iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach

Bin Liu\(^{a,b,c}\), Longyun Fang\(^{a}\), Fule Liu\(^{a}\), Xiaolong Wang\(^{a,b}\) and Kuo-Chen Chou\(^{c,d}\)

\(^{a}\)School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China; \(^{b}\)Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China; \(^{c}\)Gordon Life Science Institute, Belmont, MA 02478, USA; \(^{d}\)Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

Communicated by Ramaswamy H. Sarma

(Received 17 October 2014; accepted 29 January 2015)

A microRNA (miRNA) is a small non-coding RNA molecule, functioning in transcriptional and post-transcriptional regulation of gene expression. The human genome may encode over 1000 miRNAs. Albeit poorly characterized, miRNAs are widely deemed as important regulators of biological processes. Aberrant expression of miRNAs has been observed in many cancers and other disease states, indicating that they are deeply implicated with these diseases, particularly in carcinogenesis. Therefore, it is important for both basic research and miRNA-based therapy to discriminate the real pre-miRNAs from the false ones (such as hairpin sequences with similar stem-loops). Particularly, with the avalanche of RNA sequences generated in the post-genomic age, it is highly desired to develop computational sequence-based methods for effectively identifying the human pre-miRNAs. Here, we propose a predictor called “iMiRNA-PseDPC”, in which the RNA sequences are formulated by a novel feature vector called “pseudo distance-pair composition” (PseDPC) with 10 types of structure statuses. Rigorous cross-validations on a much larger and more stringent newly constructed benchmark data-set showed that our approach has remarkably outperformed the existing ones in either prediction accuracy or efficiency, indicating the new predictor is quite promising or at least may become a complementary tool to the existing predictors in this area. For the convenience of most experimental scientists, a user-friendly web server for the new predictor has been established at http://bioinformatics.hitsz.edu.cn/iMiRNA-PseDPC/, by which users can easily get their desired results without the need to go through the mathematical details. It is anticipated that the new predictor may become a useful high throughput tool for genome analysis particularly in dealing with large-scale data.

**Keywords:** pre-miRNA; Chou’s PseAAC approach; local structure status; free energy; iMiRNA-PseDPC; large-scale analysis

1. **Introduction**

MicroRNAs (miRNAs) are small single-strand and non-coding RNAs, which play important roles in gene regulation by targeting messenger RNAs (mRNAs) for cleavage or translational repression. Their length is about 17–25 nt (Lopes, Schliep, & Carvalho, 2014). The miRNAs are also involved in many important biological processes, such as affecting stability, translation of mRNAs, and negatively regulating gene expression in post-transcriptional processes. Because using traditional experimental techniques to timely and systematically detect miRNAs from a genome is difficult (Xuan et al., 2011), it is highly demanded to develop computational methods for identifying miRNAs based on their sequence information.

In the last decade or so, some efforts have been made by using different features and machine learning techniques to identify miRNAs. Features derived from RNA sequences or their predicted secondary structures were proposed to capture the characteristics of miRNAs, for example, k-mer (sub-sequence of RNAs) is one of the main sequence-based features reflecting the local sequence composition of RNAs (Wei et al., 2014). Because most of the pre-miRNAs have the characteristic of stem-loop hairpin structures (Xue et al., 2005), some features were constructed based on the predicted secondary structures so as to reflect this characteristic; e.g. a set of 32 local triplet sequence-structure features were used in the Triplet-SVM (Xue et al., 2005) to predict the human miRNAs (Xue et al., 2005). The minimum of free energy of the secondary structure and the P-value of randomization test were efficient features used in this field, which are based on the fact that miRNAs in the folding state have lower free energies than their random sequences (Jiang et al., 2007). Meanwhile, various computational predictors were constructed based on the above mentioned features and some well-known machine learning techniques, such as support vector machine (SVM)
(Helvik, Snove, & Saetrom, 2007; Huang et al., 2007; Nam et al., 2005; Wang et al., 2011; Wu, Wei, Liu, Li, & Rayner, 2011; Xue et al., 2005), random forest (Jiang et al., 2007), hidden Markov model (Agarwal, Vaz, Bhattacharya, & Srinivasan, 2010), naive Bayes (Yousef et al., 2006), and linear genetic programming (Brameier & Wiuf, 2007). Recently, using feature mining and AdaBoost algorithms, Zhong et al. (2013) developed an efficient method called MtrID for miRNA identification.

Previous studies showed that the global or long-range structural effects and the sequence-order information of RNA are important for improving the prediction quality. This is because in the folding state, the nucleobases far away from each other along an RNA sequence may be quite close according to their spatial distance. However, it is by no means an easy job to incorporate this kind of information into a predictor. This is because the number of different sequence-order combinations is extremely high as can be shown with the following illustration, which is quite similar to the one in (Chou, 1996). Take an RNA with 70 nt formed by four types of nucleobases (A, C, G, U) as an example, the number of its total possible different sequence-order combinations will be $4^{70} = 1.394 \times 10^{42}$. For such an astronomical number, it would be impracticable to construct a reasonable training data-set to statistically cover all the possible different sequence-order patterns. Besides, miRNA sequences have different lengths, and hence posing an additional difficulty for constructing a benchmark data-set able to incorporate the sequence-order information. Furthermore, almost all the existing machine learning algorithms were developed and used to predict DNA recombination spots (Chen, Feng, & Lin, 2013; Qiu & Xiao, 2014), DNA splicing sites (Chen, Feng, & Lin, 2014), nucleosome-positioning in genomes (Guo, Deng, Xu, Ding, & Lin, 2014), translation initiation site in human genes (Chen, Deng, & DENG, 2014), and sigma-54 promoters (Lin, Deng, & Ding, 2014).

Encouraged by the successes of using PseAAC and PseKNC approaches to deal with various problems in proteomics, genomics, and genome analysis, as well as the drug target area (Chou, 2015; Zhong & Zhou, 2014), here we are to propose a new predictor for identifying the miRNA precursors. In the new predictor, an RNA sequence is formulated by a novel feature vector called “pseudo distance-pair composition” (PseDPC), where its global or long-range sequence-order effects is partially incorporated via the secondary structure chain.

