Citation: Molecular Therapy—Nucleic Acids (2016) 5, e332; doi:10.1038/mtna.2016.37
Official journal of the American Society of Gene & Cell Therapy
www.nature.com/mtna

iRNA-PseU: Identifying RNA pseudouridine sites

Wei Chen[1,2], Hua Tang[3], Jing Ye[1], Hao Lin[2,4] and Kuo-Chen Chou[2,5]

As the most abundant RNA modification, pseudouridine plays important roles in many biological processes. Occurring at the uridine site and catalyzed by pseudouridine synthase, the modification has been observed in nearly all kinds of RNA, including transfer RNA, messenger RNA, small nuclear or nucleolar RNA, and ribosomal RNA. Accordingly, its importance to basic research and drug development is self-evident. Despite some experimental technologies have been developed to detect the pseudouridine sites, they are both time-consuming and expensive. Facing the explosive growth of RNA sequences in the postgenomic age, we are challenged to address the problem by computational approaches: For an uncharacterized RNA sequence, can we predict which of its uridine sites can be modified as pseudouridine and which ones cannot? Here a predictor called "iRNA-PseU" was proposed by incorporating the chemical properties of nucleotides and their occurrence frequency density distributions into the general form of pseudo nucleotide composition (PseKNC). It has been demonstrated via the rigorous jackknife test, independent dataset test, and practical genome-wide analysis that the proposed predictor remarkably outperforms its counterpart. For the convenience of most experimental scientists, the web-server for iRNA-PseU was established at *http://lin.uestc.edu.cn/server/iRNA-PseU*, by which users can easily get their desired results without the need to go through the mathematical details.

Molecular Therapy—Nucleic Acids (2016) **5**, e332; doi:10.1038/mtna.2016.37; published online 5 July 2016
Subject Category: Bioinformatics

Introduction

Pseudouridine (5-ribosyluracil, abbreviated by the Greek letter Ψ) is the most prevalent RNA (ribonucleic acid) modification and has been found in virtually all kingdoms of life.[1] Recent findings have demonstrated that Ψ is present in various categories of RNAs, such as tRNA (transfer RNA), mRNA (messenger RNA), snRNA (small nuclear RNA), snoRNA (small nucleolar RNA), and rRNA (ribosomal RNA).[2] As shown in **Figure 1**, Ψ is the isomer of uridine and is catalyzed by highly conserved pseudouridine synthase that detaches the uridine residue's base from its sugar, followed by "rotating" it 180° along the N3-C6 axis, and subsequently by reattachment of the base's 5-carbon to the 1'-carbon of the sugar.[3]

The molecular functions of Ψ modification have just been revealed in recent years. For example, Ψ modification plays an integral part in the stabilization of tRNA structure,[2–4] and it also has a prominent role in spliceosomal RNA responsible for gene regulation. The Ψ modification is present in the regions involved with RNA-RNA or RNA-protein interactions to promote the assembly and reaction of a spliceosome to yield viable mRNA such as in AU/AC intron splicing.[2,3,5] Moreover, it has been demonstrated that incorporation of Ψ into mRNA may increase the translation efficiency and reduce the RNA-elicited innate immune responses.[6] Despite great progresses have been made in uncovering the roles of Ψ modification, its biological functions and action mechanisms remain elusive for most RNA systems. Therefore, the information of the Ψ modification sites during transcriptome is crucial for in-depth revealing the biological principle concerned.

By using high-throughput techniques such as Ψ-Seq,[7] the distribution of Ψ modification has been characterized for the transcriptome in *H. sapiens*, *M. musculus*, and *S. cerevisiae*.[7–10] But these techniques are time-consuming and costly for genome-wide analysis. Facing the rapidly increasing number of sequenced genomes, it is highly desired to develop computational methods for timely acquiring this kind of information.

Actually, an effort has been made by Li *et al.*[11] recently in this regard. These authors proposed a predictor called PPUS for identifying PUS-specific pseudouridine sites. The PPUS predictor,[11] however, is only able to identify Ψ modification sites in *H. sapiens* and *S. cerevisiae*. Besides, its accuracy definitely needs to be improved, which can be realized by incorporating the nucleotide chemical property into consideration.

