BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 10 2011, pages 1368–1376
doi:10.1093/bioinformatics/btr153

Structural bioinformatics

Advance Access publication March 26, 2011

PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs

Ping Xuan[1], Maozu Guo[1,*], Xiaoyan Liu[1], Yangchao Huang[1], Wenbin Li[2] and Yufei Huang[3,*]

[1]Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, [2]Soybean Research Institute (Key Laboratory of Soybean Biology of Chinese Education Ministry), Northeast Agricultural University, Harbin 150030, P.R. China and [3]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249-0669, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MicroRNAs (miRNAs) are a set of short (21–24 nt) non-coding RNAs that play significant roles as post-transcriptional regulators in animals and plants. While some existing methods use comparative genomic approaches to identify plant precursor miRNAs (pre-miRNAs), others are based on the complementarity characteristics between miRNAs and their target mRNAs sequences. However, they can only identify the homologous miRNAs or the limited complementary miRNAs. Furthermore, since the plant pre-miRNAs are quite different from the animal pre-miRNAs, all the *ab initio* methods for animals cannot be applied to plants. Therefore, it is essential to develop a method based on machine learning to classify real plant pre-miRNAs and pseudo genome hairpins.

Results: A novel classification method based on support vector machine (SVM) is proposed specifically for predicting plant pre-miRNAs. To make efficient prediction, we extract the pseudo hairpin sequences from the protein coding sequences of *Arabidopsis thaliana* and *Glycine max*, respectively. These pseudo pre-miRNAs are extracted in this study for the first time. A set of informative features are selected to improve the classification accuracy. The training samples are selected according to their distributions in the high-dimensional sample space. Our classifier *PlantMiRNAPred* achieves >90% accuracy on the plant datasets from eight plant species, including *A.thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Physcomitrella patens, Medicago truncatula*, *Sorghum bicolor*, *Zea mays* and *G.max*. The superior performance of the proposed classifier can be attributed to the extracted plant pseudo pre-miRNAs, the selected training dataset and the carefully selected features. The ability of *PlantMiRNAPred* to discern real and pseudo pre-miRNAs provides a viable method for discovering new non-homologous plant pre-miRNAs.

Availability: The web service of *PlantMiRNAPred*, the training datasets, the testing datasets and the selected features are freely available at http://nclab.hit.edu.cn/PlantMiRNAPred/.

Contact: maozuguo@hit.edu.cn; yufei.huang@utsa.edu

Received on October 29, 2010; revised on February 28, 2011; accepted on March 23, 2011

*To whom correspondence should be addressed.

1 INTRODUCTION

Derived from hairpin precursors (pre-miRNAs), mature microRNAs (miRNAs) are non-coding RNAs that play important roles in gene regulation by targeting mRNAs for cleavage or translational repression (Bartel, 2004; Chatterjee and Grobhans, 2009). MiRNAs are involved in many important biological processes including plant development, signal transduction and protein degradation (Zhang *et al.*, 2006b).

However, systematically detecting miRNAs from a genome by experimental techniques is difficult (Bartel, 2004; Berezikov *et al.*, 2006). As an alternative, the computational prediction methods are used to analyze the genomic DNA and to obtain the putative candidates for experimental verification. Since many miRNAs are evolutionarily conserved in multiple species, methods that use comparative genomics to identify putative miRNAs have been presented. MirFinder (Bonnet *et al.*, 2004) predicted the potential miRNA of in the *Arabidopsis thaliana* genome. It is based on the conservation of short sequences between the genomes of *Arabidopsis* and *Oryza sativa*. MicroHARVESTOR (Dezulian *et al.*, 2006) can identify candidate plant miRNA homologs for a given query miRNA. The approach is based on a sequence similarity search step followed by a set of structural filters. The miRNAs of *Vigna unguiculata* are identified through homology alignment of highly conserved miRNAs in multiple species (Lu and Yang, 2010). Although these methods are effective in identifying new conserved miRNAs, they cannot discover novel miRNAs with less homology. On the other hand, since miRNAs in plants often have near perfect matches to their target mRNAs, methods that are based on the complementarity characteristics have also been proposed. MIRcheck (Jones-Rhoades and Bartel, 2004) searches for the new miRNAs whose target sites are conserved in *A.thaliana* and *O.sativa*. FindMiRNA (Adai *et al.*, 2005) predicts potential *Arabidopsis* miRNAs from candidate pre-miRNAs that have corresponding target sites within transcripts. In order to decrease the number of miRNA candidates, the predicted candidates are required to have orthologs in rice. MiMatcher (Lindow and Krogh, 2005) also predicts the plant miRNAs through exploiting the complementarity of plant miRNAs. However, since the candidates that are complementary to target mRNAs are enormous within intergenic regions and introns, rigorous criteria such as conservation among multiple species are introduced to significantly reduce the

number of candidates. This practice also considerable reduces the chance of discovering new miRNAs.

As an alternative, the *ab initio* methods have been developed to distinguish real pre-miRNAs from pseudo pre-miRNAs. The real pre-miRNAs and pseudo pre-miRNAs are used to train the classification models including support vector machines (SVM) (Batuwita and Palade, 2009; Ng and Mishra, 2007; Sewer *et al.*, 2005; Xue *et al.*, 2005), probabilistic co-learning model (Nam *et al.*, 2005), naïve Bayes (Yousef *et al.*, 2006), random forest (Jiang *et al.*, 2007) and kernel density estimation (Chang *et al.*, 2008). These trained classification models can be then used to classify real pre-miRNAs from pseudo pre-miRNAs. However, they all are trained by the human pre-miRNAs and pseudo pre-miRNAs, and mainly used to identify human pre-miRNAs. *Triplet-SVM* (Xue *et al.*, 2005) is the only one that has been tested on the pre-miRNAs from *A.thaliana* and *O.sativa*. As the plant pre-miRNAs differ greatly from the animal pre-miRNAs, *triplet-SVM* cannot achieve high classification accuracy for the plant species including *A.thaliana* and *O.sativa*.

Since nearly all the pre-miRNAs in plants and animals have the stem–loop hairpin structures, this characteristic is widely used as an important feature in the *ab initio* methods. However, the plant pre-miRNAs usually have more complex secondary structures than the animal pre-miRNAs and the structures have not been considered in existing methods. We propose a computational classification approach that considers the unique characteristics of plant pre-miRNAs. To construct a comprehensive training dataset, the new pseudo plant pre-miRNAs are extracted from the protein coding regions of the *A.thaliana* and *Glycine max* genomes, based on which an efficient SVM classifier is constructed to classify the real/pseudo plant pre-miRNAs.