As shown by a series of recent publications (Chen et al., 2013; Chen, Feng, & Deng, 2014; Fan, Xiao, & Min, 2014; Guo et al., 2014; Liu, Xu, et al., 2014; Qiu & Xiao, 2014; Qiu, Xiao, & Lin, 2014; Xu, Ding, & Wu, 2013; Xu, Wen, Wen, & Wu, 2014; Xu, Zhou, et al., 2014) in response to the call from (Chou, 2011), it would make the development of new predictor logically more clear and practically more useful if it can be documented according to the following procedures: (1) construct or select a valid benchmark data-set to train and test the predictor; (2) formulate the samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (5) establish a user-friendly web server for the predictor that is accessible to the public. Below, we are to present our new predictor according to such logical procedures.

2. Materials and method
2.1. Benchmark data-set
The pre-miRNAs or positive samples were acquired from the miRNABase (release 20: June 2013) (Ambros, 2003; Kozomara & Griffiths-Jones, 2011), which contains 1872
experiment-confirmed sapiens pre-miRNA entries. The false pre-miRNAs or negative samples were obtained from the data constructed by Xue et al. (2005), which contains 8489 false pre-miRNA samples. They are sequence segments collected from the protein coding regions (CDSs) and have similar stem-loop structures as genuine pre-miRNAs but have not been reported as pre-miRNAs. Since all reported miRNAs are located in the un-translated regions or intergenic regions, it is reasonable to take the hairpins collected from CDS sequences as examples of pseudo pre-miRNAs. Both the false pre-miRNAs and the real pre-miRNAs share the following common features which ensure that the extracted pseudo pre-miRNAs are similar to real pre-miRNAs: (1) the RNA length ranges from 51 to 137 nt; (2) the stem of the hairpin structure has a minimum of 18 base pairings; and (3) the secondary structure has a maximum of −15 kal/mol free energy.

To get a high quality of benchmark data-set by getting rid of the redundancy and avoiding homology bias, the CD-HIT software (Li, Feng, Coukos, & Zhang, 2009; Li & Godzik, 2006) with the cutoff threshold set at 80% (note that the most stringent cutoff threshold for DNA sequences by CD-HIT is 75%) was utilized to winnow those samples that have ≥80 sequence similarity to any others in a same subset. By doing so, we obtained 1612 positive samples.

Also, to avoid the subset size imbalance problem, we randomly picked 1612 samples from the 8489 false pre-miRNAs. Again, none of the 1612 samples has ≥80 sequence identity to any other in a same subset. As mentioned in a comprehensive review (Chou & Shen, 2007), it is unnecessary to separate a benchmark data-set into a training data-set and a testing data-set 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 data-set tests. Therefore, the benchmark data-set \( S \) can be formulated as

\[
S = S^+ \cup S^-
\]

where the positive subset \( S^+ \) contains 1612 human pre-miRNAs, the negative subset \( S^- \) contains 1612 false pre-miRNAs, and the symbol \( \cup \) represents the “union” in the set theory. The detailed sequences of the 1,612 × 2 = 3,224 RNA samples are given in the Supplementary material, which is so far the most stringent and largest benchmark data-set in this area.

### 2.2. Pseudo distance structure status pair composition

Stimulated by the PseAAC approach (Chou, 2001a, 2005) in computational proteomics, here we are to propose a novel feature vector called the pseudo distance structure status pair composition or just PseDPC to incorporate the global or long-range structure-order information so as to improve the prediction quality in identifying the pre-miRNAs. The detailed procedures are as follows.

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

\[
R = B_1B_2B_3B_4B_5B_6B_7 \ldots B_L
\]

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

\[
B_i \in \{ \text{adenine (A), cytosine (C), guanine (G), uracil (U)} \}
\quad i = 1, 2, \ldots, L
\]

By using the Vienna RNA software package (released 2.1.6) (Hofacker, 2003), the secondary structure of RNA sequence \( R \) can be identified and can be represented as \( R' \):

\[
R' = \Psi_1\Psi_2\Psi_3\Psi_4\Psi_5 \ldots \Psi_L
\]

where \( \Psi_1 \) denotes the structure status of \( B_1 \), \( \Psi_2 \) denotes the structure status of \( B_2 \), and so forth. They can be any of the 10 structure statuses; i.e.

\[
\Psi_i \in \{ \text{A, C, G, U, A-U, U-A, G-C, C-G, G-U, U-G} \}
\quad i = 1, 2, \ldots, L
\]

where \( A, C, G, \) and \( U \) represent the structure statuses of the four kinds of unpaired nucleobases, while A-U, U-A, G-C, C-G, G-U, and 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 is true to G-C, C-G, G-U, and U-G.

In order to capture the sequence-order information for the RNA sequence \( R \) in Equation (2), we introduce a new concept called “distance structure status pair” or just “distance-pair” \( D(\Psi_i, \Psi_j|d) \) and its occurrences can be represented as:

\[
f(D(\Psi_i, \Psi_j|d)), \quad 0 \leq d \leq L
\]

where \( \Psi_i \) and \( \Psi_j \) can be any of the 10 structure statuses of a RNA chain \( R' \) (cf. Equation (4)), and \( d \) represents the distance between structure statuses \( \Psi_i \) and \( \Psi_j \) along the RNA chain \( R' \).