The present study was initiated in an attempt to develop a new and more powerful predictor for identifying the Ψ modification sites with higher success rates and being able to cover more species.

In order to develop a predictor with crystal-clear logic and widely useful value, let us follow the five-step guidelines[12] as done by a series of recent publications (see, *e.g.*, refs. 13–21): (i) how to construct or select a valid benchmark dataset to train and test the predictor; (ii) how to formulate the biological sequence samples with a valid mathematical

[1]Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, China; [2]Gordon Life Science Institute, Boston, Massachusetts, USA; [3]Department of Pathophysiology, Southwest Medical University, Luzhou, China; [4]Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China; [5]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia. Correspondence: Wei Chen, Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China. E-mail: chenweiimu@gmail.com or Hao Lin, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054 China. E-mail: hlin@uestc.edu.cn or Kuo-Chen Chou, Gordon Life Science Institute, Boston, Massachusetts 02478, USA. E-mail: kcchou@gordonlifescience.org
Keywords: Ψ site; iRNA-PseU; nucleotide chemical property; nucleotide frequency; pseudouridine; Web-server
Received 4 March 2016; accepted 20 April 2016; published online 5 July 2016. doi:10.1038/mtna.2016.37

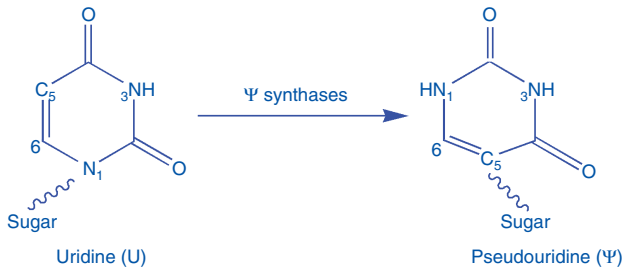

Figure 1 Illustration to show the pseudouridine (Ψ) modification. Its formation is catalyzed by the Ψ synthase.

expression that can truly reflect their essential correlation with the target to be predicted; (iii) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) how to appropriately conduct cross-validation to objectively estimate its anticipated accuracy; (v) how to establish a user-friendly web-server that is accessible to the public. Below, we are to address the aforementioned five steps one-by-one.

Results

As mentioned in Introduction, among the five important steps for developing a useful predictor, one of them is how to objectively evaluate its anticipated success rates.[12] To address this, the following two considerations are needed: one is what metrics should be adopted to reflect the predictor's success rates; the other is what test method should be used to derive the metrics rates. Below, we are to address the two problems.

Metrics for quantitatively measuring the predictor's quality

The following four metrics are usually used to measure the quality of a predictor: (i) overall accuracy or Acc; (ii) Mathew's correlation coefficient or MCC; (iii) sensitivity or Sn; and (iv) specificity or Sp.[22] Unfortunately, the conventional formulations for the four metrics are not intuitive, and most experimental scientists feel difficult to understand them, particularly for the MCC metrics. It is interesting, however, that if using the Chou's symbols and derivation for studying the signal peptides,[23] the above four metrics can be formulated as follows[13,24]:

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

where N^+ represents the total number of true Ψ-site-containing RNA samples investigated, whereas N_-^+ the number of true Ψ-site-containing RNA samples incorrectly predicted to be of false Ψ-site-containing RNA sample; N^- the total number of false Ψ-site-containing RNA samples, whereas N_+^- the number of false Ψ-site-containing RNA samples incorrectly predicted to be of true Ψ-site-containing RNA sample.

