miRBase 22 [34] (released on March 2018) which contains 1917 homo sapiens pre-microRNA sequences, while miRBase 20 (released on June 2013) contains 1872 homo sapiens pre-microRNA sequences. We selected 78 new homo sapiens pre-microRNA sequences as our independent test set from the latest version of miRBase 22, and those sequences are not in the released miRBase 20.

Feature extraction methods

In this study, we use three different methods to extract different features of pre-microRNAs from PSI-BLAST profiles [35, 36], parameters of networks [37] and spectrum analysis based on the Hilbert-Huang transform [25].

PSI−BLAST profile−based features

The PSI-BLAST profile is represented as a so-called position specific score matrix (PSSM), which is acquired through aligning a query amino acid sequence to the NCBI's nonredundant (NR) database by using PSI-BLAST [35]. In this work, we apply this idea to the nucleotide sequences.

First, we build a new database, which is composed of all the pre-microRNA sequences in the miRBase [38] and 8494 human pseudo pre-microRNAs in the work of Xue et al. [19] and 410 non-coding RNAs in the work of Batuwita et al. [18].

Second, we use PSI-BLAST to align a query nucleotide sequence in the dataset to the newly built database and to get the PSSM for the sequence. The PSSM is a matrix of size $L \times 5$, where L is the length of the query sequence and 5 is due to the 4 nucleotide symbols (A, C, G, U) and the symbol $-$. Its elements are $10\times$ $\log_e$ of the ratios between the observed base frequencies and the background base frequencies, and rounded down to the nearest integer.

Third, our feature extraction method also starts by transforming each element s_{ij} of the PSSM into $s_{ij}^{'}$ using

$$s_{ij}^{'} = 2^{0.1\times s_{ij}}. \tag{1}$$

The resulting value $s_{ij}^{'}$ is guaranteed to be non-negative even when s_{ij} is negative. We further apply the normalization to the values $s_{ij}^{'}$ so that each row sums to one. Let f_{ij} denote the normalized value of $s_{ij}^{'}$. All the values f_{ij} form a matrix, which we called the frequency matrix (FM).

Fourth, to extract PSI-BLAST profile features, a so-called concensus sequence (CS) [39] is first constructed from the FM as follows:

$$\mu(i) = \arg\max\{f_{ij} :\ 1 \le j \le 4\}, 1 \le i \le L. \tag{2}$$

The i-th base $CS(i)$ of the consensus sequence is set to be the $\mu(i)$-th nucleotide in the nucleotide alphabet. It can be seen that a consensus sequence retains the most valuable evolutionary information from the PSSM.

Fifth, we compute

$$\mathrm{NCCS}(j) = \frac{n(j)}{L}, 1 \le j \le 4, \tag{3}$$

where $n(j)$ is the number of the nucleotide j occurring in the CS. It gives 4 features corresponding to the nucleotide of the CS. Moreover, we also include the entropy into our feature set, that is,

$$\mathrm{ECS} = -\sum_{j=1}^{5} \mathrm{NCCS}(j) \log_e \mathrm{NCCS}(j). \tag{4}$$

Another entropy-based feature is directly computed from FM to reflect the global characteristic of the PSSM:

$$\mathrm{EFM} = -\frac{1}{L} \sum_{i=1}^{L} \sum_{j=1}^{5} f_{ij} \log_e f_{ij}. \tag{5}$$

Most of the extracted features of k-mer features shown in many articles are based on the original sequences. In this study, we extract their features from the CS of the original nucleotide sequences. Since a pre-microRNA sequence is too short (about 60bp-130bp), longer k are less likely to be exactly conserved among species. So, we computed k-mers with $k = 2, 3$ resulting in 80 (16+64) different features. At the same time, we also calculate the content of GC from consensus sequences.

In summary, for each query sequence, a total of 87 features are extracted from its PSI-BLAST profile. Our experimental results show that the features extracted from CS are more effective to discriminate than those from the original nucleotide sequences.

