PLOS ONE

RESEARCH ARTICLE

Identification of Real MicroRNA Precursors with a Pseudo Structure Status Composition Approach

Bin Liu[1,2,3]*, Longyun Fang[1], Fule Liu[1], Xiaolong Wang[1,2], Junjie Chen[1], Kuo-Chen Chou[3,4]

1 School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China, **2** Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China, **3** Gordon Life Science Institute, Belmont, Massachusetts, United States of America, **4** Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

* bliu@insun.hit.edu.cn

Citation: Liu B, Fang L, Liu F, Wang X, Chen J, Chou K-C (2015) Identification of Real MicroRNA Precursors with a Pseudo Structure Status Composition Approach. PLoS ONE 10(3): e0121501. doi:10.1371/journal.pone.0121501

Academic Editor: Hikmet Budak, Sabanci University, TURKEY

Received: October 17, 2014

Accepted: January 31, 2015

Published: March 30, 2015

Data Availability Statement: Data are available via the web-server at http://bioinformatics.hitsz.edu.cn/iMcRNA/. The data has also been uploaded into figshare.com. The DOI is doi: http://dx.doi.org/10.6084/m9.figshare.1289312.

Funding: This work was supported by the National Natural Science Foundation of China (No. 61300112, 61272383), the Scientific Research Innovation Foundation in Harbin Institute of Technology (Project No. HIT.NSRIF.2013103), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, and State Education Ministry. The funders had no role in study design, data collection and

Abstract

Containing about 22 nucleotides, a micro RNA (abbreviated 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 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 postgenomic age, it is highly desired to develop computational sequence-based methods in this regard. Here two new predictors, called "iMcRNA-PseSSC" and "iMcRNA-ExPseSSC", were proposed for identifying the human pre-microRNAs by incorporating the global or long-range structure-order information using a way quite similar to the pseudo amino acid composition approach. Rigorous cross-validations on a much larger and more stringent newly constructed benchmark dataset showed that the two new predictors (accessible at http://bioinformatics.hitsz.edu.cn/iMcRNA/) outperformed or were highly comparable with the best existing predictors in this area.

Introduction

MicroRNAs (miRNAs) are small single-strand, non-coding RNAs about 22 nucleotides (nt) in length, which play important roles in gene regulation by targeting messenger RNAs (mRNAs) for cleavage or translational repression. The miRNAs are also involved in many important biological processes, such as affecting stability, translation of mRNAs and negatively regulating

analysis, decision to publish, or preparation of this manuscript.

Competing Interests: The authors have declared that no competing interests exist.

gene expression in post-transcriptional processes. In animals, the biogenesis of miRNA is shown in Fig. 1, and can be divided into the following steps: (i) The genes of miRNA are transcribed by RNA polymerase II [1,2], resulting in the primary transcripts termed as pri-miRNAs, which are typically 60–70 nt. (ii) The pre-miRNAs are processed by the enzyme Drosha to release the hairpin-shaped intermediates (pre-miRNAs) [3]. (iii) The pre-miRNAs are then exported into the cytoplasm by Exportin V and Ran-GTP cofactor [4–6] and cleaved by the enzyme Dicer to yield miRNA/miRNA* duplexes [7–11].

Owing to the difficulty of systematically detecting miRNAs from a genome by existing experiment techniques, computational methods have been indispensable tools in miRNA studies [12]. Various computational methods have been proposed to predict pre-miRNAs. Most of these methods employed the machine learning techniques to build their prediction models, which treated this problem as a binary classification task to discriminate the real pre-miRNAs from false pre-miRNAs. These methods are different in the feature selections and machine learning algorithms or operation engines. The machine learning algorithms usually used in this field include Support Vector Machine (SVM) [11,13–20], Random Forest (RF) [21–23], Hidden Markov Model (HMM) [24], Covariant Discrimination (CD) [25] or Naive Bayes (NB) [26], and Linear Genetic Programming (LGP) [27].

The secondary structure is an important feature used in the computational methods, because most of the pre-miRNAs have the characteristic of stem-loop hairpin structures [16]. Mir-abela [28] is an SVM-based method trained with 16 statistic features computed from the entire hairpin structure. Triplet-SVM [16] employed a SVM classifier to train 32 local triplet sequence-structure features. Later, MiPred [21] improved Triplet-SVM [16] by employing the Random Forest classifier trained with the local triplet sequence-structure features, minimum of free energy (MFE), and *P*-values. MiRFinder [29] is a high-throughput pre-miRNA prediction

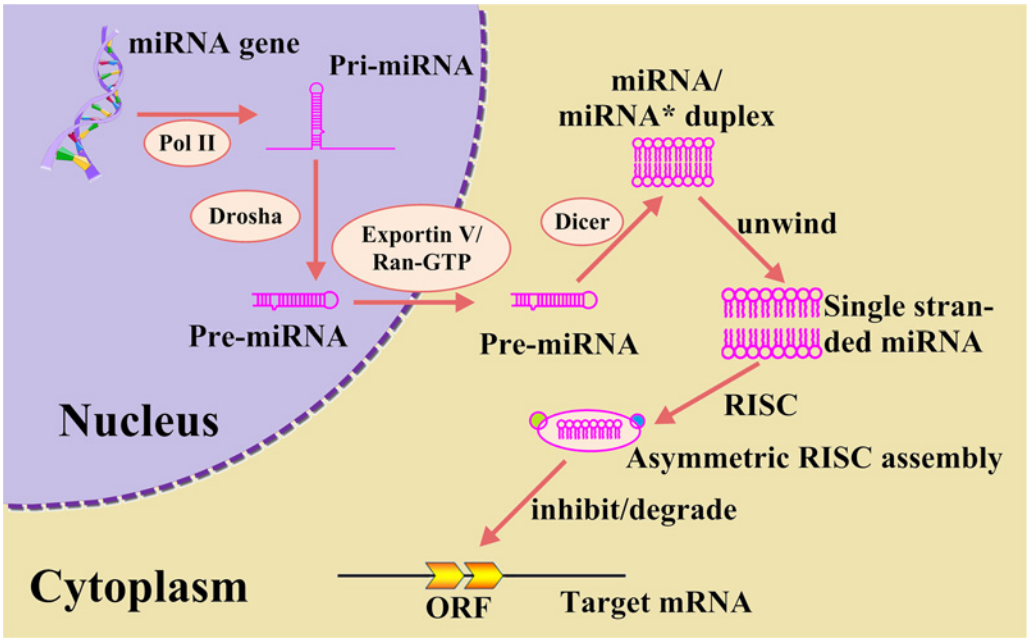

Fig 1. An illustration to show biogenesis of miRNAs and model of miRNA-mediated translational repression or mRNA degradation. MiRNA genes are transcribed by RNA polymerase II [2,90], resulting in the primary transcripts termed as pri-miRNAs, which are typically 60–70 nucleotides. The pri-miRNAs are processed by the enzyme Drosha to release the hairpin-shaped intermediates (pre-miRNAs) [3], followed by being exported into the cytoplasm by Exportin V and Ran-GTP cofactor [4–6], and then cleaved by the enzyme Dicer to yield miRNA/miRNA* duplexes [7–11].

doi:10.1371/journal.pone.0121501.g001

method, which consists of two steps: a search for hairpin candidates and exclusion of the non-robust structures based on the analysis of 18 parameters by the SVM.

All these computational methods could yield quite encouraging results, and each of them did play a role in simulating the development of pre-miRNA identification. However, further work is needed due to the following reasons: (i) The datasets constructed in those methods were too small to reflect the statistical profile of human pre-miRNAs. Most of these methods were trained and tested with a dataset containing only several hundreds of human pre-miRNA samples or pseudo pre-miRNA samples. (ii) No cutoff threshold was imposed to rigorously exclude the redundant samples or those with high sequence similarity with others in a same benchmark dataset. (iii) Most of these methods only consider the local structure or sequence order information of RNA sequences, and all the global or long range structure or sequence order effects were ignored.

In this study, we attempted to improve the accuracy for human pre-miRNA identification from the above three aspects; especially, we focused on how to incorporate the global structure-order effects into the predictor. However, it is difficult to incorporate this kind of information into a statistical predictor because the RNA sequences have different lengths with extremely large number of possible structure patterns. To overcome this difficulty, is it possible to find an approximate way to take the structure-order effects into account?