According to Equation 1, it is crystal clear to see the following. When $N_-^+ = 0$ meaning none of the true Ψ-site-containing RNA samples is incorrectly predicted to be of false one, we have the sensitivity Sn = 1. When $N_-^+ = N^+$ meaning that all the true Ψ-site-containing RNA samples are incorrectly predicted to be of false one, we have the sensitivity Sn = 0. Likewise, when $N_+^- = 0$ meaning none of the false Ψ-site-containing RNA samples is incorrectly predicted to be of true one, we have the specificity Sp = 1; whereas $N_+^- = N^-$ meaning that all the false Ψ-site-containing RNA samples are incorrectly predicted to be of true one, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that none of the true Ψ-site-containing RNA samples in the positive dataset and none of the false Ψ-site-containing RNA samples in the negative dataset was incorrectly predicted, we have the overall accuracy Acc = 1 and MCC = 1; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the true Ψ-site-containing RNA samples in the positive dataset and all the false Ψ-site-containing RNA samples in the negative dataset were incorrectly predicted, we have the overall accuracy Acc = 0 and MCC = −1; whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have Acc = 0.5 and MCC = 0 meaning no better than random guessing. As we can see from the above discussion, the formulation of Equation 1 has made the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient much more intuitive and easier-to-understand, particularly for the meaning of MCC, as concurred and adopted by many investigators in a series of recent publications (see, e.g., refs. 14,17,25–30). Note that, of the four metrics in Equation 1, the most important are the Acc and MCC since the former reflects the overall accuracy of a predictor while the latte its stability. The metrics Sn and Sp are used to measure a predictor from two different angles, and they are constrained with each other.[31]

It should be pointed out, however, the set of equations defined in Equation 1 is valid only for the single-label systems. For the multi-label systems whose emergence has become more frequent in system biology[32–34] and system medicine,[35] a completely different set of metrics is needed as elucidated in ref. 36.

Validation by Jackknife tests

With a good set of evaluation metrics defined, the next thing is what validation method should be used to derive the metrics values.

In statistical prediction, the following three cross-validation methods are often used to derive the metrics values for a predictor: independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test.[37] Of these three, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset as elucidated in ref. 12 and demonstrated by Equations 28–32 therein. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g., refs. 38–46). Accordingly, the jackknife test was also used to examine the performance of the model proposed in the current study.

During the jackknife test, each RNA sample in the benchmark dataset was in turn singled out as an independent test sample and all the rule-parameters were calculated without including the one being identified.

The result obtained by the jackknife test on the benchmark datasets $\mathbb{S}(1)$ for *H. sapiens* (see Equation 1 as well as **Supplementary Information S1**) are given by

$$\begin{cases} S_n = 61.01\% \\ S_p = 59.80\% \\ Acc = 60.40\% \\ MCC = 0.21 \end{cases} \quad (2)$$

that on $\mathbb{S}(2)$ for *S. cerevisiae* (see **Supplementary Information S2**) by

$$\begin{cases} S_n = 64.65\% \\ S_p = 64.33\% \\ Acc = 64.49\% \\ MCC = 0.29 \end{cases} \quad (3)$$

And that on $\mathbb{S}(3)$ for *M. musculus* (see **Supplementary Information S3**) given by

$$\begin{cases} S_n = 73.31\% \\ S_p = 64.83\% \\ Acc = 69.07\% \\ MCC = 0.38 \end{cases} \quad (4)$$

Discussion

Comparison with the existing predictor

To our best knowledge, **PPUS**[11] is so far the only existing predictor available for identifying the Ψ sites in RNA sequences. It should be pointed out that the results given in Equation 4 are beyond the reach of **PPUS**[11] because it can be used to identify the Ψ sites in the RNA sequences from *H. sapiens* and *S. cerevisiae* species but not from *M. musculus*.

For the cases of *H. sapiens* and *S. cerevisiae* species, however, it is also hard to give the corresponding jackknife results without the program code of **PPUS**. Fortunately, like the **iRNA-PseU** predictor, **PPUS** also has a web-server predictor, which will make it possible to compare the two predictors via their performances on a same independent dataset.

To realize this, we constructed two independent datasets $\mathbb{S}(4)$ and $\mathbb{S}(5)$ for *H. sapiens* and *S. cerevisiae*, respectively.

