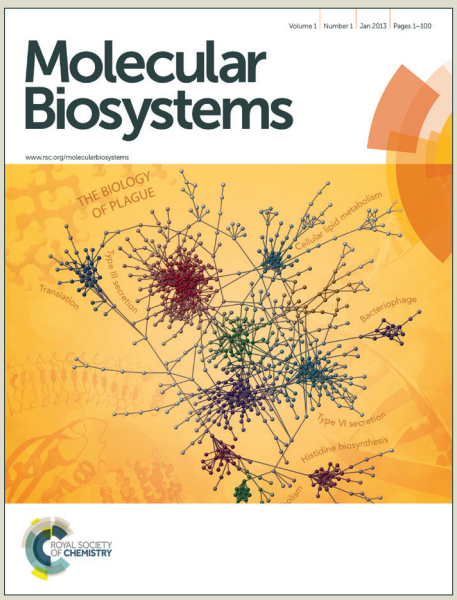
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

miRNA-dis: microRNA precursor identification based on distance structure status pairs

Bin Liu[1,2*], Longyun Fang[1], Junjie Chen[1], Fule Liu[1], Xiaolong Wang[1,2]

1 School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China
2 Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China

* Corresponding authors

E-mail addresses of all authors
BL: bliu@insun.hit.edu.cn
LYF: dragoncloudest@gmail.com
JJC: chenjunjie@hitsz.edu.cn
FLL: liufule12@gmail.com
XLW: wangxl@insun.hit.edu.cn

Mail addresses of the corresponding authors
Bin Liu: Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen, 518055, China; Phone: (+86) 0755-86011630

Abstract

MicroRNA precursor identification is an important task in bioinformatics. Support Vector Machine (SVM) is one of the most effective machine learning methods used in this field. The performance of SVM-based methods depends on the vector representations of RNAs. However, the discriminative power of the existing feature vectors is limited, and many methods lack an interpretable model for analysis of characteristic sequence features. Prior studies have demonstrated that sequence or structure order effects were relevant for discrimination, but little work has explored how to use this kind of information for human pre-microRNA identification. In this study, in order to incorporate the structure-order information into the prediction, a method called "miRNA-dis" was proposed, in which the feature vector was constructed by the occurrence frequency of "distance structure status pair" or just "distance-pair". Rigorous cross-validations on a much larger and more stringent newly constructed benchmark dataset showed that the miRNA-dis outperformed some state-of-the-art predictors in this area. Remarkable, miRNA-dis trained with human data can correctly predict 87.02% of the 4022 pre-miRNAs from 11 different species ranging from animals, plants and virus. miRNA-dis would be a useful high throughput tool for large-scale analysis of microRNA precursors. In addition, the learnt model can be easily analyzed in terms of discriminative features, and some interesting patterns were discovered, which could reflect the characteristics of microRNAs. A user-friendly web-server of miRNA-dis was constructed, which is freely accessible to the public at the web-site on http://bioinformatics.hitsz.edu.cn/miRNA-dis/.

Index Terms: microRNA precursor identification, structure status, distance-pair, Support Vector Machine

1. Introduction

MicroRNAs (abbreviated miRNA) are small single-strand, non-coding RNA molecules (about 22-nucleotides-long) found in plants, animals, and some viruses, which function in transcriptional and post-transcriptional regulation of gene expression (1). Aberrant expression of miRNAs has been implicated in numerous disease states, and miRNA-based therapies are under investigation (2-4). Therefore, it is fundamentally important to classify real vs. false pre-miRNAs.

Sequencing techniques such as RNA-seq can accurately identify expressed miRNA genes, and inexpensive direct sequencing of small RNA molecules (5) is used by some biologists to discover new microRNAs. However, there are still plenty of rooms for computational methods to be involved in. Most of the computational methods treated this problem as a binary classification task to discriminate the real pre-miRNAs from false pre-miRNAs and built their predictors by adopting the machine learning techniques. These methods are different in feature extraction and machine learning algorithms. The machine learning algorithms widely used in this field include Support Vector Machine (SVM)(6-11), Random Forest (RF) (12), Hidden Markov Model (HMM) (13), Naive Bayes (NB) (14), Linear Genetic Programming (LGP) (15), etc.

Because most of the pre-miRNAs have the characteristics of stem-loop hairpin structures, which play an important role during the biogenesis procedure of a mature miRNA (6), the secondary structure is thereby an important feature used in the computational methods (16). For example Triplet-SVM (6) employed a SVM classifier trained with 32 local triplet sequence-structure features. Later, MiPred (12) improved Triplet-SVM by adopting the Random Forest classifier trained with the local triplet sequence-structure features, minimum of free energy (MFE), and P-values. MiRFinder (8) is a high-throughput pre-miRNA prediction method consisting of two steps: a search for hairpin candidates and an exclusion of the non-robust structures based on the analysis of 18 parameters by the SVM. Recently, Zou et al (17) found that the negative samples had significantly impact on the computational predictors, and constructed a new benchmark dataset with high quality negative samples. Various experiments showed that this benchmark dataset could improve the performance of different methods.

All these computational methods could yield quite encouraging results, but most of them only considered the local structure-order information of RNAs, and therefore, all the global or long range structure-order information was ignored. As shown in the literature, the sequence or structure order information showed strong discriminative power for many tasks in bioinformatics, for example, in the field of proteomics, the Pseudo Amino Acid Composition (PseAAC) (18)was proposed to incorporate the long-range or global sequence order information of protein sequences; Physicochemical Distance Transformation (PDT) (19)was able to incorporate the global physicochemical properties of amino acids. In the field of genomics, the concept of PseAAC was applied to the DNA recombination spot identification (20-22) and the nucleosome position prediction (23), which considers the global sequence-order information of DNA sequences. These

computational methods outperformed those methods only using local sequence or structure order information. However, in the field of microRNA precursor prediction, it is difficult to incorporate the global or long range structure-order effects into a predictor, because the different length of RNA sequences and the high number of different structure statuses and their combinations. If all these structure statuses and their combinations were considered, the dimension of the feature vector is high, which would result in the "curse of dimensionality" and high computational cost.

In this study, we proposed a new predictor "miRNA-dis" with an intuitively interpretable feature space to represent RNA sequences called "distance structure status pair" or just "distance-pair" for pre-miRNA identification. In this method, the long range structure-order information was approximately represented by the occurrences of distance-pairs.

2. Materials and Method

2.1. Benchmark Dataset

The pre-miRNAs or positive samples were acquired from the miRBase (release 20: June 2013) (24,25), which contains 1,872 experiment-confirmed sapiens pre-miRNA entries. The false pre-miRNAs or negative samples were obtained from the data constructed by Xue et al. (6), which contains 8,489 false pre-miRNA samples. These false pre-miRNAs are similar to the real pre-miRNAs according to some widely accepted characteristics (6): (i) the RNA length ranges from 51 nt to 137 nt; (ii) a minimum of 18 base pairings on the stem of the hairpin structure; (iii) a maximum of -15 kal/mol free energy of the secondary structure.

In order to get rid of the redundancy and avoid homology bias, in the current study, the CD-HIT software(26) with the cutoff threshold set as 80% (note that the most stringent cutoff threshold for DNA sequences by CD-HIT is 75%) was employed to kick out those samples having $\geq 80\%$ sequence similarity to any others in a same subset.

We constructed the negative dataset by randomly picking 1,612 samples from the 8,489 false pre-miRNAs so as to avoid the imbalance problem caused by different size of positive and negative sample sets. None of the samples included had $\geq 80\%$ sequence similarity to any other in a same subset.

