Molecular Therapy
Nucleic Acids
Original Article

AMERICAN SOCIETY of GENE & CELL THERAPY

2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function

Bin Liu,[1,2,3] Fan Yang,[1] and Kuo-Chen Chou[3,4,5]

[1]School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China; [2]Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China; [3]Gordon Life Science Institute, Belmont, MA 02478, USA; [4]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China; [5]Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Involved with important cellular or gene functions and implicated with many kinds of cancers, piRNAs, or piwi-interacting RNAs, are of small non-coding RNA with around 19–33 nt in length. Given a small non-coding RNA molecule, can we predict whether it is of piRNA according to its sequence information alone? Furthermore, there are two types of piRNA: one has the function of instructing target mRNA deadenylation, and the other does not. Can we discriminate one from the other? With the avalanche of RNA sequences emerging in the postgenomic age, it is urgent to address the two problems for both basic research and drug development. Unfortunately, to the best of our knowledge, so far no computational methods whatsoever could be used to deal with the second problem, let alone deal with the two problems together. Here, by incorporating the physicochemical properties of nucleotides into the pseudo K-tuple nucleotide composition (PseKNC), we proposed a powerful predictor called 2L-piRNA. It is a two-layer ensemble classifier, in which the first layer is for identifying whether a query RNA molecule is piRNA or non-piRNA, and the second layer for identifying whether a piRNA is with or without the function of instructing target mRNA deadenylation. Rigorous cross-validations have indicated that the success rates achieved by the proposed predictor are quite high. For the convenience of most biologists and drug development scientists, the web server for 2L-piRNA has been established at http://bioinformatics.hitsz.edu.cn/2L-piRNA/, by which users can easily get their desired results without the need to go through the mathematical details.

INTRODUCTION

With a length of around 19–33 nt, piRNAs (piwi-interacting RNAs) distinctly belong to the largest class of small non-coding RNA molecules in animal cells.[1–4] They are involved with many cellular or gene functions including the transposon silencing, specific protein translation, gene expression regulation, and the formation and maintenance of germ cells.[5–7] Moreover, many studies (see, e.g., Mei et al.,[8] Cheng et al.,[9] Moyano and Stefani,[10] and Hashim et al.[11]) have shown that piRNAs have been implicated with many kinds of cancers. Therefore, knowledge about piRNAs and their functions is very important for drug development, as well as for RNA biology and many other relevant areas.

Given an RNA molecule, can we identify whether it belongs to piRNA? Lee et al.[12] and Nishibu et al.[13] had developed some experimental methods to address this problem, greatly stimulating the development of this area. But purely using experimental methods alone to do the sequence analyses is not only inefficient and expensive, but also insensitive for many cases (e.g., it is difficult to get sufficient quantity of samples for observation). Facing the explosive growth of RNA sequences in the postgenomic age, to make the piRNA analysis in a more efficient way, as well as in a faster pace and at a deeper level, we could not help but resort to the computational approach.

Actually, several computational methods have been proposed for classifying piRNA from non-piRNA sequences. For instance, by combining the k-mer scheme and support vector machine (SVM), Zhang et al.[14] proposed a model called piRNApredictor. Three years later, Wang et al.[15] proposed a different model for predicting piRNAs by using the transposon interaction and SVM. Recently, two more papers were published for identifying piRNAs. One was authored by Luo et al.,[16] who considered the physicochemical properties of RNA sequences, and the other was authored by Li et al.,[17] who used the powerful ensemble approach. Both methods were quite powerful, reaching the state-of-the-art performance.

Received 16 January 2017; accepted 7 April 2017;
http://dx.doi.org/10.1016/j.omtn.2017.04.008.

Correspondence: Bin Liu, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen, Guangdong 518055, China.
E-mail: bliu@hit.edu.cn
Correspondence: Kuo-Chen Chou, Gordon Life Science Institute, Belmont, MA 02478, USA.
E-mail: kcchou@gordonlifescience.org

Table 1. A Comparison of the Proposed Predictor with the Existing State-of-the-Art Methods in Identifying piRNAs, First Layer, and Their Functional Types, Second Layer

Method	Sn (%)[a]	Sp (%)[a]	Acc (%)[a]	MCC[a]
First Layer				
2L-piRNA[b]	88.3	83.9	86.1	0.723
Accurate piRNA prediction[c]	83.1	82.1	82.6	0.651
GA-WE[d]	90.6	78.3	84.4	0.694
Second Layer				
2L-piRNA[b]	79.1	76.0	77.6	0.552
Accurate piRNA prediction[c]	N/A	N/A	N/A	N/A
GA-WE[d]	N/A	N/A	N/A	N/A

All of the data listed were obtained by the 5-fold cross-validation on the same benchmark dataset (Supplemental Information). N/A means "not available," namely, the corresponding method fails to yield any result for the second-layer prediction.

[a] See Equation 15 for the metrics' definition.

[b] The new method presented in this paper.

[c] The existing state-of-the-art method proposed by Luo et al.[16]

[d] The existing state-of-the-art method proposed by Li et al.[17]

It is instructive to point out, however, that there are two types of piRNA in the real world. One has the function of instructing target mRNA deadenylation[18] and the other does not. But none of the aforementioned methods has the function to distinguish these two types.

The present study was initiated in an attempt to fill in this empty area by developing a new predictor that not only can be used to identify piRNAs, but also can be used to identify their functional types.

RESULTS AND DISCUSSION

Listed in Table 1 are the success rates measured by the four metrics of Equation 15 that have been achieved by the proposed two-layer predictor 2L-piRNA on the benchmark datasets $\mathbb{S}$ and $\mathbb{S}^+$ of Equation 1, respectively. For facilitating comparison, listed in the table are also the corresponding results obtained by the powerful existing state-of-the-art methods[16,17] published very recently. From Table 1, we can clearly see the following: (1) for the first-layer prediction, the new predictor 2L-piRNA is superior to the existing state-of-the-art methods in both accuracy (Acc) and Matthews correlation coefficient (MCC), the two most important metrics; the former reflects the overall accuracy of a predictor, and the latter reflects its stability; (2) it is slightly better or comparable with the existing state-of-the-art methods in Sn (sensitivity) and Sp (specificity); and (3) for the second-layer prediction, 2L-piRNA is overwhelmingly better because the existing state-of-the-art methods simply did not have the function to yield any results at this step. Accordingly, the significance of the newly proposed predictor is self-evident.

To further show the advantage of the current 2L-piRNA in using the ensemble classifier approach, we adopted the graphic analysis because it is particularly useful for studying complicated biological systems, as demonstrated by a series of previous studies in many different fields (see, e.g., Jiang et al.,[19] Chou and Forsén,[20] Zhou and Deng,[21] Chou,[22] Althaus et al.,[23,24] Wu et al.,[25] Chou et al.,[26] Zhou,[27] and Zhou et al.[28]). Shown in Figure 1 is the graph of receiver operating characteristic (ROC).[29,30] As we can see from the figure, the area under the ROC curve (AUC) for the ensemble classifier is remarkably larger than any of the individual ones in both the first-layer case (Figure 1A) and second-layer case (Figure 1B), once again demonstrating the merit of 2L-piRNA via the intuitive graphical approach.

Conclusions

It is anticipated that the 2L-piRNA will become a very useful high-throughput tool in genome analysis and drug development, particularly in those areas involved with non-coding RNAs.