Table 1 A comparison of the new predictor with the existing predictor when performed on the independent dataset of *H. sapiens* (**Supplementary Information S4**) and that of *S. cerevisiae* (**Supplementary Information S5**), respectively

Species	Predictor	Acc (%)[c]	MCC[c]	Sn (%)[c]	Sp (%)[c]
H. sapiens	PPUS[a]	52.50	0.13	6.0	99.00
	iRNA-PseU[b]	65.00	0.30	60.00	70.00
S. cerevisiae	PPUS[a]	71.00	0.44	56.00	86.00
	iRNA-PseU[b]	73.00	0.46	81.00	65.00

[a]The predictor developed by Li *et al.*,[11] which is available at http://lyh.pkmu.cn/ppus/. [b]The predictor proposed in this paper. [c]See Equation 1 for the definition of metrics.

None of the samples in $\mathbb{S}(4)$ occurs in the benchmark dataset $\mathbb{S}(1)$; none of the samples in $\mathbb{S}(5)$ occurs in the benchmark dataset $\mathbb{S}(2)$. For the detailed sequences in the two independent datasets, see **Supplementary Information S4** and **Supplementary Information S5**, respectively.

Listed in **Table 1** are the results obtained by using the web-server of **PPUS**[11] and that of **iRNA-PseU** on the two independent datasets for the species of *H. sapiens* and *S. cerevisiae*, respectively. From the table we can see the following. (i) The rates of both Acc and MCC achieved by **iRNA-PseU** are remarkably higher than those by **PPUS**, indicating that the proposed predictor is not only more accurate but also more stable in comparison with its counterpart. (ii) The gap between Sn and Sp yielded by **PPUS**[11] is much larger than that by **iRNA-PseU**. This kind of extremely skewed profile generated by **PPUS** implies its predicted results contain many false positive or negative as well as a lot of noise. As mentioned in the section "Metrics for quantitatively measuring the predictor's quality", Sn and Sp are mutually restrained.[31] Accordingly, it is meaningless to use only one of the two for comparison. A meaningful comparison should be based on the result of their combination, which is none but MCC.

To further demonstrate its power in practical application, the genome-wide analysis by **iRNA-PseU** was performed on the chromosome XII of the *S.cerevisiae* genome. The results thus obtained on such an independent RNA sequence are given in **Figure 2**, where for facilitating comparison the corresponding experimental results[7] obtained by the Pseudo-Seq technique are also shown. As can be seen from the figure, of the six known Ψ sites, five were correctly identified by **iRNA-PseU**, demonstrating once again that the **iRNA-PseU** is indeed quite promising for Ψ site identification.

Graphical analysis

Why could the proposed method be so successful? It is not easy to give a simple answer to address this problem. Fortunately,

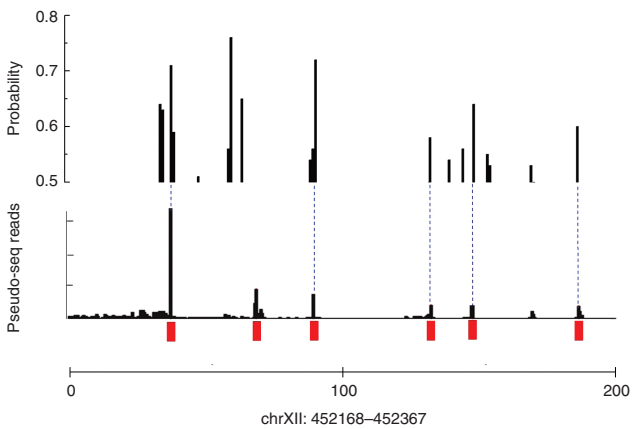

Figure 2 A comparison between predicted results of **iRNA-PseU** and experimental results on a 200-nt (from 452168 to 452367) genomic region of chromosome XII from *S. cerevisiae*. The top panel shows the probability values calculated by **iRNA-PseU**. The middle panel shows the experimental results determined by using the Pseudo-Seq technique, where the six known Ψ sites are highlighted with red rectangles.[7] The dashed blue line shows the consistency between the predicted result and the experimental one. The lower panel shows the relative genomic coordinate.

many biological systems and the complicated relations therein could be revealed via the intuitive graphical approaches, such as in studying enzyme-catalyzed reactions,[47–49] protein folding kinetics and folding rates,[50] inhibition of HIV-1 reverse transcriptase,[51,52] drug metabolism systems,[53] analyzing large-scale biological sequences,[54] and recently using wenxiang diagrams or graphs[55] to analyze protein-protein interactions.[56]