2 METHODS

2.1 Features of plant pre-miRNAs

Recent research indicates that pre-miRNAs in animals and plants have many features in both primary sequence and secondary structure. These features can be used to classify real plant pre-miRNAs and pseudo hairpins with an *ab initio* method.

miPred (Ng and Mishra, 2007) extracted 29 global and intrinsic folding features from human real and pseudo pre-miRNAs. These features include (i) 17 base composition variables: 16 dinucleotide frequencies, that is *%XY* where $X,Y \in \{A,C,G,U\}$ and *%G+C* content; (ii) six folding measures: adjusted base pairing propensity *dP* (Schultes *et al.*, 1999), adjusted minimum free energy (MFE) of folding denoted as *dG* (Freyhult *et al.*, 2005; Seffens and Digby, 1999), adjusted base pair distance *dD* (Freyhult *et al*, 2005; Moulton *et al*., 2000), adjusted Shannon entropy *dQ* (Freyhult *et al.*, 2005), MFE Index 1 $MFEI_1$ (Zhang *et al.*, 2006a) and MFE Index 2 $MFEI_2$; (iii) one topological descriptor which is the degree of compactness *dF* (Gan *et al.*, 2004); (iv) five normalized variants of *dP*, *dG*, *dQ*, *dD* and *dF*: *zP*, *zG*, *zQ*, *zD* and *zF*.

In addition to the 29 features listed above, *microPred* (Batuwita and Palade, 2009) extracted 19 new features. These features are (i) two MFE-related features: MFE Index 3 $MFEI_3$ and MFE Index 4 $MFEI_4$; (ii) four RNAfold-related features: normalized ensemble free energy (NEFE), frequency of the MFE structure Freq, structural diversity denoted as Diversity and a combined feature Diff; (iii) six thermodynamical features: structure entropy *dS* and *dS/L*, structure enthalpy *dH* and *dH/L*, melting energy of the structure T_m and T_m/L, where *L* is the length of pre-miRNA sequence; (iv) seven base pair-related features: $|A-U|/L$, $|G-C|/L$, $|G-U|/L$, average base pairs per stem *Avg_BP_Stem*, $\%(A-U)$*/n_stems*, $\%(G-C)$*/n_stems*

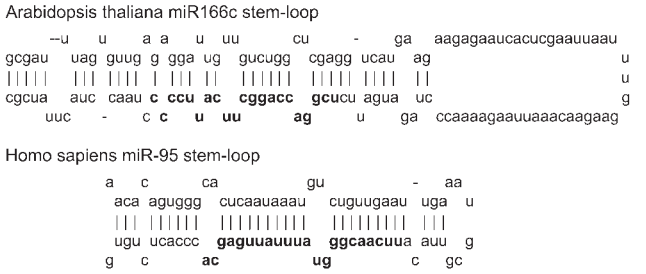

Fig. 1. *Arabidopsis* pre-miRNA ath-miR166c and human pre-miRNA hsa-mir-95. Their secondary structures are obtained from miRBase. The nucleotides of the mature miRNA are displayed in bold.

and $\%(G-U)$*/n_stems*, where *n_stems* is the number of stems in the secondary structure.

The plant pre-miRNAs have more diversities than the animal pre-miRNAs. First, the pre-miRNAs in animals are typically 60–80 nt (Ambros *et al.*, 2003), whereas the length of pre-miRNAs in plants ranges from 60 to >400 nt (Smalheiser and Torvik, 2005). Secondly, the molecular characteristics of plant pre-miRNAs are also different from the animal pre-miRNAs. The former have great varieties in secondary structures. For instance, like *Homo sapiens* miR-95, the central loops of the pre-miRNAs in animals are typically 3–20 nt in length (Nam *et al.*, 2005). The loops of plant pre-miRNAs have great varieties in length, such as the loop in *A.thaliana* miR166c. This is shown in Figure 1. Further, some plant pre-miRNAs contain the big bugles, e.g. *G.max* miR166b in Figure 2b. Moreover, there are big unmatched parts in some plant pre-miRNAs, e.g. *Physcomitrella patens* miR166i in Figure 2c.

Stems of plant pre-miRNAs are relatively stable and conserved. Central loops, big bugles and big unmatched parts have great diversities and are not conserved. Therefore, they are removed to obtain the stems. Two novel MFE-related features are proposed and calculated for stems, since they are more stable. (i) MFE Index 5: $MFEI_5 = MFE/\%G+C\_S$, where *%G+C_S* is the GC content in the stems. (ii) MFE Index 6: $MFEI_6 = MFE/stem\_tot\_bases$, where *stem_tot_bases* is the number of base pairs in the stems. (iii) Average number of mismatches per 21-nt window: *Avg_mis_num* = *tot_mismatches/n_21nts*, where *tot_mismatches* is the total number of mismatches in the 21-nt sliding windows (which is roughly the length of a mature miRNA region and naturally has fewer than four successive mismatches) and *n_21nts* is the number of sliding windows in a stem.

For the 48 existing features and 3 new features, the 17 dinucleotide frequencies features describe the sequential characteristic. The remaining 31 features are mainly related to the thermodynamics and stability of the secondary structures of the hairpins. The current research indicates that the structural characteristic is also significant for distinguishing the hairpins of real/pseudo pre-miRNAs. Therefore, 32 structured triplet composition features [frequencies of "U(((","A((.", etc., which were defined in *triplet-SVM* (Xue *et al.*, 2005)] are extracted from the pre-miRNAs. Since nearly all mature miRNAs are located in stems, these 32 features are extracted again from stems and denoted as "U(((_S", "A((._S", etc.

The transformation of secondary structure of a pre-miRNA is shown in Figure 2. First, loops, big bugles and big unmatched parts of pre-miRNAs are removed, in order to capture the features of stable stems. Then, the 5′-arm and 3′-arm are connected by a linker sequence, 'LLLLLLLL'. It is helpful to unify the length of loops in all the real/pseudo pre-miRNAs. Since 'L' is not an RNA nucleotide, it does not match with any nucleotide and prevents nucleotides in 5′-arm and 3′-arm from binding with sequence-specific linker sequences. Finally, the three new features ($MFEI_5$, $MFEI_6$ and *Avg_mis_num*) and the 32 structure-related features are extracted from the linked stems.

