Exploring Essential Attributes For Detecting MicroRNA Precursors From Background Sequences

Yun Zheng, Wynne Hsu, Mong Li Lee and Lim Soon Wong

Department of Computer Science, School of Computing
National University of Singapore, Singapore 117543
{zhengy, whsu, leeml, wongls}@comp.nus.edu.sg

Abstract. MicroRNAs (miRNAs) have been shown to play important roles in post-transcriptional gene regulation. The hairpin structure is a key characteristic of the microRNAs precursors (pre-miRNAs). How to encode their hairpin structures is a critical step to correctly detect the pre-miRNAs from background sequences, i.e., pseudo miRNA precursors. In this paper, we have proposed to encode the hairpin structures of the pre-miRNA with a set of features, which captures both the global and local structure characteristics of the pre-miRNAs. Furthermore, we find that four essential attributes are discriminatory for classifying human pre-miRNAs and background sequences with an information theory approach. The experimental results show that the number of conserved essential attributes decreases when the phylogenetic distance between the species increases. Specifically, one A-U pair, which produces the U at the start position of most mature miRNAs, in the pre-miRNAs is found to be well conserved in different species for the purpose of biogenesis.

1 Introduction

MicroRNAs (miRNAs) are small non-coding RNAs of about 22 nucleotides long. More and more evidences show that miRNAs play important roles in gene regulation and various biological processes, as reviewed in [1–3]. MicroRNAs transcripts, which may be produced by RNA polymerase II or III [3], often fold to form stem loop structures, and become what are called primary miRNAs, or pri-miRNAs. In the nucleus, the Drosha RNase III endonuclease cleavages both strands of the stem at the base of the primary stem loop [4], and produce the pre-miRNAs. Then, in cytoplasm, a second RNase III endonuclease, Dicer, together with its dsRNA-binding partner protein makes a second pair of cuts and defines the other end of the mature miRNAs (see example in Figure 1), which produces the miRNA:miRNA* duplex. Finally, the miRNA stand is separated from the duplex by the helicase and form the mature miRNA molecules [2–4]. The mature miRNAs are then loaded to RNA-induced silencing complex (RISC), which binds the $3^{'}$ untranslate region of messenger RNAs of the miRNA target genes to repress the production of related proteins [3, 5].

The hairpin structures of the pre-miRNAs are highly conserved in different species [6, 7]. Thus, how to convert the hairpin structures into informative features is a critical step to correctly identify the pre-miRNAs against the background sequences, i.e., pseudo pre-miRNAs.

There have been some endeavors for this purpose. The MirScan relied on the observation that the known miRNAs derive from phylogenetically conserved stem loop precursor RNAs with characteristic features [7]. The MiRseeker has been used to identify miRNA genes from insect DNA sequences [6]. It uses the hairpin structure predicted with the Mfold [8] as the primary criteria, but also takes into account the nucleotide divergence of miRNA candidates. The phylogenetic shadowing is a new method to find miRNA genes by comparing DNA sequences of different species [9, 10]. Xue *et al.* [11] proposed a triplet-SVM classifier which encoded the hairpin structures with local structure features and obtained good sensitivity for both human data sets and data sets of other species. Bentwich *et al.* [12] proposed to score the pre-miRNAs with thermodynamical stability and structural features, which mainly capture the global properties of the hairpin structures, to classify the pre-miRNA. Sewer *et al.* [13] proposed an SVM-based method to find clustered pre-miRNAs. Yang *et al.* [14] proposed to encode the pre-miRNAs with their secondary structures, the upstream and downstream sequences.

However, first and foremost, few endeavors have been given to exploring the essential attributes for classifying pre-miRNAs and finding the biological roles of these essential attributes. Second, pre-miRNAs have been phylogenetically conserved not only for the whole hairpin structures but also for local properties at the level of nucleotides and their secondary structures, as shown in [6, 15]. Finally, the specificities of the existing methods [13, 11] still need to be improved.

In this research, we propose to encode the hairpin structures with the combination of global and local characteristics. In our approach, the pre-miRNA sequences and negative samples are first analyzed using the RNAfold software [16]. Second, the global characteristics of the hairpins, such as the number of base pairs and GC content, and the local structure triplet elements are used to encode the pre-miRNAs and background sequences that are predicted to contain hairpin structures with 43 features. Third, the resulted data sets are used to build classification models, which display better performance, especially specificity, for new data sets. Finally, we investigate the phylogenetically conserved essential attributes of the pre-miRNAs with the Discrete Function Learning (DFL) algorithm [17]. These features found by the DFL algorithm are accurate in predicting the pre-miRNAs, which is discussed with respect to the biogenesis mechanism of the miRNAs.

The rest of the paper is organized as follows. In Section 2, we introduce the features for encoding the pre-miRNAs and the classification algorithms for evaluating the separation ability of the generated features. In Section 3, we briefly review the DFL algorithm. In Section 4, we introduce the data sets and show the experimental results. In Section 5, we summarize the paper and discuss some future directions.