Topological parameters of small molecular networks constructed from secondary structures

The pre-microRNA has a very significant secondary structure in the hairpin shape. There are many machine-learning based methods to identify pre-microRNAs which take advantage of the hairpin shape, so that the prediction accuracy has been greatly improved. There are more representative Triplet-SVM [19], iMiRNA-PseDpc [27], and properties based on networks [37] in these methods. In Refs. [40, 41], the authors have verified that the features based on networks have higher prediction accuracies. Meanwhile, in Ref. [37], Childs et al. further discussed the topological properties of the networks, which can reflect more essential characteristics of the pre-microRNAs. Therefore, in this work, we also extract features based on networks constructed from the secondary structure, and the process is as follows:

Firstly, each nucleotide sequence of positive and negative samples is folded into a stem-loop secondary structure by RNAfold [42]. Secondly, we use a two-dimensional network (graph) to represent the RNA secondary structure, with all nucleotides converted to nodes and all bonds between nucleotides converted to edges. Network elements, including nodes and edges, can be defined by the network itself or parameters which may relate to limited or full knowledge of the network. Based on these criteria Childs et al. classified the network parameters into three types: local, local-global and global structural properties that can be used as a method in identification of RNA family [37]. Here we use the summary statistics for the local-global properties, since they provide insight not only on the global level of the graph itself,

but also on the level of its nodes and edges. Thirdly, all properties were calculated using the *igraph R* package [43] for complex networks. In this study, 24 network parameters are extracted to describe the stem-loop structure of pre-microRNAs based on previous works and experimental criteria [37] although a number of network parameters are available. We also choose the following features: degree, path length, shortest path, graph motifs, articulation point, modularity, graph density, coreness, closeness, centrality, bibliographic coupling, transitivity, cocitation coupling, diameter, node betweenness, edge betweenness, grith, constraint, hub score, and so on. A brief definition of all graph properties used in this study is provided in [37].

Extraction of sequence-order features based on the Hilbert-Huang transform

The features of the pre-microRNAs based on k-mers, with k small, they can only describe the short-range relationship between the nucleotide sequences. When k is larger, they can describe the long-range relationship of the nucleotide sequences, but the dimension of extracted feature vector is too large, which leads to the curse of dimensionality, and the classifier's performance will be reduced. Since most of the previous methods extracted k-mer composition information from a nucleotide sequence (for pre-microRNAs, k generally takes the values 2, 3, 4), the sequence-order information is missing. Although Chen and Li [44] considered local sequence-order information based on Chou's concept of pseudo amino acid composition, the overall prediction accuracy was not significantly improved. In order to depict the long range relationship and order information of the sequence, we introduce the Hilbert-Huang transform [25] based on the physical and chemical properties of the known dinucleotides.

The Hilbert-Huang transformation consists of two parts: empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA). The empirical mode decomposition is a time-frequency analysis and was originally proposed by Huang et al. [25] for the study of ocean waves. The EMD method has been used by our group to simulate geomagnetic field data [45] and to predict protein subnuclear localization [46]. In EMD, the base functions, which are called intrinsic mode functions (IMFs), are obtained adaptively from the original signal. The principle and details of Hilbert spectral analysis can be found in [25, 46]. Combining the sequences of the pre-microRNAs and the physical and chemical characteristics of the dinucleotides, the feature extraction method based on the Hilbert-Huang transform is described as follows:

1. According to the physical and chemical properties of dinucleotides and the intrinsic characteristics of Hilbert-Huang transform, we selected 15 physical and chemical properties for RNAs from the database [47], including: enthalpy, enthalpy2, entropy, entropy2, free energy, free energy2, hydrophilicity, hydrophilicity2, rise, roll, shift, slide, stackingenergy, tilt, twist.

2. According to the physical and chemical properties of dinucleotides, the sequence of each pre-microRNA was converted into 30 time series by sliding a window along the sequence.