MATERIALS AND METHODS

Benchmark Dataset

According to Chou's five-step rule[31] for developing a really useful statistical predictor, the first important thing is to construct or select a reliable benchmark dataset. In literature the benchmark dataset usually consists of a training dataset and a testing dataset: the former is for the usage of training a model, whereas the latter is for testing the model. But as elucidated in a comprehensive review,[32] there is no need to artificially separate a benchmark dataset into the aforementioned two parts if the prediction model is examined by the jackknife test or subsampling (K-fold) cross-validation because the outcome thus obtained is actually from a combination of many different independent dataset tests. Thus, the benchmark datasets $\mathbb{S}$ for the current study can be formulated as

$$\begin{cases} \mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \\ \mathbb{S}^+ = \mathbb{S}^+_{\text{inst}} \cup \mathbb{S}^+_{\text{non-inst}} \end{cases}, \tag{1}$$

where $\mathbb{S}^-$ is the negative subset that contains the non-piRNA samples only, $\cup$ is the symbol for union in the set theory, $\mathbb{S}^+$ is the subset that contains the piRNA samples only, $\mathbb{S}^+_{\text{inst}}$ is the sub-subset that contains piRNA samples having the function of instructing target mRNA deadenylation,[18] whereas $\mathbb{S}^+_{\text{non-inst}}$ is the sub-subset without such function.

The concrete procedures to construct the benchmark dataset of Equation 1 are as follows: (1) The piRNA sequences were taken from piRBase;[33] (2) collected for $\mathbb{S}^+_{\text{inst}}$ are only those samples that were annotated with piRNA having the function of instructing target mRNA deadenylation; (3) collected for $\mathbb{S}^+_{\text{non-inst}}$ are only those samples that were annotated with piRNA without the function of instructing target mRNA deadenylation; (4) the corresponding non-piRNA sequences for the negative subset S^- were taken from Bu et al.;[34] (5) the CD-HIT software[3] with the cutoff threshold 0.8 was used to remove the redundancy for each of the aforementioned subsets; and (6) to minimize the negative effect caused by the skewed benchmark dataset,[35–38] the random sampling method was applied to balance out each of the subsets with its counterpart. The final benchmark dataset obtained by strictly following the above procedures contains 2,836 samples, of which 709 belong to $\mathbb{S}^+_{\text{inst}}$, 709 to $\mathbb{S}^+_{\text{non-inst}}$, and 1,418 to $\mathbb{S}^-$.

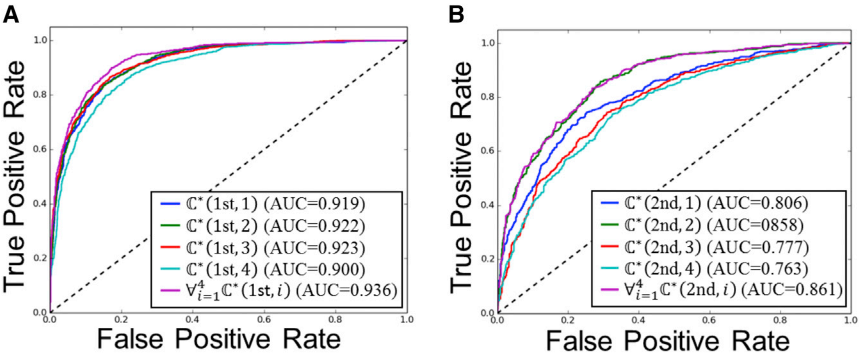


Figure 1. The Performances of the First- and Second-Layer Ensemble Sub-predictors in Comparison with Their Respective Individual Four Basic Predictors

(A and B) A graphical illustration to show the performances of (A) the first-layer ensemble sub-predictor and (B) the second ensemble sub-predictor predictor in comparison with their respective individual four basic predictors (cf. Equation 14). The performances are illustrated by means of the ROC curves.[29,30] The greater the area under the ROC curve (AUC) value is, the better the performance will be.

Shown in Figure 2 is the sequence length distribution of the samples in the benchmark dataset; their detailed sequences and the relevant codes are given in the Supplemental Information.

Pseudo K-Tuple Nucleotide Composition

With a good benchmark dataset, the next thing we need to consider is how to formulate the samples therein. Actually, this is one of the most challenging problems in computational biology. This is because all the existing machine learning algorithms were designed to handle the discrete models or vectors only.[39] But a biological sequence expressed by a vector may completely miss its sequence order or pattern,[40] so as to limit the prediction quality. The pseudo amino acid composition (PseAAC) was proposed to deal with such a dilemma.[40–45] Ever since the concept of PseAAC was introduced, it has been rapidly and widely used in nearly all the areas of computational proteomics (see, e.g., Du et al.,[45] Lin and Lapointe,[46] Chou,[47] Khan et al.,[48] and Meher et al.[49] and a long list of references cited in these papers). Inspired by the great successes of using PseAAC to represent protein-peptide sequences, the PseKNC (pseudo K-tuple nucleotide composition) was introduced to represent DNA/RNA sequences.[50–54] Likewise, since its introduction, PseKNC has also been increasingly applied in many areas of genome analysis.[37,55–68]

For an RNA sample with L nucleotide, its sequence expression is generally given by

$$\mathbf{R} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\cdots\mathrm{N}_i\cdots\mathrm{N}_L, \tag{2}$$

where

$$\mathrm{N}_i \in \{\ \mathrm{A\ (adenine)},\quad \mathrm{C\ (cytosine)},\quad \mathrm{G\ (guanine)},\quad \mathrm{U\ (uracil)}\ \} \tag{3}$$

denotes the nucleotide at the i-th sequence position, and $\in$ is the a symbol in the set theory meaning "member of." According to a recent review paper,[69] the general form of PseKNC for **R** of Equation 2 can be formulated as

$$\mathbf{R} = [\,\phi_1 \quad \phi_2 \quad \cdots \quad \phi_u \quad \cdots \quad \phi_Z\,]^{\mathbf{T}}, \tag{4}$$

where the components $\phi_u (u = 1,\ 2,\ \cdots)$ and Z is an integer; their values will depend on how the desired features are extracted from the RNA sample; and **T** is the transposing operator to a matrix or vector.

In this study, we take

$$\phi_u = \begin{cases} \dfrac{f_u^{\mathrm{K}-tuple}}{\sum_{i=1}^{4^{\mathrm{K}}} f_u^{\mathrm{K}-tuple} + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 4^{\mathrm{K}}) \\ \dfrac{w\theta_{u-4^{\mathrm{K}}}}{\sum_{i=1}^{4^{\mathrm{K}}} f_u^{\mathrm{K}-tuple} + w\sum_{j=1}^{\lambda}\theta_j} & (4^{\mathrm{K}}+1 \le u \le 4^{\mathrm{K}}+\lambda) \end{cases}, \tag{5}$$

where $f_u^{\mathrm{K-tuple}}$ is the u-th component of the K-tuple nucleotide composition for the RNA sample sequence, and

$$\theta_j = \frac{1}{L-\mathrm{K}-(\lambda-1)} \sum_{i=1}^{L-\mathrm{K}-(\lambda-1)} C_{i,i+j}\ (j = 1,\ 2, \cdots, \lambda;\ \lambda < L-K). \tag{6}$$

In Equation 6, the correlation function or coupling factor is given by

$$C_{i,i+j} = \frac{1}{\mu}\sum_{\xi=1}^{\lambda} \times \left[H_\xi(\mathrm{N}_i\mathrm{N}_{i+1}\cdots\mathrm{N}_{i+\mathrm{K}-1}) - H_\xi\left(\mathrm{N}_{i+j}\mathrm{N}_{i+j+1}\cdots\mathrm{N}_{i+j+\mathrm{K}-1)}\right)\right]^2, \tag{7}$$

where μ is the number of physicochemical properties considered, whereas $H_\xi(\mathrm{N}_i\mathrm{N}_{i+1}\cdots\mathrm{N}_{i+\mathrm{K}-1)})$ is the numerical value of the ξ-th physicochemical property for the K-mer $\mathrm{N}_i\mathrm{N}_{i+1}\cdots\mathrm{N}_{i+\mathrm{K}-1}$ in the RNA sequence, and so forth.