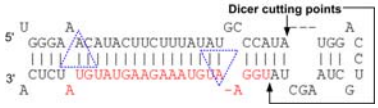


Fig. 1. The secondary structure of *Homo sapiens miR-1-1* pre-miRNA. The red nucleotides represents the mature miRNA. The two triangles represent the happening of the local feature A·((at position 7 and 50. The two positions pointed by the two arrows are the cutting points of Dicer [5], which produces the miRNA:miRNA* duplex by cutting off the central loop on the right side at the two positions.

2 Methods

In this section, we show how to encode the pre-miRNA sequences and their secondary structures with a set of features which captures both their global and local characteristics. Then, we briefly review the classification algorithms used in this research.

2.1 Encoding the Hairpin Structures of Pre-miRNAs

In our approach, the secondary structures of the pre-miRNAs, as well as the candidates, are predicted with the RNAfold [16]. Then, we propose to encode the nucleotide sequences and secondary structure sequences of the pre-miRNAs with 43 features, which consist of 11 global features and 32 local features.

We first talk about the global features. The 11 global features of the pre-miRNAs are symmetric difference, number of basepairs, GC content, length basepair ratio (length of the sequence/the number of basepairs), sequence length, length of central loop, free energy per nucleotide, bulge size, bulge number, tail length and the number of tail(s).

The definitions of these features are given as follows with an example in Figure 1. To be convenient, the pre-miRNA hairpin is divided into two arms. The left arm is from 5' end (upper in Figure 1) to the center of the central loop, and the rest nucleotides form the right arm. The symmetric difference is defined as the difference of length of the two arms. For example, the symmetric difference of the *hsa-miR-1-1* precursor in Figure 1 is 2. The bulge size is defined as the size of the largest mismatch region in either of two arms. As shown in Figure 1, the largest mismatch region is the three consecutive mismatch nucleotides in the right arm. Thus, the bulge size is 3 for the *hsa-miR-1-1* precursor. The bulge number is the defined as the larger number of bulge in the two arms. Similarly, there are 3 bulges in the right arm of the *hsa-miR-1-1* precursor. Thus, its bulge number is 3. The tail length is defined as the the length of the longer free tail of the two arms. The free energy per nucleotide is obtained by dividing the free energy given by the RNAfold program with the number of nucleotide in the pre-miRNAs. For *hsa-miR-1-1* precursor, the free energy given by the RNAfold program is -30 kcal/mol. Then, the energy per nucleotide is -30/71 = -0.42 kcal.

In summary, for the *hsa-miR-1-1* pre-miRNA in Figure 1, the values of the eleven global features are 2, 28, 0.37, 2.54, 71, 5, -0.42 (kcal), 3, 3, 1 and 2 respectively.

We choose these global features of the hairpins based on the following considerations. First, the pre-RNAs sequences have lower GC content than background sequences [12]. Second, the pre-miRNAs have lower folding energy than background sequences [18]. Third, we noticed that the lengths of the two arms of the pre-miRNAs are often equal or approximately equal. But the lengths of the two arms of the pseudo pre-miRNAs may be quite different. Fourth, the length of pre-miRNAs has a stable distribution [19]. Fifth, the number of basepairs and length basepair ratio are important features to decide the free energy [20]. Sixth, the number of bulge and the size of bulges are related to length basepair ratio. Seventh, we also noticed that the pre-miRNAs have no or shorter free tails at the ends of two arms than background sequences.

The local features is defined with the triplet elements proposed by Xue *et al.* [11]. One triplet is defined by one nucleotide and the secondary structure of its -1,0,+1 positions. There are 4 nucleotide, A, C, G, U, and 2 possible secondary structures, match '(' and mismatch '·'. Thus, there are totally $4 \times 2^3 = 32$ possible triplet elements. The count values of them are used as the 32 local features of our data sets. For example, for the *hsa-miR-1-1* pre-miRNA in Figure 1, the value of the feature "A·((" is 2, since it has happened at position 7 and 50, as indicated by the two dotted triangles.

2.2 The Classification Algorithms

In prior section, we demonstrate how to transform the pre-miRNA into a set of features, which carries the information of the class value of the sequences. The converted data sets are used by different algorithms to build predictors (classifiers).

In this study, we use four classification algorithms to demonstrate the value of encoding the pre-miRNA with both the global and local structural properties. The selected algorithms are the Support Vector Machines (SVM) algorithm [21], the C4.5 algorithm [22], the k-Nearest-Neighbors (kNN) algorithm [23] and the RIPPER algorithm [24].

3 The Discrete Function Learning Algorithm

To find which subset of features are relatively more important when used to predict the samples of different species, we use the Discrete Function Learning algorithm [17] to find the *essential attributes* (EAs) that contribute most to the class distinctions between samples. As to be introduced, there are two parameters for the DFL algorithm, the expected cardinality K and the ϵ value. The choice of parameters of the DFL algorithm is available in our early work [17] or at the supplementary website [1] of this paper.

[1] The supplements of this paper are available at http://www.comp.nus.edu.sg/∼zhengy/vldb2006.htm.

We will first introduce some notation. We use capital letters to represent discrete random variables, such as X and Y; lower case letters to represent an instance of the random variables, such as x and y; bold capital letters, like $\mathbf{X}$, to represent a vector; and lower case bold letters, like $\mathbf{x}$, to represent an instance of $\mathbf{X}$. In the remainder parts of this paper, we denote the attributes except the class attribute as a set of discrete random variables $\mathbf{V} = \{X_1, \ldots, X_n\}$, the class attribute as variable Y. The entropy of X is represented with $H(X)$, and the mutual information between X and Y is represented with $I(X;Y)$.