Suppose \( \Psi_i \) is A, \( \Psi_j \) is C, and \( d = 3 \), then \( f(D(A, C|3)) \) means the occurrence frequency of the
A-C structure status pair with its two counterparts separated by two bases along the RNA chain. Thus, when \( d = 0 \), Equation (6) is reduced to

\[
f(D(\Psi_i, \Psi_j|0)) = f^0(u), \quad (1 \leq u \leq 10)
\]

(7)

meaning the occurrence frequencies of the 10 structure status in the RNA; when \( d = 1 \), we have

\[
f(D(\Psi_i, \Psi_j|1)) = f^1(u), \quad (11 \leq u \leq 110)
\]

(8)

meaning the occurrence frequencies of the 10 \( \times 10 = 100 \) nearest structure status pairs (Liu & Chou, 1999); when \( d = 2 \), we have

\[
f(D(\Psi_i, \Psi_j|2)) = f^2(u), \quad (111 \leq u \leq 210)
\]

(9)

meaning the occurrence frequencies of the 10 \( \times 10 = 100 \) second nearest structure status pairs (Xu, Shao, Wu, & Deng, 2013), and so forth.

According to (Chou, 2011; Du et al., 2014), the feature vector of any biological sequence (including RNA sequence) can be formulated by the form of Chou’s general PseAAC; thus, Equation (2) can be written as

\[
R = [\Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_u \quad \cdots \quad \Psi_\Omega]^T
\]

(10)

where \( T \) is the transpose operator, while \( \Omega \) is an integer to reflect the vector’s dimension. The value of \( \Omega \) as well as the components \( \Psi_u (u = 1, 2, \ldots, \Omega) \) in Equation (10) will depend on how to extract the desired information from an RNA sequence. Similar to Equation (12) of (Liu, Xu, et al., 2014), the dimension of the feature vector for the RNA sample (Equation (10)) is \( \Omega = 10 + 100d \) and its components are given by

\[
\Psi_u = \begin{cases}
  f^0(u) & \text{if } 1 \leq u \leq 10 \\
  f^1(u) & \text{if } 11 \leq u \leq 110 \\
  f^2(u) & \text{if } 111 \leq u \leq 210 \\
  \vdots \\
  f^d(u) & \text{if } 11 + 100(d-1) \leq u \leq 10 + 100d \\
  (0 \leq d \leq L)
\end{cases}
\]

(11)

The process of generating the feature vector based on distance-pairs described above with the structure statuses of the RNA sequence \( R \) is shown in the Figure 1.

On the other hand, in a way parallel to the formulation in (Chou, 2001a), the global structure-order information for the RNA structure status sequence of Equation (4) can be reflected by a series correlation factors as given by

\[
\theta_1 = \frac{1}{L} \sum_{i=1}^{L-1} \Theta(\Psi_i, \Psi_{i+1})
\]

\[
\theta_2 = \frac{1}{L} \sum_{i=1}^{L-2} \Theta(\Psi_i, \Psi_{i+2})
\]

(12)

\[
(\lambda < L)
\]

\[
\theta_3 = \frac{1}{L} \sum_{i=1}^{L-3} \Theta(\Psi_i, \Psi_{i+3})
\]

\[
\theta_4 = \frac{1}{L} \sum_{i=1}^{L-4} \Theta(\Psi_i, \Psi_{i+4})
\]

\[
\vdots
\]

where \( \lambda \) is an integer, representing the highest counted rank (or tier) of the structural correlation along an RNA chain; \( \theta_1 \) is the first-tier correlation factor reflecting the structure-order information between all the most contiguous nucleobases along an RNA chain (Figure 2(a)); \( \theta_2 \) is the second-tier correlation factor between all the second most contiguous nucleobases (Figure 2(b)); \( \theta_3 \) the third-tier correlation factor between all the third most contiguous nucleobases (Figure 2(c)); \( \theta_4 \) the fourth-tier correlation factor between all the fourth most contiguous nucleobases (Figure 2(d)), and so forth. These correlation factors combine all the local structure-order information at given distances along the RNA chain, which approximately reflects the global structure-order information. In Equation (12), the correlation function is given by

\[
\Theta(\Psi_i, \Psi_j) = \left[F(\Psi_i) - F(\Psi_j)\right]^2
\]

(13)

where \( F(\Psi_i) \) is the free energy of the structure status \( \Psi_i \) of the nucleobase at position \( i \), and \( F(\Psi_j) \) is the free energy of the structure status \( \Psi_j \) of the nucleobase at position \( j \). For the base pairs A-U and U-A, since they have two hydrogen bonds, their free energy values could be set as \(-2\) kcal/mol; for the base pairs G-C or C-G, they have three hydrogen bonds, and hence, their free energy values were set as \(-3\) kcal/mol; for the wobble base pairs G-U and U-G, their free energy values were set as \(-1\) kcal/mol; for the four unpaired nucleobases, their free energy values were each set as 0 kcal/mol.

Similar to Equation (5) of (Chou, 2001a), by incorporating the factors of Equation (12) for the global sequence-order effect, the local structure status composition vector of Equation (10) can be augmented to PseDPC; i.e.

\[
R = [\Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_u \quad \cdots \quad \Psi_\Omega]^T
\]

(14)

where \( \Omega = 10 + 100d + \lambda \), and

\[
\Psi_u = \begin{cases}
  \psi & \text{if } 1 \leq u \leq 10 + 100d \\
  \psi_0 \sum_{i=1}^{\lambda} \theta_i & \text{if } 11 + 100d \leq u \leq \Omega
\end{cases}
\]

(15)
where $\Psi_u$ is given by Equation (11), $\theta_j$ is the $j$-tier sequence correlation factor computed according to Equations (12) and (13) for the RNA sequence, and $w$ is the weight factor used to adjust the effect of the correlation factors.

In order to help the readers to understand the process of generating the feature vector of PseDPC, an example of converting a real miRNA precursor named has-mir-3713_ss into a PseDPC feature vector is given. As shown in the Figure 1, the has-mir-3713_ss sequence

Figure 1. An example for the process of calculating the occurrences of distance-pairs.

Notes: This figure shows the process of calculating the occurrences of distance-pairs of a real miRNA precursor named has-mir-3713_ss when $d$ is set 2. The structure status sequence is displayed in red color, while its complementary sequence is in blue color.