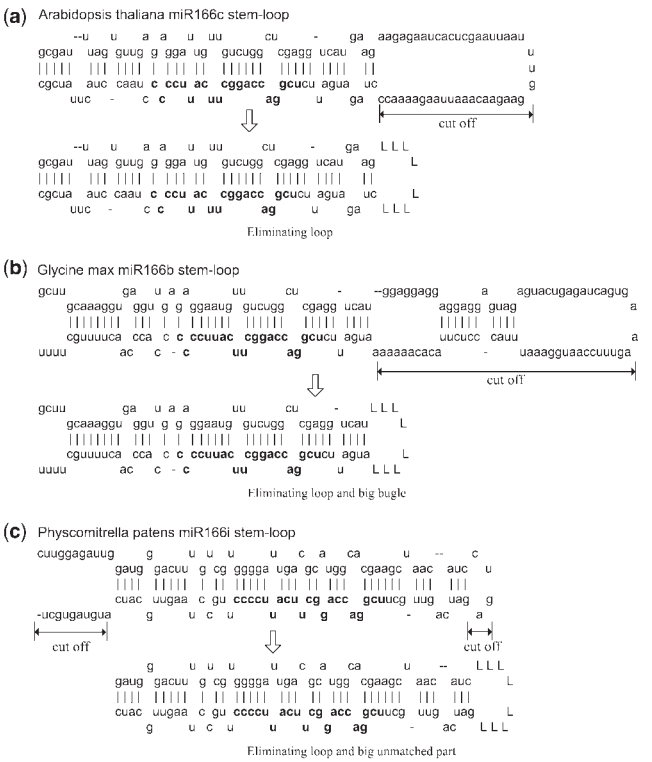

Fig. 2. Transforming secondary structures of ath-miR166c, gma-miR166b and ppt-miR166i. (**a**) The big loop of ath-miR166c is removed and replaced with 'LLLLLLLL'. (**b**) Some pre-miRNAs in plants, such as gma-miR166b, have big bugles which are near the loops. The big bugles are removed and the loop is replaced with 'LLLLLLLL'. (**c**) The big unmatched part in the left end of the stem is removed.

In total, 115 features are obtained from each real/pseudo plant pre-miRNAs. They include redundant features that do not contribute to classification. The algorithm based on graph is presented to eliminate the redundant features. In the end, the discriminative feature subset is selected to achieve the best classification accuracy.

2.2 SVM

Due to the excellent generalization ability of SVM, we use it to classify real/pseudo plant pre-miRNAs with high-dimensional (115-dimensional) feature vectors. Given a training dataset S, each $x_i \in S$ ($i=1,\ldots,N$) is a feature vector of real/pseudo pre-miRNA with corresponding labels z_i ($z_i=+1$ or -1, real pre-miRNA or pseudo pre-miRNA, respectively). SVM constructs a decision function (classification of unknown sequence x with stem–loop structure),

$$g(x)=\mathrm{sgn}\left(\sum_{i=1}^{N} z_i \alpha_i k(x,x_i)+w_0\right) \quad (1)$$

where α_i is the coefficient to be learned ($0 \le \alpha_i \le C$) and k is a kernel function. In our study, a radial basis function (RBF) kernel is used, where the parameter γ determines the similarity level of the features so that the classifier becomes optimal.

$$k(x,x_i)=\exp(-\gamma \|x-x_i\|) \quad (2)$$

The penalty parameter C and the RBF kernel parameter γ are tuned based on the training dataset using the grid search strategy in libSVM (version 2.9).

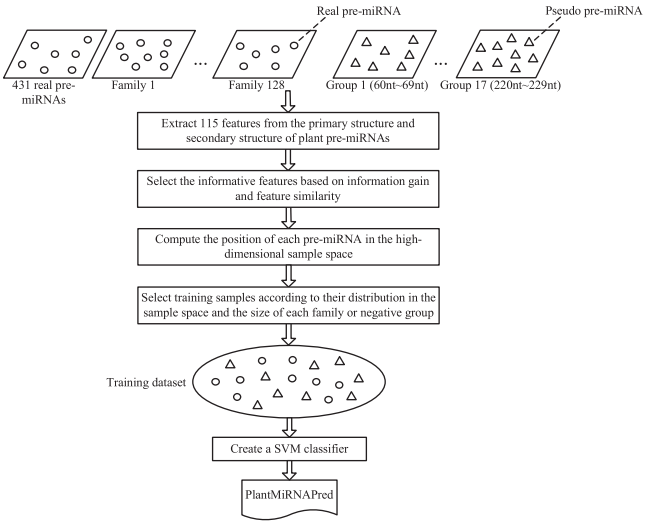

Fig. 3. Construction of SVM classifier based on feature selection and sample selection. Each circle/triangle represents a real/pseudo pre-miRNA.

2.3 Classification based on SVM

Both the features of real/pseudo pre-miRNAs and the training samples are important for constructing an efficient SVM classifier. As shown in Figure 3, we propose a classification method based on SVM. All the 2043 plant pre-miRNAs from miRBase 14 (covers 29 plant species) are collected as positive datasets. The homologous pre-miRNAs among different species are gathered into the same miRNA gene family by RFam (Gardner *et al.*, 2009). Most of the pre-miRNAs in the same families are similar. Therefore, the representative pre-miRNAs are selected from the 128 plant miRNA families of miRBase 14 (including 1612 real pre-miRNAs), as the positive training samples. Since the remaining 431 pre-miRNAs do not belong to any of miRNA families, they are used as the positive training samples. The negative dataset consists of the 17 groups. Each group is composed of pseudo pre-miRNAs from the genome segments of *A.thaliana* and *G.max* (see Section 3 for details). (i) The 115 features are extracted from the primary sequence and secondary structure of pre-miRNAs and their stems. (ii) The redundant features are eliminated and the informative feature subset is selected through calculating the information gain of features and the similarity between any two features. (iii) The positive/negative training samples are selected according to their distribution in the high-dimensional sample space, the size of each family and the size of each negative sample group. (iv) An SVM-based classifier named *PlantMiRNAPred* is trained by using these samples to classify real pre-miRNAs and pseudo pre-miRNAs. The feature selection and sample selection modules are implemented in Java. The web service of classifying plant pre-miRNAs is developed in PHP on the Linux platform.

2.4 Feature subset selection

Feature selection aims to select a group of informative features which can retain most information of original data and distinguish each sample in the dataset. The feature selection method considers information gain and feature redundancy.

Information gain: since all the features of pre-miRNAs are discrete, the discrimination ability of a feature is measured by information gain based on Shannon entropy. Suppose a feature of pre-miRNAs is x and the entropy of x is denoted as $H(x)$. When the value of feature y is given, the conditional entropy is $H(x|y)$. IG(c,x) is the information gain of feature x relative to the class attribute c (Quinlan, 1993). Since classification of real or pseudo pre-miRNAs is binary classification problem, c is assigned to 1 (real pre-miRNA)

or −1 (pseudo pre-miRNA).

$$\begin{aligned} \mathrm{IG}(c,x) &= H(c) - H(c|x) \\ &= \sum_{c,x} p(cx) \log_2 \frac{p(cx)}{p(c)p(x)} \end{aligned} \quad (3)$$

Suppose that the complete feature set is $X = \{x_1, x_2, \ldots, x_{115}\}$ and the class attribute is c. The values of 115 features are obtained from each real pre-miRNA (1906 real pre-miRNAs in total) and c is set to 1. Also, the values of 115 features are obtained from each pseudo hairpin (2122 pseudo hairpins in total) and c is set to −1. The information gain of feature $x_i (1 \le i \le 115)$ is calculated and denoted as $\mathrm{IG}(c,x_i)$. The features with greater information gain are given higher preference.