In this section, we start with a the theoretic background of information theory. Then, we introduce the motivation of the DFL algorithm. Finally, we briefly describe the DFL algorithm.

3.1 Theoretic Background

The entropy of a discrete random variable X is defined in terms of probability of observing a particular value x of X as [25]:

$$H(X) = -\sum_{x} P(X = x) log P(X = x).$$

The entropy is used to describe the diversity of a variable or vector. The more diverse a variable or vector is, the larger entropy it will have. Hereafter, for the purpose of simplicity, we represent $P(X = x)$ with $p(x)$, $P(Y = y)$ with $p(y)$, and so on. The mutual information between a vector $\mathbf{X}$ and Y is defined as [25]:

$$I(\mathbf{X};Y) = H(Y) - H(Y|\mathbf{X}) = H(\mathbf{X}) - H(\mathbf{X}|Y) = H(\mathbf{X}) + H(Y) - H(\mathbf{X},Y) \quad (1)$$

Basically, the stronger the relation between two variables, the larger mutual information they will have. Zero mutual information means the two variables are independent or have no relation.

The conditional mutual information $I(X;Y|Z)$ [26](the mutual information between X and Y given Z) is defined by

$$I(X;Y|Z) = \sum_{x,y,z} p(x,y,z) \frac{p(x,y|z)}{p(x|z)p(y|z)}.$$

The chain rule for mutual information is give by Theorem 1, for which the proof is available in [26].

Theorem 1. $I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1)$.

3.2 Motivation

$I(\mathbf{X};Y)$ is evaluated with respect to $H(Y)$ in the DFL algorithm, which is different from those in existing methods, as shown in Equation 2. Suppose that

$\mathbf{U}_{s-1}$ is the already selected feature subset at the step $s-1$, and the DFL algorithm is trying to add a new feature $X_i \in \mathbf{V} \setminus \mathbf{U}_{s-1}$ to $\mathbf{U}_{s-1}$. Specifically, $X_{(1)} = argmax_i I(X_i; Y)$, and

$$X_{(s)} = argmax_i I(\mathbf{U}_{s-1}, X_i; Y), \tag{2}$$

where $\forall s$, $1 < s \leq k$, $\mathbf{U}_1 = \{X_{(1)}\}$, and $\mathbf{U}_s = \mathbf{U}_{s-1} \cup \{X_{(s)}\}$. From Theorem 1, we have

$$I(\mathbf{U}_{s-1}, X_i; Y) = I(\mathbf{U}_{s-1}; Y) + I(X_i; Y|\mathbf{U}_{s-1}). \tag{3}$$

In Equation 3, note that $I(\mathbf{U}_{s-1}; Y)$ does not change when trying different $X_i \in \mathbf{V} \setminus \mathbf{U}_{s-1}$. Hence, the maximization of $I(\mathbf{U}_{s-1}, X_i; Y)$ in the DFL algorithm is actually maximizing $I(X_i; Y|\mathbf{U}_{s-1})$, the conditional mutual information of X_i and Y given the already selected features $\mathbf{U}_{s-1}$, i.e., the information of Y not captured by $\mathbf{U}_{s-1}$ but carried by X_i.

To measure which subset of features is optimal, we restate the following theorem, which is the theoretical foundation of our algorithm. It has been proved that if $H(Y|X) = 0$, then Y is a function of X [26]. Since $I(X;Y) = H(X) - H(Y|X)$, it is immediate to obtain Theorem 2.

Theorem 2. *If the mutual information between* $\mathbf{X}$ *and* Y *is equal to the entropy of* Y*, i.e.,* $I(\mathbf{X}; Y) = H(Y)$*, then* Y *is a function of* $\mathbf{X}$*.*

The entropy $H(Y)$ represents the diversity of the variable Y. The mutual information $I(\mathbf{X}; Y)$ represents the relation between vector $\mathbf{X}$ and Y. From this point of view, Theorem 2 actually says that the relation between vector $\mathbf{X}$ and Y are very strong, such that there is no more diversity for Y if $\mathbf{X}$ has been known. In other words, the value of $\mathbf{X}$ can fully determine the value of Y.

3.3 Training Methods

A classification problem is trying to learn or approximate a function, which takes the values of attributes (except the class attribute) in a new sample as input and output a categorical value which indicates the class of the sample under consideration, from a given training data set. The goal of the training process is to obtain a function which makes the output value of this function be the class value of the new sample as accurately as possible. From Theorem 2, the problem is converted to finding a subset of attributes $\mathbf{U} \subseteq \mathbf{V}$ whose mutual information with Y is equal to the entropy of Y. The $\mathbf{U}$ is the EAs that we are trying to find from the data sets. Here, we will briefly describe the main steps of the DFL algorithm as shown in the following.