As mentioned in a comprehensive review (27), it is unnecessary to separate a benchmark dataset into a training dataset and a testing dataset for validating a prediction method if it is tested by the jackknife or subsampling (*K*-fold) cross-validation because the outcome thus obtained is actually from a combination of many different independent dataset tests. Therefore, the benchmark dataset S can be formulated as

$$\mathrm{S} = \mathrm{S}^{+} \bigcup \mathrm{S}^{-} \quad (1)$$

where the subset S^+ contains 1,612 human pre-miRNAs, the subset S^- contains1,612 false pre-miRNAs, and the symbol $\cup$ represents the "union" in the set theory. The detailed sequences are given in the **Supplementary Information S1**, which is the largest and most stringent benchmark dataset in this area.

2.2. Cross species test set

11 species, including animals, plants and virus, were selected from miRBase (release 21: June 2014) (24,25) to investigate if the miRNA-dis trained with human pre-miRNAs can be used to predict pre-miRNAs of other species. There are 4607 samples from 11 species, including Caenorhabditis elegans, Caenorhabditis briggsae, Drosophila melanogaster, Drosophila pseudoobscura, Dnio rerio, Gallus gallus, Mus musculusi, Rattus norvegicus, Arabidopsis thaliana, Oryza sativa and Epstein Barr Virus. The pre-miRNAs having higher than 80% sequence similarities with the human pre-miRNAs were removed so as to avoid biased evaluation of the model trained with human data. The similarity is calculated by using CD-HIT with c = 0.8, n = 5. Finally, 4022 non-redundant pre-miRNAs were obtained.

2.3. Distance Structure Status Pairs

Suppose an RNA sequence **R** with L nucleobases (nitrogenous bases or nucleic acid residues) i.e.

$$\mathbf{R} = \mathrm{B}_1\mathrm{B}_2\mathrm{B}_3\mathrm{B}_4\mathrm{B}_5\mathrm{B}_6\mathrm{B}_7\cdots\mathrm{B}_L \tag{2}$$

where B_1 denotes the nucleobase at sequence position 1, B_2 denotes the base at position 2, and so forth. They can be any of the four nucleobases; i.e.,

$$\mathrm{B}_i \in \{\text{adenine(A)}, \text{cytosine(C)}, \text{guanine(G)}, \text{uracil(U)}\} \quad i = 1, 2, \ldots, L \tag{3}$$

If the RNA sequence **R** is formulated according to its secondary structure derived from the Vienna RNA software package (released 2.1.6) (28), we have

$$\mathbf{R} = \Psi_1\Psi_2\Psi_3\Psi_4\Psi_5\ldots\Psi_L \tag{4}$$

where Ψ_1 denotes the structure status of B_1, Ψ_2 the structure status of B_2, and so forth. They can be any of the 10 structure statuses; i.e.,

$$\Psi_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}, \mathrm{A}\text{-}\mathrm{U}, \mathrm{U}\text{-}\mathrm{A}, \mathrm{G}\text{-}\mathrm{C}, \mathrm{C}\text{-}\mathrm{G}, \mathrm{G}\text{-}\mathrm{U}, \mathrm{U}\text{-}\mathrm{G}\} \quad i = 1, 2, \cdots, L \tag{5}$$

where A, C, G, U represent the structure statuses of the four kinds of unpaired nucleobases, while A-U, U-A, G-C, C-G, G-U, U-G represent the structure statuses of the six kinds of paired bases. Note that A-U means the base A located near the 5'-end paired with its complementary base U near the 3'-end. Therefore, A-U and U-A represent two different structure statuses. The same applies to G-C, C-G, G-U, U-G.

In order to capture the structure-order information of the RNA sequence **R** in **Eq.2**, we proposed a new concept called "the distance structure status pair" or just "distance-pair" $D\left(\Psi_i, \Psi_j \mid d\right)$, as formulated by

$$\begin{cases} D\left(\Psi_i, \Psi_j \mid 0\right) & \text{if } d = 0 \text{ then } i = j \\ D\left(\Psi_i, \Psi_j \mid 1\right) & \text{if } d = 1 \\ D\left(\Psi_i, \Psi_j \mid 2\right) & \text{if } d = 2 \\ \vdots & \vdots \\ D\left(\Psi_i, \Psi_j \mid L-1\right) & \text{if } d = L-1 \end{cases} \tag{6}$$

where $0 \le d \le L\text{-}1$, Ψ_i and Ψ_j can be any of the 10 structure statuses of an RNA chain **R** (cf. **Eq.4**), and d represents the value counted by the distance between structure statuses Ψ_i and Ψ_j along the RNA chain **R**. Suppose Ψ_i is A-U, Ψ_j is U-G, and d=3, then D(A-U,U-G|3) means the structure status pair (A-U, U-G) with its two counterparts separated by 2 nucleotides along the RNA chain **R**.

As we can see from **Fig. 1**, the structure status order effects of an RNA chain **R** can be, to some extent, reflected through the distance-pairs as defined by **Eq.6**. The feature vector can be uniquely defined as an Ω-dimensional ($\Omega = 10 + 100d$) vector as:

$$\left[f_1^0 \quad f_2^0 \quad f_3^0 \quad \cdots \quad f_u^k \quad \cdots \quad f_\Omega^d \right]^{\mathrm{T}} \tag{7}$$

where

$$f_u^k = \begin{cases} f_u^0 & \text{if } 1 \le u \le 10 \\ f_u^1 & \text{if } 11 \le u \le 110 \\ f_u^2 & \text{if } 111 \le u \le 210 \\ \vdots & \vdots \\ f_u^d & \text{if } 11+100(d\text{-}1) \le u \le 10+100d \end{cases} \tag{8}$$

where

$$f_u^0 = f\left(D\left(\Psi_i, \Psi_j \mid 0\right)\right), \quad (1 \le u \le 10) \tag{9}$$

meaning the occurrence frequencies of the 10 distance-pairs $D(\Psi_i,\Psi_j|0)$ in **R** (**Fig. 1(a)**);

$$f_u^1 = f\left(D\left(\Psi_i, \Psi_j \mid 1\right)\right), \quad (11 \le u \le 110) \tag{10}$$

meaning the occurrence frequencies of the distance-pairs $D(\Psi_i,\Psi_j|1)$ of the nearest structure status pairs in **R** (**Fig. 1(b)**);

$$f_u^2 = f\left(D\left(\Psi_i, \Psi_j \mid 2\right)\right), \quad (111 \le u \le 210) \tag{11}$$

meaning the occurrence frequencies of the distance-pairs $D(\Psi_i,\Psi_j|2)$ of the second nearest structure status pairs in **R** (**Fig. 1(c)**), and so forth.

The process of generating the feature vector based on "distance-pairs" described above with the structure statuses sequence of the RNA sequence **R** is shown in the **Fig. 1**.

2.4. Support vector machine (SVM)

Support Vector Machines (SVMs) (29) are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis. Given a set of training samples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new samples into one category or the other, making it a non-probabilistic binary linear classifier. A SVM model is a representation of the samples as points in space, where the samples of the separate categories are divided by a clear gap as wide as possible. New samples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to perform linear classification, SVMs can efficiently perform a non-linear classification by using the so called kernel trick, implicitly mapping the inputs into high-dimensional feature spaces.

In the current study, the LIBSVM algorithm (30) was employed, which is software for SVM classification and regression. The kernel function was set as Radial Basis Function (RBF), which is defined as

$$K\left(X_i, X_j\right) = \exp\left(-\gamma \left\| X_i - X_j \right\|^2\right) \qquad (12)$$

The two parameters C and γ were optimized on the benchmark dataset by adopting the grid tool provided by LIBSVM (30), and their actual values in this study will be given later.