3. At first, we need to get the intrinsic mode functions of each time series by empirical mode decomposition. And then we applied Hilbert spectral analysis to every intrinsic mode function to obtain the analysis signals. Finally, we obtained

16 features for each time series. The specific signal analysis process can be found in [46].

In this study, we firstly transformed all the RNA sequences into time series according to 15 physical and chemical properties of dinucleotides. Finally, we extracted 480 Hilbert-Huang features.

Feature selection method

After the feature extraction is completed for a sequence such as pre-microRNA, some extracted features may be redundant, some may not be related to a class. Therefore, before prediction is carried out, it may be necessary to remove some features according to certain rules. There are many ways to remove redundant or useless features (in the sense that they have no significant relation to a class), such as mRMR [26], FOCUS [48], Wrapper [49], and so on. In this work, we choose the mRMR method as our feature selection method, now described:

Let Ω be the whole feature space which contains all of the aforementioned 591 features in this work; each sequence is represented by a vector consisting of the values of these 591 features. We assume that E and F are two disjoint subsets of Ω and $\Omega{=}E{\cup}F$. In order to select a feature f_j in E with maximum relevance and minimum redundancy in F, we use the following formula:

$$\max_{f_j \in E}[I(f_j, \theta) - D(f_j, F)],\ j = 1, 2, \ldots, \sharp E, \tag{6}$$

where θ is a vector characterizing the class of all nucleotide sequences in the sample set, $\sharp E$ denotes the cardinality of the subset E.

In the actual computation process, we regard E as a feature set to be selected, and F as an already selected feature set. At the beginning, E is the feature space, F is the null space, the process of the mRMR method is as follows: First, we select a feature that is most relevant to the class vector in E, then remove it from E and add it to F. Second, according to the mRMR function, repeat the first step. After $\sharp\Omega$ cycles, E is null, F is the entire feature set. According to the order in which the feature is added to F, the features in the whole feature set are reordered, and we use S to represent the ordered feature set:

$$S = \{f_{i_1}, f_{i_2}, f_{i_3}, \ldots, f_{i_{\sharp\Omega}}\}. \tag{7}$$

After all features are ranked, we can determine the optimal feature components by an incremental feature selection (IFS) method [50]. For the ranked feature set S, we can construct the feature component sets by adding one component at a time in an ascending order as follows:

$$S_k = \{f_{i_1}, f_{i_2}, f_{i_3}, \ldots, f_{i_k}\}\ (1 \leqq k \leqq \sharp\Omega). \tag{8}$$

For each feature component set, a predictor is constructed and the accuracy is obtained by the rigorous jackknife validation. Finally, we choose the feature component set for the best jackknife validation accuracy as the optimal feature set.

Support vector machine

A Support Vector Machine (SVM) is a class of supervised learning algorithms first introduced in [51]. It is based on statistical theory, and has a good general application. In this work, we use an SVM as a classifier to identify the real and pseudo pre-microRNAs.

Given a set of labelled training vectors (positive and negative input samples), SVM learns a linear decision boundary from both positive and negative training samples to discriminate between the unknown RNA sequences. A key feature of SVM is that it needs a fixed length of the input vector. The pre-microRNAs in the training set and the test set are transformed into fixed-dimension feature vectors following the process introduced above, and then the training vectors are input into SVM to construct the classifier. The SVM gives a predicted class for each sample in the test set.

The LIBSVM algorithm [52] was employed, which is a type of software for SVM classification and regression. The radial basis function (RBF) defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma(\parallel \mathbf{x}_i - \mathbf{x}_j \parallel)^2), \gamma > 0 \tag{9}$$

is used as the kernel function $k(\mathbf{x}, \mathbf{y})$ in the SVM. Here, $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ is a given dataset. For a Gaussian RBF, γ is parametrized as $\gamma = \frac{1}{2\sigma^2}$. The parameter γ and the soft margin parameter C are optimized on the benchmark dataset by adopting the grid tool provided by LIBSVM [52]. More details are provided in [53].