1. $\forall X_i \in \mathbf{V}$, compute $I(X_i; Y)$;
2. add $A = argmax_i I(X_i; Y)$ to the EA set $\mathbf{U}_1$;
3. $\forall X_i \in \mathbf{V} \setminus \mathbf{U}_{s-1}$, compute $I(\mathbf{U}_{s-1}, X_i; Y)$;
4. add $B = argmax_i I(\mathbf{U}_{s-1}, X_i; Y)$ to the EA set $\mathbf{U}_{s-1}$;
5. repeat 3-4, until find $\mathbf{U}$ so that $I(\mathbf{U}; Y) = H(Y)$.

The DFL algorithm will find the most informative feature A in the first step. Then, the DFL algorithm will try every subsets with A and another remaining feature in $\mathbf{V}$, and find the most informative feature subset $\{A, B\}$ in the second step. Next, the similar calculation will be done until the target combination $\mathbf{U}$, which satisfies the criterion of Theorem 2, is found.

To prevent exhaustive search of all subsets of $\mathbf{V}$, one parameter called the expected cardinality K of the EAs is introduced to restrict the searching space to subsets with $\leq K$ features.

After $\mathbf{U}$ is found, the DFL algorithm will stop its searching process, and obtain the classifiers by deleting the non-essential attributes and duplicate rows in the training data sets.

3.4 The ϵ Value Method

We also introduce a method called ϵ value to overcome the noisy problems [17]. Theorem 2, the exact functional relation demands the strict equality between the entropy of Y, $H(Y)$ and the mutual information of $\mathbf{X}$ and Y, $I(\mathbf{X}; Y)$. However, this equality is often ruined by the noisy data, like microarray gene expression data. In these cases, we have to relax the requirement to obtain a best estimated result. By defining a significant factor ϵ, if the difference between $I(\mathbf{X}; Y)$ and $H(Y)$ is less than or equal to $\epsilon \times H(Y)$, then the DFL algorithm will stop the searching process, and build the classifier for Y with $\mathbf{X}$ at the significant level ϵ. The ϵ is the second parameter of the DFL algorithm.

3.5 Prediction Method

After the DFL algorithm obtaining the classifiers as function tables of the pairs $(\mathbf{u}, y)$, the most reasonable way to use such function tables is to check the input values $\mathbf{u}$, then find the corresponding output values y. Therefore, we perform predictions in the space defined by the EAs $\mathbf{U}$, the *EA space*, with the 1-Nearest-Neighbor (1NN) algorithm [23] based on the Hamming distance defined as follows.

Definition 1. *Let* $1(a, b)$ *be an indicator function, which is 0 if and only if* $a = b$*, otherwise is 1. The Hamming distance between two arrays* $\mathbf{A} = [a_1, \ldots, a_n]$ *and* $\mathbf{B} = [b_1, \ldots, b_n]$ *is* $Dist(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^{n} 1(a_i, b_i)$.

Note that the Hamming distance [27] is dedicated to binary arrays, however, we do not differentiate between binary or non-binary cases in this paper. We use the Hamming distance as a criterion to decide the class value of a new sample, since we believe that the rule with minimum Hamming distance to the EA values of a sample contains the maximum information of the sample. Thus, the class value of this rule is the best prediction for the sample.

In the prediction process, if a new sample has same distance to several rules, we choose the rule with the biggest count value happened in the training data set.

Table 1. The summary of data sets.

	Data Set	Sample #	Class
0	TR-C (training)	163/168	pre-miRNAs/background
1	TE-C1	30	pre-miRNAs
2	TE-C2	1000	background
3	CONSERVED-HAIRPIN(T3)	2444	background
4	UPDATED(T4)	39	pre-miRNAs
5	*Mus musculusi*(mmu)	36	pre-miRNAs
6	*Rattus norvegicus*(rno)	25	pre-miRNAs
7	*Gallus gallus*(gga)	13	pre-miRNAs
8	*Danio rerio*(dre)	6	pre-miRNAs
9	*Caenorhabditis briggsae*(cbr)	73	pre-miRNAs
10	*Caenorhabditis elegans*(cel)	110	pre-miRNAs
11	*Drosophila pseudoobscura*(dps)	71	pre-miRNAs
12	*Drosophila melanogaster*(dme)	71	pre-miRNAs
13	*Oryza sativa*(osa)	96	pre-miRNAs
14	*Arabidopsis thaliana*(ath)	75	pre-miRNAs
15	*Epstein Barr Virus*(ebv)	5	pre-miRNAs
	total (1 to 15)	4094	

4 Results

In this section, we first introduce the data sets used. Then, we show the experimental results. All data sets and software used in this study are available at the supplementary website of this paper.

4.1 Data Sets and Preprocessing

In this research, we use the data sets in literature [11] to validate our approach, since it is valuable to compare the published results. These data sets are summarized in Table 1. Data set 0 to 4 is from human, and data set 5 to 15 is from other species, as indicated by their names. Data set 0 is used as the training data set, and data set 1 to 15 are used as testing data sets. There are totally 4094 samples used as testing data sets, with 3444 background sequences and 650 pre-miRNAs.

The sequences of human pre-miRNAs are obtained from miRNA registry database (release 5.0) [28]. The secondary structures of these 207 pre-miRNA sequences are predicted with the RNAfold [16]. Then, 193 sequences with only 1 loop are chosen. Next, 163 of them are randomly selected to be positive samples of the training data set, i.e., TR-C in Table 1. The rest 30 samples are used as TE-C1 testing data set.