However, the 115 features still include redundant features, inclusion of which will not improve the classification accuracy. To identify the redundant features, the feature similarity is used to measure the similarity between two features.

Feature similarity: let $\mathrm{Sim}(x,y)$ represent the similarity between features x and y and it is defined as,

$$\mathrm{Sim}(x,y) = 2\left[\frac{\mathrm{IG}(x,y)}{H(x)+H(y)}\right] \quad (4)$$

where $\mathrm{IG}(x,y)$ denote the information gain of y respect to x. $\mathrm{Sim}(x,y)$ ranges from 0 to 1. $\mathrm{Sim}(x,y)$ equal to 0 means that x and y are completely irrelevant. $\mathrm{Sim}(x,y)$ equal to 1 means that x and y are completely relevant. When $\mathrm{Sim}(x,y)$ is greater than a threshold ε, x or y is a redundant feature. Keeping both features does not improve the classification performance. In such situation, the feature with greater information gain is kept and the other one is dropped. In order to effectively eliminate the redundant features when multiple features are correlated, a redundant feature graph G is constructed. We propose an algorithm based on graph. Suppose that the redundant feature graph is $G = (V, E)$. Each node v_i ($v_i \in V$) represents the feature x_i. The weight of node v_i is the $\mathrm{IG}(c,x_i)$. If the similarity between two nodes v_i and v_j is more than the threshold $\varepsilon(\varepsilon = 0.49)$, x_i or x_j is redundant. Then a new edge is added to connect the two nodes. The weight of the edge between v_i and v_j is $\mathrm{Sim}(x_i,x_j)$. Here, ε is determined by the experiments and the prior experience (Ng and Mishra, 2007).

The process of eliminating redundant features is illustrated by an example. As shown in Figure 4, a redundant feature graph G consists of eight features, including $x_1, x_2, \ldots$ and x_8. Suppose that groups of redundant features exist, where features within a group are redundant to one another but are independent to the remaining features. If we assume that there are three groups, then the graph G is composed of three subgraphs: SG_1, SG_2 and SG_3. Feature selection weight (FSW) of each feature is first calculated. Suppose that k edges are adjacent to v_i. FSW of v_i is defined as $\mathrm{FSW}_{vi} = \mathrm{Sum}$ of weights of the k edges (that are connected with v_i) + weight of v_i. The feature node v_x with the most FSW is selected. The nodes adjacent to v_x are the redundant features and should be deleted. In the subgraph SG_2, FSW of $x_3 = (0.71 + 0.51) + 0.42 = 1.64$. FSWs of x_4, x_5, x_7 and x_8 are 1.48, 0.8, 1.52 and 1.15, respectively. Therefore, x_3 is selected, and the adjacent x_5 and x_7 are deleted. The bolded nodes in Figure 4 are the selected feature nodes. Next, the FSWs of the remaining nodes x_4 and x_8 in the current SG_2 are calculated again. Since x_8 has greater FSW, it is selected. In the same way, in SG_1, x_1 is selected and x_2 is deleted. In SG_3, x_6 is an independent node. As no other node can represent the independent node, x_6 is also selected. At last, a feature subset $\{x_1, x_3, x_6, x_8\}$ is obtained and the redundant features x_2, x_4, x_5 and x_7 are eliminated. In addition, if two nodes (v_y and v_z) are with the maximum FSW in a subgraph, the node weights of v_y and v_z are compared and the node with greater weight is selected. The algorithm of eliminating redundant features is described in Figure 5.

The proposed algorithm is applied to eliminating redundant features among the 115 features. We found that two pairs of attributes are strongly correlated: *dQ* versus *dD*, and *dG* versus *NEFE*. This observation is consistent with the result in *miPred*, which indirectly confirms the result of eliminating features. In addition, we found two new strongly correlated pairs of attributes: *dH* versus *dS*, and *dH/L* versus *dS/L*, and *dQ*, *dG*, *dH* and *dH/L* are selected due to their higher selection weights.

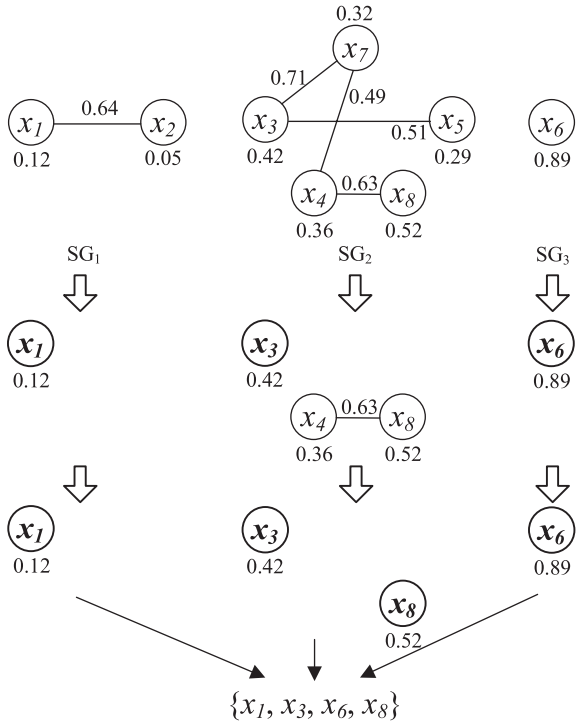

Fig. 4. Eliminating redundant features based on the graph G.

ALGORITHM: Graph based algorithm for eliminating redundant features
Input: $G(V,E)$ composed of n subgraphs $SG_1, SG_2, \ldots, SG_n$
Output: Feature subset S without redundant features
1 While any subgraph of SG_i still has edges
2 do
3 Calculate the *FSW* of each node v_i in the SG_i ($1 \le i \le n$)
4 If there is only one node v_x with maximum *FSW* in the SG_i
5 Add the node v_x into S and v_x is removed from SG_i
6 End if
7 If there are multiple nodes (≥2) with equivalent maximum *FSW* in the SG_i
8 Compare the node weight of these nodes
9 Add the node v_x with the greatest node weight into S and v_x is removed from SG_i
10 End If
11 Delete the nodes which are adjacent to v_x and the edges connected to the deleted nodes
12 End While

Fig. 5. Algorithm of eliminating redundant features.

Initially, the 115 features are categorized into three feature subsets: (i) primary sequence-related feature subset $S_1 = \{\%G+C, \%XY | X,Y \in \{A,C,G,U\}\}$ (17 features); (ii) secondary structure-related feature subset S_2 = {"A(((",…"U…","A(((_S",…,"U…_S"} (64 features); (iii) energy- and thermodynamics-related feature subset S_3 = {*dP*, *dG*, …, *zF*, $MFEI_5$, $MFEI_6$, *Avg_mis_num*} (34 features). Supplementary Table S1 illustrates the name and the classification of 115 features. After eliminating the redundant features, the three subsets are denoted as S'_1, S'_2, and S'_3. For each subset, the remaining features are sorted by information gain in descending order. The features with information

gain greater than a threshold λ ($\lambda_1=0.136$, $\lambda_2=0.083$, $\lambda_3=0.0159$) are selected for the classification. λ is determined by the experiments and the prior experience (Ng and Mishra, 2007). In the end, a total of 68 features are selected and listed in Section 3.