Actually, similar problems were also encountered in computational proteomics and genomics. To incorporate the long-range or global sequence order information for protein/peptide sequences, the pseudo amino acid composition [30,31] or Chou's PseAAC [32] was proposed. Ever since the concept of PseAAC was proposed in 2001[30], it has been penetrating into almost all the fields of protein attribute predictions (see, e.g., [33–44], as well as a Wikipedia article at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition and a long list of papers cited in [45]) and some fields of drug development and biomedicine [46]. Recently, the concept of PseAAC has also been further extended to the field of genomics by using different modes of pseudo K-tuple nucleotide composition or PseKNC [47–49] to predict the recombination spots of DNA [19,50], the nucleosome positions [20], sigma-54 promoters [51], and DNA methylation sites [52]. For more information about this, see a recent review [53].

Encouraged by the successes of PseAAC and PseKNC approaches in the fields of proteomics and genomics, we proposed a feature vector called "pseudo structure status composition (PseSSC)" to represent RNA sequences by incorporating the structure-order effects so as to improve the prediction quality in identifying human pre-miRNA. The detailed approach is elaborated as follows.

As pointed out in a comprehensive review [54] and carried out in a series of recent publications (see, e.g., [19,20,50,55–57]), to develop a really useful statistical predictor or model for a biological system, one needs to engage the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the model; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these procedures one-by-one.

Materials and Method

1. Benchmark Dataset

The pre-miRNAs or positive samples were downloaded from the latest version (release 20: June 2013) of miRNABase [58,59], which contained 1,872 experiment-confirmed sapiens

pre-miRNA entries. The false pre-miRNAs or negative samples were taken from the data constructed by Xue et al. [16], which contained 8,489 false pre-miRNA samples. These false pre-miRNAs are similar to the real pre-miRNAs according to the following widely accepted characteristics [16]: (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.

To get rid of the redundancy and avoid homology bias, the CD-HIT software [60] with the cutoff threshold set at 80% (note that the most stringent cutoff threshold for DNA sequences by CD-HIT is 75%) was used to winnow those samples which had ≥80% sequence identity to any other in a same subset. After such a screening procedure, we obtained 1,612 human pre-miRNAs, which formed the positive dataset in the current study.

To avoid imbalance problem caused by different number of positive and negative samples, we randomly picked 1,612 samples from the 8,489 false pre-miRNAs to form the negative dataset. Again, none of the samples included had ≥80% sequence identity to any other in a same subset.

As pointed out by a review [25], there is no need 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 $\mathbb{S}$ can be formulated as

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \tag{1}$$

where the subset $\mathbb{S}^+$contains 1,612 human pre-miRNAs, the subset $\mathbb{S}^-$ contains 1,612 false pre-miRNAs, and the symbol represents the "union" in the set theory.

The detailed sequences are given in S1 Dataset that is not only the largest but also most stringent benchmark dataset in this area.

2. Pseudo Structure Status Composition (PseSSC)

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

$$\mathrm{R} = \mathrm{B}_1\mathrm{B}_2\mathrm{B}_3\mathrm{B}_4\mathrm{B}_5\ldots\mathrm{B}_L \tag{2}$$

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.,

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

If the RNA sequence is formulated according to its secondary structure derived from the Vienna RNA software package (released 2.1.6) [61], we have

$$\mathrm{R} = \Psi_1\Psi_2\Psi_3\Psi_4\Psi_5\cdots\Psi_L \tag{4}$$

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.,

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

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

G-C, C-G, G-U, U-G. Therefore, we have additional six different structure statuses of the paired bases in RNA (Fig. 2).

Based on the ten structure statuses, if the RNA sequence is represented by the structure statuses of its n adjacent nucleotides, or the so-called "n-tuple nucleobase composition" [47], the corresponding feature vector will contain 10^n components as given by (cf. Fig. 3)

$$\mathbf{R} = [f_1 \quad f_2 \quad f_3 \quad f_4 \quad \cdots \quad f_{10^n}]^{\mathbf{T}} \tag{6}$$

where $f_i = (i = 1,2,\ldots,10^n)$ represents the normalized occurrence frequency of the structure status combination of n adjacent nucleobases. As indicated by the above equation, with the increase of n, the structure-order information within a local or short-range scope could be incorporated, but none of the global or long-range structure information would be reflected.

Stimulated by the PseAAC approach [30,31] in computational proteomics, here we are to propose a novel feature vector called the pseudo structure status composition (PseSSC) 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 described as follows.

In a way parallel to the formulation in [30], 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

$$\begin{cases} \theta_1 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1}\Theta(\Psi_i, \Psi_{i+1}) \\ \theta_2 = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2}\Theta(\Psi_i, \Psi_{i+2}) \\ \theta_3 = \dfrac{1}{L-3}\displaystyle\sum_{i=1}^{L-3}\Theta(\Psi_i, \Psi_{i+3}) \qquad (\lambda < L) \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda}\Theta(\Psi_i, \Psi_{i+\lambda}) \end{cases} \tag{7}$$

where λ is an integer, representing the highest counted rank (or tier) of the structural correlation along a RNA chain; θ_1 is the first-tier correlation factor reflecting the structure-order information between all the most contiguous bases along a RNA chain (Fig. 4a); θ_2 the second-tier correlation factor between all the second most contiguous nucleobases (Fig. 4b); θ_3 the third-tier correlation factor between all the third most contiguous nucleobases (Fig. 4c); and so forth. In Equation 8 the correlation function is given by

$$\Theta(\Psi_i, \Psi_j) = [F(\Psi_i) - F(\Psi_j)]^2 \tag{8}$$

where $F(\Psi_i)$ is the free energy of the structure status Ψi of the nucleobase at position i, and $F(\Psi_j)$ is the free energy of the structure status Ψ_j of the nucleobase at position j. As mentioned above, if we distinguish the nucleobase near 5' end and 3' end, there are 6 different structure statuses for the paired nucleobases (Fig. 2). For the base pairs A-U and U-A, since they have 2 hydrogen bonds, their free energy values could be set as-2 kcal/mol; for the base pairs G-C or C-G, they have 3 hydrogen bonds (Fig. 2b) and hence their free energy values were set as-3 kcal/mol; for the wobble base pairs G-U and U-G (Fig. 2c), 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.