2.5. Jackknife test approach

Among the three often used cross-validation methods, i.e., independent dataset test, sub-sampling (or *K*-fold cross-validation) test, and jackknife test, the jackknife test is deemed the least arbitrary and most objective as elucidated in (20), and hence has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors. Therefore, the jackknife test was used to evaluate the performance of the model proposed in the current study. In the jackknife test, each sequence in the benchmark dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the one being identified.

2.6. Criteria for performance evaluation

We adopted sensitivity (Sn), specificity (Sp), accuracy (Acc), and Mathew's Correlation Coefficient (Mcc) to measure the performance of different methods. The calculating formulae are listed below,

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP+FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN+FP}} \\ \mathrm{Acc} = \dfrac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \\ \mathrm{Mcc} = \dfrac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP+FP})(\mathrm{TP+FN})(\mathrm{TN+FP})(\mathrm{TN+FN})}} \end{cases} \quad (13)$$

where TP represents the true positive, FP represents the false positive, TN represents the true negative and FN represents the false negative.

3. Results and Discussion

3.1 Influence of *d* on the predictive performance of miRNA-dis

There is a parameter *d* ($0 \le d < L$, where *L* is the length of the longest RNA sequence in the dataset) in the proposed method miRNA-dis (see method section for details), which would affect the method's predictive performance. We optimized this parameter by using the 5-fold cross validation so as to reduce the computational cost. The Acc values with different *d* values were shown in **Fig.2**, from which we can see that miRNA-dis achieved the best performance when *d* = 7 (Acc = 88.68%), the dimension of the corresponding feature vector is 10+100×7=710. Hence, the parameter *d* was set as 7 in the current study.

3.2. Comparison with other existing related predictor

In the current study, we adopted the benchmark dataset S (cf. **Eq.1**) to evaluate the predictive performance of various methods, which contains 1,612 human pre-miRNAs, 1,612 false pre-miRNAs, and none of the samples had $\ge 80\%$ sequence similarity to any others. The predictive results of miRNA-dis and two other state-of-the-art methods, Triplet-SVM (6) and MiPred (12), were tested by using jackknife validation and the results were shown in the **Table.1**.

To provide an intuitional display of the performance of the three predictors, the corresponding ROC (receiver operating characteristic) curves were drawn in **Fig. 3**, where the horizontal coordinate X is for the false positive rate or 1-Sp, and the vertical coordinate Y is for the true positive rate or Sn. The best possible predictor should yield a point with the coordinate (0, 1) meaning 0 false positive rate (or 100% specificity), and 100% true positive rate or sensitivity Sn. Therefore, the (0, 1) point is also called a perfect classification. A completely random guess would give a point along a diagonal from the point (0, 0) to (1, 1). The area under the ROC curve is called AUC, which is often used to indicate the performance quality of a binary classification predictor: the larger the area, the better the prediction quality is.

From **Table.1** and **Fig. 3** we observed that the predictor miRNA-dis achieved the best performance, outperforming two methods: Triplet-SVM (6) and MiPred (12). Triplet-SVM (6) was a predictor whose features were also derived from the predicted secondary structure. miRNA-dis obviously outperformed Triplet-SVM. The main reason is that Triplet-SVM only considered the local structure status information, while miRNA-dis incorporated the long range or global structure-order effects into the prediction. The good results of miRNA-dis indicate that this is a suitable approach for microRNA precursor identification. miRNA-dis also outperformed MiPred, and it is more efficient than MiPred (see computational efficiency section).

3.3. Computational efficiency

In order to identify microRNA precursors for a large-scale database, methods with low computational cost are required. As discussed above, miRNA-dis outperformed the Triplet-SVM and MiPred. Next, let us investigate the computational efficiency of these methods. In this regard, the computational cost of the vectorization step of these methods, converting the RNA sequences into fixed length vectors, is the bottleneck preventing their widespread application to large databases, for example, the MiPred requires a time consuming P-value feature calculation step. For each query RNA sequence, in order to calculate the P-value feature, the secondary structures of its 1,000 shuffled sequences need to be predicted via running Vienna RNA software (12). By contrast, our method (miRNA-dis) doesn't require any computational expensive step for generating the feature vectors, and therefore, lower computational cost is required.

In order to further illustrate the efficiency of the miRNA-dis approach, its time complexity of vectorization step is analysed. In this approach, each feature vector element f_u^k of a query sequence can be calculated by **Eq.8** with a time complexity of $\mathrm{O}(L)$, where L is the length of the sequence. The total number of f_u^k is $(10+100\times d)\times N$, where N is the total number of samples in the benchmark dataset S, here the number is 3,224 (see method section). The optimal value of d is 7 as illustrated above. Therefore, the time complexity of vectorization step for miRNA-dis is $\mathrm{O}(NdL)$. All the 3,224 RNA sequences in the benchmark dataset S can be converted into fixed length vectors via miRNA-dis in 28 seconds, while for the same task, 35 seconds and 48,712 seconds were required for Triplet-SVM and MiPred, respectively. These experiments were performed on a Linux server with CPU of 2.2 GHz and memory of 94 GB. We conclude that the proposed miRNA-dis approach is able to achieve higher predictive performance with lower computational cost than Triplet-SVM and MiPred.

3.4. Discriminant features' visualization and interpretation

In order to further investigate the discriminant power of the features and reveal the biological meaning of the feature space in miRNA-dis, we followed the

study (31) to calculate the discriminant weight vector in the feature space. The sequence-specific weight obtained from the SVM training process can be used to calculate the discriminant weight of each feature. Given the weight vector $\alpha = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_N]$ of the training set with *N* samples obtained from the kernel-based training, the discriminant weight vector **W** in the feature space can be calculated by the following equation:

$$\mathbf{W} = \alpha^{\mathrm{T}} \cdot \mathbf{M} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1\Omega} \\ m_{21} & m_{22} & \cdots & m_{2\Omega} \\ \vdots & \vdots & \vdots & \vdots \\ m_{N1} & m_{N2} & \cdots & m_{N\Omega} \end{bmatrix} \tag{14}$$

where **M** is the matrix of sequence representatives. The element in **W** represents the discriminative power of the corresponding feature. The discriminant weight vector **W** can be written as:

$$[w_1^0 \quad w_2^0 \quad w_3^0 \quad \ldots \quad w_u^k \quad \ldots \quad w_\Omega^d]^{\mathrm{T}} \tag{15}$$

where

$$w_u^k = w\left(D\left(\Psi_i, \Psi_j \mid k\right)\right) = \begin{cases} w\left(D\left(\Psi_i, \Psi_j \mid 0\right)\right) & \text{if } 1 \le u \le 10 \\ w\left(D\left(\Psi_i, \Psi_j \mid 1\right)\right) & \text{if } 11 \le u \le 110 \\ w\left(D\left(\Psi_i, \Psi_j \mid 2\right)\right) & \text{if } 111 \le u \le 210 \\ \vdots & \vdots \\ w\left(D\left(\Psi_i, \Psi_j \mid d\right)\right) & \text{if } 10+100(d-1) \le u \le 10+100d \end{cases} \tag{16}$$

where $0 \leqslant k \leqslant d$.