The proposed identification method

Fig.1 illustrates the overall architecture of our proposed method which is called *MicroRNA-NHPred*. Firstly, the query nucleotide (RNA) sequences are input into PSI-BLAST to obtain PSSM, and entropy of sequences and consensus sequences (CS) [39]. We then obtain k-mer composition of CS. The query nucleotide sequence is submitted to RNAfold software to generate a secondary structure.

We build a single molecule network from the secondary structure, then extract network topological parameters. Each pre-microRNA molecule is represented by the topological parameters of a single molecule network.

On the other hand, the query nucleotide sequence is converted into a time series based on the physicochemical properties of the RNA. The obtained time series are transformed and 480 characteristics are obtained. Ultimately, we get 591 features in total. These features are finally put into an SVM-based classifier for pre-microRNA classifier recognition.

Performance evaluation

The performance of the predictor should be objectively evaluated. In statistical prediction, three cross-validation tests are often used to evaluate the prediction performance: independent dataset test, sub-sampling (or K-fold crossover validation) test and jackknife test. Only the jackknife test is the least arbitrary that can always yield a unique result for a given benchmark dataset [54, 55]. That is why researchers have a preference for the jackknife test for examining the quality of various machine learning based predictors such as [30, 31, 46]. Hence, we also

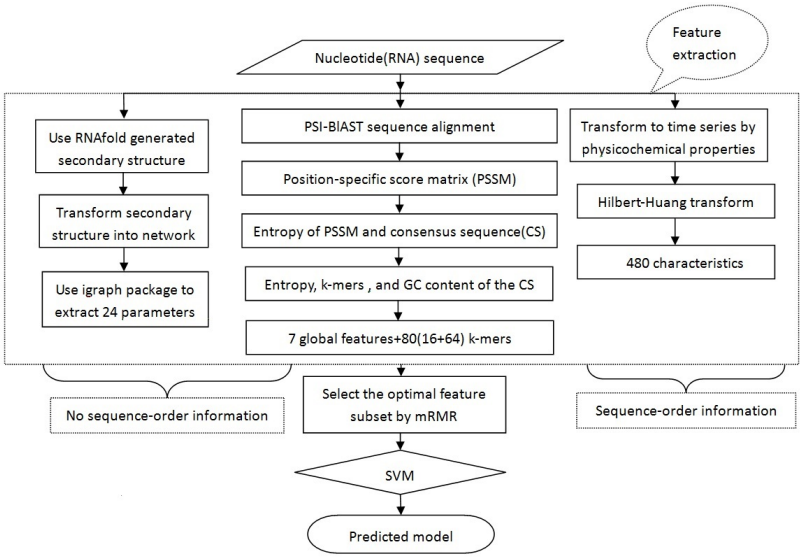

Fig 1 Flow chart of the identification method in this study.

use the jackknife test and independent dataset test to evaluate the accuracy of the current predictor in this work. In the jackknife test, each sequence in the samples is singled out in turn as a test sample and the remaining sequences are used as training samples. Although the jackknife test consumes more computing resources, it is worthwhile to have a single output for a given set of samples.

When the cross-validation method is selected, we need to choose the performance metrics of the predictor. The identification of pre-microRNAs is a binary classification problem. For this problem, we select the following indicators to evaluate our predictor: S_n (sensitivity), S_p (specificity), Acc (overall accuracy), Mcc (Mathew correlation coefficient) [56], calculated by $S_n = TP/(TP + FN)$, $S_p = TN/(TN + FP)$, $Acc = (TP + TN)/(TP + TN + FP + FN)$, and

$$Mcc = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

In the above formulas, TP means the true positive, TN the true negative, FP the false positive and FN the false negative. The sensitivity denotes correct identification of positive pre-microRNAs by avoiding false negative, while the specificity denotes correct identification of negative pre-microRNAs by avoiding false positive. The sensitivity and the specificity range between 0 and 1, the bigger the value, the better the predictor. The Mathew correlation coefficient (Mcc) ranges between -1 and 1, the overall accuracy (Acc) ranges between 0 and 1.