In this study, we consider pseudo dinucleotide composition. Thus, we can substitute K = 2 into Equations 5, 6, and 7. Also, we used the values of the following six RNA dimer's physicochemical properties: rise, roll, shift, slide, tilt, and twist (Table 2). Thus, we can substitute μ = 6 as well as Rise ($\mathrm{N}_i\mathrm{N}_{i+1}$), Slide ($\mathrm{N}_i\mathrm{N}_{i+1}$), Shift ($\mathrm{N}_i\mathrm{N}_{i+1}$), Twist

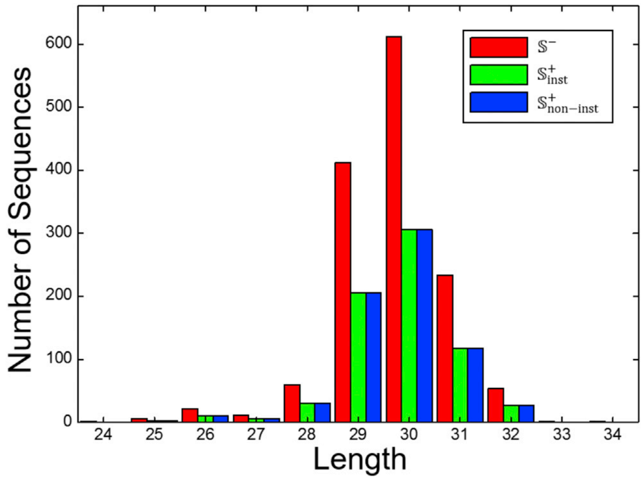


Figure 2. Length Distribution of the Sequences in the Benchmark Dataset

(N_iN_{i+1}), Roll (N_iN_{i+1}), Tilt(N_iN_{i+1}), and so forth into Equation 7 to get the coupling factors.

Note that before substituting them into Equation 7, all the original values in Table 2 were subjected to a standard conversion,[41] as described by the following equations:

$$\begin{cases} \text{Rise}(N_iN_{i+1}) \Leftarrow \dfrac{\text{Rise}(N_iN_{i+1}) - \langle\text{Rise}\rangle}{\text{SD}(\text{Rise})} \\ \text{Slide}(N_iN_{i+1}) \Leftarrow \dfrac{\text{Slide}(N_iN_{i+1}) - \langle\text{Slide}\rangle}{\text{SD}(\text{Slide})} \\ \text{Shift}(N_iN_{i+1}) \Leftarrow \dfrac{\text{Shift}(N_iN_{i+1}) - \langle\text{Shift}\rangle}{\text{SD}(\text{Shift})} \\ \text{Twist}(N_iN_{i+1}) \Leftarrow \dfrac{\text{Twist}(N_iN_{i+1}) - \langle\text{Twist}\rangle}{\text{SD}(\text{Twist})} \\ \text{Roll}(N_iN_{i+1}) \Leftarrow \dfrac{\text{Roll}(N_iN_{i+1}) - \langle\text{Roll}\rangle}{\text{SD}(\text{Roll})} \\ \text{Tilt}(N_iN_{i+1}) \Leftarrow \dfrac{\text{Tilt}(N_iN_{i+1}) - \langle\text{Tilt}\rangle}{\text{SD}(\text{Tilt})} \end{cases}, \tag{8}$$

where the symbol "< >" means taking the average of the quantity therein over 16 different dinucleotides. The converted values obtained by Equation 8 will have a zero mean value over the 16 different dinucleotides and will remain unchanged if going through the same conversion procedure again. Listed in Table 3 are the corresponding values obtained via the standard conversion of Equation 8 from those of Table 2.

Operation Engine

Below, let us consider the third step of the five-step rule,[31] i.e., what kind of algorithms should be used to operate the training and predicting.

Support Vector Machine

Being widely used in many different areas of computational biology (see, e.g., Feng et al.[70] Han et al.,[71] Liu et al.,[72] Qumar et al.,[73] Kiu et al.,[74] Liu et al.,[75,76] Rahimi et al.,[77] and Chen et al.[78]), SVM is a powerful algorithm in cluster analysis. Its basic idea has been elaborated in the aforementioned papers, and hence there is no need to repeat it here. For those who are interested in knowing more about SVM, refer to the previous papers[79,80] or a monograph.[81]

Table 2. The Original Values of Rise, Roll, Shift, Slide, Tilt, and Twist for the 16 Dinucleotides in RNA[51,142]

	Physicochemical Property					
Dimer	Rise	Roll	Shift	Slide	Tilt	Twist
AA	3.18	7.0	−0.08	−1.27	−0.8	31
AC	3.24	4.8	0.23	−1.43	0.8	32
AG	3.30	8.5	−0.04	−1.50	0.5	30
AU	3.24	7.1	−0.06	−1.36	1.1	33
CA	3.09	9.9	0.11	−1.46	1	31
CC	3.32	8.7	−0.01	−1.78	0.3	32
CG	3.30	12.1	0.30	−1.89	−0.1	27
CU	3.30	8.5	−0.04	−1.50	0.5	30
GA	3.38	9.4	0.07	−1.70	1.3	32
GC	3.22	6.1	0.07	−1.39	0.0	35
GG	3.32	12.1	−0.01	−1.78	0.3	32
GU	3.24	4.8	0.23	−1.43	0.8	32
UA	3.26	10.7	−0.02	−1.45	−0.2	32
UC	3.38	9.4	0.07	−1.70	1.3	32
UG	3.09	9.9	0.11	−1.46	1.0	31
UU	3.18	7.0	−0.08	−1.27	−0.8	31

In this study, we used the Scikit-learn[82] as the implementation of the LIBSVM[83] with the radial basis function (RBF) kernel.

Two-Layer Classification Framework

Inspired by the recent study,[76] we constructed a two-layer classification framework as done in Chou and Shen,[84–86] Wang et al.,[87] Xiao et al.,[88–90] and Shen and Chou.[91–93] The SVM model in the first-layer classifier was trained with S of Equation 1, serving to predict a query RNA sample as of piRNA or non-piRNA; the SVM model in the second layer was trained with $\mathbb{S}^+$ of Equation 1 to further identify whether the predicted piRNA sample is with the function of instructing target mRNA deadenylation. Shown in Figure 3 is a flowchart to illustrate how the two-layer classifier is working.

Ensemble Learning

As we can see from Equations 5 and 6, the RNA sample defined by the PseKNC approach in this study contains three uncertain parameters: K, λ, and *w*. In this study, the ranges considered for these parameters are

$$\begin{cases} 1 \le K \le 6 & \text{with step gap } \Delta K = 1 \\ 1 \le \lambda \le 17 & \text{with step gap } \Delta\lambda = 4 \\ 0.1 \le w \le 0.9 & \text{with step gap } \Delta w = 0.2 \end{cases}. \tag{9}$$

Table 3. The Normalized Values Obtained from Table 2 via the Standard Conversion of Equation 8

	Physicochemical Property					
Dimer	Rise	Roll	Shift	Slide	Tilt	Twist
AA	−0.862	−0.689	−1.163	1.386	−1.896	−0.270
AC	−0.149	−1.698	1.545	0.510	0.555	0.347
AG	0.565	0.000	−0.813	0.127	0.096	−0.888
AU	−0.149	−0.643	−0.988	0.894	1.015	0.965
CA	−1.931	0.643	0.497	0.346	0.862	−0.270
CC	0.802	0.092	−0.551	−1.407	−0.211	0.347
CG	0.565	1.652	2.156	−2.009	−0.823	−2.741
CU	0.565	0.000	−0.813	0.127	0.096	−0.888
GA	1.515	0.413	0.147	−0.969	1.321	0.347
GC	−0.386	−1.102	0.147	0.729	−0.670	2.201
GG	0.802	1.652	−0.551	−1.407	−0.211	0.347
GU	−0.149	−1.698	1.545	0.510	0.555	0.347
UA	0.089	1.010	−0.639	0.401	−0.977	0.347
UC	1.515	0.413	0.147	−0.969	1.321	0.347
UG	−1.931	0.643	0.497	0.346	0.862	−0.270
UU	−0.862	−0.689	−1.163	1.386	−1.896	−0.270