2.5 Training samples selection

The classification performance of SVM is highly dependent on the selection of training dataset. First, we noted that the real pre-miRNAs from the same species and the same miRNA families are highly similar to one another. These redundant positive samples should be removed from training samples to avoid over-fitting. Secondly, current research has shown that training an SVM classifier with an imbalance positive and negative dataset would result in poor classification performance with respect to the minority class (Weiss, 2004). So we select the appropriate proportion of representative real/pseudo pre-miRNAs to construct positive/negative training dataset.

The sample selection method (miSampleSelection) selects the positive/negative training samples according to the sample distribution in the positive families/negative groups. As shown in Figure 3, the real pre-miRNAs are selected based on the sample distribution in the 128 families. Since 68 informative features have been selected, each sample is denoted as a 68-dimensional feature vector. Suppose the feature vector of a sample is v and there are M ($M=128$) families. The vector set of central points is $C=\{c_1,c_2,\ldots,c_M\}$, where c_i represents the feature vector of the central point in the i-th family. The sample selection process of the i-th family is as follows.

(1) Assume that the number of samples in the i-th family is N_i. v_k is the feature vector corresponding to the k-th sample. c_i is then calculated as,

$$c_i=\frac{1}{N_i}\sum_{k=1}^{N_i}v_k \tag{5}$$

(2) The distance between the k-th sample (real pre-miRNA) v_k and the central point c_i is denoted as $d_{v_kc_i}$. v_k^t means the transpose of vector v_k. Then, the radius of the i-th family is r_i, where $r_i=\max(d_{v_kc_i})(1\le k\le N_i)$.

$$d_{v_kc_i}=1-\frac{v_k^t\cdot c_i}{v_k^t\cdot v_k+c_i^t\cdot c_i-v_k^t\cdot c_i} \tag{6}$$

(3) Suppose that the selection rate of sample space is $1/n$. That is, N_i/n samples in the i-th family are selected. The number of the selected samples is denoted as $P_i=N_i/n$.

(4) Suppose that c_i is the center of a circle, draw two circles with radius $0r_i$ and $(1/P_i)r_i$, respectively. The region between these two circles is denoted as A_0. The degree of coverage for each sample s in A_0 is calculated and denoted as $C(s)$. $C(s)$ represents the number of samples in A_0 whose nearest neighbor sample is s. The sample s with the greatest $C(s)$ value is selected as a training sample.

(5) We set $(1/P_i)r_i$ as the step length and compute the degree of sample coverage in the region A_k between two circles with the radius $(1/P_i)kr_i$ and $(1/P_i)(k+1)r_i(1\le k\le P_i-1)$, respectively. The sample in A_k with the largest degree of coverage is selected.

The positive training dataset is composed of the samples selected from the 128 families. For 431 pre-miRNAs that do not belong to any of miRNA families, they are added into the positive training dataset. The process of selecting negative training samples is similar. The negative samples are composed of pseudo hairpins grouped by length. There are 17 groups in total, where the 60nt_Group refers to pseudo hairpins of length from 60 to 69 nt, the 220nt_Group refers to pseudo hairpins of length from 220 to 229 nt, etc. Seventeen groups of the negative dataset correspond to the families of the positive dataset. The negative training samples are selected in the same way as that of the positive samples.

3 RESULTS AND DISCUSSION

3.1 Data collection

A classifier of pre-miRNAs should distinguish real plant pre-miRNAs from pseudo plant hairpins. The positive dataset is composed of known plant pre-miRNAs, while the negative dataset is composed of both pseudo *A.thaliana* hairpins and pseudo *G.max* hairpins.

Positive dataset: there are 2043 known plant pre-miRNAs in the miRNA database miRBase 14 (http://www.mirbase.org/). Rfam (http://rfam.janelia.org/) grouped the available real pre-miRNAs into a set of families by means of multiple sequence alignments. A miRNA gene family is composed of the homologous pre-miRNAs from different species. One thousand six hundred and twelve pre-miRNAs belong to 128 miRNA families and 431 pre-miRNAs do not belong to any of miRNA families. Two thousand forty-three real pre-miRNAs are chosen as the positive sample dataset. They are from *A.thaliana*, *O.sativa*, *Populus trichocarpa*, *P.patens*, *Medicago truncatula*, *Sorghum bicolor*, *Zea mays*, *G.max* and other 21 plant species. As shown in Figure 6, the top 22 families consist of 1066 plant pre-miRNAs. Supplementary Table S2 shows the pre-miRNAs distribution and the species distribution in all the 128 plant families. Each family contains at least two pre-miRNAs and covers at least one plant species. After eliminating the special sequences with complex secondary structures, 1906 real pre-miRNAs remain in the positive dataset.

Negative dataset: the complete genome sequence of *A.thaliana* was released in 2000 (AGI, 2000). The sequences of 20 chromosomes in *G.max* genome are released in 2010 (Schmutz *et al.*, 2010). *Arabidopsis thaliana* is a typically model plant and *G.max* is one of the most important crop. The negative samples are extracted from these two species. Since almost all reported miRNAs are located in the untranslated regions or intergenic regions, the pseudo hairpins are extracted from protein coding sequences (CDSs) of *A.thaliana* and *G.max*. The CDSs of *A.thaliana* and *G.max* are downloaded from the plant database Phytozome 6 (http://www.phytozome.net/).

The ratios of pre-miRNAs with different length in the 2043 real pre-miRNAs are listed in Figure 7. It is found that most of known plant pre-miRNAs in length ranges from 60 to 220 nt. Thus, a sliding window of width ranging from 60 to 220 nt is used to scan the CDSs to produce sequence segments. The secondary structures of the sequence segments are predicted by RNAfold from the Vienna

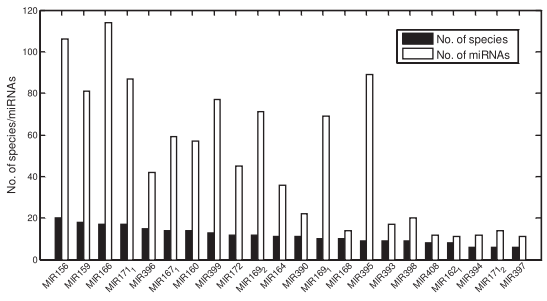

Fig. 6. The 128 families are ranked by the size of species that a miRNA gene family covers. The distribution of the top 22 miRNA families (containing 1066 pre-miRNAs) is shown. The names of miRNA familes are listed in the x-axis. The y-axis represents the number of species/miRNAs.