The background sequences in data set 2 are collected from protein coding regions (CDSs) according to the UCSC refGene annotation tables [11]. The length of these sequences has the same distribution of human pre-miRNAs. The RNAfold is also used to predict the secondary structure of them. Then, the sequences with multiple loops, the sequences with less than 18 base pairs and the

sequences with larger than -15kcal/mol free energy are removed. Finally, there are 8494 sequences in this data sets. Among them, 168 are randomly selected as the negative samples of the TR-C data set, and 1000, different from the 168 used, are randomly chosen as the TE-C2 testing data set.

The data set 3 also consists of background sequences, which are retrieved from the genome region from position 56,000,001 to 57,000,000 on the human chromosome 19 with the UCSC database [29]. A window of 100 nucleotides is used to scan the region and those sequences with a predicted hairpin secondary structure by the RNAfold [16] are selected. This produces 2444 background sequences in data set 3. Unlike data set 2, some sequences on data set 3 are likely to be the true pre-miRNAs. Actually, there are 3 known miRNAs (*hsa-mir-99b*, *hsa-let-7e* and *hsa-mir-125a*) in data set 3 [11].

Bentwich *et al.* [12] reported 89 new pre-miRNAs, of which 1 has multiple loops and is removed. To further remove the similar sequences, BLASTCLUST with S = 80, L = 0.5 and W = 16 is applied to the remaining 88 sequences. Only one representative sequence in each cluster is selected to remove the closely related sequences. This produces 40 non-redundant sequences, which are further checked with respect to the training data set. One of the 40 sequences that has high similarity to the training data set is removed. Finally, only 39 sequences are chosen as the data set 4.

The sequences from other species are chosen from the release 5.0 of the miRNA registry [28]. 581 out of 1138 pre-miRNAs are remained after removing the sequences with high similarity with the pre-miRNAs in the training data set. The similarity of the sequences is also calculated with BLASTCLUST with S = 80, L = 0.5 and W = 16. These pre-miRNAs form the data set 5 to 15.

4.2 Experimental Results

We have developed a software, *miREncoding*, to encode the pre-miRNA sequences, together with their secondary structure sequences, into the 43 proposed features with the Java language. We use the *Weka* software (version 3.4) [30] to evaluate the performance of the selected classification algorithms. For the SVM algorithm, polynomial kernels are used. All selected algorithms are applied to the data sets with the default settings of the *Weka* software.

To demonstrate the advantage of using both the global and local structural characteristics, we generate three data sets for each data set in Table 1 with the *miREncoding* program. The first one contains both the global and local features, the second one contains only the 32 local features, and the third one contains only the 11 global features. Then, we apply the selected algorithms to the three data sets to compare their prediction accuracies, which are shown in Figure 2.

As shown in Figure 2, the SVM, C4.5, and kNN algorithms show large improvements of accuracy for data set 2 and 3 when applied to data sets with all features. This suggests that the combination of global and local characteristics are critical in removing false positives, since data sets 2 and 3 are negative samples, i.e., the background sequences. For the remaining data sets, the four algorithms demonstrate stable prediction accuracies when applied to data sets

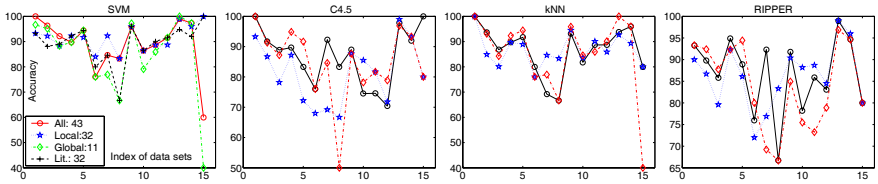


Fig. 2. The prediction accuracies of difference classification algorithms, where the detailed values are available in Supplementary Table S2 to S4. (a) SVM. The curve marked with pluses represents the results of the triplet-SVM classifier, on 32 local features, in literature [11]. (b) C4.5. (c) kNN ($k = 5$). (d) RIPPER.

Table 2. The summary of prediction performance of the classification algorithms. The values shown present the performance of the classification algorithms on data sets with 32 local/11 global/all 43 features respectively. The best value for each measure (MS) is shown in bold face. The SS, SP and AC in measure (MS) column stand for sensitivity, specificity and accuracy respectively.

	MS	SVM	C4.5	kNN	RIPPER	Tri-SVM[1]
human	SS	92.8/92.8/94.2	89.9/**97.1**/94.2	94.2/95.7/94.2	91.3/92.8/94.2	92.8
(D1-4)	SP	89.5/90.3/**93.3**	80.7/88.4/89.7	81.5/86.8/88.8	81.7/89.1/87.0	88.7
	AC	89.6/90.3/**93.3**	80.8/88.6/89.8	81.7/87.0/88.9	81.8/89.2/87.1	88.8
other species	SS[2]	**91.9**/89.8/91.7	84.7/85.5/83.3	87.2/90.4/87.8	89.7/84.0/87.8	90.9
total	SS	**92.0**/90.2/**92.0**	85.2/86.8/84.5	88.9/90.9/89.4	89.9/84.9/88.5	91.1
	SP	89.5/90.3/**93.3**	80.7/88.4/89.7	81.5/86.8/88.8	81.7/89.1/87.0	88.7
	AC	89.9/90.3/**93.1**	81.4/88.1/88.9	85.9/87.5/88.9	83.0/88.4/87.2	89.1

[1] This column shows the results of the triplet-SVM classifier [11]. [2] The sensitivity equals to the accuracy, since there are only positive samples for data set 5 to 15.

with all features. When applied to only the global or local features, the prediction accuracies of the algorithms fluctuate intensively. This suggests that the combination of global or local features carries more information of the class attribute than only the global or local features.