Discussion and results

Parameter selection by mRMR

We use three different methods to extract 591 features. Since some of these features are not essential and may not be significantly related to the classes of pre-microRNAs, we used the method in subsection "Feature selection method " to sort

the features first and used the increment feature selection method to select the optimal feature set. For each feature subset, we constructed a classifier and derived its jackknife validation accuracy. Finally, we obtained the best feature subset corresponding to the best jackknife validation accuracy as the optimal feature subset. We used all the feature sets to construct the predictor, whose jackknife validation accuracy turns out to be 89.73%. We used the optimal feature subset to construct a predictor with a jackknife validation accuracy of 94.83% being achieved.

Performance of predictor on different feature sets

As shown in subsection "Feature extraction methods", we used 3 different methods to extract 3 different feature sets. In order to study the effect of different feature sets on the performance of the predictor, we tested the single feature set and different feature combinations respectively on prediction performance, as shown in Table 1. We can see that the three feature sets have different contributions to the recognition of pre-microRNAs, of which the contribution of the network feature set is the most significant and the accuracy of the predictor is 87.85%.

Table 1 The performance of different feature sets.

Method	Mcc	Accuracy	Precision	Recall
PSI-BLAST	0.5129	0.7564	0.7681	0.7446
HHT	0.4887	0.7440	0.7731	0.7148
Network	0.7589	0.8785	0.9144	0.8425
PSI-BLAST+Network	0.7707	0.8853	0.8909	0.8797
Network+HHT	0.7212	0.8802	0.8783	0.8841
PSI-BLAST+HHT+Network	0.7850	0.8973	0.9028	0.8718

We firstly introduced PSI-BLAST to the prediction of pre-microRNAs. In order to verify the performance contribution of the k-mers from CS, we separately extracted k-mers (k=2, 3) from the original sequence and the CS for jackknife test verification. The result of the test is shown in Table 2. The accuracy of jackknife test validation shows that the contribution of k-mers from CS is more significant.

Table 2 The performance of different k-mers: ($k = 2, 3$).

Predictors	Mcc	Accuracy	S_n	S_p
PSI-BLAST-K-mer	0.5129	0.7404	0.7501	0.7218
K-mer	0.5010	0.7201	0.69	0.6901

Secondary structure features have a variety of different representations, e.g, triplet-SVM [19], iMcRNA-PseSSC [27], network [37], and so on. To verify the effect of three secondary structure features on the problem of pre-microRNA classification, we used the jackknife test on the same benchmark dataset. As shown in Table 3, we found that the parameters of networks reflect the pre-microRNA secondary structure. So, we used the parameters of networks to depict the secondary structure of pre-microRNAs in this work.

Table 3 The performance of different features of secondary structure.

Predictors	Mcc	Accuracy	S_n	S_p
Network	0.7589	0.8785	0.9144	0.8425
Triplet-SVM [19]	0.64	0.8185	0.7847	0.8520
IMcRNA-PseSSC [27]	0.72	0.8576	0.8836	0.8350

Comparison with other methods

We compared our predictor with the best and most accurate predictors in this field, triplet-SVM [19], miPred [24], iMcRNA-EXPseSSC [27], microR-Pred (SVM) [31]. The comparison indicates that the accuracy of our predictor is higher than other predictors in the same larger and more stringent benchmark dataset using rigorous jackknife tests. As can be seen from Table 4, we have the highest prediction accuracy on Mcc, Accuracy and S_n, and only S_p is lower than miPred [24] and microR-Pred (SVM) [31], but also higher than 90%.

Table 4 The performance of different methods on the same benchmark dataset.