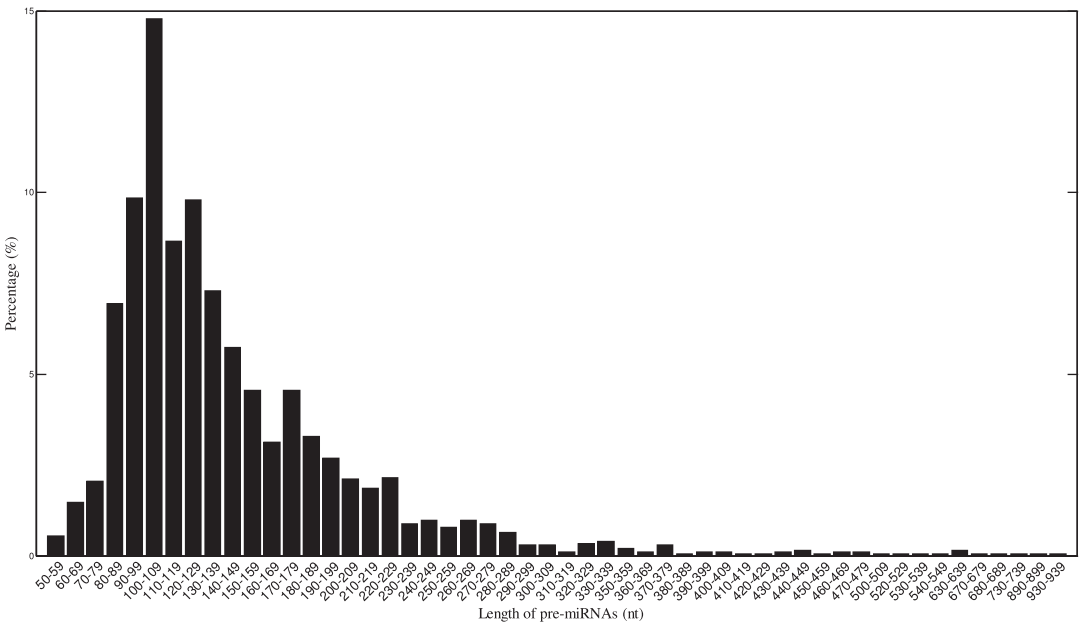

Fig. 7. Ratio of plant pre-miRNAs in different length. The *x*-axis represents the length of real pre-miRNAs. The *y*-axis represents the proportion of the pre-miRNAs in length among 60–69 nt, 70–79 nt, …, 220–229 nt accounting for the sum of real pre-miRNAs.

package (Hofacker *et al.*, 1994). The sequence segments should be folded into stem–loop structures. Further, they should satisfy three criteria on the number of base pairs in hairpins, *%G+C* and *MFEI*. The criteria are determined by observing real plant pre-miRNAs in length among 60–69 nt, 70–79 nt, etc., till 220–229 nt. For instance, the criteria for selecting the pseudo miRNAs in length from 60 to 69 nt are: minimum of 19 base pairings in hairpin structure, *%G+C* >0.242 and <0.825 and *MFEI* >0.522 and <1.39. The length of the sliding windows changes from 60 to 69 randomly. Supplementary Table S3 listed all the criteria for different lengths. Therefore, the extracted pseudo pre-miRNAs are similar to the real pre-miRNAs.

The negative samples (pseudo pre-miRNAs) are collected according to the proportion of the real pre-miRNAs of different lengths. For example, suppose the ratio of real pre-miRNAs in length 70–79, 80–89, 90–99 and 100–109 nt is 0.02:0.08:0.12:0.20. Then the negative samples in different length are added into the negative dataset in corresponding proportion. In total, 2122 pseudo pre-miRNAs are collected as negative dataset.

Positive and negative training dataset: the 980 real pre-miRNAs and 980 pseudo pre-miRNAs are selected by the sample selection algorithm. The final training dataset includes a total of 1960 samples. It is denoted as 1960 training dataset.

Positive and negative testing dataset: *A.thaliana*, *O.sativa*, *P.trichocarpa*, *P.patens* and *M.truncatula* are typical model plants. *Sorghum bicolor*, *Z.mays* and *G.max* are important crops. To date, relatively more miRNAs are identified from the eight species listed above. Thus eight groups of testing datasets are created to evaluate our classifier. The first group is composed of all the 180 *A.thaliana* (ath) pre-miRNAs, referred to as ath dataset. The 397 *O.sativa* (osa) pre-miRNAs, 233 *P.trichocarpa* (ptc) pre-miRNAs, 211 *P.patens* (ppt) pre-miRNAs, 106 *M.truncatula* (mtr) pre-miRNAs, 131 *S.bicolor* (sbi) pre-miRNAs, 97 *Z.mays* (zma) pre-miRNAs and 83 *G.max* (gma) pre-miRNAs are used to construct the osa dataset, ptc dataset, ppt dataset, mtr dataset, sbi dataset, zma dataset and gma dataset, respectively. The remaining 1142 pseudo pre-miRNAs from 2122 pseudo pre-miRNAs (excluding the 980 pseudo pre-miRNAs) are used as *1142 negative testing dataset*. The 191 *A.lyrata* (*updated aly dataset*) and 118 *G.max* pre-miRNAs (*updated gma dataset*) were newly reported by miRBase 15–16 when this work was nearly completed.

3.2 Evaluation method

The informative feature subset and the training samples were used to construct the classifier *PlantMiRNAPred*. The prediction result of *PlantMiRNAPred* can be either one of the following four outcomes: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). The sensitivity (SE), specificity (SP), geometric mean (Gm) and total prediction accuracy (Acc) for assessment of the prediction system are as follows,

$$\mathrm{SE}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\quad \mathrm{SP}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\quad \mathrm{Gm}=\sqrt{\mathrm{SE}\times\mathrm{SP}},\quad \mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{7}$$

where SE is the proportion of the positive samples (real pre-miRNAs) correctly classified, and SP is the proportion of the negative samples (pseudo pre-miRNAs) correctly classified.

3.3 Feature subset selection result

The selected 68 informative features by feature selection process and the corresponding information gain are listed in Table 1. They are ranked by their normalized information gain.