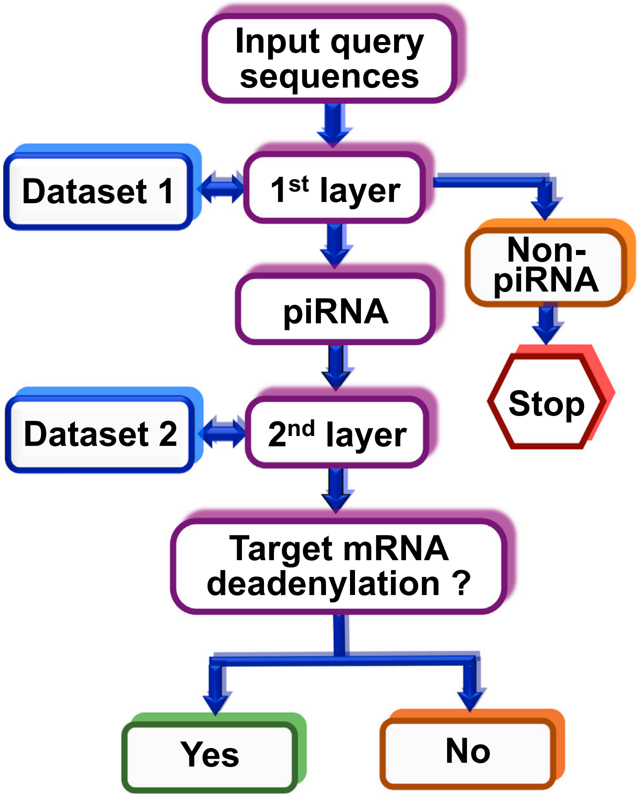


Figure 3. A Flowchart to Show How the 2L-piRNA Predictor Is Working

The input query sequences are first identified by the first-layer sub-predictor as of piRNA or non-piRNA. Subsequently, the predicted or asserted piRNAs are further identified by the second-layer sub-predictor because they have the function to instruct target mRNA deadenylation or not. Dataset 1 and dataset 2 refer to S and S^+ of Equation 1, respectively.

In other words, K may be 1, 2, 3, 4, 5, and 6; λ may be 1, 5, 9, 13, 17, and 19; w may be 0.1, 0.3, 0.5, 0.7, and 0.9. Accordingly, there are a total of 5 × 6 × 5 = 150 individual classifiers for each layer.

Suppose each of these individual classifiers is expressed by $\mathbb{C}(i)$ $(i = 1,\ 2,\ \cdots,\ 150)$, their ensemble classifier $\mathbb{C}^E$ can be formulated as

$$\mathbb{C}^E = \mathbb{C}(1)\forall\mathbb{C}(2)\mathbb{C}(3)\forall\cdots\forall\mathbb{C}(150) = \forall_{i=1}^{150}\mathbb{C}(i), \tag{10}$$

where the symbol ∀ denotes the fusing operator.[32] The ensemble predictor formed by fusing an array of individual predictors via a voting system can yield much better prediction quality, as demonstrated by a series of previous studies including signal peptide prediction,[86,92] membrane protein type classification,[84,94] protein subcellular location prediction,[95,96] protein fold pattern recognition,[97] enzyme functional classification,[98] protein-proteins interaction prediction,[99] protein-protein binding site identification,[100] and DNA recombination spot identification.[68]

Unfortunately, if all of the 150 classifiers in Equation 10 were directly used to form an ensemble predictor by the voting approach, it would be not only computationally inefficient, but also might reduce the success rate because of too much noise. One of the effective approaches is to select some key classifiers from them. To realize this, let us introduce the concept of "complementing degree" between two individual classifiers, C(i) and C(j), or their "mutually strengthening degree," D(i,j), as defined below:

$$D(i,j) = 1 - \frac{1}{2m}\sum_{t=1}^{m} C_t(i,j) \quad (0 \le D(i,j) \le 1)\ , \tag{11}$$

where m represents the number of training samples, and

$$\mathrm{C_t(i,j)} = \begin{cases} \mathrm{p_t(i) + p_t(j),} & \text{if both fail} \\ 0, & \text{otherwise} \end{cases} . \tag{12}$$

In Equation 12, $p_t(i)$ denotes the probability or output when applying the classifier C(i) on the t-th sample, $p_t(j)$ the corresponding output for $C(j)$, and "both fail" means that both predicted results are incorrect.

By means of Equations 11 and 12, all of the 150 classifiers in each layer were clustered with the AP (affinity propagation) clustering algorithm[101] using the default parameters. Four clusters were thus obtained for each of the two layers. Subsequently, the classifiers in the four cluster centers were selected as the representative classifiers, respectively, that have the highest complementing/strengthening degrees, as illustrated by the flowchart in Figure 4. Suppose the four

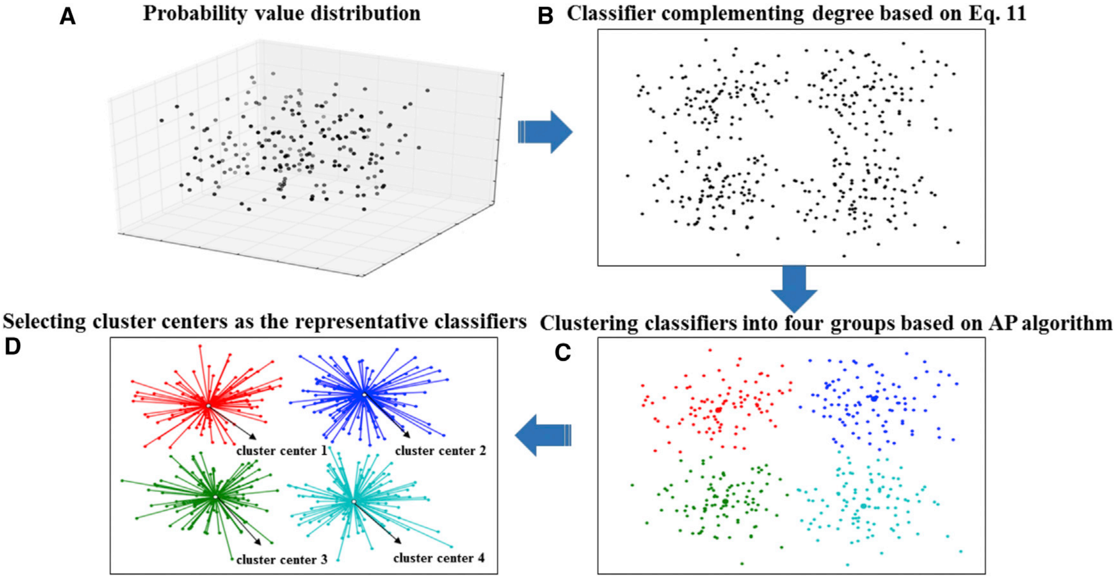

A
Probability value distribution
B
Classifier complementing degree based on Eq. 11
Clustering classifiers into four groups based on AP algorithm
C
Selecting cluster centers as the representative classifiers
D
cluster center 1
cluster center 2
cluster center 3
cluster center 4

Figure 4. A Flowchart to Show the Process of How to Select the Four Representative Classifiers in Equation 13 from the 150 Individual Basic Classifier in Equation 10 for the First and Seconds Layers, Respectively

representative classifiers thus selected for the first and second layers are denoted by

$$\begin{cases} \mathbb{C}^*(1st,\ 1),\ \mathbb{C}^*(1st,\ 2), \mathbb{C}^*(1st,\ 3), \mathbb{C}^*(1st,\ 4) \\ \mathbb{C}^*(2nd,\ 1),\ \mathbb{C}^*(2nd,\ 2), \mathbb{C}^*(2nd,\ 3), \mathbb{C}^*(2nd,\ 4) \end{cases}. \tag{13}$$

Listed in Table 4 are the detailed values of their parameters for the first and second layers, respectively. Thus, instead of Equation 10, the final ensemble classifier should be formulated as

$$\mathbb{C}^{\mathrm{E}} = \begin{cases} \forall_{\mathrm{i}=1}^{4} \mathbb{C}^*(1\mathrm{st},\ \mathrm{i}), & \text{for the 1st layer} \\ \forall_{\mathrm{i}=1}^{4} \mathbb{C}^*(2\mathrm{nd},\ \mathrm{i}), & \text{for the 2nd layer} \end{cases}. \tag{14}$$

Note that different from the ensemble classifiers formed in Chou and Shen[102–105] and Qiu et al.,[106,107] the voting weighted factors V_w were included during the fusion process for each layer, and their optimal values can be easily derived by optimizing success rates during the validation process as shown in Table 4 (Voting Weighted Factor V_w column).