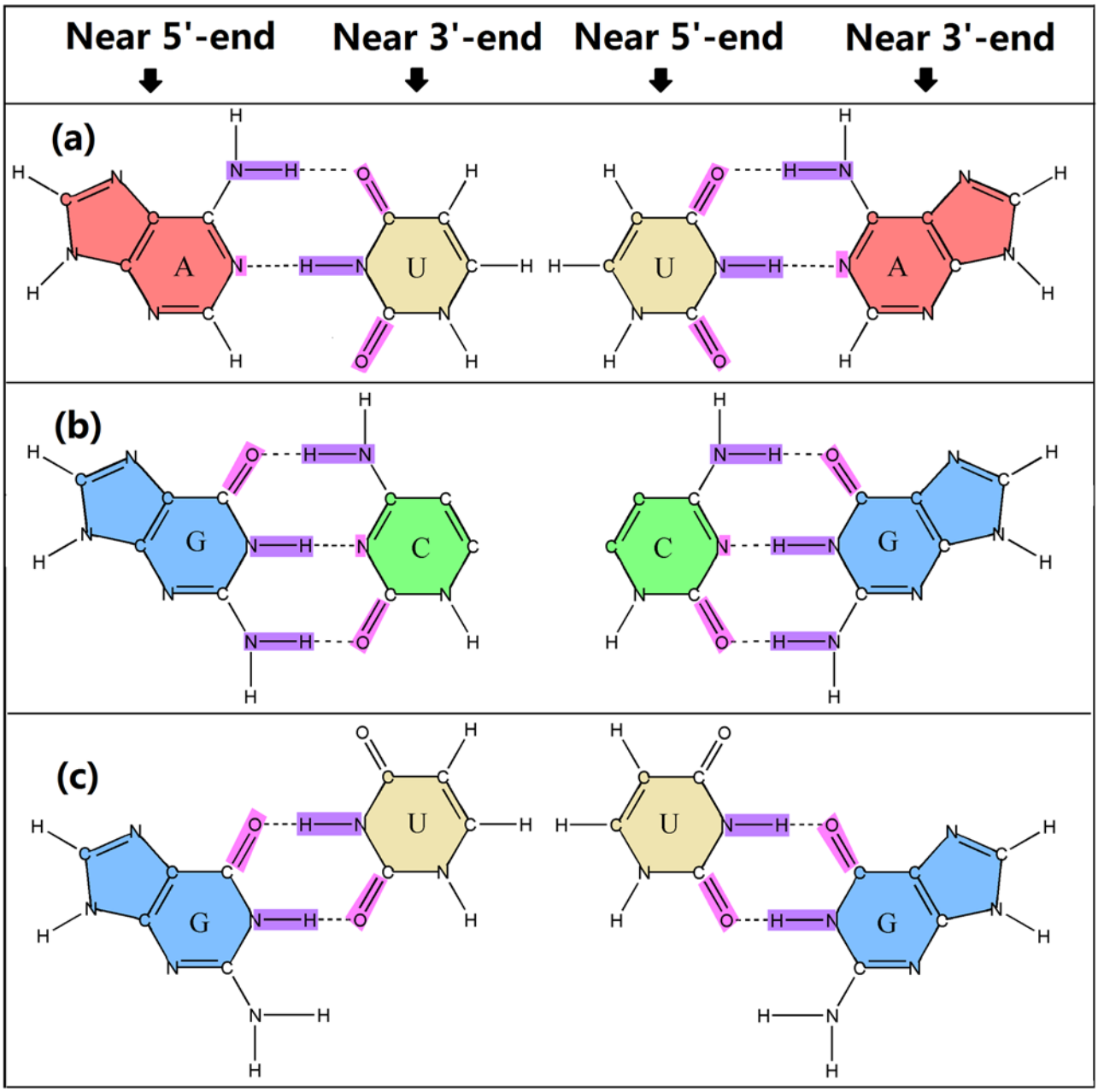

Fig 2. Illustration to show the 6 structure statuses of paired nucleic acid residues. Note that the nucleotide near 5' end is different with the one near 3'end: (a) the base pairs A-U or U-A has 2 hydrogen bonds; (b) the base pair G-C or C-G has 3 hydrogen bonds; and (c) the wobble base pair G-U or U-G has 2 weaker hydrogen bonds. See the main text for further explanation.

doi:10.1371/journal.pone.0121501.g002

After incorporating the correlation factors, the original Equation 6 for the n-tuple nucleobase composition of RNA is augmented to

$$\mathbf{R} = [f_1^* \quad f_2^* \quad f_3^* \quad \cdots \quad f_{10^n}^* \quad f_{10^n+1}^* \quad \cdots \quad f_{10^n+\lambda}^*]^{\mathrm{T}} \tag{9}$$

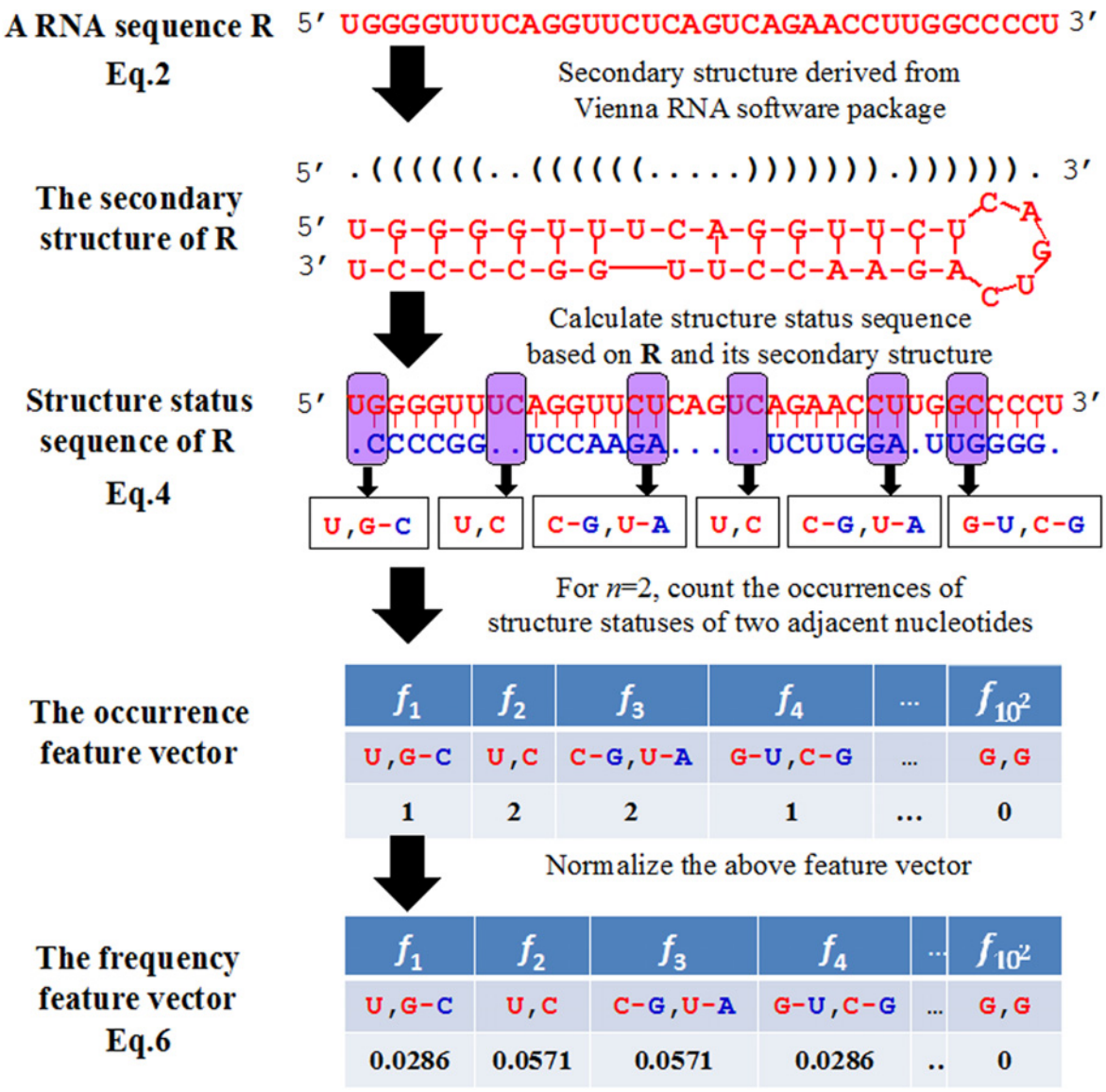

Fig 3. A flowchart to show the process of generating the feature vector for a RNA sequence by its structure status composition. Given a RNA sequence R (cf. Equation 2), its secondary structure sequence was derived from Vienna RNA software package, as formulated in Equation 4. According to the definition in that package, there are two types of status for each of the nucleotides: unpaired or paired. The former is denoted by a dot "." and the latter by the symbol "(" or ")". The left bracket "(" stands for a nucleotide near the 5'-end while the right bracket for the one near the 3'-end. Since the number of different structure elements in the RNA sequence thus obtained is 10 (cf. Equation 5), its n-tuple element composition will contain 10^n components (cf. Equation 6). For simplicity, however, shown here is only for the case of $n = 2$; i.e., the 2-tuple element composition that contains $10^2 = 100$ components formed by different pairs of the most contiguous secondary structure status elements.

doi:10.1371/journal.pone.0121501.g003

where

$$f^* = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{10^n} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 10^n) \\ \dfrac{w\theta_{u-10^n}}{\sum_{i=1}^{10^n} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (10^n + 1 \le u \le 10^n + \lambda) \end{cases} \quad (10)$$