Table 1. Selected features ranked by their information gain

Rank	AttrName	IG(c, attr)	Rank	AttrName	IG(c, attr)
1	Tm	1.0	35	G.((_S	0.1365
2	$MFEI_1$	0.6362	36	%GG	0.1358
3	$MFEI_4$	0.4837	37	*zF*	0.1325
4	$MFEI_6$	0.3959	38	G(.(	0.1302
5	$MFEI_3$	0.3759	39	%G+C	0.1223
6	*dG*	0.3598	40	G((.	0.1218
7	A(((	0.2964	41	G((._S	0.1148
8	A(((_S	0.2804	42	U...	0.1106
9	%(G-C)/stems	0.2653	43	G(._S	0.1061
10	C(((	0.2411	44	*dP*	0.1032
11	G(((	0.2274	45	A...	0.1014
12	U(((	0.2269	46	G(..	0.1009
13	U(((_S	0.2187	47	G..(	0.0989
14	C(((_S	0.2154	48	C(..	0.0982
15	G(((_S	0.2096	49	*dH*	0.0979
16	C((.	0.1979	50	*zP*	0.0971
17	$MFEI_5$	0.1948	51	Avg_mis_num	0.0967
18	C.((	0.1922	52	U..._S	0.0860
19	*dH*/*L*	0.1822	53	C..(	0.0855
20	C(.(	0.1822	54	G(.._S	0.0852
21	%GC	0.1769	55	U..(	0.0832
22	%UA	0.1767	56	$MFEI_2$	0.0774
23	%AU	0.1742	57	*dQ*	0.0662
24	%AA	0.1727	58	Avg_Bp_Stem	0.0662
25	%CG	0.1678	59	Diff	0.0618
26	C(._S	0.1663	60	Freq	0.0607
27	Tm/*L*	0.1618	61	Diversity	0.0606
28	*zG*	0.1592	62	|G-C|/*L*	0.0597
29	G.((	0.1558	63	*zD*	0.0498
30	C((._S	0.1556	64	*dF*	0.0376
31	%UU	0.1554	65	|G-U|/*L*	0.0374
32	%(A-U)/stems	0.1536	66	|A-U|/*L*	0.0305
33	%CC	0.1528	67	%(G-U)/stems	0.0171
34	C.((_S	0.1491	68	*zQ*	0.0159

It has been well studied that the stem–loop structures of plant pre-miRNAs is thermodynamically stable. Most of the selected features are related to the thermodynamic stability of the secondary structures. It indirectly confirms the effectiveness of the selected features. There are some features with suffix _S and three new features ($MFEI_5$, $MFEI_6$, *Avg_mis_num*) in the selected feature subset. It shows the significance of extracting the new features for the stems. In addition, some features and the corresponding features obtained from stems appear in pairs, such as A(((and A(((_S, U(((and U(((_S, etc. It indicates that there is an obvious difference between two features in any pair of the features listed above. Also, both features are important for the classification of real/pseudo pre-miRNAs.

In order to validate the efficiency of the feature selection method, we tested the classification accuracies of 68 features, 80 features (containing no features of stems), 51 features (containing no structural features) and all 115 features, respectively. For each feature subset, 980 real pre-miRNAs and 980 pseudo pre-miRNAs were selected by the sample selection method to train an SVM classifier. These four SVM classifiers were tested by performing five-fold cross-validation. With five-fold cross-validation, all pre-miRNAs in the training dataset were divided into five equal subsets, four of which were used for training the classifier, while the left out subset was used for validation. We performed 10 repeated evaluations for each testing dataset and averaged the results.

Table 2. Classification results on different feature subsets

Feature subset	Classification results (%)			
	SE	SP	G_m	Acc
68 features	91.93	97.84	94.84	94.39
80 features	89.08	92.82	90.93	90.63
51 features	88.78	92.96	90.85	90.51
All 115 features	90.31	94.54	92.40	92.06

Table 3. Classification results with different sample selection methods

Sample Selection methods	Dataset	Classification results (%)			
		SE	SP	G_m	Acc
miSampleSelection	1960 training dataset	91.93	97.84	94.84	94.39
Random Selection	1960 random dataset	89.69	93.25	91.45	91.17

The classification results are summarized in Table 2. The classification performances of 80 features and 51 features are worse than that of 115 features. It indicates that the stem-related features and the structural features are absolutely necessary. Obviously, the classifier trained by the selected 68 features achieves the best classification performance. It shows the importance of feature selection during construction of the efficient classifier.

3.4 Training sample selection result

The 980 positive samples and 980 negative samples with 68 features were selected by our sample selection method miSampleSelection to construct the classifier *PlantMiRNAPred*. Moreover, the equal number of real/pseudo pre-miRNAs were randomly selected from the positive/negative dataset, referred to as 1960 random dataset. The performance of *PlantMiRNAPred* was compared with the classifier trained with 1960 random dataset. As shown in Table 3, five-fold cross-validation was performed on each training dataset.

The classifier trained by 1960 training dataset achieves much higher sensitivity and specificity. It demonstrates that miSampleSelection is effective for improving the classification accuracy. In addition, the classifier which was trained by the 1960 random dataset achieves excellent classification accuracy. It further confirms that the selected 68 features are sufficient to ensure the classification performance.

3.5 Comparison with *triplet-SVM* and *microPred*

Triplet-SVM is the only *ab initio* method that has been tested with the pre-miRNAs from *A.thaliana* and *O.sativa*. Therefore, we compared *PlantMiRNAPred* with *triplet-SVM*. The program of *triplet-SVM* was downloaded from Xue's web site (http://bioinfo.au.tsinghua.edu.cn/mirnasvm/). The eight testing datasets composed of known pre-miRNAs from eight species were tested to evaluate the ability of identifying the real pre-miRNAs.

Table 4. Classification results on different testing datasets

Testing dataset	Type	Size	Accuracy (%)		
			PlantMiRNA Pred	*Triplet-SVM*	*micro Pred*
ath dataset	Real	180	92.22	76.06	89.44
osa dataset	Real	397	94.21	75.54	90.43
ptc dataset	Real	233	91.85	75.21	84.98
ppt dataset	Real	211	92.42	71.49	89.57
mtr dataset	Real	106	100	80.18	95.28
sbi dataset	Real	131	98.47	69.51	94.66
zma dataset	Real	97	97.94	66.97	93.81
gma dataset	Real	83	98.31	74.12	86.75
1,142 negative testing dataset	Pseudo	1142	98.59	86.34	93.61
updated aly dataset	Real	191	97.91	70.98	91.62
updated gma dataset	Real	118	98.31	79.66	93.22

The 1142 negative testing dataset was tested to evaluate the ability of identifying the pseudo hairpins. The Updated dataset was also tested to observe the ability of discovering new plant pre-miRNAs.

We performed evaluations for all the testing datasets and illustrated the results in the Table 4. *PlantMiRNAPred* is nearly 18% better than *triplet-SVM* in overall accuracy. SE increased by 22.19% and SP increased by 12.25% on average. As many plant pre-miRNAs contain multiple loops, *triplet-SVM* cannot classify them correctly. Almost all the plant pre-miRNAs with multiple loops in the testing dataset are classified by *PlantMiRNAPred* correctly. This indicates that our method is sensitive enough to identify pre-miRNAs with multi-loops.