In Figure 2 (a), we also compare the prediction performance of the SVM algorithm with the triplet-SVM classifiers in literature [11]. The SVM algorithm performs better than the triplet-SVM classifiers with the more information given by the global feature of the pre-miRNAs. Especially, the total specificity is improved from 88.7% of the triplet-SVM classifiers to 93.3% of the SVM algorithm in our study, which has totally reduced 40.6% (or 158 samples) false positives in literature [11]. For the local data sets, the prediction accuracies of the SVM algorithm in our study are slightly better than those in literature [11]. We attribute this to the encoding region in our research. We encode the triplet local features for the whole pre-miRNAs, except the first and last nucleotide. However, only the paired regions of the pre-miRNAs are encoded into triplet features in [11].

The prediction performance of all selected algorithms is summarized in Table 2. As shown in Table 2, the SVM algorithm performs best for these data sets among all selected algorithms and method in literature [11]. From table 2, it is also shown that the performance of the algorithms generally becomes better when applied to the data sets with all 43 features. Especially for the specificity, the SVM, C4.5 and kNN algorithms show large improvements when applied to data sets with all 43 features. For instance, the specificity of the SVM algorithm has been dramatically improved from 89.5% for local features and 90.3% for global features to 93.3%, as shown in Table 2, which means total reduction of 131 and 103 false positive predictions respectively.

4.3 Investigating The Essential Attributes

In this section, we investigate the essential attributes for classifying pre-miRNAs and background sequences with the DFL algorithm, which has been implemented with the Java language [17]. The DFL algorithm is not designed for continuous features. Hence, we discretize the continuous features with an entropy-based discretization method [31], which has been implemented in the *Weka* software, before performing feature selection with the DFL algorithm. The discretization is carried out in such a way that the training data set is first discretized. Then the testing data set is discretized according to the cutting points of variables determined in the training data set. After that, the original continuous values of the selected features are used by other algorithms.

To find the optimal subset of EAs for the data sets, we first set the expected cardinality of the EAs K as 10. Next, we use the DFL algorithm to perform leave-one-out cross validation (LOOCV) on the training set with different ϵ values, from 0 to 0.8 with a step of 0.01. Then, we find that the DFL algorithm reaches its best prediction performance in the LOOCV when $\epsilon \in [0.12, 0.13]$ (see Supplementary Figure S2). When using all samples in training data set, the distributions of attributes are slightly different from those in LOOCV. Thus, we try the DFL classifiers obtained from a wider region of $\epsilon \in [0.1, 0.15]$. Finally, we choose the DFL classifier obtained when $\epsilon = 0.11$ because it shows overall better prediction accuracies for data set 1 to 4. In this way, a subset of 4 features, {A(((, G·((, length basepair ratio, energy per nucleotide}, is chosen as EAs for the human data sets, D1 to D4. For other data sets, the performance of the DFL algorithm is not as good as for data set 1 to 4. We attribute this to the phylogenetic distance between the species. Because fewer and fewer characteristics of the pre-miRNAs are conserved when the species become more distantly related. Thus, we try the subsets of these 4 EAs and choose those subsets on which the 1NN algorithm introduced in Section 3.5 produces the best prediction performances. The selected EAs are shown in Figure 3 (a).

We examine the phylogenetic relations between the species of the selected data sets with miRBase [32]. As shown in Figure 3 (b), data set 5 and 6 are from the Rodentia which has the closest relation with the species, *Homo sapiens*, of the training data set, D0. Then, 3 out of the 4 EAs are conserved for data set 5 and 6. For other data sets, only the 1 of the 4 EAs, A(((or length basepair

a

	Data Set	Features Chosen By The DFL Algorithm
1	TE-C1	A(((, G-((, LenBPRatio, EnergyPerNucleotide
2	TE-C2	A(((, G-((, LenBPRatio, EnergyPerNucleotide
3	T3	A(((, G-((, LenBPRatio, EnergyPerNucleotide
4	T4	A(((, G-((, LenBPRatio, EnergyPerNucleotide
5	mmu	A(((, LenBPRatio, EnergyPerNucleotide
6	mo	A(((, LenBPRatio, EnergyPerNucleotide
7	gga	A(((
8	dre	A(((
9	cbr	A(((
10	cel	A(((
11	dps	A(((
12	dme	A(((
13	osa	A(((
14	ath	A(((
15	ebv	LenBPRatio

b

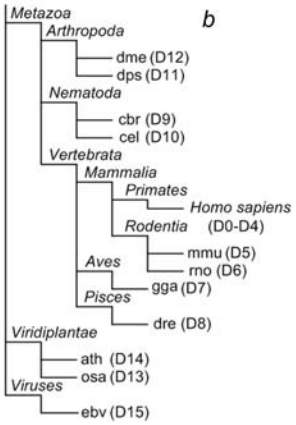
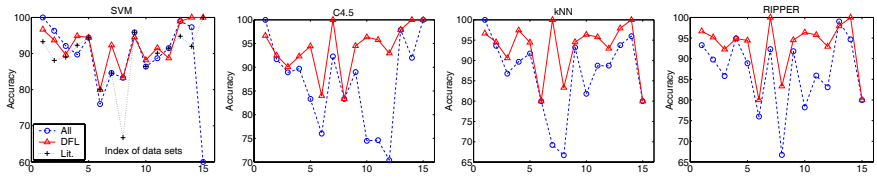