In order to reveal the biological meaning of the proposed feature space, the sum score of the discriminant weights for each distance-pair $D(\Psi_i,\Psi_j|k)$ was calculated by the following equation:

$$\begin{cases} S^+\left(\Psi_i, \Psi_j\right) = \sum_{k=0}^{d} w\left(D\left(\Psi_i, \Psi_j \mid k\right)\right) & \text{if } w\left(D\left(\Psi_i, \Psi_j \mid k\right)\right) \ge 0 \\ S^-\left(\Psi_i, \Psi_j\right) = \sum_{k=0}^{d} w\left(D\left(\Psi_i, \Psi_j \mid k\right)\right) & \text{if } w\left(D\left(\Psi_i, \Psi_j \mid k\right)\right) < 0 \end{cases} \tag{17}$$

Fig. 4 (a) and **(b)** showed the positive discriminative power $S^+\left(\Psi_i, \Psi_j\right)$ and negative discriminative power $S^-\left(\Psi_i, \Psi_j\right)$ of all the 100 possible structure status pairs (Ψ_i,Ψ_j) in miRNA-dis approach, respectively. According to the darkest spots in the two figures, structure status pairs (A-U, A-U) and (U-A, U-A) showed the highest positive discriminative power, while structure status pairs (A, A) and (C, C) showed the highest negative discriminative power, indicating these four structure status pairs are very important for identifying human

microRNA precursors. **Fig. 5** showed the specific discriminant weights of the distance-pairs with different *k* values for each of the above four structure status pairs. As can be seen from this figure, the distance-pairs $D(\Psi_i,\Psi_j|0)$ showed the highest discriminant weights for each structure status pair, and the other distance-pairs showed lower values, indicating the distance-pairs $D(\Psi_i,\Psi_j|0)$ are the most important features.

In order to investigate the reason why these four structure status pairs showed strong discriminative power, we calculated their average occurrence frequencies in the subset S^+ and the subset S^- in benchmark dataset S (cf. **Eq.1**), respectively, and the results were shown in **Fig. 6**. Comparing the different frequency distribution of the same distance-pair in the two subsets, we found that for the two distance-pairs *D*(A-U,A-U|*k*) and *D*(U-A,U-A|*k*) with the highest positive discriminative power as shown in **Fig. 4(a)**, their occurrence frequencies on S^+ are much higher than those on S^- (**Fig. 6 (a)** and **(b)**), which could explain the reason why these two distance-pairs showed positive discriminative power, and indicate they might contain important characteristics of human pre-miRNAs. Similar patterns were also observed for the two distance-pairs *D*(A,A|*k*), *D*(C,C|*k*) with highest negative discriminative power (**Fig. 4(b)**), they tend to occur in the false pre-miRNAs as shown in **Fig. 6(c)** and **(d)**, which is the reason why they show negative discriminative power.

Two RNA sequences were selected from the benchmark dataset as two examples (**Fig. 7**) so as to show the distribution of the four kinds of important distance-pairs (*D*(A-U,A-U|*k*), *D*(U-A,U-A|*k*), *D*(A,A|*k*), *D*(C,C|*k*), *k*=0, 1, 2, 3, 4, 5, 6, 7) in these two RNA sequences. One selected sample is a real human microRNA precursor (positive sample, ID: has-mir-3713_ss), another one is a false microRNA precursor (negative sample, ID: random_seq_from_cds_NO_6886_ss). We clearly observed the following. (i) Most of the distance-pairs *D*(A,A|*k*) and *D*(C,C|*k*) occur at the loop of the hairpin structure, while almost all the distance-pairs *D*(A-U,A-U|*k*) and *D*(U-A,U-A|*k*) occur at the stem of the hairpin; (ii) The distance-pairs *D*(A,A|*k*) and *D*(C,C|*k*) are abundant in this negative sample, while the distance-pairs *D*(A-U,A-U|*k*) and *D*(U-A,U-A|*k*) tend to occur in this positive sample, which is fully consistent with the experimental results shown in **Fig. 4** and **Fig. 6** that the *D*(A-U,A-U|*k*) and *D*(U-A,U-A|*k*) have positive discriminative power, and the *D*(A,A|*k*) and *D*(C,C|*k*) have negative discriminative power.

Independent test

The benchmark dataset was constructed based on miRBase release 20 (June 2013). At the time this paper was being written, 16 new human pre-miRNAs were reported by the latest miRBase release 21 (June 2014). These 16 pre-miRNAs were treated as an independent test set to further evaluate the performance of the proposed miRNA-dis. miRNA-dis trained with the benchmark date set can correctly predict 15 testing samples in the independent data set as true human pre-miRNAs. Its overall accuracy is 93.75%, which demonstrates the stable predictive performance of miRNA-dis for predicting human pre-miRNAs.

Cross species experiments.

After years of studies, we have known almost all the human pre-miRNAs. Therefore, if one computational predictor can only predict human pre-miRNAs, its value is limited. It is interesting to investigate if the miRNA-dis trained with the human pre-miRNAs can be used to predict the pre-miRNAs of other species. We applied the miRNA-dis trained only with human data to predict the 4022 pre-miRNAs from 11 other species ranging from animals, plants, and virus. These 11 species were also used to evaluate the performance of Triplet-SVM for cross species prediction (6). The predictive results of miRNA-dis on the 11 species were shown in **Table.2**. The miRNA-dis achieved an overall accuracy of 87.02%, which is highly comparable with the accuracy reported by Triplet-SVM (6) (90.0%). However, Triplet-SVM was only tested with 581 samples, while miRNA-dis was extensively evaluated with 4022 samples extracted from the latest miRBase (release 21: June 2014). The high performance of miRNA-dis across the wide range of different species is not surprising. Previous studies (6) showed that the structure-sequence features were conserved in pre-miRNAs of different species. The miRNA-dis can efficiently extract these structure-sequence characteristics, which would be the main reason for its stable performance for predicting pre-miRNAs from various species.

3.5. Web-Server Guide

As mentioned in (31) and implemented in a series of recent publications (see, e.g., (32-34)), user-friendly and publicly accessible web-servers are indispensable for developing practically more useful predictors, a web-server for the current predictor miRNA-dis has also been established. Furthermore, for the convenience of the vast majority of experimental scientists, a step-by-step guide on how to use the web-server to get their desired results without the need to follow the complicated mathematic equations is given below.

Step 1. Visit the web-server by clicking the link at http://bioinformatics.hitsz.edu.cn/miRNA-dis/ and you will see its top page as shown in **Fig.8**. Click on the Read Me button to see a brief introduction about the server.

Step 2. You can either type or copy and paste the query RNA sequence into the input box at the center of **Fig. 8**, or directly upload your input data by the Browse button. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with the symbol, > , in the first column, followed by lines of sequence data in which nucleotides or amino acids are represented using single-letter codes. Except for the mandatory symbol >, all the other characters in the single initial line are optional and only used for the purpose of identification and description. The sequence ends if another line starting with the symbol >appears; this indicates the start of another sequence. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box.

Step 3. Click on the Submit button to see the predicted result. For example, if you use the four query RNA sequences in the Example window as the input and then click the Submit button, you will see on your screen that the predicted results for the 1st and 2nd query RNA sequences are "**Real Pre-miRNA**", and that for the 3rd and 4th ones are "**False Pre-miRNA**". All these predicted results are fully consistent with the experimental observations.

4. Conclusion

In this study, we proposed a computational method "miRNA-dis" for human microRNA precursor identification, in which the feature vector was constructed based on the occurrences of "distance structure status pairs" $D(\Psi_i,\Psi_j|k)$, $0 \le k \le d$ .By this way, long range or global structure-order information was approximately incorporated into the predictor. It was observed via the rigorous cross-validation on a larger and more stringent newly constructed benchmark dataset that the new predictor outperformed two state-of-art predictors Triplet-SVM (6) and MiPred (12) and it was more efficient than MiPred, because in MiPred, for each query RNA sequence, the secondary structures of its 1000 shuffled sequences need to be predicted via running Vienna RNA software in order to calculate the P-value feature. Another important advantage of our approach arises from the explicit feature space representation: the possibility to calculate the discriminant weight vector in feature space, which allows the users to analyse the learnt model for identification of the most discriminative features. These features, which correspond to distance-pairs, may in turn reveal biologically relevant properties of human microRNA precursors. Cross species experiments showed that the current predictor can be easily used to identify the pre-microRNAs of other species, indicating that miRNA-dis would be a useful high throughput tool for large-scale analysis of microRNA precursors of various species.