The predictor developed via the above procedures is called 2L-piRNA, where 2L represents the two-layer ensemble classifier and piRNA represents the piwi-interacting RNA and its function.

Prediction Quality Measurement

How to measure the prediction quality is one of the five indispensable steps[31] in developing a new prediction method for a biological system. It consists of two issues: What scales should be used to measure the predictor's quality? And what test method should be adopted to score them? Below, let us address the two problems one by one.

Formulation of Measurement Scales

The following metrics were widely used in the literature to measure the prediction quality from four different aspects: (1) Acc that was used for checking the overall accuracy of a predictor, (2) MCC for its stability, (3) Sn for its sensitivity, and (4) Sp for its specificity.[108] Unfortunately, the four metrics' original formulations copied directly from mathematical books are difficult to understand for most biologists due to lack of intuitiveness. Fortunately, by using the scales defined by Chou[109] in studying signal peptides, Xu et al.[110] and Chen et al.[55] had successfully converted them into a set of intuitive equations that are much easier for most biologists to understand, as given below:

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases}, \tag{15}$$

where N^+ represents the total number of the positive samples investigated, whereas N_-^+ is the number of the positive samples incorrectly predicted to be negative, and N^- is the total number of the negative samples investigated, whereas N_+^- is the number of the negative samples incorrectly predicted to be positive.

Table 4. List of the Four Individual Representative Base Classifiers Selected by Using the Affinity Propagation Clustering Algorithm[101] for Each of the Two Layers Concerned

Base Classifier	Feature	Dimension	Voting Weighted Factor V_w	Acc (%)
First Layer				
$\mathbb{C}^*(1st,\ 1)$	PseKNC[a]	17	0.200	84.1
$\mathbb{C}^*(1st,\ 2)$	PseKNC[b]	21	0.100	84.0
$\mathbb{C}^*(1st,\ 3)$	PseKNC[c]	69	0.300	84.6
$\mathbb{C}^*(1st,\ 4)$	PseKNC[d]	257	0.400	82.1
Second Layer				
$\mathbb{C}^*(2nd,\ 1)$	PseKNC[e]	17	0.100	73.8
$\mathbb{C}^*(2nd,\ 2)$	PseKNC[f]	69	0.800	77.0
$\mathbb{C}^*(2nd,\ 3)$	PseKNC[g]	4,101	0.000	71.4
$\mathbb{C}^*(2nd,\ 4)$	PseKNC[h]	4,113	0.100	70.1

[a] The optimal parameters were K = 2, $\lambda = 1$, $w = 0.1$, $C = 2^7$, $\gamma = 2$.
[b] The optimal parameters were K = 2, $\lambda = 5$, $w = 0.3$, $C = 2^{15}$, $\gamma = 2^{-1}$.
[c] The optimal parameters were K = 3, $\lambda = 5$, $w = 0.1$, $C = 2^{13}$, $\gamma = 2^{-1}$.
[d] The optimal parameters were K = 4, $\lambda = 1$, $w = 0.3$, $C = 2^{13}$, $\gamma = 2^{-1}$.
[e] The optimal parameters were K = 2, $\lambda = 1$, $w = 0.9$, $C = 2^{13}$, $\gamma = 2$.
[f] The optimal parameters were K = 3, $\lambda = 5$, $w = 0.1$, $C = 2^9$, $\gamma = 2$.
[g] The optimal parameters were K = 6, $\lambda = 5$, $w = 0.7$, $C = 2^7$, $\gamma = 2^3$.
[h] The optimal parameters were K = 6, $\lambda = 17$, $w = 0.9$, $C = 2^{11}$, $\gamma = 2^3$.

Based on the definition of Equation 15, the meanings of Sn, Sp, Acc, and MCC have become much more intuitive and easier to understand, as discussed and used in a series of recent studies in various biological areas (see, e.g., Jia et al.,[35,36,99,100,111–113] Liu et al.,[37,75,114,115] Xiao et al.,[38] Lin et al.,[59] Chen et al.,[61,116] Qiu et al.,[106,107,117,118] Xu et al.,[119–122] and Ding et al.[123]).

It should be pointed out, however, that for the multi-label systems (see, e.g., Xiao et al.,[90] Qiu et al.,[118] Xiao et al.,[124] Chou et al.,[125] Lin et al.,[126] and Cheng et al.[127]), a much more sophisticated set of scales is needed as elaborated by Chou.[128]

Cross-Validation

There are three different cross-validation methods[129] that are widely used in literature: (1) jackknife test, (2) subsampling (or K-fold cross-validation) test, and (3) independent dataset test. Of these three, however, the jackknife is the least arbitrary that can always yield a unique outcome for a given benchmark dataset, as elaborated by Chou[31] and widely recognized and increasingly adopted by researchers to analyze the quality of various predictors (see, e.g., Kabir and Hayat,[64] Kumar et al.,[73] Chen et al.,[78] Ali and Hayat,[130] Khan et al.,[131] Mondal and Pai,[132] Dehzangi et al.,[133] Ahmad et al.,[134] Ju et al.,[135] and Behbahani et al.[136]). In this study, however, to reduce the computational time, we adopted the 5-fold cross-validation method for each layer in 2L-piRNA, as done by many investigators with SVM as the prediction engine. For each layer, the benchmark dataset was divided into five subsets; for each run, four subsets were used as the training set, and the remaining one was used as the test set to evaluate the performance. This process was repeated five times until each subset was used as a test set once. To do this, we first randomly divided the benchmark datasets in Equation 1 into five subsets with approximately the same size. For instance, for the first benchmark dataset in Equation 1, we have

$$\begin{cases} \mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 = \cup_{i=1}^{5} \mathbb{S}_i \\ \varnothing = \mathbb{S}_1 \cap \mathbb{S}_2 \cap \mathbb{S}_3 \cap \mathbb{S}_4 \cap \mathbb{S}_5 = \cap_{i=1}^{5} \mathbb{S}_i \end{cases}, \tag{16}$$

where $\cup$, $\cap$, and $\varnothing$ represent the symbols for union, intersection, and

empty set in the set theory,[95,137] respectively, and

$$\mathbb{S}_i = \mathbb{S}_i^+ \cup \mathbb{S}_i^- \quad (i = 1,\ 2,\ \cdots,\ 5) \tag{17}$$

with

$$\begin{cases} \left|\mathbb{S}_1^+\right| \approx \left|\mathbb{S}_2^+\right| \approx \left|\mathbb{S}_3^+\right| \approx \left|\mathbb{S}_4^+\right| \approx \left|\mathbb{S}_5^+\right| \\ \left|\mathbb{S}_1^-\right| \approx \left|\mathbb{S}_2^-\right| \approx \left|\mathbb{S}_3^-\right| \approx \left|\mathbb{S}_4^-\right| \approx \left|\mathbb{S}_5^-\right| \end{cases}, \tag{18}$$

where $\left|\mathbb{S}_1^+\right|$ denotes the number of samples (or cardinalities) in $\mathbb{S}_1^+$, and so forth.