Figure 2. A schematic drawing to show the correlations of structure statuses along a RNA sequence. (a) The first-tier correlation reflects the structure-order mode between all the most contiguous nucleotides. (b) The second-tier correlation reflects the structure-order mode between all the second-most contiguous nucleotides. (c) The third-tier correlation reflects the structure-order mode between all the third-most contiguous nucleotides. (d) The fourth-tier correlation reflects the structure-order mode between all the fourth-most contiguous nucleotides. As we can see, the global or long-range sequence-order information of RNA can, thus, be approximately and indirectly incorporated into the current prediction model.
contains 45 nucleotides, and its secondary structure is derived from Vienna RNA software package. The occurrences of distance-pairs are generated based on the secondary structure. When \( d \) is set as 2, \( f(D(\Psi_{i}, \Psi_{j}/0)) \), \( f(D(\Psi_{i}, \Psi_{j}/1)) \), and \( f(D(\Psi_{i}, \Psi_{j}/2)) \) are calculated, and the feature vector of Equation (10) can be obtained, whose dimension is \( 10 + 100 \times 2 = 210 \). This process is shown in Figure 1, where the distance pairs \( D(\Psi_{i}, \Psi_{j}/0) \), \( D(\Psi_{i}, \Psi_{j}/1) \), and \( D(\Psi_{i}, \Psi_{j}/2) \) are labeled with pink rectangles. In order to incorporate the global structure-order information, this feature vector is extended according to the Equations (12)–(15) so as to generate the PseDPC feature vector (Equation (14)), and its dimension is \( 210 + \lambda \).

### 2.3. Support vector machine

SVMs (Cortess & Vapnik, 1995) are supervised learning models with associated learning algorithms that analyze data and recognize patterns used for classification and regression analysis. Given a set of training samples, each of which is labeled with one of the two categories, an SVM training algorithm builds a non-probabilistic binary linear classifier to assign new samples into which category they should belong to. In a SVM model, the samples are mapped to points in space, where the separate categories are divided by a clear wide gap. New samples are then mapped into that same space and their categories are predicted according to 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 (Liu et al., 2015) was employed, which is a type of software for SVM classification and regression. The software contains two parameters \( C \) and \( \gamma \), which will be optimized on the benchmark data-set using the grid tool provided in LIBSVM (Liu et al., 2015), as will be discussed later.

For a brief formulation of SVM and how it works, see the papers (Cai & Zhou, 2003; Chou & Cai, 2002); for more details about SVM, see a monograph (Cristianini & Shawe-Taylor, 2000).

The predictor thus obtained is called iMiRNA-PseDPC.

### 2.4. Jackknife test validation

How to properly examine the prediction quality is a key for developing a new predictor and estimating its potential application value. Generally speaking, to avoid the “memory effect” of the re-substitution test in which a same data-set was used to train and test a predictor (Chou & Cai, 2003, 2005), the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent data-set test, subsampling or K-fold (such as 5-fold, 7-fold, or 10-fold) test, and jackknife test (Chou, 2011). However, as elaborated by a penetrating analysis in (Chou, 2011), considerable arbitrariness exists in the independent data-set test. Also, as demonstrated by Equations (28)–(30) in (Chou, 2011), the subsampling test (or K-fold cross validation) cannot avoid arbitrariness either. Only the jackknife test is the least arbitrary that can always yield a unique result for a given benchmark data-set. Therefore, the jackknife test has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors (see, e.g. Du, Jiang, He, & Li, 2006; Hajisharifi, Piryaiee, Mohammad Beigi, Behbahani, & Mohabatkar, 2014; Kong et al., 2014; Mohabatkar, Beigi, Abdolahi, & Mohsenzadeh, 2013; Mondal & Pai, 2014; Nanni et al., 2014; Shen & Chou, 2010; Shen & Yang, 2007; Xiao & Wu, 2011). Accordingly, in the current study, the quality of the proposed predictor was also examined by the jackknife test, during which each of the samples in the benchmark data-set is in turn singled out as an independent entity and identified by the model that is trained with the remaining samples without including the one being tested. The overall jackknife success rate is determined by combining the scores thus obtained for all these independent samples. Therefore, the jackknife test can exclude the memory effect, and its outcome is always unique for a given benchmark data-set.

### 2.5. The process of identifying pre-miRNAs with iMiRNA-PseDPC

Here, we briefly summarize the process of how to use the proposed iMiRNA-PseDPC predictor to identify the pre-miRNAs. For given query RNA sequences, the predictor will first convert them into length-fixed feature vectors via the aforementioned PseDPC approach, followed by inputting into the SVM operation engine trained with the samples in the benchmark data-set. Subsequently, the trained SVM engine will generate a probability value for each of the query RNA sequences. If the probability value is greater than .5, the corresponding query RNA sequence will be assigned as a true pre-miRNA; otherwise, false pre-miRNA.

### 2.6. Metrics for validation and evaluation

The following four indexes are generally used for the problem studied here: (1) overall accuracy or Acc; (2) Mathews correlation coefficient or MCC; (3) sensitivity or Sn; and (4) specificity or Sp (see, e.g. Chen, Liu, & Yang, 2007). However, the conventional formulations for the four metrics are not easy to be understood by most experimental scientists, particularly the one for MCC.
As used by Chou (Chou, 2001b) in studying signal peptides, the aforementioned four metrics can be formulated by a set of equations given below (see, e.g. Chen et al., 2013; Guo et al., 2014; Lin et al., 2014; Liu, Xu, et al., 2014; Qiu & Xiao, 2014)

$$\begin{align*}
\text{Sn} &= 1 - \frac{N^+}{N^-} \\
\text{Sp} &= 1 - \frac{N^-}{N^-} \\
\text{Acc} &= \Lambda = 1 - \frac{N^+ + N^-}{N^+ + N^-} \\
\text{MCC} &= \frac{1 - \frac{N^+}{N^-}}{\sqrt{\left(\frac{N^- - N^+}{N^-}\right)\left(\frac{N^- - N^-}{N^-}\right)}}
\end{align*}$$

where $N^+$ represents the total number of true pre-miRNAs investigated, while $N^+$ represents the number of true pre-miRNAs incorrectly predicted as the false pre-miRNAs; $N^-$ represents the total number of false pre-miRNAs investigated, while $N^-$ represents the number of false pre-miRNAs incorrectly predicted as true pre-miRNAs.