Acknowledgements

We would like to thank the two reviewers for their constructed comments, which are very helpful for further strengthening the presentation of this paper. This work was supported by the National Natural Science Foundation of China (61300112 and 61272383), the Scientific Research Innovation Foundation in Harbin Institute of Technology (Project No. HIT.NSRIF.2013103), the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and Shenzhen Municipal Science and Technology Innovation Council (Grant No. CXZZ20140904154910774).

Table 1. Results on benchmark dataset (**Appendix A**) for different methods through jackknife validation

Method	Acc(%)	Mcc	Sn(%)	Sp(%)	AUC
Triplet-SVM[a]	81.85	0.64	78.47	85.20	0.894
MiPred[b]	87.30	0.75	84.00	90.60	0.941
miRNA-dis[c]	88.92	0.78	89.39	88.47	0.953

[a]Results computed by in-house implementation of Triplet-SVM (6) on the benchmark dataset, the parameters used: $C = 2^{11}$, and $\gamma = 2^{-3}$.

[b]Results computed by in-house implementation of MiPred (12) on the benchmark dataset.

[c]The parameters used: $d = 7$, $C = 2$, and $\gamma = 2^{-11}$

Table 2. Results on cross species test set for miRNA-dis trained with human data

Species	Number of samples	Acc(%)
Mus musculus	962	85.55
Rattus norvegicus	277	87.00
Gallus gallus	659	74.66
Danio rerio	291	93.47
Caenorhabditis briggsae	175	91.43
Caenorhabditis elegans	250	85.60
Drosophila pseudoobscura	210	87.14
Drosophila melanogaster	256	85.94
Oryza sativa	592	94.09
Arabidopsis thaliana	325	96.62
Epstein Barr Virus	25	96.00
Total	4022	87.02

Figures Legends

Figure 1. The process of generating the feature vector based on distance-pairs. Given an RNA sequence **R** with L nucleobases, the structure status sequence of **R** can be derived by Vienna RNA software package. Suppose d = 2, the $f(D(\Psi_i,\Psi_j|0)),f(D(\Psi_i,\Psi_j|1)),f(D(\Psi_i,\Psi_j|2))$ are calculated based on the structure status sequence of **R**, and then the feature vector is generated according to the **Eq. 7-11**.

Figure 2. **The overall Acc values achieved by miRNA-dis with different *d* values based on the benchmark dataset through five-fold cross validation.**

Figure 3. **A graphical illustration to show the performance of different methods on the benchmark dataset using the jackknife tests by means of the receiver operating characteristic (ROC) curves.** The areas under the ROC curves or AUC are 0.953, 0.894, and 0.941 for miRNA-dis, Triplet-SVM, and MiPred, respectively.

Figure 4. **An illustration for the discriminant visualization.** The structure statuses labelled by y-axis and x-axis indicate the first structure status and the second structure status in the distance-pairs, respectively. The adjacent colour bar shows the mapping of sum score values. (a) The positive discriminative power $S^+\left(\Psi_i,\Psi_j\right)$ of all the 100 possible structure status pairs (Ψ_i,Ψ_j) in miRNA-dis. (b) The negative discriminative power $S^-\left(\Psi_i,\Psi_j\right)$ of all the 100 possible structure status pairs (Ψ_i,Ψ_j) in miRNA-dis.

Figure 5. **The discriminant weights of *D*(A-U,A-U|*k*), *D*(U-A,U-A|*k*), *D*(A,A|*k*), *D*(C,C|*k*) with** $0 \le k \le 7$ **.**

Figure 6. **The average occurrence frequencies of the distance-pairs for the four structure status pairs (A-U, A-U), (U-A, U-A), (A, A) and (C, C) in the two classes (real pre-miRNAs vs. false pre-miRNAs).**

Figure 7. **Examples of the distribution of *D*(A-U,A-U|*k*), *D*(U-A,U-A|*k*), *D*(A,A|*k*), *D*(C,C|*k*) in RNA sequences.** The distance-pairs (D(A-U,A-U|k), D(U-A,U-A|k)) with positive discriminative power were labeled in pink rectangles, and the distance-pairs (D(A,A|k), D(C,C|k)) with negative discriminative power were labeled in grey blue rectangles. The distribution of these distance-pairs in a real microRNA precursor (ID: has-mir-3713_ss) and a false microRNA precursor (ID: random_seq_from_cds_NO_6886_ss) were given in subfigure (a) and (b), respectively. For each subfigure, the first figure shows the predicted secondary structure of the target RNA, and the second figure shows its structure status sequence (cf. **Eq.4**).

Figure 8. A semi-screenshot to show the top page of the web-server miRNA-dis, which is available at http://bioinformatics.hitsz.edu.cn/miRNA-dis/