Predictors	Mcc	Accuracy	S_n	S_p
Triplet-SVM [19]	0.64	0.8185	0.7847	0.8520
MiPred [24]	0.75	0.8730	0.84	0.9060
IMcRNA-EXPseSSC [27]	0.80	0.8986	0.8993	0.8978
MicroR-Pred(SVM) [31]	0.88	0.9390	0.93	**0.9470**
MicroRNA-NHPred	**0.8965**	**0.9483**	**0.9420**	0.9010

Validation based on an independent test set

The benchmark dataset was constructed based on miRBase released 20 (June 2013). At present, compared with miRBase released 20, the latest miRBase released 22 reports 78 new homo sapiens pre-microRNAs, which were treated as an independent test set to further evaluate the performance of the proposed *MicroRNA-NHPred*. The test results are shown in Table 5. This method trained with the benchmark dataset can correctly predict 75 testing samples in the independent dataset as true sapiens pre-microRNAs. The accuracy of the proposed method can reach 96.15%, which demonstrates the stable prediction performance of *microRNA-NHPred* for predicting sapiens pre-microRNAs.

MicroR-Pred (SVM) [31] and iMcRNA-EXPseSSC [27], which are the most accurate predictors in this field as we know, were also tested on the same independent test set. It is worth noting that microR-Pred (SVM) [31] and iMcRNA-EXPseSSC [27] correctly identified 71 and 67 homo spaeins pre-microRNAs with an accuracy of 91.03% (71/78) and 85.90% (67/78) respectively.

Table 5 The result of different methods on an independent test set.

Method	Accuracy	Pre-microRNAs without the correct identification
IMcRNA-EXPseSSC [27]	0.8590(67/78)	hsa-mir-8069-2, hsa-mir-1843, hsa-mir-10393, hsa-mir-10394, hsa-mir-10395, hsa-mir-10400, hsa-mir-10527, hsa-mir-11401, hsa-mir-12115, hsa-mir-12128, hsa-mir-9500
MicroR-Pred(SVM) [31]	0.9103(71/78)	hsa-mir-10395, hsa-mir-9500, hsa-mir-8069-2, hsa-mir-12115, hsa-mir-10400, hsa-mir-11401, hsa-mir-12128,
MicroRNA-NHPred	0.9615(75/78)	hsa-mir-1843, hsa-mir-12115, hsa-mir-11401.

Conclusion

We used three different methods to extract different level features of pre-microRNAs and used SVM to classify positive and negative samples. The extracted features can describe sequence and the secondary structure characteristics of pre-microRNAs.

1. We firstly introduced PSI-BLAST into the analysis of pre-microRNAs, extracted the consensus sequence of every sample and the entropy of the PSSM, the entropy of the consensus sequence, the k-mers and G+C content of the consensus

sequence. The PSI-BLAST profile describes the local or short-range sequence order information and evolution information of pre-microRNAs.

2. We transformed the sequence of positive and negative samples into a secondary structure, transformed the secondary structure into a single molecule network. The network parameters were extracted and each sample was represented by network parameters. These network parameters can describe more completely the local and global characteristics of RNAs. Under the same benchmark dataset, the accuracy of network parameters can reach 87.85%; The well-known triplet-SVM can only reach 81.85%.

3. We introduced the Hilbert-Huang transform into pre-microRNA identification for the first time, used it to describe the local and long-range relationship between sequence bases.

Finally, we combined these features, and then selected the optimal 268 features by mRMR. Compared with the most accurate predictor, MicroR-Pred (SVM) [31], in this field, the accuracy is not improved significantly, but the results of the experiment show that the extracted features are related to pre-microRNAs. We believe that the features extracted from this method are relevant and useful for further works by biologists.

Acknowledgements
Yuanlin Ma would like to express her gratitude to Dr. Jianyi Yang in Nankai University for useful discussion.