According to Equation (16), it is crystal clear now to see the following. When $N^+ = 0$ meaning none of the true pre-miRNAs are incorrectly predicted to be false pre-miRNAs, we have the sensitivity $\text{Sn} = 1$. When $N^- = N^+$ meaning that all the true pre-miRNAs are incorrectly predicted to be false pre-miRNAs, we have the sensitivity $\text{Sn} = 0$. Likewise, when $N^- = 0$ meaning none of the false pre-miRNAs were incorrectly predicted to be the true pre-miRNAs, we have the specificity $\text{Sp} = 1$; whereas $N^+ = N^-$ meaning that all the false pre-miRNAs were incorrectly predicted as true pre-miRNAs, we have the specificity $\text{Sp} = 0$. When $N^+ = N^- = 0$ meaning that none of true pre-miRNAs in the positive data-set and none of the false pre-miRNAs in the negative data-set were incorrectly predicted, we have the overall accuracy $\text{Acc} = 1$ and $\text{MCC} = 1$; when $N^+ = N^+$ and $N^- = N^-$ meaning that all the true pre-miRNAs in the positive data-set and all the false pre-miRNAs in the negative data-set were incorrectly predicted, we have the overall accuracy $\text{Acc} = 0$ and $\text{MCC} = -1$; whereas when $N^+ = N^+/2$ and $N^- = N^-/2$, we have $\text{Acc} = 0.5$ and $\text{MCC} = 0$ meaning no better than random prediction. As we can see from the above discussion based on Equation (16), the meanings of sensitivity, specificity, overall accuracy, and Mathew’s correlation coefficient have become much more intuitive and easier to understand.

It is instructive to point out that the set of metrics as defined in Equation (16) is valid only for the single-label systems. For the multi-label systems, whose emergence has become more frequent in system biology (Lin, Fang, & Xiao, 2013; Xiao & Wu, 2011) and system medicine (Chen, Zeng, Cai, & Feng, 2012; Xiao, Wang, Lin, & Jia, 2013), a completely different set of metrics as defined in (Chou, 2013) is needed.

### 3. Results and discussion

#### 3.1. Parameter optimization of iMiRNA-PseDPC

As we can see from Equations (6)–(15), the predictor iMiRNA-PseDPC contains three uncertain parameters, namely $d$, $\lambda$, and $w$, where $d$ reflects the local or short-range structure status order effects, $\lambda$ reflects the global or long-range structure status order effect, and $w$ is the factor to adjust the weight between the local and global effects. Generally speaking, the greater the values of $d$ and $\lambda$ are, the more the structure status order information will be incorporated. However, if $d$ or $\lambda$ is too large, it would reduce the cluster-tolerant capacity (Chou, 1999) and cause the “overfitting” or “high dimension disaster” (Wang, Yang, & Shen, 2008) problem, so as to reduce the prediction accuracy. Accordingly, in the current study, their optimal values were determined within the ranges as defined below

$$\begin{align*}
1 &\leq d \leq 7 \quad \text{with step } \Delta = 1 \\
1 &\leq \lambda \leq 20 \quad \text{with step } \Delta = 1 \\
0 &\leq w \leq 1 \quad \text{with step } \Delta = 0.1
\end{align*}$$
It can be seen from Equation (17) that, to determine the optimal values for the three parameters, a total of $7 \times 20 \times 11 = 1540$ different combination cases would need to be considered. To reduce the computational time, we adopted the fivefold cross-validation approach for the predictor on the benchmark data-set. As shown in Figure 3, the overall accuracy of iMiRNA-PseDPC is the most sensitive to the parameter $d$, followed by the parameter $\lambda$ and the least sensitive to the parameter $w$. The final optimal values for the three parameters along with the two parameters $C$ and $\gamma$ in SVM (cf. Section 2.3) were determined by taking the highest overall accuracy among the 1540 combination cases. The values for the five parameters, thus, obtained are given below

$$
\begin{align*}
&d = 7 \\
&\lambda = 15 \\
&w = 1 \\
&C = 2 \\
&\gamma = 2^{-7}
\end{align*}
(18)
$$

The parameter values in Equation (18) will be used to carry out the rigorous jackknife test for the iMiRNA-PseDPC predictor on the benchmark data-set to examine its success rates according to the four metrics defined in Equation (16).

### 3.2. Comparison with other methods

Listed in Table 1 are the rates of the four metrics (cf. Equation (16)) obtained by the new predictor via the jackknife test on the benchmark data-set (Supplementary material). For facilitating comparison, listed there are also the corresponding rates by Triplet-SVM (Xue et al., 2005) and MiPred (Jiang et al., 2007), the two most popular and widely used predictors in this area. To make the comparison absolutely fair, all the rates listed there for the three different methods were obtained via the same rigorous jackknife test on exactly the same benchmark data-set.

Also, since graphic approaches can provide intuitive insights and are particularly useful for dealing with complicated systems (see, e.g. Althaus et al., 1993; Chou, 2010; Liu, Zhang, Xu, Xu, & Wang, 2014; Zhou, 2011), here we also provide a graphic comparison (Figure 4) of the current predictor with their counterparts via the receiver operating characteristic (ROC) plot (Fawcett, 2005). According to ROC (Fawcett, 2005), the larger the area under the curve, the better the corresponding predictor is.