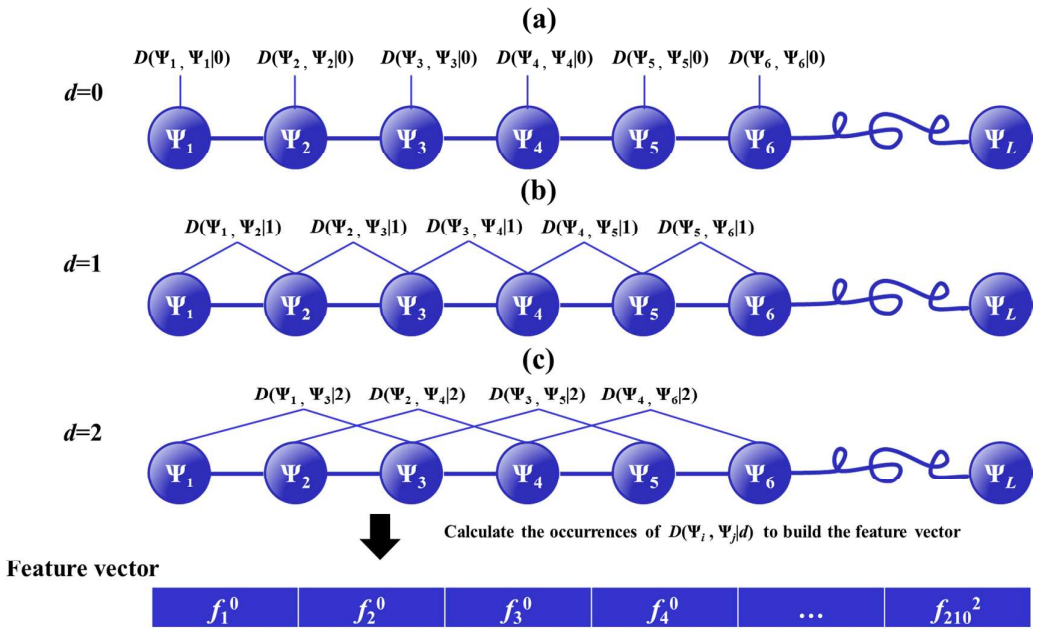


Figure 1

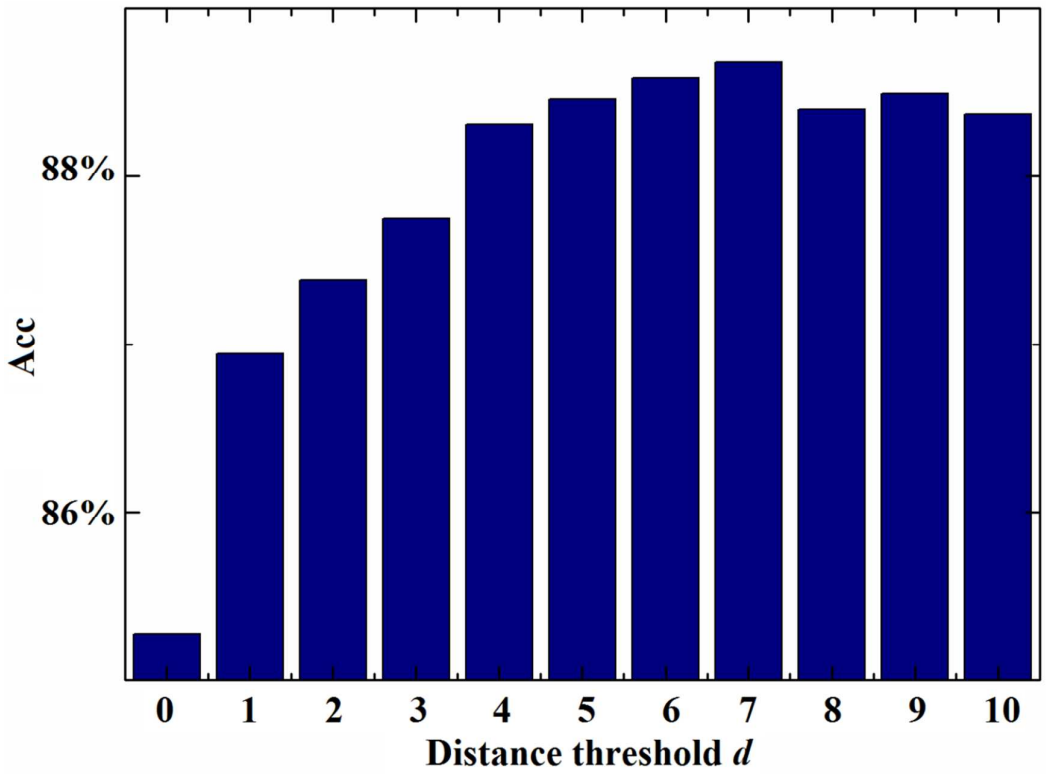


Figure 2

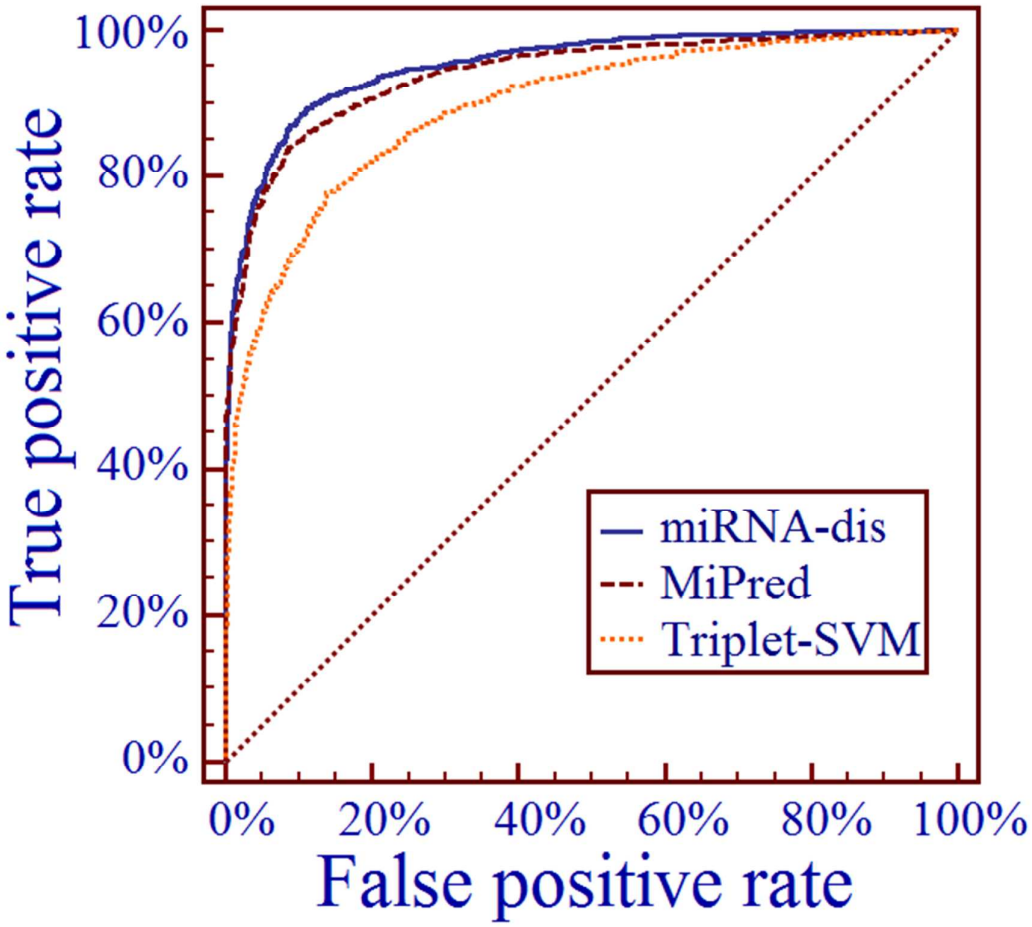


Figure 3

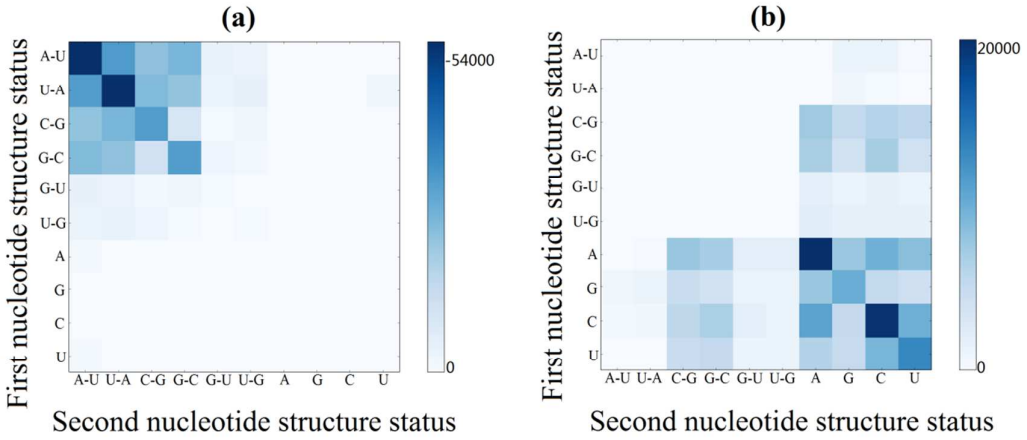


Figure 4

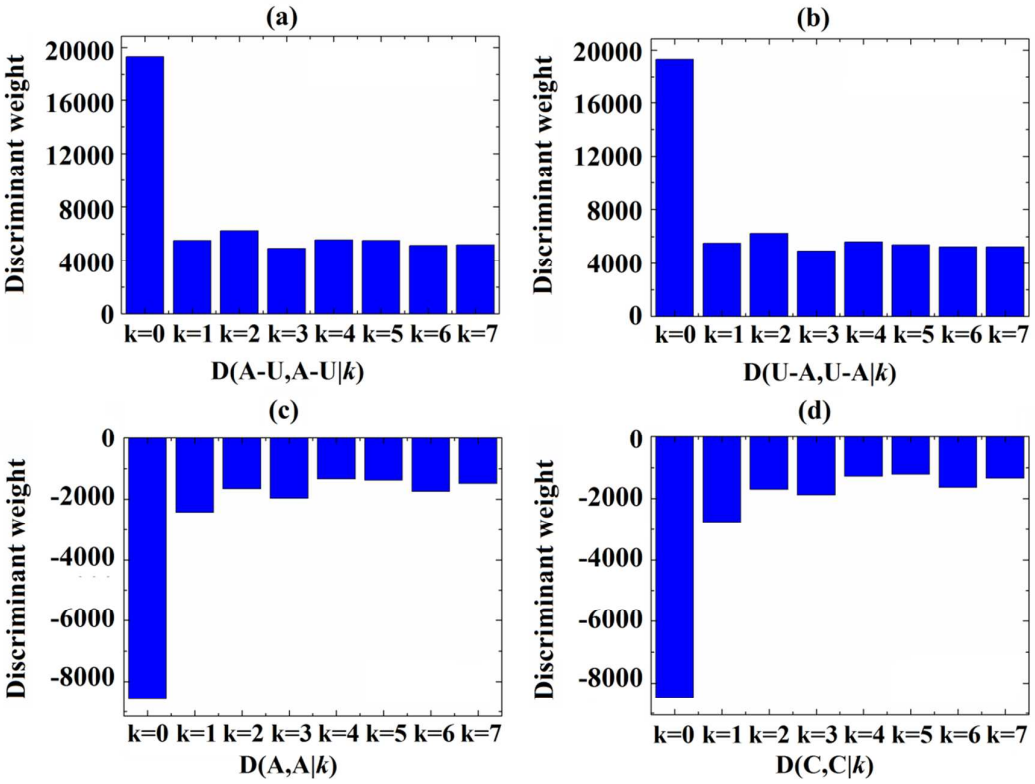


Figure 5

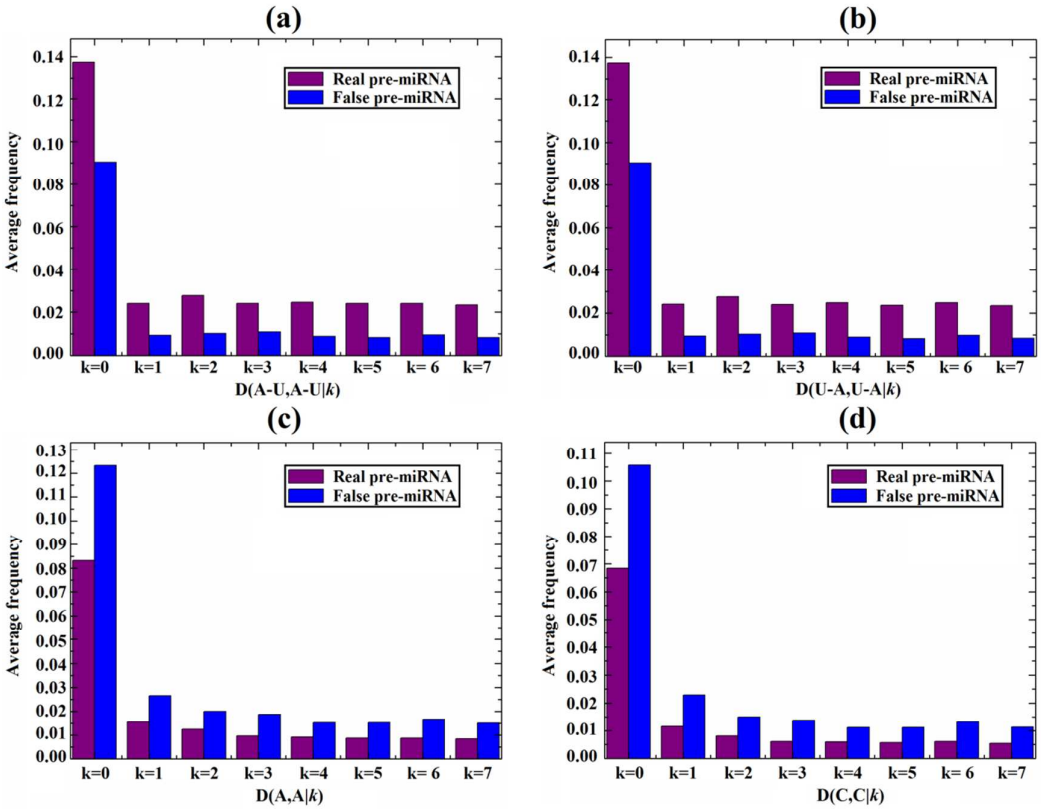


Figure 6

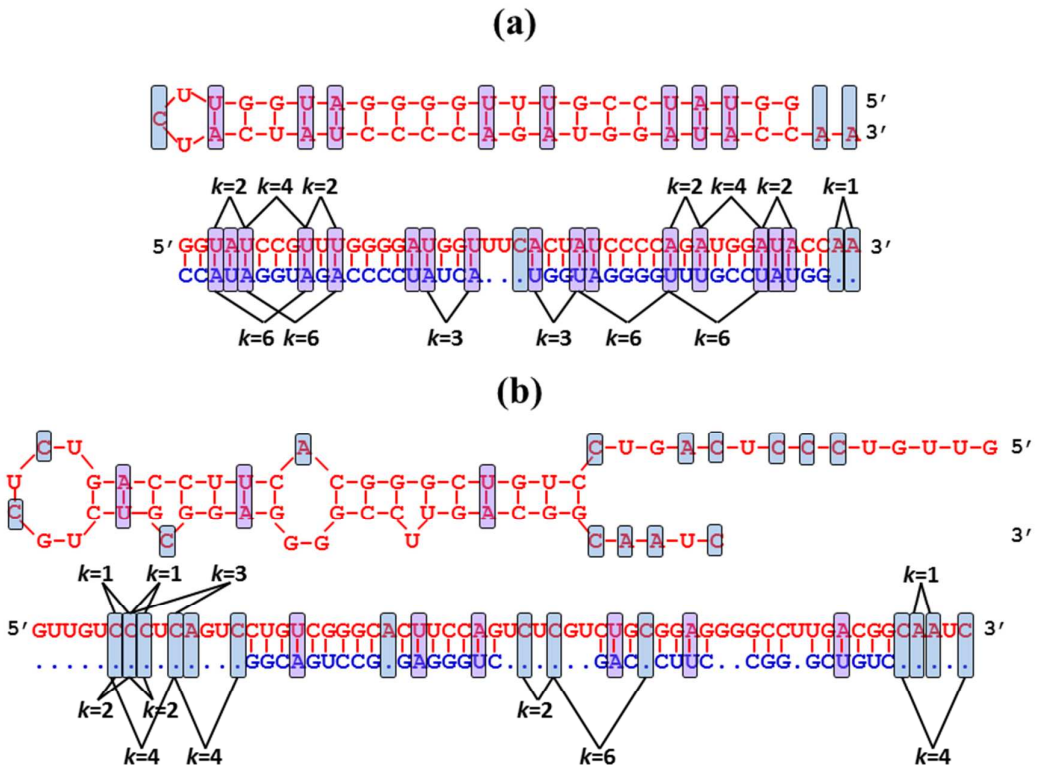


Figure 7

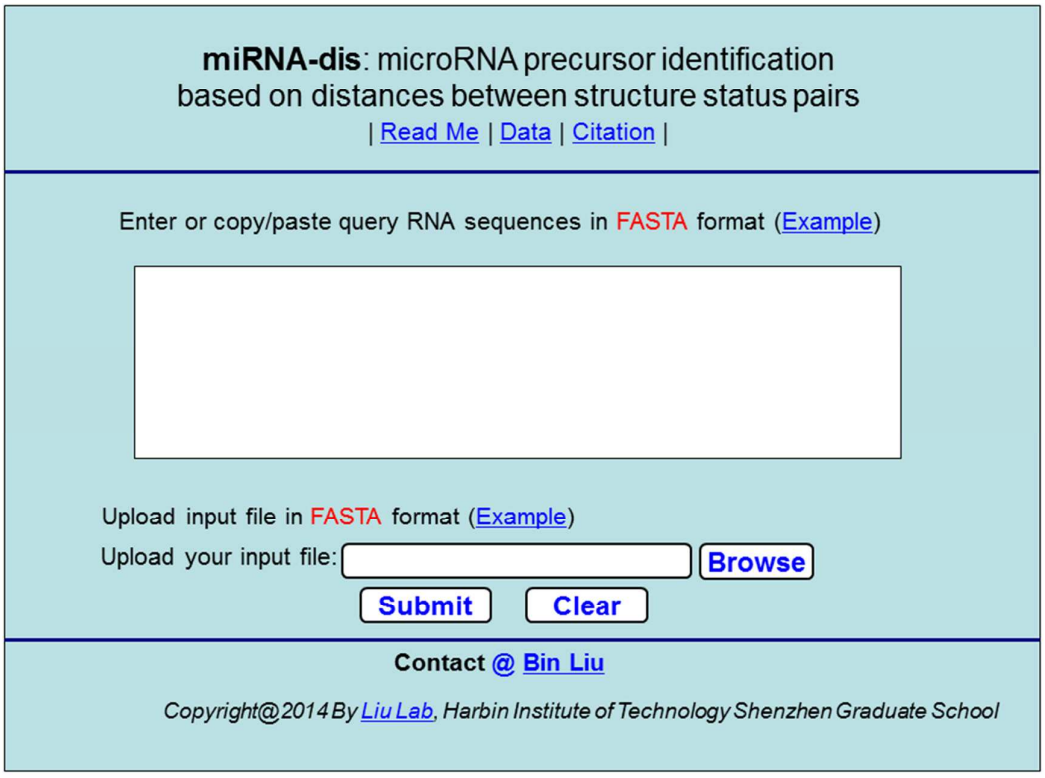


Figure 8

References

1. Chen, K. and Rajewsky, N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nature reviews. Genetics*, **8**, 93-103.
2. Fasanaro, P., Greco, S., Ivan, M., Capogrossi, M.C. and Martelli, F. (2010) microRNA: emerging therapeutic targets in acute ischemic diseases. *Pharmacology & therapeutics*, **125**, 92-104.
3. Trang, P., Weidhaas, J.B. and Slack, F.J. (2008) MicroRNAs as potential cancer therapeutics. *Oncogene*, **27 Suppl 2**, S52-57.
4. Li, C., Feng, Y., Coukos, G. and Zhang, L. (2009) Therapeutic microRNA strategies in human cancer. *The AAPS journal*, **11**, 747-757.
5. Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S. and Rajewsky, N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotech*, **26**, 407-415.
6. Xue, C., Li, F., He, T., Liu, G.P., Li, Y. and Zhang, X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC bioinformatics*, **6**, 310.
7. Nam, J.W., Shin, K.R., Han, J., Lee, Y., Kim, V.N. and Zhang, B.T. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic acids research*, **33**, 3570-3581.
8. Huang, T.H., Fan, B., Rothschild, M.F., Hu, Z.L., Li, K. and Zhao, S.H. (2007) MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC bioinformatics*, **8**, 341.
9. Wu, Y., Wei, B., Liu, H., Li, T. and Rayner, S. (2011) MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC bioinformatics*, **12**, 107.