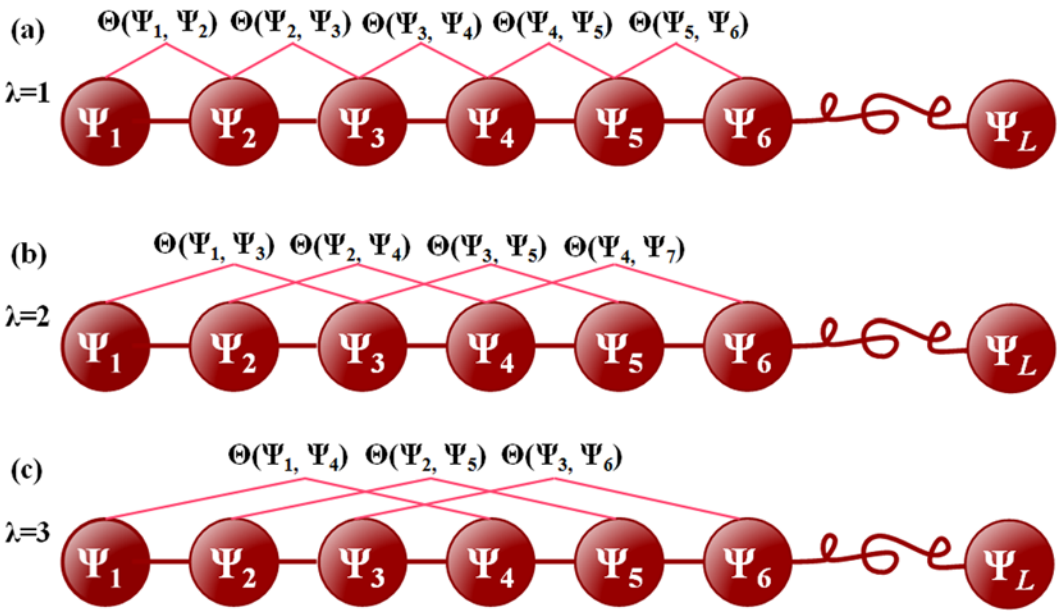

Fig 4. A schematic illustration to show the correlation of structure statuses along a RNA sequence. (a) The first-tier correlation reflects the structure-order mode between all the most contiguous nucleotides. (b) The 2nd-tier correlation reflects the structure-order mode between all the second-most contiguous nucleotides. (c) The 3rd-tier correlation reflects the structure-order mode between all the third-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 as done by the PseAAC approach for proteins [30].

doi:10.1371/journal.pone.0121501.g004

where $f_i = (i = 1,2,\ldots,10^n)$ are the same as in Equation 6, θ_j the j-tier sequence correlation factor computed according to Equations 7–8 for the RNA sequence, and w is the weight factor used to adjust the effect of the correlation factors.

As shown in Equations 9 and 10, the first 10^ncomponents reflect the effect of the n-tuple structure status composition, whereas the components from 10^n+1 to 10^n+λ reflect the effect of structure order. A vector formed with such 10^n+λ components is called pseudo structure status composition or PseSSC for the RNA sequence with L nucleobases.

Finally, the PseSSC vector of Equation 9 was further augmented to

$$\tilde{\mathbf{R}} = [f_1^* \quad f_2^* \quad \cdots \quad f_{10^n}^* \quad f_{10^n+1}^* \quad \cdots \quad f_{10^n+\lambda}^* \quad a \quad b \quad c_1 \quad \cdots \quad c_{64}]^{\mathrm{T}} \tag{11}$$

where $\tilde{\mathbf{R}}$ is the augmented PseSSC, a is the minimum of free energy (MFE) derived from the Vienna RNAsoftware package (released 2.1.6) [61], b the P-value of randomization test feature calculated by using the Monte Carlo randomization test [62], and $c_i(i = 1,2,\ldots,64)$ the occurrence frequencies of the tri-nucleobases in the RNA sequence. A feature vector formed with such 10^n+λ+66 components is called extended pseudo structure status composition or ExPseSSC for the RNA sequence with L nucleobases.

3. Support Vector Machine

Support Vector Machine (SVM) is a class of supervised learning algorithms first introduced by Vapnik [63]. Given a set of labelled training vectors (positive and negative input samples), SVM learns a linear decision boundary from both positive and negative training samples to discriminate between the unseen protein sequences. A key feature of SVM is that it needs fixed length of the input vector. The proteins in the training set and test set were transformed into fixed-dimension feature vectors following the process introduced above, and then the training

vectors were input into SVM to construct the classifier. The SVM gives a predicted class for each sample in the test set.

In the current study, the LIBSVM algorithm [64] was employed, which is a type of software for SVM classification and regression. The kernel function was set as Radial Basis Function (RBF), which is defined as

$$K(X_i, X_j) = \exp(-\gamma \|X_i - X_j\|^2) \tag{12}$$

The two parameters C and γ were optimized on the benchmark dataset by adopting the grid tool provided by LIBSVM [64], and their actual values in this study will be given later. For a brief formulation of SVM and how it works, see the paper [65]; for more details about SVM, see a monograph [66].

Finally, we obtain two predictors, one is based on Equation 9, and the other based on Equation 11, as formulated below

$$\begin{cases} \text{iMcRNA} - \text{PseSSC}, & \text{if use } \mathbf{R} \text{ of Eq.9 to represent RNA samples} \\ \text{iMcRNA} - \text{ExPseSSC}, & \text{if use } \tilde{\mathbf{R}} \text{ of Eq.11 to represent RNA samples} \end{cases} \tag{13}$$

where "i" stands for "identifying", "McRNA" for "microRNA", "Pse" for "pseudo", "SS" for "structure status", "C" for "composition", and "Ex" for extended.

4. Cross Validation

In examining the accuracy of a statistical predictor, it is very important to choose an objective method to perform the test. In literature, the following three cross-validation methods are often used to examine the quality of a predictor and its effectiveness in practical application: independent dataset test, subsampling or K-fold (such as 5-fold, 7-fold, or 10-fold) crossover test, and jackknife test. However, as elucidated by a penetrating analysis in [54], considerable arbitrariness exists in the independent dataset test. Also, as demonstrated by Eqs.28–32 of [54], the subsampling (or K-fold crossover validation) test cannot avoid arbitrariness either. Only the jackknife test is the least arbitrary that can always yield a unique result for a given benchmark dataset. Therefore, the jackknife test has been widely recognized and increasingly utilized by investigators to examine the quality of various predictors (see, e.g., [20,41,57,67–72]). Accordingly, in this study we also use the jackknife test to evaluate the accuracy of the current predictor. During the jackknife test, each of the samples in the benchmark dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the sample being identified. Although the jackknife test may take more computational time, it is worthwhile because it will yield a unique outcome for a given benchmark dataset.

5. Metrics for Measuring Prediction Quality

After choosing the cross validation method, the next important thing is how to quantitatively measure the prediction quality. To introduce a more intuitive and easier-to-understand method for scoring the prediction quality, the following set of metrics based on the formulation used by Chou [73] in predicting signal peptides was adopted. According to the formulation, the sensitivity Sn, specificity Sp, overall accuracy Acc, and Matthews correlation coefficient

MCC can be respectively expressed as [19,20,50]

$$\begin{cases} Sn = 1 - \dfrac{N_-^+}{N^+}, & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-}, & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ - N^-}, & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- + N_-^+}{N^+}\right)\left(1 + \dfrac{N_+^- - N_-^+}{N^-}\right)}}, & -1 \le \mathrm{MCC} \le 1 \end{cases} \quad (14)$$

where $N^+$is the total number of the pre-miRNAs investigated whereas N_-^+ the number of the pre-miRNAs incorrectly predicted as false pre-miRNAs; N^- the total number of the false pre-miRNAs investigated whereas N_+^- the number of the false pre-miRNAs incorrectly predicted as the real pre-miRNAs.

According to Equation 14 we can easily see the following. When $N_-^+ = 0$ meaning none of the pre-miRNAs was mispredicted to be a false pre-miRNAs, we have the sensitivity Sn = 1; while $N_-^+ = N^+$ meaning that all the real pre-miRNAs were mispredicted to be the false pre-miRNAs, we have the sensitivity Sn = 0. Likewise, when $N_+^- = 0$ meaning none of the false pre-miRNAs was mispredicted, we have the specificity Sp = 1; while $N_+^- = N^-$ meaning all the false pre-miRNAs were incorrectly predicted as real pre-miRNAs, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that none of the pre-miRNAs in the dataset $\mathbb{S}^+$ and none of the false pre-miRNAs in $\mathbb{S}^-$ was incorrectly predicted, we have the overall accuracy Acc = 1; while $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the real pre-miRNAs in the dataset $\mathbb{S}^+$ and all the false pre-miRNAs in $\mathbb{S}^-$ were mispredicted, we have the overall accuracy Acc = 0. The Matthews correlation coefficient (MCC) is usually used for measuring the quality of binary (two-class) classifications. When $N_-^+ = N_+^- = 0$ meaning that none of the real pre-miRNAs in the dataset $\mathbb{S}^+$ and none of the false pre-miRNAs in $\mathbb{S}^-$ was mispredicted, we have MCC = 1; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have MCC = 0 meaning no better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$ we have MCC = -1 meaning total disagreement between prediction and observation. As we can see from the above discussion, it is much more intuitive and easier to understand when using Equation 14 to examine a predictor for its four metrics, particularly for its Mathew's correlation coefficient. It is instructive to point out that the metrics as defined in Equation 14 are valid for single label systems; for multi-label systems, a set of more complicated metrics should be used as given in [74].