To provide an intuitive graph about the performance of the newly proposed method, the receiver operating characteristic (ROC)[57,58] was utilized. In the ROC graph, the vertical coordinate is for the true positive rate (sensitivity) while the horizontal coordinate for the false positive rate (1-specificity). The best possible prediction method would yield a point with the coordinate (0, 1) representing 100% sensitivity with 0 false positive rate or 100% specificity.[57,58] 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, also called AUROC, is often used to indicate the performance quality of a binary classifier: the value 0.5 of AUROC is equivalent to random prediction, while 1 of AUROC represents a perfect one. Accordingly, in order to objectively evaluate the overall performance of **iRNA-PseU** for identifying Ψ sites, we plotted the ROC curves and reported the AUROCs in **Figure 3**. As shown from the figure, the AUROC scores for **iRNA-PseU** in identifying Ψ sites are 0.64, 0.75, and 0.81 for *H. sapiens*, *M. musculus*, and *S. cerevisiae* genomes, respectively.

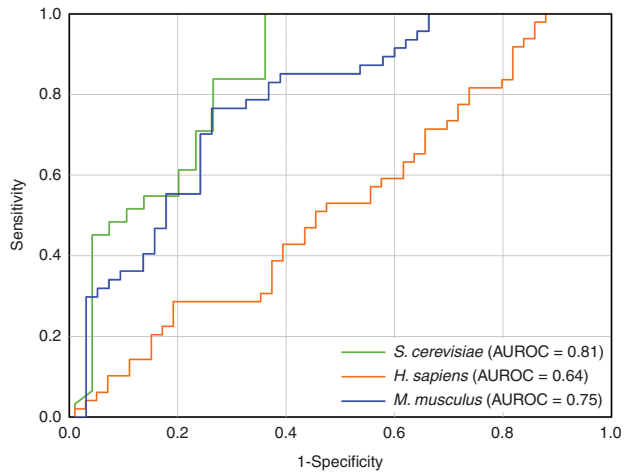

Figure 3 A graphical illustration to show the performance of **iRNA-PseU** by means of the receiver operating characteristic curve.

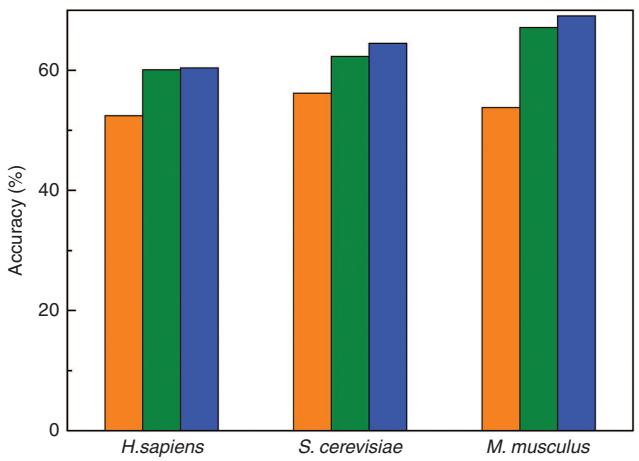

Figure 4 An in-depth analysis into the contributions of three models: the orange histogram stands for the accuracy score obtained by the model trained based on the nucleotide density in identifying Ψ sites; the green one for that based on the nucleotide chemical properties; and the blue for that by combing the above two models. See the text for more explanation.

Furthermore, for in-depth analyzing the contributions from different features to the Ψ site identification, we had built two models: one was based on nucleotide chemical property and the other based on the nucleotide density. The validated results are shown in **Figure 4**, where the orange, green and blue histograms denote the accuracy scores for the models trained based on nucleotide density, nucleotide chemical properties and their combinations, respectively. As shown from the figure, the nucleotide chemical property (green) had greater contribution than the nucleotide density (orange) for Ψ site identification, but the latter did play the complementary role in the prediction, as reflected by the blue histogram that is higher than both the blue and orange ones. Since pseudouridine is catalyzed by Ψ synthases that need to recognize and bind with specific genomic regions, the above findings suggest that nucleotide chemical properties may closely correlate with the interactions between the synthases and RNA sequence.