Fig. 3. The EAs chosen by the DFL algorithm. (a) The EAs chosen by the DFL algorithm. (b) The phylogenetic tree of the species of data sets from the miRBase [32].



(a) (b) (c) (d)

Fig. 4. The prediction accuracies of difference classification algorithms. (a) SVM. The curve marked with pluses represents the results of the triplet-SVM classifier, on 32 local features, in literature [11]. (b) C4.5. (c) kNN ($k = 5$). (d) RIPPER.

ratio, is conserved. This reduction of EAs suggests that less characteristics of the pre-miRNAs are conserved when the species of data sets and *Homo sapiens* of the training data set become more distantly related.

Then, we further run the selected algorithms on the features chosen by the DFL algorithm. The prediction performance the classification algorithms is shown in Figure 4 (details available in Supplementary Table S2 and S5) and summarized in Table 3. In Figure 4 and Table 3, it is shown that the selected algorithms, except the SVM algorithm, demonstrate large improvements of prediction performance on the EAs in Figure 3 (a). For instance, the C4.5 algorithm reaches overall sensitivity of 95.5% for the EAs chosen by the DFL algorithm, as shown in Table 3. However, the C4.5 algorithm only obtains overall sensitivity of 84.5% for all features. The RIPPER algorithm reaches best overall accuracy of 93.5% on DFL features, which is slightly better than the 93.1% achieved by the SVM algorithm on all features. These results suggest that the EAs shown in Figure 3 (a) are critical for classifying the pre-miRNAs against background

Table 3. The summary of prediction performance of the classification algorithms on the DFL features. The best value for each measure is shown in bold face.

	Measures	SVM		C4.5		kNN		RIPPER		Tri-SVM[1]
features		All	DFL	All	DFL	All	DFL	All	DFL	
human	sensitivity	94.2	95.7	94.2	94.2	94.2	**97.1**	94.2	95.7	92.8
(D1-D4)	specificity	**93.3**	91.5	89.7	90.8	88.8	91.7	87.0	93.1	88.7
	accuracy	**93.3**	91.6	89.8	90.9	88.9	91.8	87.1	93.2	88.8
other species	sensitivity[2]	91.7	91.9	83.3	**95.7**	87.8	95.5	87.8	95.4	90.9
total	sensitivity	92.0	92.3	84.5	**95.5**	89.4	**95.5**	88.5	95.4	91.1
	specificity	**93.3**	91.5	89.7	90.8	88.8	91.7	87.0	93.1	88.7
	accuracy	93.1	91.6	88.9	91.6	88.9	92.3	87.2	**93.5**	89.1

[1] This column shows the results of the triplet-SVM classifier [11]. [2] The sensitivity equals to the accuracy, since there are only positive samples for data set 5 to 15.

sequences. Although the prediction accuracies of the SVM algorithm slightly decreases for the DFL features, the SVM classifiers are much less complex than the models for all features.

5 Discussions

From Figure 3 (a) and Figure 4, it is shown that the classification algorithms are accurate on one local feature, A(((, for data sets whose species are distantly related to the species of the training data set. The A(((feature is actually originating from the A-U pairs in the pre-miRNAs. By examining the distribution of A(((in the training data set, it is known that there tend to be more A-U pairs in the pre-miRNAs than in background sequences. We attribute this higher frequency of A-U pair to two reasons. First, we consider the biogenesis process of miRNAs. It is reported that most known miRNAs begin with a U [15, 19], which is originally coming from an A-U pair in the pre-miRNAs, as shown in Figure 1. In the biogenesis of the mature miRNAs, the Dicer recognizes the A-U pair in the pre-miRNAs, and performs the second cut in the biogenesis of mature miRNAs exactly at the A-U pair to produce the miRNA:miRNA* duplex [5]. This indicates that the A(((feature found in this study is critical for the biogenesis of the mature miRNAs. The high accuracies shown in Table 3 suggest that this A-U pair is well conserved in different species, even those distantly related in the phylogenetic tree, for the biogenesis of miRNAs. Second, the lower GC content in pre-miRNAs [12] partially contributes to the higher frequency of the A-U pair in the pre-miRNAs.

We have displayed how to encode the hairpin structures of pre-miRNAs with a set of features, which captures both their global and local structural properties. Different classification algorithms have shown large improvements of their prediction performance, especially the specificity, when applied to these features. This suggests that the proposed set of features have captured more information

about characteristics of the hairpin structures of the pre-miRNAs than only the local features or the global features.