Results and Discussion

1. Performance of iMcRNA-PseSSC and iMcRNA-ExPseSSC

As we can see from Equation 9–11, both the iMcRNA-PseSSC and iMcRNA-ExPseSSC predictors contain three uncertain parameters, namely n,λ, and w, where n reflects the local or short-range structure-order effect, λ reflects the global or long-range structure-order effect, and 0077 is the factor to adjust the weight between the local and global effects. Generally speaking, the greater the n is, the more local structure-order information is incorporated. And the greater the λ is, the more global structure-order information is taken into account. However, if n or λ is too large, it would reduce the cluster-tolerant capacity [75] and cause the "overfitting" or

"high dimension disaster" [76] problem, so as to reduce the prediction accuracy. Accordingly, in the current study, their optimal values were determined within the ranges as defined below

$$\begin{cases} 1 \le n \le 4 & \text{with step } \Delta = 1 \\ 1 \le \lambda \le 20 & \text{with step } \Delta = 1 \\ 0 \le w \le 1 & \text{with step } \Delta = 0.1 \end{cases} \quad (15)$$

It can be seen from Equation 15 that, to determine the optimal values for the three parameters, 4×20×10 = 800 different combination cases need to be considered. To reduce the computational time, we adopted the 5-fold cross-validation approach on the benchmark dataset. The final optimal values for the three parameters along with the two parameters C and γ in SVM (see Equation 12) were defined by the highest overall accuracy after trying all the 800 combination cases for each of the two predictors in Equation 13, as given by

$$\begin{cases} n = 2, \lambda = 13, w = 0.5, C = 8, \gamma = 2^{-5} & \text{for iMcRNA} - \text{PseSSC} \\ n = 1, \lambda = 17, w = 0.2, C = 128, \gamma = 2^{-7} & \text{for iMcRNA} - \text{ExPseSSC} \end{cases} \quad (16)$$

Thus, the parameters in Equation 16 were used to perform the rigorous jackknife test on the benchmark dataset to calculate the metrics defined in Equation 14.

The results thus obtained by the two new predictors are given in Table 1, from which we can see that the overall accuracy (Acc) achieved by iMcRNA-PseSSC was 85.76% with the Matthews correlation coefficient (MCC) equal to 0.72. The corresponding rates achieved by iMcRNA-ExPseSSC were even better; i.e., 89.86% and 0.80 for Acc and MCC, respectively. It is not surprising because the additional features counted in Equation 11 play a complementary role to the feature in Equation 9. In other words, all these features are complementary with each other: PseSSC is a structure-based feature reflecting the global or long-range structure-order effects; MFE and *P*-value are for the secondary structure state of minimum free energy; and trinucleobase composition is for the local or short-range sequence order information [47].

2. Comparison with Other Methods

We have also made a comparison of the current iMcRNA-PseSSC and iMcRNA-ExPseSSC (Equation 13) with Triplet-SVM [16] and MiPred [21], two of the best existing predictors in this area. As mentioned in the Introduction section, the accuracy rates by the two predictors as originally reported [16,21] were based on small benchmark datasets without removing high similarity or redundant RNA sequences, and hence the rates thus obtained might be over-estimated.

For instance, Triplet-SVM [16] was trained with 163 human pre-miRNAs and 168 false pre-miRNAs, and tested with only 30 human pre-miRNAs and 1,000 false pre-miRNAs. Also, MiPred [21] was trained using the same dataset as used by Triplet-SVM [16] and tested with 263 human pre-miRNAs and 265 false pre-miRNAs. In contrast, the current predictors iMcRNA-PseSSC and iMcRNA-ExPseSSC were trained and tested on a much larger and more stringent benchmark dataset that contained 1,612 human pre-miRNAs and 1,612 false pre-miRNAs in which none had more than 80% pairwise sequence identity to any other.

If using the larger and more stringent benchmark dataset (S1 Dataset) to examine the two predictors via the rigorous jackknife tests, we obtained the corresponding results as given in Table 1

Furthermore, to provide a graphic illustration to show the performances of the four predictors, the corresponding ROC (receiver operating characteristic) curves were drawn in Fig. 5, 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

Table 1. Comparison of different predictors by the jackknife tests on a same benchmark dataset (S1 Dataset).

Method	Acc (%)	MCC	Sn (%)	Sp (%)
iMcRNA-PseSSC[a]	85.76	0.72	88.36	83.50
iMcRNA-ExPseSSC[b]	89.86	0.80	89.93	89.78
Triplet-SVM[c]	81.85	0.64	78.47	85.20
MiPred[d]	87.30	0.75	84.00	90.60

[a]The parameters used: $n = 2$, $\lambda = 13$, $w = 0.5$, $C = 8$, and $\gamma = 2$. $\gamma = 2^{-5}$.

[b]The parameters used: $n = 1$, $\lambda = 17$, $w = 0.2$, $C = 128$, and $\gamma = 2^{-7}$.

[c]Results obtained by in-house implementation from [16].

[d]Results obtained by in-house implementation from [21].

doi:10.1371/journal.pone.0121501.t001

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. 5 we can clearly observe the following. (i) The predictor iMcRNA-PseSSC outperformed Triplet-SVM [16] and was highly comparable with MiPred [21], meaning that the prediction quality can be enhanced to the level of the existing best predictor by only taking into account the long-range or global secondary structure sequence order information. (ii) The predictor iMcRNA-ExPseSSC outperformed all its counterparts, meaning that the prediction quality can be further enhanced by combing the aforementioned long-range information with the local features as used in the existing predictors [16,21].

3. Discriminant Visualization and Interpretation

Why was the current approach able to enhance the success rates so remarkably? To address this problem, we are to carry out a graphical analysis. It can provide an intuitive picture or useful insights for helping understand varieties of complicated relations, as demonstrated by many previous studies on a series of important biological topics, such as using graphical rules to study enzyme-catalyzed reactions [77,78], inhibition of HIV-1 reverse transcriptase [79], and drug metabolism systems [80]; using the "cellular automaton image" [81] to study hepatitis B viral infections [82] and HBV virus gene missense mutation [83]; and using wenxiang diagram or graph [84,85] to study protein-protein interactions [86,87]. Here, we used the heat map [88] to present an intuitive analysis. Similar to the approach in [89], we calculated the discriminant weight vector in the feature space of iMcRNA-PseSSC. The results thus obtained are illustrated in Fig. 6a, where the darker the spot is, the more discriminative power the corresponding structure status has. Thus, according to the degree of dark colour in the subfigure, we can see that the statuses of the four structures (A-U, U-A, C-G, G-C) are more important than the others in identifying human microRNA precursors because they have stronger discriminative power. Moreover, the discriminative powers of the 13 features incorporating the structure-order effects are shown in Fig. 6b, from which we can see that the discriminative power for miRNAs tends to be stronger with the increasing λ in value, indicating that the long-range or global structure-order effect do have considerable impacts upon the discrimination. That is the main reason why iMcRNA-PseSSC can remarkably outperform its counterparts.