Then, each of the five sub-benchmark datasets was singled out one by one and tested by the model trained with the remaining four sub-benchmark datasets. The cross-validation process was repeated five times, with their average as the final outcome. In other words, during the process of 5-fold cross-validation, both the training dataset and testing dataset were actually open, and each sub-benchmark dataset was in turn moved between the two. The 5-fold cross-validation test can exclude the "memory" effect, just like conducting five different independent dataset tests.

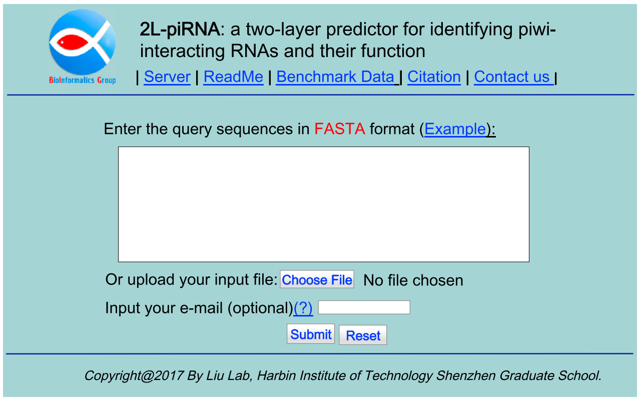
Web Server and User Guide

In Chou's five-step rule[31] for developing a useful predictor, the last one is to establish a user-friendly web server. This not only represents the future direction for developing any computational methods,[138] but is also particularly important for most experimental scientists working in drug development.[39] Accordingly, as done in a series of recent studies,[63,66,67,107,112,117,127,139–141] the web server for 2L-piRNA has been established as well. Moreover, to maximize users' convenience, a step-by-step guide is provided below.

Step 1. Open the web server at http://bioinformatics.hitsz.edu.cn/2L-piRNA/ 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 and the caveat when using it.

Step 2. You can either type or copy/paste the query RNA sequence into the input box. You can also directly upload your input data via the Browse button. The input sequence should be in the FASTA format. For the examples of sequences in the FASTA format, click 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



2L-piRNA: a two-layer predictor for identifying piwi-interacting RNAs and their function

Enter the query sequences in FASTA format (Example):

Or upload your input file: Choose File No file chosen

Input your e-mail (optional)(?)

Copyright@2017 By Liu Lab, Harbin Institute of Technology Shenzhen Graduate School.

Bioinformatics Group
| Server | ReadMe | Benchmark Data | Citation | Contact us |
Submit
Reset

Figure 5. A Semi-screen Shot to Show the Top Page of the Web Server 2L-piRNA

Its website address is http://bioinformatics.hitsz.edu.cn/2L-piRNA/.

window as the input, you will see on your computer screen that the first and second query sequences are of non-piRNA. The third one is of piRNA with the function for instructing target mRNAs deadenylation. The fourth one is of piRNA, but without that function. All these predicted results are fully consistent with the experimental observations as reported in Gou et al.[18]

SUPPLEMENTAL INFORMATION

Supplemental Information includes one data file and can be found with this article online at http://dx.doi.org/10.1016/j.omtn.2017.04.008.

AUTHOR CONTRIBUTIONS

B.L. conceived of the study and designed the experiments, participated in drafting the manuscript and performing the statistical analysis. F.Y. participated in coding the experiments and drafting the manuscript. K.-C.C. participated in revising the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors wish to thank the two anonymous reviewers, whose constructive comments were very helpful for strengthening the presentation of this paper. This work was supported by the National Natural Science Foundation of China (grant No. 61672184), the Natural Science Foundation of Guangdong Province (grant No. 2014A030313695), Guangdong Natural Science Funds for Distinguished Young Scholars (grant No. 2016A030306008), and Scientific Research Foundation in Shenzhen (grant No. JCYJ20150626110425228).

REFERENCES

1. Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203–207.

2. Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199–202.

3. Grivna, S.T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. Genes Dev. *20*, 1709–1714.

4. Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363–367.

5. Zhang, P., Kang, J.-Y., Gou, L.-T., Wang, J., Xue, Y., Skogerboe, G., Dai, P., Huang, D.W., Chen, R., Fu, X.D., et al. (2015). MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. Cell Res. *25*, 193–207.

6. Klattenhoff, C., and Theurkauf, W. (2008). Biogenesis and germline functions of piRNAs. Development *135*, 3–9.

7. Beyret, E., and Lin, H. (2011). Pinpointing the expression of piRNAs and function of the PIWI protein subfamily during spermatogenesis in the mouse. Dev. Biol. *355*, 215–226.

8. Mei, Y., Clark, D., and Mao, L. (2013). Novel dimensions of piRNAs in cancer. Cancer Lett. *336*, 46–52.

9. Cheng, J., Deng, H., Xiao, B., Zhou, H., Zhou, F., Shen, Z., and Guo, J. (2012). piR-823, a novel non-coding small RNA, demonstrates in vitro and in vivo tumor suppressive activity in human gastric cancer cells. Cancer Lett. *315*, 12–17.

10. Moyano, M., and Stefani, G. (2015). piRNA involvement in genome stability and human cancer. J. Hematol. Oncol. *8*, 38.

11. Hashim, A., Rizzo, F., Marchese, G., Ravo, M., Tarallo, R., Nassa, G., Giurato, G., Santamaria, G., Cordella, A., Cantarella, C., and Weisz, A. (2014). RNA sequencing identifies specific PIWI-interacting small non-coding RNA expression patterns in breast cancer. Oncotarget *5*, 9901–9910.

12. Lee, E.J., Banerjee, S., Zhou, H., Jammalamadaka, A., Arcila, M., Manjunath, B.S., and Kosik, K.S. (2011). Identification of piRNAs in the central nervous system. RNA *17*, 1090–1099.

13. Nishibu, T., Hayashida, Y., Tani, S., Kurono, S., Kojima-Kita, K., Ukekawa, R., Kurokawa, T., Kuramochi-Miyagawa, S., Nakano, T., Inoue, K., and Honda, S. (2012). Identification of MIWI-associated Poly(A) RNAs by immunoprecipitation with an anti-MIWI monoclonal antibody. Biosci. Trends *6*, 248–261.

14. Zhang, Y., Wang, X., and Kang, L. (2011). A k-mer scheme to predict piRNAs and characterize locust piRNAs. Bioinformatics *27*, 771–776.

15. Wang, K., Liang, C., Liu, J., Xiao, H., Huang, S., Xu, J., and Li, F. (2014). Prediction of piRNAs using transposon interaction and a support vector machine. BMC Bioinformatics *15*, 419.

16. Luo, L., Li, D., Zhang, W., Tu, S., Zhu, X., and Tian, G. (2016). Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features. PLoS ONE *11*, e0153268.

17. Li, D., Luo, L., Zhang, W., Liu, F., and Luo, F. (2016). A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. BMC Bioinformatics *17*, 329.

18. Gou, L.-T., Dai, P., Yang, J.-H., Xue, Y., Hu, Y.P., Zhou, Y., Kang, J.Y., Wang, X., Li, H., Hua, M.M., et al. (2014). Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res. *24*, 680–700.

19. Chou, K.-C., Jiang, S.-P., Liu, W.-M., and Fee, C.-H. (1979). Graph theory of enzyme kinetics: 1. Steady-state reaction system. Sci. Sin. *22*, 341–358.