10. Wang, Y., Chen, X., Jiang, W., Li, L., Li, W., Yang, L., Liao, M., Lian, B., Lv, Y., Wang, S. *et al.* (2011) Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM. *Genomics*, **98**, 73-78.
11. Helvik, S.A., Snove, O., Jr. and Saetrom, P. (2007) Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, **23**, 142-149.
12. Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X. and Lu, Z. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic acids research*, **35**, W339-344.
13. Agarwal, S., Vaz, C., Bhattacharya, A. and Srinivasan, A. (2010) Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). *BMC bioinformatics*, **11 Suppl 1**, S29.
14. Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L.C. and Showe, M.K. (2006) Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, **22**, 1325-1334.
15. Brameier, M. and Wiuf, C. (2007) Ab initio identification of human microRNAs based on structure motifs. *BMC bioinformatics*, **8**, 478.
16. Li, L., Xu, J., Yang, D., Tan, X. and Wang, H. (2010) Computational approaches for microRNA studies: a review. *Mammalian genome : official journal of the International Mammalian Genome Society*, **21**, 1-12.
17. Wei, L.Y., Liao, M.H., Gao, Y., Ji, R.R., He, Z.Y. and Zou, Q. (2014) Improved and Promising Identification of Human MicroRNAs by Incorporating a

High-quality Negative Set. *Computational Biology and Bioinformatics*, **11**.
18. Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *PROTEINS: Structure, Function, and Genetics*, **43**, 246-255.
19. Liu, B., Wang, X., Chen, Q., Dong, Q. and Lan, X. (2012) Using amino acid physicochemical distance transformation for fast protein remote homology detection. *PloS one*, **7**, e46633.
20. Chen, W., Feng, P.M., Lin, H. and Chou, K.C. (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic acids research*, **41**, e68.