4. Web-Server Guide

We have also established a web-server for the two predictors as formulated in Equation 13. Furthermore, for the convenience of the vast majority of experimental scientists, below let us

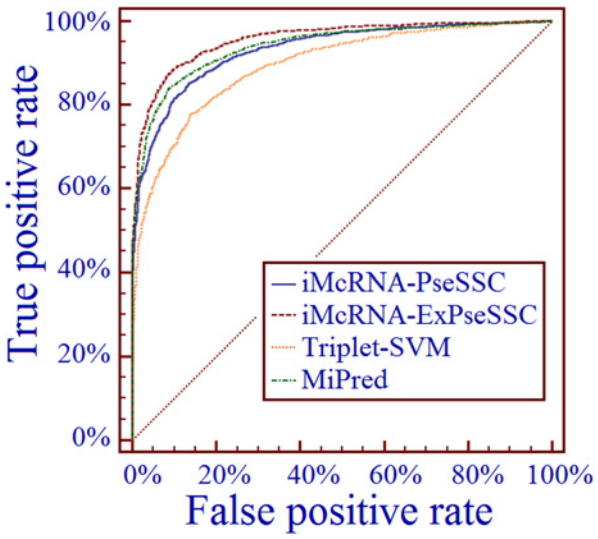

Fig 5. A graphical illustration to show the performance of different methods by means of the receiver operating characteristic (ROC) curves. The areas under the ROC curves, or AUC are 0.93, 0.96, 0.90, and 0.94 for iMcRNA-PseSSC, iMcRNA-ExPseSSC, Triplet-SVM, and MiPred, respectively. See section "Comparison with Other Methods" for further explanation.

doi:10.1371/journal.pone.0121501.g005

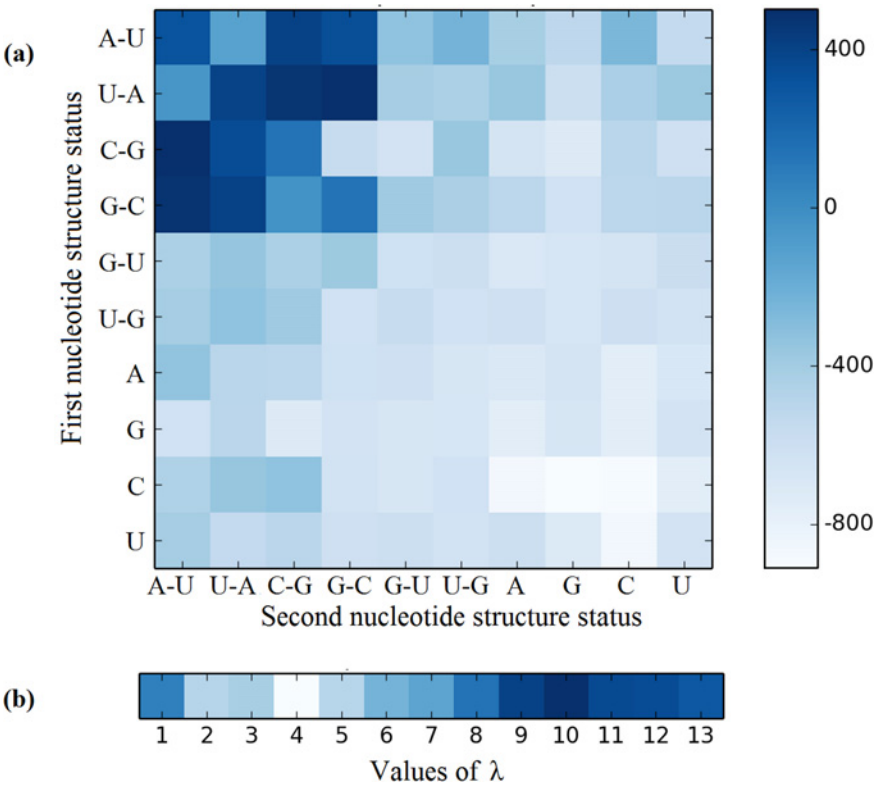

Fig 6. Visualizing the discriminative power with a heat map. (a) The discriminative power of the 100 local structure status compositions. The structure statuses marked on the vertical and horizontal axes indicate the first structure status and the second structure status in the local structure status compositions. (b) The discriminative power of the 13 features incorporating the structure-order effect. The λ values are marked on horizontal axis.

doi:10.1371/journal.pone.0121501.g006

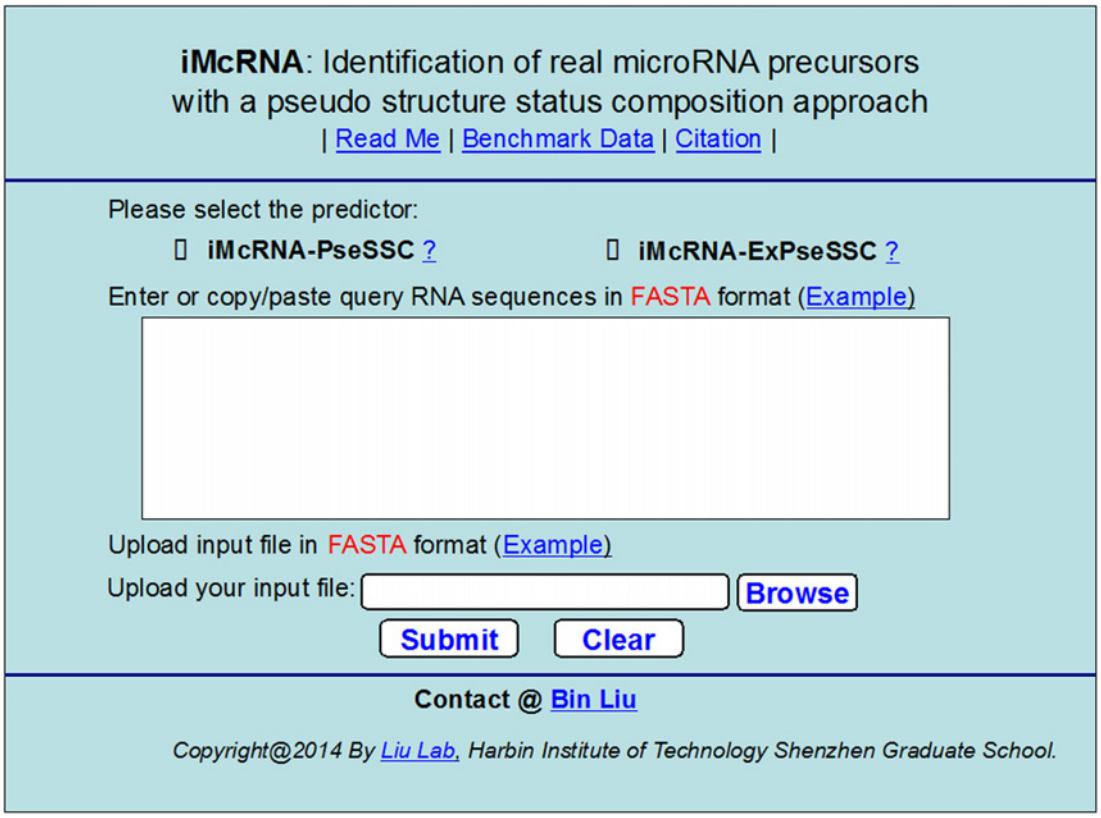

Fig 7. A semi-screenshot to show the top page of the web-server iMcRNA. Its website address is at http://bioinformatics.hitsz.edu.cn/iMcRNA/.

doi:10.1371/journal.pone.0121501.g007

give 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.

Step 1. Open the web-server by clicking the link at http://bioinformatics.hitsz.edu.cn/iMcRNA/ and you will see its top page as shown in Fig. 7. Click on the Read Me button to see a brief introduction about the server that contains two predictors: iMcRNA-PseSSC and iMcRNA-ExPseSSC.

Step 2. Check the open circle right in front of iMcRNA-PseSSC or iMcRNA-ExPseSSC to choose which of the two predictors you are to use for prediction.

Step 3. You can directly enter the query RNA sequences into the input box at the center of Fig. 7, or use the Browse button to upload them via a file. All the input sequences 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 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.