20. Chou, K.C., and Forsén, S. (1980). Graphical rules for enzyme-catalysed rate laws. Biochem. J. *187*, 829–835.

21. Zhou, G.P., and Deng, M.H. (1984). An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem. J. *222*, 169–176.

22. Chou, K.C. (1989). Graphic rules in steady and non-steady state enzyme kinetics. J. Biol. Chem. *264*, 12074–12079.

23. Althaus, I.W., Gonzales, A.J., Chou, J.J., Romero, D.L., Deibel, M.R., Chou, K.C., Kezdy, F.J., Resnick, L., Busso, M.E., So, A.G., et al. (1993). The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J. Biol. Chem. *268*, 14875–14880.

24. Althaus, I.W., Chou, J.J., Gonzales, A.J., Deibel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Palmer, J.R., Thomas, R.C., Aristoff, P.A., et al. (1993). Kinetic studies

with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry *32*, 6548–6554.

25. Wu, Z.C., Xiao, X., and Chou, K.C. (2010). 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J. Theor. Biol. *267*, 29–34.
26. Chou, K.-C., Lin, W.Z., and Xiao, X. (2011). Wenxiang: a web-server for drawing wenxiang diagrams. Nat. Sci. *3*, 862–865.
27. Zhou, G.P. (2011). The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. J. Theor. Biol. *284*, 142–148.
28. Zhou, G.P., Chen, D., Liao, S., and Huang, R.B. (2016). Recent progresses in studying helix-helix interactions in proteins by incorporating the Wenxiang diagram into the NMR spectroscopy. Curr. Top. Med. Chem. *16*, 581–590.
29. Fawcett, J.A. (2006). An introduction to ROC analysis. Pattern Recognit. Lett. *27*, 861–874.
30. Davis, J., and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning (ACM), pp. 233–240.
31. Chou, K.C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. *273*, 236–247.
32. Chou, K.C., and Shen, H.B. (2007). Recent progress in protein subcellular location prediction. Anal. Biochem. *370*, 1–16.
33. Zhang, P., Si, X., Skogerbø, G., Wang, J., Cui, D., Li, Y., Sun, X., Liu, L., Sun, B., Chen, R., et al. (2014). piRBase: a web resource assisting piRNA functional study. Database (Oxford) *2014*, bau110.
34. Bu, D., Yu, K., Sun, S., Xie, C., Skogerbø, G., Miao, R., Xiao, H., Liao, Q., Luo, H., Zhao, G., et al. (2012). NONCODE v3. 0: integrative annotation of long noncoding RNAs. Nucleic Acids Res. *40*, D210–D215.
35. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. Molecules *21*, E95.
36. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal. Biochem. *497*, 48–56.
37. Liu, Z., Xiao, X., Qiu, W.R., and Chou, K.C. (2015). iDNA-methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal. Biochem. *474*, 69–77.
38. Xiao, X., Min, J.L., Lin, W.Z., Liu, Z., Cheng, X., and Chou, K.C. (2015). iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach. J. Biomol. Struct. Dyn. *33*, 2221–2233.
39. Chou, K.C. (2015). Impacts of bioinformatics to medicinal chemistry. Med. Chem. *11*, 218–234.
40. Chou, K.C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins *43*, 246–255.
41. Chou, K.C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics *21*, 10–19.
42. Shen, H.B., and Chou, K.C. (2008). PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. *373*, 386–388.
43. Du, P., Wang, X., Xu, C., and Gao, Y. (2012). PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal. Biochem. *425*, 117–119.
44. Cao, D.S., Xu, Q.S., and Liang, Y.Z. (2013). propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics *29*, 960–962.
45. Du, P., Gu, S., and Jiao, Y. (2014). PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. Int. J. Mol. Sci. *15*, 3495–3506.
46. Lin, S.X., and Lapointe, J. (2013). Theoretical and experimental biology in one—a symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers. J. Biomed. Sci. Eng. *6*, 435–442.
47. Chou, K.C. (2009). Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr. Proteomics *6*, 262–274.
48. Khan, M., Hayat, M., Khan, S.A., and Iqbal, N. (2017). Unb-DPC: identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. J. Theor. Biol. *415*, 13–19.
49. Meher, P.K., Sahu, T.K., Saini, V., and Rao, A.R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci. Rep. *7*, 42362.
50. Chen, W., Lei, T.Y., Jin, D.C., Lin, H., and Chou, K.C. (2014). PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. *456*, 53–60.
51. Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.C. (2015). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics *31*, 119–120.
52. Liu, B., Liu, F., Fang, L., Wang, X., and Chou, K.C. (2015). 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 *31*, 1307–1309.
53. Liu, B., Liu, F., Fang, L., Wang, X., and Chou, K.C. (2016). repRNA: a web server for generating various feature vectors of RNA sequences. Mol. Genet. Genomics *291*, 473–481.
54. Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. *43* (W1), W65–W71.
55. Chen, W., Feng, P.M., Lin, H., and Chou, K.C. (2013). iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. *41*, e68.
56. Qiu, W.R., Xiao, X., and Chou, K.C. (2014). iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int. J. Mol. Sci. *15*, 1746–1766.
57. Chen, W., Feng, P.M., Lin, H., and Chou, K.C. (2014). iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. BioMed Res. Int. *2014*, 623149.
58. Guo, S.H., Deng, E.Z., Xu, L.Q., Ding, H., Lin, H., Chen, W., and Chou, K.C. (2014). iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics *30*, 1522–1529.
59. Lin, H., Deng, E.Z., Ding, H., Chen, W., and Chou, K.C. (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. *42*, 12961–12972.
60. Chen, W., Feng, P.M., Deng, E.Z., Lin, H., and Chou, K.C. (2014). iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal. Biochem. *462*, 76–83.
61. Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.C. (2015). iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. *490*, 26–33.
62. Liu, B., Fang, L., Liu, F., Wang, X., and Chou, K.C. (2016). iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J. Biomol. Struct. Dyn. *34*, 223–235.
63. Xiao, X., Ye, H.X., Liu, Z., Jia, J.H., and Chou, K.C. (2016). iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. Oncotarget *7*, 34180–34189.
64. Kabir, M., and Hayat, M. (2016). iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. Mol. Genet. Genomics *291*, 285–296.
65. Tahir, M., and Hayat, M. (2016). iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC. Mol. Biosyst. *12*, 2587–2593.

66. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.C. (2017). iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget *8*, 4208–4217.

67. Zhang, C.J., Tang, H., Li, W.C., Lin, H., Chen, W., and Chou, K.C. (2016). iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. Oncotarget *7*, 69783–69793.

68. Liu, B., Wang, S., Long, R., and Chou, K.C. (2017). iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics *33*, 35–41.

69. Chen, W., Lin, H., and Chou, K.C. (2015). Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol. Biosyst. *11*, 2620–2634.

70. Feng, P.M., Chen, W., Lin, H., and Chou, K.C. (2013). iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal. Biochem. *442*, 118–125.

71. Han, G.S., Yu, Z.G., and Anh, V. (2014). A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC. J. Theor. Biol. *344*, 31–39.

72. Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., Dong, Q., and Chou, K.C. (2014). Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics *30*, 472–479.

73. Kumar, R., Srivastava, A., Kumari, B., and Kumar, M. (2015). Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. J. Theor. Biol. *365*, 96–103.

74. Qiu, W.R., Xiao, X., Lin, W.Z., and Chou, K.C. (2015). iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. J. Biomol. Struct. Dyn. *33*, 1731–1742.

75. Liu, B., Fang, L., Wang, S., Wang, X., Li, H., and Chou, K.C. (2015). Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J. Theor. Biol. *385*, 153–159.

76. Liu, B., Fang, L., Long, R., Lan, X., and Chou, K.C. (2016). iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics *32*, 362–369.

77. Rahimi, M., Bakhtiarizadeh, M.R., and Mohammadi-Sangcheshmeh, A. (2017). OOgenesis_Pred: a sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition. J. Theor. Biol. *414*, 128–136.