We have found that four EAs, with both global and local features, are critical for classifying the testing samples from the same species as the training data set. But when the phylogenetic distance between the species of the testing data sets and training data set increases, the number of EAs is reducing gradually. The selected classification algorithms generally show better prediction performance when applied to these EAs. This indicates that the pre-miRNAs of distantly related species share less common characteristics than closely related species. Therefore, to obtain better prediction performance, it is better to use the samples from the same species or closely related species as the training data set.

The false positives provide a valuable source for finding new pre-miRNAs. The improvement of specificities of the classification algorithms when applied to the combination of global and local features, as well as the EAs, can help to significantly reduce the number of putative pre-miRNA candidates, thus to save much resource for validating them.

The pre-miRNAs with multiple loops are not considered in this research. How to encode them is a valuable future direction.

Acknowledgement

We thank Xue *et al.* [11] for their generous sharing of their data sets. This research was supported by the research grant, R-252-000-172-593, of the Institute of Infocomm Research in Singapore.

References

1. Alvarez-Garcia, I., Miska, E.A.: Microrna functions in animal development and human disease. Development **132** (2005) 4653–62
2. Ambros, V.: The functions of animal microrna. Nature **431** (2004) 350–5
3. Bartel, D.P.: Micrornas: Genomics, biogenesis, mechanism, and function. Cell **116** (2004) 281–297
4. Lee, Y., Ahn, C. Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., Kim, V.: The nuclear rnase iii drosha initiates microrna processing. Nature **425** (2003) 415–419
5. Zamore, P.D., Haley, B.: Ribo-gnome: The Big World of Small RNAs. Science **309**(5740) (2005) 1519–1524
6. Lai, E.C., Tomancak, P., Williams, R.W., Rubin, G.M.: Computational identification of drosophila microrna genes. Genome Biol **4** (2003) R42
7. Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., Bartel, D.P.: Vertebrate MicroRNA Genes. Science **299**(5612) (2003) 1540–
8. Zuker, M.: Mfold web server for nucleic acid folding and hybridization prediction. Nucl. Acids Res. **31**(13) (2003) 3406–3415
9. Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H., Cuppen, E.: Phylogenetic shadowing and computational identification of human microrna genes. Cell **120** (2005) 21–24

10. Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., Rubin, E.M.: Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. Science **299**(5611) (2003) 1391–1394
11. Xue, C., Li, F., He, T., Liu, G.P., Li, Y., Zhang, X.: Classification of real and pseudo microrna precursors using local structure-sequence features and support vector machine. BMC Bioinformatics **6**(1) (2005) 310
12. Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., Bentwich, Z.: Identification of hundreds of conserved and nonconserved human microrna. Nature Genetics **37**(7) (2005) 766–70
13. Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M., Tuschl, T., van Nimwegen, E., Zavolan, M.: Identification of clustered microrna using an ab initio prediction method. BMC Bioinformatics **6**(1) (2005) 267
14. Yang, L., Hsu, W., Lee, M., Wong, L.: Identification of microrna precursors via svm. In: Proc. of the 4th Asia-Pacific Bioinformatics Conference. (2006) 267–276
15. Lewis, B.P., Burge, C.B., Bartel, D.P.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. Cell **120** (2005) 15–20
16. Hofacker, I.L.: Vienna RNA secondary structure server. Nucl. Acids Res. **31**(13) (2003) 3429–3431
17. Zheng, Y., Kwoh, C.K.: Identifying simple discriminatory gene vectors with an information theory approach. In: Proceedings of the 4th Computational Systems Bioinformatics Conference, CSB 2005, Stanford, CA (2005) 12–23
18. Bonnet, E., Wuyts, J., Rouze, P., Van de Peer, Y.: Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics **20**(17) (2004) 2911–2917
19. Wang, X.J., Reyes, J., Chua, N.H., Gaasterland, T.: Prediction and identification of arabidopsis thaliana microrna and their mrna targets. Genome Biology **5**(9) (2004) R65
20. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucl. Acids Res. **9**(1) (1981) 133–148
21. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Advances in kernel methods: support vector learning. MIT Press, Cambridge, MA (1999) 185–208
22. Quinlan, J.R.: C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco, CA (1993)
23. Aha, D., Kibler, D., Albert, M.: Instance-based learning algorithms. Machine Learning **6** (1991) 37–66
24. Cohen, W.W.: Fast effective rule induction. In Prieditis, A., Russell, S., eds.: Proc. of the 12th International Conference on Machine Learning, Tahoe City, CA, Morgan Kaufmann (1995) 115–123
25. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL (1963)
26. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, Inc., New York, NY (1991)
27. Hamming, R.: Error detecting and error correcting codes. Bell System Technical Jounral **9** (1950) 147–160
28. Griffiths-Jones, S.: The microRNA Registry. Nucl. Acids Res. **32**(90001) (2004) D109–111

29. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., Kent, W.J.: The UCSC Genome Browser Database. Nucl. Acids Res. **31**(1) (2003) 51–54
30. Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.: Data mining in bioinformatics using Weka. Bioinformatics **20**(15) (2004) 2479–2481
31. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence, IJCAI-93, Chambery, France (1993) 1022–1027
32. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., Enright, A.J.: miRBase: microRNA sequences, targets and gene nomenclature. Nucl. Acids Res. **34**(supp1) (2006) D140–144