Funding
This project was supported by the Natural Science Foundation of China (Grant No. 11371016), the Chinese Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT)(Grant No. IRT_15R58), the Research Foundation of Education Commission of Hunan Province of China (Grant No. 17K090), the Natural Science Foundation of China (Grant No. 11401503), the Natural Science Foundation of Hunan Province of China (Grant No. 2016JJ3116), the Outstanding Youth Foundation of Hunan Educational Committee (Grant No. 16B256), the innovation project of Hunan Province of China (Grant No. Cx2016B252), and partially by the Australian Research Council Grant DP160101366.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Yuanlin Ma contributed to the conception and design of the study, downloaded the datasets, analyzed the results and has been involved in programming. Zuguo Yu gave the ideas and supervised the project. Guosheng Han has been discussing on the results. Vo Anh revised the manuscript. All authors contributed to the writing. All authors read and approved the final manuscript.

Author details
[1]Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Hunan 411105 P.R. China. [2]School of Electrical Engineering and Computer Science, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia. [3]School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia.

References

1. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281-297.
2. Chatterjee S, Grobhans H. Active turnover modulates mature microRNA activity in Caenorhabditis elegans. Nature. 2009;461(7263):546-549.
3. Wang Y, Chen X, Jiang W. Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM. Genomics. 2011; 98(2):73-78.
4. Cai R, Zhang Z, Hao Z. BASSUM: A Bayesian semi-supervised method for classification feature selection. Pattern Recognition. 2011; 44(4):811-820.
5. Weber MJ. New human and mouse microRNA genes found by homology search. Febs Journal. 2005;272(1):59-73.
6. Dezulian T, Remmert M, Palatnik JF, Huson DH. Identification of plant microRNA homologs. Bioinformatics. 2006;22(3):359-360.
7. Legendre M, Lambert A, Gautheret D. Profile-based detection of microRNA precursors in animal genomes. Bioinformatics. 2005;21(7):841-845.

8. Gautheret D, Lambert A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. Journal of Molecular Biology. 2001;313(5):1003.
9. 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.
10. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. The microRNAs of Caenorhabditis elegans. Genes & Development. 2003;17(8):991-1008.
11. Ohler U, Yekta S, Lim LP, Bartel DP, Burge CB. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. Rna-a Publication of the Rna Society. 2004;10(9):1309-1322.
12. Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of Drosphila microRNA genes. Genome Biology. 2003;4(7):R42.
13. Wang XJ, Reyes JL, Chua NH, Gaasterland T. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. Genome Biology. 2004;5(9):R65.
14. Jonesrhoades MW, Bartel DP. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Molecular Cell. 2004;14(6):787-799.
15. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 2005;434(7031):338-345.
16. Adai A, Johnson C, Mlotshwa S, Sundaresan V. Computational prediction of miRNAs in Arabidopsis thaliana. Genome Research. 2005;15(1):78-91.
17. Ng KL, Mishra SK. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics. 2007;23(11):1321-1330.
18. Batuwita R, Palade V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics. 2009;25(8):989-995.
19. Xue C, Li F, He T, Liu GP, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics. 2005;6(1):310.
20. Ding J, Zhou S, Guan J. MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. BMC Bioinformatics. 2010;11Suppl 11(Suppl 11):S11.
21. Nam JW, Shin KR, Han J, Lee Y, Kim VN, Zhang BT. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Research. 2005;33(11):3570-3581.
22. Yousef M, Khalifa W, ilhan Erkin Acar, Allmer J. MicroRNA categorization using sequence motifs and k-mers. BMC Bioinformatics. 2017;18(1):170.
23. Lopes IDO, Schliep A, Carvalho ACDLD. The discriminant power of RNA features for pre-miRNA recognition. BMC Bioinformatics. 2014;15(1):1-11.
24. 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 Research. 2007;35(Web Server issue):W339-344.
25. Huang NE, Shen Z, Long SR, Wu M, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings Mathematical Physical snd Engineering Sciences. 1998;454(1971):903-995.
26. 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.
27. Liu B, Fang L, Liu F, Wang X, Chen J, Chou KC. Identification of real microRNA precursors with a pseudo structure status composition approach. PLos One, 2015;10(3):e0121501.
28. Liu B, Fang L, Chen J, Liu F, Wang X. miRNA-dis: microRNA precursor identification based on distance structure status pairs. Molecular Biosystems. 2015;11(4):1194-1204.
29. Liu B, Fang L, Wang S, Wang X, Li H, Chou KC. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. Journal of Theoretical Biology. 2015;385(21):153-159.
30. Liu B, Fang L, Liu F, Wang X, Chou KC. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. Journal of Biomolecular Structure and Dynamics. 2016; 34(1):223-235.
31. Khan A, Shah S, Wahid F, Khan FG, Jabeen S. Identification of microRNA precursors using reduced and hybrid features. Molecular Biosystems. 2017;13(8):1640-1645.
32. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Research. 2011;39(Database issue):D152-D157.
33. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-3152.
34. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence miRNAs using deep sequencing data. Nucleic Acids Research. 2014;42(Database issue):D68-D73.
35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids research. 1997;25(17):3389-3402.
36. Yang J Y, Chen X. Improving taxonomy-based protein fold recognition by using global and local features. Proteins-structure Function & Bioinformatics. 2011;79(7):2053-2064.
37. Childs L, Nikoloski Z, May P, Walther D. Identification and classification of ncRNA molecules using graph properties. Nucleic Acids Research. 2009;37(9):e66.
38. http://www.mirbase.org/
39. Patthy L. Detecting homology of distantly related proteins with consensus sequences. Journal of Molecular Biology. 1987;198(4):567-577.
40. Fera D, Kim N, Shiffeldrim N, Zorn J, Laserson U, Gan HH, Schlick T. RAG: RNA-As-Graphs web resource. BMC Bioinformatics. 2004;5(1):1-9.
41. Gan HH, Fera D, Zorn J. RAG: RNA-As-Graphs database-concepts, analysis, and features. Bioinformatics. 2004;20(8):1285-1291.
42. Lorenz R, Bernhart SH, Zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL, Siederdissen C.