78. Chen, J., Long, R., Wang, X.L., Liu, B., and Chou, K.C. (2016). dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. Sci. Rep. *6*, 32333.

79. Chou, K.C., and Cai, Y.D. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. *277*, 45765–45769.

80. Cai, Y.D., Zhou, G.P., and Chou, K.C. (2003). Support vector machines for predicting membrane protein types by using functional domain composition. Biophys. J. *84*, 3257–3263.

81. Cristianini, N., and Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods (Cambridge University Press).

82. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

83. Chang, C.C., and Lin, C.J. (2011). LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. *2*, 1–27.

84. Chou, K.C., and Shen, H.B. (2007). MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem. Biophys. Res. Commun. *360*, 339–345.

85. Chou, K.C., and Shen, H.B. (2008). ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem. Biophys. Res. Commun. *376*, 321–325.

86. Chou, K.C., and Shen, H.B. (2007). Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem. Biophys. Res. Commun. *357*, 633–640.

87. Wang, P., Xiao, X., and Chou, K.C. (2011). NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. PLoS ONE *6*, e23505.

88. Xiao, X., Wang, P., and Chou, K.C. (2011). GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Mol. Biosyst. *7*, 911–919.

89. Xiao, X., Wang, P., and Chou, K.C. (2011). Quat-2L: a web-server for predicting protein quaternary structural attributes. Mol. Divers. *15*, 149–155.

90. Xiao, X., Wang, P., Lin, W.Z., Jia, J.H., and Chou, K.C. (2013). iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal. Biochem. *436*, 168–177.

91. Shen, H.B., and Chou, K.C. (2009). QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. J. Proteome Res. *8*, 1577–1584.

92. Shen, H.B., and Chou, K.C. (2007). Signal-3L: a 3-layer approach for predicting signal peptides. Biochem. Biophys. Res. Commun. *363*, 297–303.

93. Shen, H.B., and Chou, K.C. (2009). Identification of proteases and their types. Anal. Biochem. *385*, 153–160.

94. Shen, H.B., and Chou, K.C. (2007). Using ensemble classifier to identify membrane protein types. Amino Acids *32*, 483–488.

95. Chou, K.C., and Shen, H.B. (2006). Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. *347*, 150–157.

96. Shen, H.B., and Chou, K.C. (2007). Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng. Des. Sel. *20*, 39–46.

97. Shen, H.B., and Chou, K.C. (2006). Ensemble classifier for protein fold pattern recognition. Bioinformatics *22*, 1717–1722.

98. Shen, H.B., and Chou, K.C. (2007). EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem. Biophys. Res. Commun. *364*, 53–59.

99. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2015). iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J. Theor. Biol. *377*, 47–56.

100. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. J. Biomol. Struct. Dyn. *34*, 1946–1961.

101. Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972–976.

102. Chou, K.C., and Shen, H.B. (2006). Large-scale predictions of gram-negative bacterial protein subcellular locations. J. Proteome Res. *5*, 3420–3428.

103. Chou, K.C., and Shen, H.B. (2007). Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome Res. *6*, 1728–1734.

104. Chou, K.C., and Shen, H.B. (2007). Large-scale plant protein subcellular location prediction. J. Cell. Biochem. *100*, 665–678.

105. Chou, K.C., and Shen, H.B. (2010). A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS ONE *5*, e9931.

106. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, D., and Chou, K.C. (2016). iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. Mol. Inform., Published online May 12, 2006. http://dx.doi.org/10.1002/minf.201600010.

107. Qiu, W.R., Xiao, X., Xu, Z.C., and Chou, K.C. (2016). iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget *7*, 51270–51283.

108. Chen, J., Liu, H., Yang, J., and Chou, K.C. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids *33*, 423–428.

109. Chou, K.C. (2001). Using subsite coupling to predict signal peptides. Protein Eng. *14*, 75–79.

110. Xu, Y., Ding, J., Wu, L.Y., and Chou, K.C. (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.

111. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J. Theor. Biol. *394*, 223–230.

112. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget *7*, 34558–34570.

113. Jia, J., Zhang, L., Liu, Z., Xiao, X., and Chou, K.C. (2016). pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics *32*, 3133–3141.

114. Liu, B., Fang, L., Liu, F., Wang, X., Chen, J., and Chou, K.C. (2015). Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE *10*, e0121501.

115. Liu, Z., Xiao, X., Yu, D.J., Jia, J., Qiu, W.R., and Chou, K.C. (2016). pRNAm-PC: predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. Anal. Biochem. *497*, 60–67.

116. Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.C. (2016). Using deformation energy to analyze nucleosome positioning in genomes. Genomics *107*, 69–75.

117. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., and Chou, K.C. (2016). iHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarget *7*, 44310–44321.

118. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., and Chou, K.C. (2016). iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics *32*, 3116–3123.

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

120. Xu, Y., and Chou, K.C. (2016). Recent progress in predicting posttranslational modification sites in proteins. Curr. Top. Med. Chem. *16*, 591–603.

121. Xu, Y., Wen, X., Shao, X.J., Deng, N.Y., and Chou, K.C. (2014). iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int. J. Mol. Sci. *15*, 7594–7610.

122. Xu, Y., Wen, X., Wen, L.S., Wu, L.Y., Deng, N.Y., and Chou, K.C. (2014). iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE *9*, e105018.

123. Ding, H., Deng, E.Z., Yuan, L.F., Liu, L., Lin, H., Chen, W., and Chou, K.C. (2014). iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Res. Int. *2014*, 286419.

124. Xiao, X., Wu, Z.C., and Chou, K.C. (2011). iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J. Theor. Biol. *284*, 42–51.

125. Chou, K.C., Wu, Z.C., and Xiao, X. (2012). iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. *8*, 629–641.

126. Lin, W.Z., Fang, J.A., Xiao, X., and Chou, K.C. (2013). iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol. Biosyst. *9*, 634–644.

127. Cheng, X., Zhao, S.G., Xiao, X., and Chou, K.C. (2016). iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics *33*, 341–346.

128. Chou, K.C. (2013). Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. *9*, 1092–1100.

129. Chou, K.C., and Zhang, C.T. (1995). Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. *30*, 275–349.

130. Ali, F., and Hayat, M. (2015). Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition. J. Theor. Biol. *384*, 78–83.

131. Khan, Z.U., Hayat, M., and Khan, M.A. (2015). Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J. Theor. Biol. *365*, 197–203.

132. Mondal, S., and Pai, P.P. (2014). Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J. Theor. Biol. *356*, 30–35.

133. Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., and Sattar, A. (2015). Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′s general PseAAC. J. Theor. Biol. *364*, 284–294.

134. Ahmad, K., Waris, M., and Hayat, M. (2016). Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition. J. Membr. Biol. *249*, 293–304.

135. Ju, Z., Cao, J.Z., and Gu, H. (2016). Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. J. Theor. Biol. *397*, 145–150.

136. Behbahani, M., Mohabatkar, H., and Nosrati, M. (2016). Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. J. Theor. Biol. *411*, 1–5.

137. Chou, K.C., and Shen, H.B. (2006). Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. J. Proteome Res. *5*, 1888–1897.

138. Shen, H.B. (2009). Review: recent advances in developing web-servers for predicting protein attributes. Nat. Sci. *1*, 63–92.

139. Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K.C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget *7*, 16895–16909.

140. Chen, W., Tang, H., Ye, J., Lin, H., and Choi, K.-C. (2016). iRNA-PseU: identifying RNA pseudouridine sites. Mol. Ther. Nucleic Acids *5*, e332.

141. Liu, B., Wu, H., Zhang, D., Wang, X., and Chou, K.C. (2017). Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods. Oncotarget *8*, 4208–4217.

142. Pérez, A., Noy, A., Lankas, F., Luque, F.J., and Orozco, M. (2004). The relative flexibility of B-DNA and A-RNA duplexes: database analysis. Nucleic Acids Res. *32*, 6144–6151.