Step 4. 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 select iMcRNA-PseSSC for prediction, after clicking the Submit button, you will see on your screen (Fig. 8) that the predicted results for the 1st and 2nd query RNA sequences are "**Real Pre-miRNA**", and that for

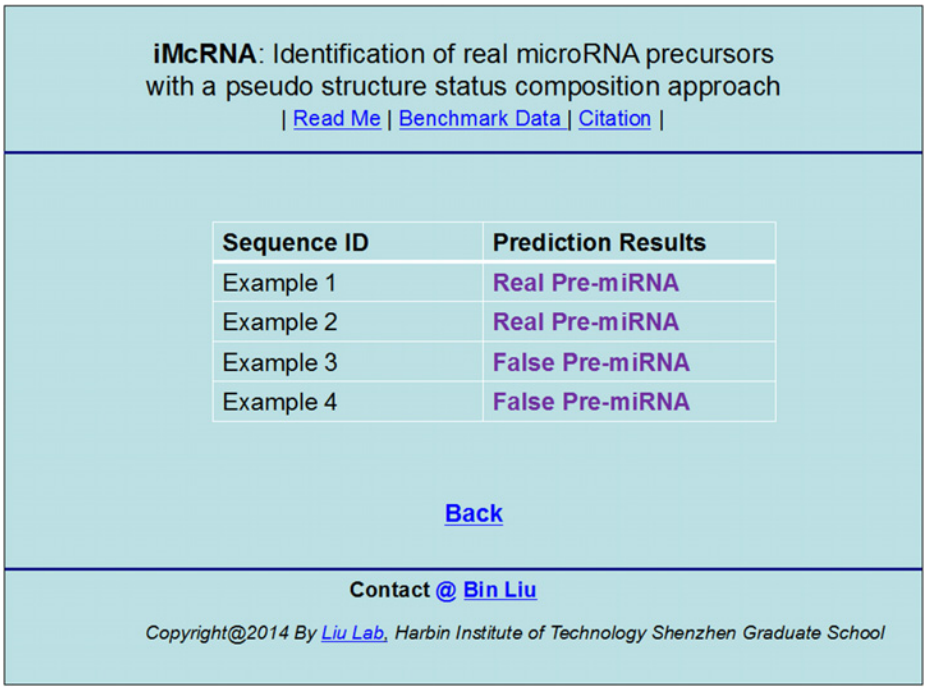

Fig 8. A semi-screenshot to show the output obtained by the web-server. See the text for further explanation.

doi:10.1371/journal.pone.0121501.g008

the 3rd and 4th ones are "**False Pre-miRNA**". **All these predicted results are fully consistent with the experimental observations**. It takes about 2 seconds for the above computation before the predicted result appears on your computer screen. If you select iMcRNA-ExPseSSC, however, for the same prediction, it may take about 20 seconds because more calculations are needed although the overall success rates thus obtained are generally higher than those by the iMcRNA-PseSSC predictor.

Conclusion

Based on the concept of pseudo amino acid composition [30] or Chou's PseAAC [32], two new predictors named iMcRNA-PseSSC and iMcRNA-ExPseSSC were proposed for identifying the human pre-micrRNAs by incorporating the global or long-range structure-order information. It was observed via the rigorous cross-validation on a larger and more stringent newly constructed benchmark dataset that the two new predictors outperformed or were highly comparable with the best existing predictor in this area. The two predictors are publically accessible via a web-server at http://bioinformatics.hitsz.edu.cn/iMcRNA/, 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 two predictors were established for identifying the human pre-micrRNAs, they can be easily used to identify the pre-micrRNAs in any other organism as well if a corresponding benchmark dataset is available.

Supporting Information

S1 Dataset. The benchmark dataset. It contains 3,224 human pre-miRNAs, of which 1,612 are real pre-miRNAs and 1,612 are false pre-miRNAs. None of the sequences included has

≥80% pairwise sequence identity with any other.
(DOC)

Acknowledgments

The authors wish to take this opportunity to thank the two anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of the paper.

Author Contributions

Conceived and designed the experiments: BL. Performed the experiments: LF FL JC BL. Analyzed the data: XW BL KCC. Contributed reagents/materials/analysis tools: XW BL. Wrote the paper: BL KCC.

References

1. Lee Y, Kim M, Han J, Yeom K- H, Lee S, et al. (2004) MicroRNAgenes are transcribed byRNApolymerase II. EMBOJ 23: 4051–4060. PMID: 15372072
2. Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA 10: 1957–1966. PMID: 15525708
3. Lee Y, Ahn C, Han J, Choi H, Kim J, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. Nature 425: 415–419. PMID: 14508493
4. Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U (2004) Nuclear export of microRNA precursors. Science 303: 95–98. PMID: 14631048
5. Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev 17: 3011–3016. PMID: 14681208
6. Bohnsack MT, Czaplinski K, Gorlich D (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA 10: 185–191. PMID: 14730017
7. Bernstein E, Caudy AA, Hammond SM, Hannon GJ (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409: 363–366. PMID: 11201747
8. Grishok A, Pasquinelli AE, Conte D, Li N, Parrish S, et al. (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. Cell 106: 23–24. PMID: 11461699
9. Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, et al. (2001) Acellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science 293: 834–838. PMID: 11452083
10. Knight SW, Bass BL (2001) Arole for theRNaseIII enzymeDCR-1 in RNA interference and germ line development in Caenorhabditis elegans. Science 293: 2269–2271. PMID: 11486053
11. Nam JW, Shin K- R, Han J, Lee Y, Kim VN, et al. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Research 33: 3570–3581. PMID: 15987789
12. Li L, Xu J, Yang D, Tan X, Wang H (2010) Computational approaches for microRNA studies: a review. Mamm Genome 21: 1–12. doi: 10.1007/s00335-009-9241-2 PMID: 20012966
13. Helvik SA, Snøve O, Sætrom P (2007) Reliable prediction of Drosha processing sites improves microRNA gene prediction. Bioinformatics 23: 142–149. PMID: 17105718
14. Huang TH, Fan B, Rothschild MF, Hu ZL, Li K, et al. (2007) MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics 8: 341. PMID: 17868480
15. Wang Y, Chen X, Jiang W, Li L, Li W, et al. (2011) Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM. Genomics 98: 73–78. doi: 10.1016/j.ygeno.2011.04.011 PMID: 21586321
16. Xue C, Li F, He T, Liu G- P, Li Y, et al. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics 6: 310. PMID: 16381612
17. 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. doi: 10.1186/1471-2105-12-107 PMID: 21504621

18. Liu B, Zhang D, Xu R, Xu J, Wang X, et al. (2014) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 30: 472–479. doi: 10.1093/bioinformatics/btt709 PMID: 24318998

19. Qiu WR, Xiao X (2014) iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci 15: 1746–1766. doi: 10.3390/ijms15021746 PMID: 24469313

20. Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, et al. (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30: 1522–1529. doi: 10.1093/bioinformatics/btu083 PMID: 24504871

21. Jiang P, Wu H, Wang W, Ma W, Sun X, et al. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic acids research 35: W339–W344. PMID: 17553836

22. Kandaswamy KK, Chou KC, Martinetz T, Moller S, Suganthan PN, et al. (2011) AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. Journal of Theoretical Biology 270: 56–62. doi: 10.1016/j.jtbi.2010.10.037 PMID: 21056045

23. Lin WZ, Fang JA, Xiao X (2011) iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. PLoS ONE 6: e24756. doi: 10.1371/journal.pone.0024756 PMID: 21935457

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

25. Chou KC, Shen HB (2007) Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry 370: 1–16. PMID: 17698024

26. Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe L, et al. (2006) Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. Bioinformatics 22: 1325–1334. PMID: 16543277

27. Brameier M, Wiuf C (2007) Ab initio identification of human microRNAs based on structure motifs. BMC Bioinformatics 8: 478. PMID: 18088431

28. Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, et al. (2005) Identification of clustered microRNAs using an ab initio prediction method. BMC Bioinformatics 6: 267. PMID: 16274478

29. Huang C, Zhang R, Chen Z, Jiang Y, Shang Z, et al. (2010) Predict potential drug targets from the ion channel proteins based on SVM. Journal of Theoretical Biology 262: 750–756. doi: 10.1016/j.jtbi.2009.11.002 PMID: 19903486

30. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid, 2001, Vol44, 60) 43: 246–255. PMID: 11288174

31. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19. PMID: 15308540

32. Lin SX, Lapointe J (2013) Theoretical and experimental biology in one. J Biomedical Science and Engineering 6: 435–442.

33. Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. Journal of Theoretical Biology 257: 17–26. doi: 10.1016/j.jtbi.2008.11.003 PMID: 19056401

34. Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. Journal of Theoretical Biology 263: 203–209. doi: 10.1016/j.jtbi.2009.11.016 PMID: 19961864

35. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17: 1207–1214. PMID: 20450487

36. Sahu SS, Panda G (2010) A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Computational Biology and Chemistry 34: 320–327. doi: 10.1016/j.compbiolchem.2010.09.002 PMID: 21106461

37. Mohammad Beigi M, Behjati M, Mohabatkar H (2011) Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. Journal of Structural and Functional Genomics 12: 191–197. doi: 10.1007/s10969-011-9120-4 PMID: 22143437

38. Nanni L, Lumini A, Gupta D, Garg A (2012) Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans Comput Biol Bioinform 9: 467–475. doi: 10.1109/TCBB.2011.117 PMID: 21860064

39. Gupta MK, Niyogi R, Misra M (2013) An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition. SAR QSAR Environ Res (SAR AND QSAR IN ENVIRONMENTAL RESEARCH) 24: 597–609. doi: 10.1080/1062936X.2013.773378 PMID: 23710804

40. Chen YK, Li KB (2013) Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. Journal of Theoretical Biology 318: 1–12. doi: 10.1016/j.jtbi.2012.10.033 PMID: 23137835

41. Hajisharifi Z, Piryaiee 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. doi: 10.1016/j.jtbi.2013.08.037 PMID: 24035842

42. Mohabatkar H, Beigi MM, Abdolahi 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. PMID: 22931491

43. Xu R, Zhou J, Liu B, He YA, Zou Q, et al. (2014) Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. Journal of Biomolecular Structure & Dynamics (JBSD).

44. Liu B, Xu J, Fan S, Xu R, Jiyun Zhou J, et al. (2015) PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation. Molecular Informatics 34: 8–17

45. 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. doi: 10.3390/ijms15033495 PMID: 24577312

46. Zhong WZ, Zhou SF (2014) Molecular science for drug development and biomedicine. Intenational Journal of Molecular Sciences 15: 20072–20078. doi: 10.3390/ijms151120072 PMID: 25375190

47. Chen W, Lei TY, Jin DC, Lin H (2014) PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Analytical Biochemistry 456: 53–60. doi: 10.1016/j.ab.2014.04.001 PMID: 24732113

48. Liu B, Liu F, Fang L, Wang X (2014) 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:.

49. Chen W, Zhang X, Brooker J, Lin H, Zhang L, et al. (2015) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 31: 119–120. doi: 10.1093/bioinformatics/btu602 PMID: 25231908

50. Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition Nucleic Acids Research 41: e68. doi: 10.1093/nar/gks1450 PMID: 23303794

51. Lin H, Deng EZ, Ding H, Chen W, Chou KC (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Research 42: 12961–12972. doi: 10.1093/nar/gku1019 PMID: 25361964

52. Liu Z, Xiao X, Qiu WR (2015) iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Analytical Biochemistry.

53. Chou KC (2015) Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry.

54. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology 273: 236–247. doi: 10.1016/j.jtbi.2010.12.024 PMID: 21168420

55. Xu Y, Ding J, Wu LY, Chou KC (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. doi: 10.1371/journal.pone.0055844 PMID: 23409062

56. Fan YN, Xiao X, Min JL (2014) iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. Intenational Journal of Molecular Sciences 15: 4915–4937. doi: 10.3390/ijms15034915 PMID: 24651462

57. Xu Y, Wen X, Shao XJ, Deng NY (2014) iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. International Journal of Molecular Sciences (IJMS) 15: 7594–7610. doi: 10.3390/ijms15057594 PMID: 24857907

58. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic acids research 39: D152–D157. doi: 10.1093/nar/gkq1027 PMID: 21037258

59. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, et al. (2003) A uniform system for microRNA annotation. RNA 9: 277–279. PMID: 12592000

60. 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. PMID: 16731699

61. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic acids research 31: 3429–3431. PMID: 12824340

62. Bonnet E, Wuyts J, Rouzé P, Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 20: 2911–2917. PMID: 15217813

63. Vapnik V (1998) Statistical Learning Theory: Wiley-Interscience, New York.

64. Chang C, Lin CJ (2009) LIBSVM—A Library for Support Vector Machines. Available: http://wwwcsientuedutw/~cjlin/libsvm/.

65. Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. Journal of Biological Chemistry 277: 45765–45769. PMID: 12186861

66. Cristianini N, Shawe-Taylor J (2000) An introduction of Support Vector Machines and other kernel-based learning methodds. Cambridge, UK: Cambridge University Press.

67. Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. PROTEINS: Structure, Function, and Genetics 50: 44–48. PMID: 12471598

68. Chen W, Lin H, Feng PM, Ding C, Zuo YC, et al. (2012) iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. PLoS ONE 7: e47843. doi: 10.1371/journal.pone.0047843 PMID: 23144709

69. 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. J Theor Biol 344: 12–18. doi: 10.1016/j.jtbi.2013.11.021 PMID: 24316044

70. Zakeri P, Moshiri B, Sadeghi M (2011) Prediction of protein submitochondria locations based on data fusion of various features of sequences. Journal of Theoretical Biology 269: 208–216. doi: 10.1016/j.jtbi.2010.10.026 PMID: 21040732

71. Hayat M, Khan A (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. Journal of Theoretical Biology 271: 10–17. doi: 10.1016/j.jtbi.2010.11.017 PMID: 21110985

72. Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, et al. (2015) Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. Journal of Theoretical Biology 364: 284–294. doi: 10.1016/j.jtbi.2014.09.029 PMID: 25264267

73. Chou KC (2001) Using subsite coupling to predict signal peptides. Protein Engineering 14: 75–79. PMID: 11297664

74. Chou KC (2013) Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. Molecular Biosystems 9: 1092–1100. doi: 10.1039/c3mb25555g PMID: 23536215

75. Chou KC (1999) A key driving force in determination of protein structural classes. Biochemical and Biophysical Research Communications (BBRC) 264: 216–224. PMID: 10527868

76. Wang T, Yang J, Shen HB (2008) Predicting membrane protein types by the LLDA algorithm. Protein & Peptide Letters 15: 915–921.

77. Chou KC, Forsen S (1980) Graphical rules for enzyme-catalyzed rate laws. Biochemical Journal 187: 829–835. PMID: 7188428

78. Zhou GP, Deng MH (1984) An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochemical Journal 222: 169–176. PMID: 6477507

79. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, et al. (1993) Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. Journal of Biological Chemistry 268: 6119–6124. PMID: 7681060

80. Chou KC (2010) Graphic rule for drug metabolism systems. Current Drug Metabolism 11: 369–378. PMID: 20446902

81. Wolfram S (1984) Cellular automation as models of complexity. Nature 311: 419–424.

82. Xiao X, Shao SH (2006) A probability cellular automaton model for hepatitis B viral infections. Biochem Biophys Res Comm (BBRC) 342: 605–610.

83. Xiao X, Shao S, Ding Y, Huang Z, Chen X, et al. (2005) An Application of Gene Comparative Image for Predicting the Effect on Replication Ratio by HBV Virus Gene Missense Mutation. Journal of Theoretical Biology 235: 555–565. PMID: 15935173

84. Chou KC, Zhang CT, Maggiora GM (1997) Disposition of amphiphilic helices in heteropolar environments. PROTEINS: Structure, Function, and Genetics 28: 99–108. PMID: 9144795

85. Chou KC, Lin WZ, Xiao X (2011) Wenxiang: a web-server for drawing wenxiang diagrams Natural Science 3: 862–865.

86. Zhou GP (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. doi: 10.1016/j.jtbi.2011.06.006 PMID: 21718705

87. Zhou GP, Huang RB (2013) The pH-Triggered Conversion of the PrP(c) to PrP(sc.). Curr Top Med Chem 13: 1152–1163. PMID: 23647538

88. Wilkinson L, Friendly M (2009) The history of the cluster heat map. The American Statistician 63: 179–184.

89. Liu B, Xu J, Lan X, Xu R, Zhou J, et al. (2014) iDNA-Prot|dis: 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. doi: 10.1371/journal.pone.0106691 PMID: 25184541

90. Lee JJ, Kim HJ, Kim YJ, Lee S, Hwang JY, et al. (2004) Imatinib induces a cytogenetic response in blast crisis or interferon failure chronic myeloid leukemia patients with e19a2 BCR-ABL transcripts. Leukemia 18: 1539–1540. PMID: 15284852