21. Qiu, W.R., Xiao, X. and Chou, K.C. (2014) iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. *International journal of molecular sciences*, **15**, 1746-1766.
22. Liu, B., Liu, F., Fang, L., Wang, X. and Chou, K.-C. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, doi: 10.1093/bioinformatics/btu1820.
23. Guo, S.H., Deng, E.Z., Xu, L.Q., Ding, H., Lin, H., Chen, W. and Chou, K.C. (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*.
24. Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, **39**, D152-157.
25. Ambros, V. (2003) A uniform system for microRNA annotation. *Rna*, **9**, 277-279.
26. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658-1659.
27. Chou, K.C. and Shen, H.B. (2007) Recent progress in protein subcellular location prediction. *Analytical biochemistry*, **370**, 1-16.
28. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic acids research*, **31**, 3429-3431.
29. CORTES, C. and VAPNIK, V. (1995) Support-Vector Networks. *Machine Leaming*, **20**, 273-297.
30. Chang, C.C. and Lin, C.J. (2009) LIBSVM A Library for Support Vector Machines.
31. Chou, K.C. and Shen, H.B. (2009) Recent advances in developing web-servers for predicting protein attributes. *Natural Science*, **1**, 63-92.
32. Xiao, X., Min, J.-L., Wang, P. and Chou, K.-C. (2013) iCDI-PseFpt: Identify the channel–drug interaction in cellular networking with PseAAC and molecular fingerprints. *Journal of theoretical biology*, **337**, 71-79.
33. Min, J.L., Xiao, X. and Chou, K.C. (2013) iEzy-drug: a web server for identifying the interaction between enzymes and drugs in cellular networking. *BioMed research international*, **2013**, 701317.
34. Xu, Y., Shao, X.J., Wu, L.Y., Deng, N.Y. and Chou, K.C. (2013) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting

cysteine S-nitrosylation sites in proteins. *PeerJ*, **1**, e171.