MicroPred is more similar to our approach as it uses the same 48 features to classify pre-miRNAs. However, it was originally developed for human pre-miRNAs. The program of *microPred* can be downloaded from the web site (http://web.comlab.ox.ac.uk/people/manohara.rukshan.batuwita/microPred.htm). In order to compare with *microPred*, the classification model of *microPred* was changed according to the plant pre-miRNAs datasets. As shown in Table 4, *PlantMiRNAPred* is nearly 5% better than *microPred* in overall accuracy. SE increased by 5.19% on average and SP increased by 4.98%. The improvement is mainly due to the additional 32 structural features extracted from the plant pre-miRNAs and the 35 features extracted from the stems.

Twenty-three of 397 *O.sativa*-positive samples, 19 of 233 *P.patens*-positive samples, two of 131 *S.bicolor*-positive samples are classified as pseudo pre-miRNAs. However, in miRBase 16, 9 of 23 *O.sativa*-positive samples, 8 of 19 *P.patens*-positive samples, 2 of 2 *S.bicolor*-positive samples are obtained by computational identification method. They are not verified by biology experiments. Despite this, the accuracies of the three testing datasets are 96.39, 95.11 and 100%, respectively.

Most of the new reported pre-miRNAs in miRBase 15–16 was correctly predicted by *PlantMiRNAPred* with an average accuracy of 98.11%. This shows that *PlantmiRNAPred* is powerful in discovering novel pre-miRNAs. In the two updated datasets, 109 of 118 *G.max* pre-miRNAs and 74 of 193 *A.lyrata* are lineage-specific pre-miRNAs. Therefore, *PlantMiRNAPred* is also shown to be able to achieve high performance in classifying lineage-specific pre-miRNAs.

In addition, 11 918 inverted repeats were also extracted from the Gm08 (the eighth chromosome of *G.max* genome) by EINVERTED (Rice *et al.*, 2000). One thousand inverted repeats (including eight real pre-miRNAs) were selected according to the proportion of the real pre-miRNAs of different lengths. Thirty-seven of 1000 are classified by *PlantMiRNAPred* as putative real pre-miRNAs, covering eight real pre-miRNAs. The FP rate of *PlantMiRNAPred* is 2.9%. *MicroPred* classified 89 inverted repeats as real pre-miRNAs, covering eight real pre-miRNAs. The FP rate of *MicroPred* is 8.1%. *Triplet-SVM* classified 184 inverted repeats as real pre-miRNAs, covering six real pre-miRNAs. The FP rate of *triplet-SVM* is 17.8%. It indicates that *PlantMiRNAPred* is more sensitive to the inverted repeats from the genome.

4 CONCLUSION

A new *ab initio* classifier (*PlantMiRNAPred*) was developed for predicting plant pre-miRNAs. We demonstrated the importance of careful feature extraction, feature selection and training sample selection in achieving efficient and effective classification result. Particularly, according to the characteristics of plant pre-miRNAs, 115 features were extracted to distinguish the hairpins of real pre-miRNAs and pseudo pre-miRNAs. After eliminating redundant features, 68 informative features were selected. Each real/pseudo pre-miRNA was mapped into the 68-dimensional space. 1960 positive/negative representative samples were selected as the training dataset.

PlantMiRNAPred has been compared with the existing pre-miRNA classification methods, *triplet-SVM* and *microPred*. The results demonstrated that *PlantMiRNAPred* has higher classification performance. Further analysis indicated that the improvement of classification accuracy was due to the informative features and the representative training samples. *PlantMiRNAPred* will be useful in generating effective hypothesis for subsequent biological testing.

ACKNOWLEDGEMENTS

We appreciate Yingpeng Han and Yongxin Liu from the soybean research institute in the Northeast Agricultural University for valuable assistance.

Funding: Natural Science Foundation of China (60932008 and 60871092); Fundamental Research Funds for the Central Universities (HIT.ICRST.2010 022); Returned Scholar Foundation of Educational Department of Heilongjiang Province (1154hz26); Natural Science Foundation (Grant CCF-0546345 to Y.H.).

Conflict of Interest: none declared.

REFERENCES

Adai,A. *et al.* (2005) Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res.*, **15**, 78–91.

Ambros,V. *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Batuwita,R. and Palade,V. (2009) MicroPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, **25**, 989–995.

Berezikov,E. *et al.* (2006) Approaches to microRNA discovery. *Nat. Genet.*, **38**, 2–7.

Bonnet,E. *et al.* (2004) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *PNAS*, **101**, 11511–11516.

Chang,D.T. *et al.* (2008) Using a kernel density estimation based classifier to predict species-specific microRNA precursors. *BMC Bioinformatics*, **9**(Suppl. 12), 2–12.

Chatterjee,S. and Grobhans,H. (2009) Active turnover modulates mature microRNA activity in caenorhabditis elegans. *Nature*, **461**, 546–549.

Dezulian,T. *et al.* (2006) Identification of plant microRNA homologs. *Bioinformatics*, **22**, 359–360.

Freyhult,E. *et al.* (2005) A comparison of RNA folding measures. *BMC Bioinformatics*, **6**, 241–248.

Gan,H.H. *et al.* (2004) RAG: RNA-as-graphs database—concepts, analysis, and features. *Bioinformatics*, **20**, 1285–1291.

Gardner, P.P. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**(Suppl. 1), 136–140.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

Jiang,P. *et al.* (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, **35**(Suppl. 2), 339–344.

Jones-Rhoades,M.W. and Bartel,D.P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell*, **14**, 787–799.

Lindow,M. and Krogh,A. (2005) Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics*, **6**, 119–127.

Lu,Y.Z. and Yang,X.Y (2010) Computatinal identification of novel microRNAs and their targets in vigna unguiculata. *Com. Funct. Genomics*, **10**, 128297–128313.

Moulton,V. *et al.* (2000) Metrics on RNA secondary structures. *J. Comput. Biol.*, **7**, 277–292.

Nam,J. *et al.* (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.*, **33**, 3570–3581.

Ng,K.L.S. and Mishra,S.K. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, **23**, 1321–1330.

Quinlan,J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco.

Rice,P. *et al.* (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Schmutz,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schultes,E.A. *et al.* (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.*, **49**, 76–83.

Seffens,W. and Digby,D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.*, 1578–1584.

Sewer,A. *et al.* (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267–281.

Smalheiser,N.R. and Torvik,V.I. (2005) Mammalian microRNAs derived from genomic repeats. *Trends Genet.*, **21**, 322–326.

Weiss,G. (2004) Mining with rarity: a unifying framework. *SIGKDD Expl.*, **6**, 7–19.

Xue,C.H. *et al.* (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310–316.

Yousef,M. *et al.* (2006) Combining multi-Species genomic data for microRNA identification using a naïve Bayes classifier machine learning for identification of microRNA genes. *Bioinformatics*, **22**, 1325–1334.

Zhang,B.H. *et al.* (2006a) Evidence that miRNAs are different from other RNAs. *Cell Mol. Life Sci.*, **63**, 246–254.

Zhang,B.H. *et al.* (2006b) Plant microRNA: a small regulatory molecule with big impact. *Dev. Biol.*, **289**, 3–16.