As we can see from Table 1, the computational cost of iMiRNA-PseDPC is comparable with Triplet-SVM, but its performance is remarkably higher than that of Triplet-SVM (Xue et al., 2005) in all the four metrics,

Table 1. Comparison of different predictors by the jackknife tests on a same benchmark data-set (Supplementary material).

| Predictor          | Acc (%) | MCC | Sn (%) | Sp (%) | Web server | CPU time (s) |
|--------------------|---------|-----|--------|--------|------------|--------------|
| Triplet-SVMa       | 81.85   | .64 | 78.47  | 85.20  | No         | 35           |
| MiPredb            | 87.30   | .75 | 84.00  | 90.60  | Yes        | 48,712       |
| iMiRNA-PseDPCc     | 87.69   | .75 | 88.87  | 86.57  | Yes        | 33           |

aResults obtained by in-house implementation of Triplet-SVM (Xue et al., 2005).
bResults obtained by in-house implementation of MiPred (Jiang et al., 2007).
cThe predictor proposed in this paper; see Equation (18) for the parameters used.
which is quite consistent with Figure 4, where the area under the ROC curve of the new predictor is remarkably greater than that of Triplet-SVM, indicating a significant improvement of the new predictor in comparison with Triplet-SVM. Meanwhile, we can also see from both Table 1 and Figure 4 that the new predictor only slightly outperformed MiPred (Jiang et al., 2007). Nevertheless, iMiRNA-PseDPC is much more efficient than MiPred (Jiang et al., 2007), as reflected by the fact that for the same load of prediction job the CPU time used by MiPred is about 48,712/33, which is 1467 folds the CPU time used by iMiRNA-PseDPC (see column 7 of Table 1). This is because MiPred requires a time-consuming $P$-value feature calculation step, in which for each of the query RNA sequences, the secondary structures of its 1000 shuffled sequences need to be derived first by running the Vienna RNA software in order to be able to calculate the $P$-value feature. In contrast to that, there is no such a requirement at all when using the current iMiRNA-PseDPC predictor. Therefore, the new predictor is particularly more useful for dealing large-scale data analysis as in genome studies.

3.3. Web server guide

As mentioned in (Chou & Shen, 2009; Lin & Lapointe, 2013) and implemented in a series of recent publications (see, e.g. Chen et al., 2013; Chen, Feng, & Deng, 2014; Fan et al., 2014; Guo et al., 2014; Liu, Wang, Chen, Dong, & Lan, 2012; Liu, Wang, Zou, Dong, & Chen, 2013; Liu, Liu, Fang, & Wang, 2015; Liu, Fang, Jie, Liu, & Wang, 2015; Liu, Xu, et al., 2014; Qiu & Xiao, 2014; Qiu et al., 2014; Xu, Ding, et al., 2013; Xu, Wen, et al., 2014; Xu, Zhou, et al., 2014), user-friendly and publicly accessible web servers are indispensable for developing practically more useful predictors, a web server for the current predictor iMiRNA-PseDPC has also been established. Furthermore, to maximize the convenience of the vast majority of experimental

Figure 4. A graphical illustration to show the performance of different methods by means of the ROC curves. The areas under the ROC curves or AUC are .894, .941, and .948 for Triplet-SVM, MiPred, and iMiRNA-PseDPC, respectively.

Figure 5. A semi-screenshot to show the top page of the web server iMiRNA-PseDPC, which is accessible at the website http://bioinformatics.hitsz.edu.cn/iMiRNA-PseDPC/.
scientists, given below is a step-by-step guide on how to use the web server to acquire their desired results without the need to follow the complicated mathematical equations that were documented in this paper just for the integrity of the new predictor.

**Step 1.** Visit the web server by clicking the link at http://bioinformatics.hitsz.edu.cn/iMiRNA-PseDPC/ and you will see its top page as shown in Figure 5. 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 Figure 5 or directly upload your input data by the Browse button. The input sequence should be in the FASTA format. An RNA 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 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 results. 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 those for the 3rd and 4th ones are “False Pre-miRNA”. All these predicted results are fully consistent with the experimental observations.

### 4. Conclusion

Illuminated by the success of introducing the pseudo amino acid composition (Chou, 2001a, 2005) or Chou’s PseAAC (Lin & Lapointe, 2013) to analyze protein/peptide sequences, we have proposed a new sequence-based method for identifying human miRNA precursors. The new predictor is called “iMiRNA-PseDPC”, in which the RNA samples are formulated by the PseDPC.

For training and testing the new model, we have constructed a new benchmark data-set, which is much larger and more stringent than any of the existing benchmark data-sets in this area.

Rigorous cross-validation tests on the new benchmark data-set have indicated that iMiRNA-PseDPC outperformed Triplet-SVM (Xue et al., 2005) and MiPred (Jiang et al., 2007), the two state-of-the-art methods widely used for identifying human miRNA precursors. Compared with the former, our predictor achieved 87.69%, .75, 88.87%, and 86.57% for Acc, MCC, Sn, and Sp, respectively, which are significantly higher than the corresponding rates by Triplet-SVM (Table 1); compared with the latter, although the aforementioned success rates are only slightly higher than those of MiPred (Jiang et al., 2007), the CPU time used by our predictor is more than 1000 times less than that by MiPred, and hence, it is much more efficient and particularly useful for dealing large-scale data analysis as often encountered in genome studies.

We have established the web server for iMiRNA-PseDPC. To access it, just click the link at http://bioinformatics.hitsz.edu.cn/iMiRNA-PseDPC/, by which users can easily get their desired results without the need to follow the complicated mathematical equations, which were presented in this paper just for the integrity of their development process.

It is instructive to point out that although the current predictor was established for identifying the human pre-miRNAs, they can be easily extended to identify the pre-miRNAs in any of other organisms as well if a corresponding benchmark data-set is available.

Although the proposed iMiRNA-PseDPC has shown promising predictive performance for pre-miRNA identification, it can be further improved from the following two aspects. (1) The performance of a predictor depends on the benchmark data-set used to train it (Chou & Shen, 2007). In this study, the negative samples in the negative subset were the segments derived from CDSs and hence devoid of their context. In the full context, they might not fold into the hairpin structures and hence might not be of good negative sample. With more experimental data available in future, it will be possible to construct a balanced benchmark data-set with higher quality negative samples as done in (Liu, Xiao, & Qiu, 2015; Xiao, Min, Lin, & Liu, 2014) so as to further enhance the prediction quality. (2) In this study, we used the free energies (Equation (13)) to represent the number of hydrogen bonds in counting the correlation of structure statuses along the RNA sequences. It is anticipated that incorporation of the physicochemical properties of RNA sequences or the physicochemical properties of their constituent nucleic acid residues will further enhance the success rates of the iMiRNA-PseDPC predictor.