ViennaRNA Package 2.0. Algorithms for Molecular Biology. 2011;6(1):26.
43. http://igraph.org.
44. Chen YL, Li QZ. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. Journal of Theoretical Biology. 2007;248(2):377-381.
45. Yu ZG, Anh V, Wang Y, Mao D, Wanliss J. Modeling and simulation of the horizontal component of the geomagnetic field by fractional stochastic differential equations in conjunction with empirical mode decomposition. Journal of Geophysical Research. 2010;115: A10219.
46. Han GS, Yu ZG, Anh V, Krishnajith D, Tian YC. An ensemble method for predicting subnuclear localizations from primary protein structures. PLos One. 2013;8(2):e57225.
47. Friedel M, Nikolajewa S, Suhnel J, Wilhelm T. DiProDB: a database for dinucleotide properties. Nucleic Acids Research. 2009;37(Database issue):D37-D40.
48. Almuallim H, Dietterich TG. Learning with many irrelevant features. National Conference on Artificial Intelligence. AAAI Press. 1991:547-552.
49. John GH, Kohavi R, Pfleger K. Irrelevant Features and the Subset Selection Problem. Eleventh International Conference on International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 1994:121-129.
50. Huang T, Shi XH, Wang P, He Z, Feng KY, Hu L, Kong X, Li YX, Cai YD, Chou KC. Analysis and Prediction of the Metabolic Stability of Proteins Based on Their Sequential Features, Subcellular Locations and Interaction Networks. PLos One. 2010;5(6):e10972.
51. Vapnik VN, Vapnik V. Statistical learning theory. New York, Wiley. 1998.
52. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011;2(3):1-27.
53. Cristianini N, Taylor JS. An introduction to support vector machines and other kernel-based methods. Cambridge University Press, Cambridge, MA. 2000.
54. Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins-structure Function & Bioinformatics. 1995;21(4):319-344.
55. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology. 2011;273(1):236-247.
56. Chen J, Liu H, Yang J, Chou KC. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids. 2007;33(3):423-428.