### Supplementary material

The benchmark data-set contains 3224 human pre-miRNAs, of which 1612 are real pre-miRNAs and 1612 are false pre-miRNAs. None of the sequences included has ≥80 pairwise sequence identity with any other. The entire data-set can be downloaded from the link at http://bioinformatics.hitsz.edu.cn/iMiRNA-PseDPC/data. The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2015.1014422.
Acknowledgments
The authors are very much indebted to the two anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper.

Funding
This work was supported by the National Natural Science Foundation of China [grant numbers 61300112 and 61272383]; the Scientific Research Innovation Foundation in Harbin Institute of Technology [Project No. HIT.NSRIF.2013103]; the Shanghai Key Laboratory of Intelligent Information Processing, China [grant number IILP-2012-002], and the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References
Agarwal, S., Vaz, C., Bhattacharya, A., & Srinivasan, A. (2010). Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). BMC Bioinformatics, 11, S29.

Althaus, I. W., Chou, J. J., Gonzales, A. J., Deibel, M. R., Chou, K. C., Kezdy, F. J., … Palmer, J. R. (1993). Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry, 32, 6548–6554.

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

Brameier, M., & Wiuf, C. (2007). Ab initio identification of human microRNAs based on structure motifs. BMC Bioinformatics, 8, 478.

Cai, Y. D., & Zhou, G. P. (2003). Support vector machines for overseas Chinese Scholars, State Education Ministry.

Chou, K. C. (2001a). Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2004, Vol. 44, 60), 43, 246–255.

Chou, K. C. (2001b). Using subsite coupling to predict signal peptides. Protein Engineering Design and Selection, 14, 75–79.

Chou, K. C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics, 21, 10–19.

Chou, K. C. (2010). Graphic rule for drug metabolism systems. Current Drug Metabolism, 11, 369–378.

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review). Journal of Theoretical Biology, 273, 236–247.

Chou, K. C. (2013). Some remarks on predicting multi-label attributes in molecular biosystems. Molecular BioSystems, 9, 1092–1100.

Chou, K. C. (2015). Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry. doi:10.2174/15734064116614129162834

Chou, K. C., & Cai, Y. D. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. Journal of Biological Chemistry, 277, 45765–45769.

Chou, K. C., & Cai, Y. D. (2003). Prediction and classification of protein subcellular location: Sequence-order effect and pseudo amino acid composition. Journal of Cellular Biochemistry (Addendum, ibid. 2004, 91, 1085), 90, 1250–1260.

Chou, K. C., & Cai, Y. D. (2005). Prediction of membrane protein types by incorporating amphiphatic effects. Journal of Chemical Information and Modeling, 45, 407–413.

Chou, K. C., & Shen, H. B. (2007). Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry, 370, 1–16.

Chou, K. C., & Shen, H. B. (2009). Review: Recent advances in developing web-servers for predicting protein attributes. Natural Science, 1, 63–92.

Cortess, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction of support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press.

Du, P., Gu, S., & Jiao, Y. (2014). PseAAC-General: Fast building various modes of general form of Chou’s pseudo-amino acid composition for large-scale protein datasets. International Journal of Molecular Sciences, 15, 3495–3506.

Du, Q. S., Jiang, Z. Q., He, W. Z., & Li, D. P. (2006). Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. Journal of Biomolecular Structure and Dynamics, 23, 635–640.

Fan, G. L., & Li, Q. Z. (2013). Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou’s pseudo amino acid composition. Journal of Theoretical Biology, 334, 45–51.

Fan, Y. N., Xiao, X., & Min, J. L. (2014). iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. International Journal of Molecular Sciences, 15, 4915–4937.

Fawcett, J. A. (2005). An introduction to ROC analysis. Pattern Recognition Letters, 27, 861–874.

Georgiou, D. N., Karakasidis, T. E., & Mearigas, A. C. (2013). A short survey on genetic sequences, Chou’s pseudo amino
acid composition and its combination with fuzzy set theory. *The Open Bioinformatics Journal*, 7, 41–48.

Guo, S. H., Deng, E. Z., Xu, L. Q., Ding, H., & Lin, H. (2014). iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, 30, 1522–1529.

Hajisharifi, Z., Piryaei, M., Mohammad Beigi, M., Behbahani, M., & Mohabatkar, H. (2014). Predicting anticancer peptides with Chou’s pseudo amino acid composition and investigating their mutagenicity via Ames test. *Journal of Theoretical Biology*, 341, 34–40.

Hayat, M., & Iqbal, N. (2014). Discriminating protein structure classes by incorporating Pseudo average chemical shift to Chou’s general PseAAC and support vector machine. *Computer Methods and Programs in Biomedicine*, 116, 184–192.

Helvík, S. A., Snove, O. Jr., & Saetrom, P. (2007). Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, 23, 142–149.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, 31, 3429–3431.

Huang, T. H., Fan, B., Rothshchild, M. F., Hu, Z. L., Li, K., & Zhao, S. H. (2007). MiRFinder: An improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics*, 8, 341.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., & Lu, Z. (2007). MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, 35, W339–W344.

Kong, L., Zhang, L., & Lv, J. (2014). Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou’s pseudo amino acid composition. *Journal of Theoretical Biology*, 344, 12–18.

Kozomara, A., & Griffiths-Jones, S. (2011). miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39, D152–D157.

Li, C., Feng, Y., Coukos, G., & Zhang, L. (2009). Therapeutic microRNA strategies in human cancer. *The AAPs Journal*, 11, 747–757.

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658–1659.

Lin, H., Deng, E. Z., & Ding, H. (2014). iPero54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Research*, 42, 12961–12972.

Lin, W. Z., Fang, J. A., & Xiao, X. (2013). iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Molecular BioSystems*, 9, 634–644.

Lin, S. X., & Lapointe, J. (2013). Theoretical and experimental biology in one. *Journal of Biomedical Science and Engineering*, 6, 435–442.

Liu, W., & Chou, K. C. (1999). Prediction of protein secondary structure content. *Protein Engineering Design and Selection*, 12, 1041–1050.

Liu, B., Fang, L., Jie, C., Liu, F., & Wang, X. (2015). miRNA-dis: microRNA precursor identification based on distance structure status pairs. *Molecular BioSystems*. doi:10.1039/C1SB00505E

Liu, B., Liu, F., Fang, L., & Wang, X. (2015). repDNA: A python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physiochemical properties and sequence-order effects. *Bioinformatics*. doi:10.1093/bioinformatics/btu1820

Liu, B., Xu, J., Fan, S., Xu, R., Zhou, J., & Wang, X. (2015). PseDNA-pro: DNA-binding protein identification by combining Chou’s PseAAC and physicochemical distance. *Molecular Informatics*, 34, 8–17.

Liu, B., Xu, J., Lan, X., Xu, R., & Zhou, J. (2014). idNA-Pro: Identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS ONE*, 9, e106691.

Liu, Z., Xiao, X., & Qu, W. R. (2015). idDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Analytical Biochemistry*. doi:10.1016/j.ab.2014.12.009

Liu, B., Wang, X., Chen, Q., Dong, Q., & Lan, X. (2012). Using amino acid physicochemical distance transformation for fast protein remote homology detection. *PLoS ONE*, 7, e46633.

Liu, B., Wang, X., Zou, Q., Dong, Q., & Chen, Q. (2013). Protein remote homology detection by combining Chou’s pseudo amino acid composition and profile-based protein representation. *Molecular Informatics*, 32, 775–782.

Liu, B., Zhang, D., Xu, R., Xu, J., & Wang, X. (2014). Combining evolutionary information extracted from sequence profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics*, 30, 472–479.

Lopes, I. O., Schliep, A., & Carvalho, A. C. (2014). The discriminant power of RNA features for pre-miRNA recognition. *BMC Bioinformatics*, 15, 124.

Mohabatkar, H., Beigi, M. M., Abdolali, K., & Mohsenzadeh, S. (2013). Prediction of allergenic proteins by means of the concept of Chou’s pseudo amino acid composition and a machine learning approach. *Medicinal Chemistry*, 9, 133–137.

Mondal, S., & Pai, P. P. (2014). Chou’s pseudo amino acid composition improves sequence-based antifreeze protein prediction. *Journal of Theoretical Biology*, 356, 30–35.

Nam, J. W., Shin, K. R., Han, J., Lee, Y., Kim, V. N., & Zhang, B. T. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Research*, 33, 3570–3581.

Nanni, L., Brahnam, S., & Lumini, A. (2014). Prediction of protein structure classes by incorporating different protein descriptors into general Chou’s pseudo amino acid composition. *Journal of Theoretical Biology*, 360, 109–116.

Qiu, W. R., & Xiao, X. (2014). iSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *International Journal of Molecular Sciences*, 15, 1746–1766.

Qiu, W. R., Xiao, X., & Lin, W. Z. (2014). iUbiqu-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model. *Journal of Biomolecular Structure and Dynamics* (JBSD). doi:10.1080/07391102.2014.968875

Shen, H. B., & Chou, K. C. (2010). Virus-mPLoc: A fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *Journal of Biomolecular Structure and Dynamics*, 28, 175–186.

Shen, H. B., & Yang, J. (2007). Euk-PLoc: An ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids*, 33, 57–67.

Wang, Y., Chen, X., Jiang, W., Li, L., Li, W., Yang, L., … Li, X. (2011). Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM. *Genomics*, 98, 73–78.
Wang, T., Yang, J., & Shen, H. B. (2008). Predicting membrane protein types by the LLDA algorithm. *Protein & Peptide Letters, 15*, 915–921.

Wei, L. Y., Liao, M. H., Gao, Y., Ji, R. R., He, Z. Y., & Zou, Q. (2014). Improved and promising identification of human micromas by incorporating a high-quality negative set. *Computational Biology and Bioinformatics, 11*, 192–201.

Wu, Y., Wei, B., Liu, H., Li, T., & 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.

Xiao, X., Min, J. L., Lin, W. Z., & Liu, Z. (2014). iDrug-Target: Predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. *Journal of Biomolecular Structure & Dynamics (JBSD)*. doi:10.1080/07391102.2014.998710

Xiao, X., Wang, P., Lin, W. Z., & Jia, J. H. (2013). iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical Biochemistry, 436*, 168–177.

Xiao, X., & Wu, Z. C. (2011). iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *Journal of Theoretical Biology, 284*, 42–51.

Xu, Y., Ding, J., & Wu, L. Y. (2013). iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS ONE, 8*, e55844.

Xu, Y., Shao, X. J., Wu, L. Y., & Deng, N. Y. (2013). iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ, 1*, e171.

Xu, Y., Wen, X., Wen, L. S., & Wu, L. Y. (2014). iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS ONE, 9*, e105018.

Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L. C., & Showe, M. K. (2006). Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics, 22*, 1325–1334.

Zhong, L., Wang, J. T., Wen, D., Aris, V., Soteropoulos, P., & Shapiro, B. A. (2013). Effective classification of microRNA precursors using feature mining and AdaBoost algorithms. *OMICS: A Journal of Integrative Biology, 17*, 486–493.

Zhong, W. Z., & Zhou, S. F. (2014). Molecular science for drug development and biomedicine. *International Journal of Molecular Sciences, 15*, 20072–20078.

Zhou, G. P. (2011). The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. *Journal of Theoretical Biology, 284*, 142–148.