Conclusion

It is anticipated that the proposed predictor will become a very useful high throughput tool for identifying the Ψ sites in genome analysis, or at the very least, play a complementary role to the existing PPUS predictor[11] for genome analysis.

Materials and methods

Benchmark dataset. For facilitating description later, we use the following scheme to represent a RNA sample

$$\mathbf{R}_{\xi}(\mathbb{U}) = \mathrm{N}_{-\xi}\mathrm{N}_{-(\xi-1)}\cdots\mathrm{N}_{-2}\mathrm{N}_{-1}\mathbb{U}\mathrm{N}_{+1}\mathrm{N}_{+2}\cdots\mathrm{N}_{+(\xi-1)}\mathrm{N}_{+\xi} \quad (5)$$

where the center $\mathbb{U}$ represents "uridine", the subscript ξ is an integer, $\mathrm{N}_{-\xi}$ represents the ξ-th upstream nucleotide from the center, the $\mathrm{N}_{+\xi}$ the ξ-th downstream nucleotide, and so forth. The $(2\xi + 1)$-tuple RNA sample $\mathbf{R}_{\xi}(\mathbb{U})$ can be further classified into the following two categories:

$$\mathbf{R}_{\xi}(\mathbb{U}) \in \begin{cases} \mathbf{R}^{+}_{\xi}(\mathbb{U}), \text{ if its center is a } \psi \text{ site} \\ \mathbf{R}^{-}_{\xi}(\mathbb{U}), \text{ otherwise} \end{cases} \quad (6)$$

where $\mathbf{R}^{+}_{\xi}(\mathbb{U})$ denotes a RNA sample whose center uridine can be converted to pseudouridine via Ψ modification as confirmed by experiments, $\mathbf{R}^{-}_{\xi}(\mathbb{U})$ a RNA sample whose center uridine cannot be so, and the symbol ϵ means "a member of" in the set theory.

In literature the benchmark dataset usually consists of a training dataset and an independent testing dataset: the former is used to train a model, while the latter used to test the model. But as pointed out in a comprehensive review,[59] there is no need at all to artificially separate a benchmark

dataset into the two parts if the model is evaluated by the jackknife test or subsampling (K-fold) cross-validation since the outcome thus obtained is actually from a combination of many different independent dataset tests. Thus, the benchmark dataset set S for the current study can be formulated as

$$\mathbb{S}_\xi = \mathbb{S}_\xi^+ \cup \mathbb{S}_\xi^- \tag{7}$$

where the positive subset $\mathbb{S}_\xi^+$ only contains the RNA samples of true Ψ site; the negative subset $\mathbb{S}_\xi^-$ only contains the RNA samples of false Ψ site; and U represents the symbol for "union" in the set theory.

Because the length of RNA sample $\mathbf{R}_\xi(\mathbb{U})$ is $2\xi + 1$ (see Equation 5), the benchmark dataset with different ξ value will contain RNA segments with different number of nucleotides, as illustrated below

$$\text{The length of RNA samples in } \mathbb{S}_\xi = \begin{cases} 17 \text{ nucleotides}, & \text{if } \xi = 8 \\ 21 \text{ nucleotides}, & \text{if } \xi = 10 \\ 26 \text{ nucleotides}, & \text{if } \xi = 13 \\ 31 \text{ nucleotides}, & \text{if } \xi = 15 \\ 41 \text{ nucleotides}, & \text{if } \xi = 20 \\ \vdots & \vdots \end{cases} \tag{8}$$

The RNA sequences with experimentally validated Ψ sites of *H. sapiens*, *M. musculus* and *S. cerevisiae* were downloaded from RMBase.[60] The detailed procedures of constructing the benchmark dataset for each of the three species are as follows: (i) As done in ref. 61, slide the $(2\xi + 1)$-tuple nucleotide window along each of the RNA sequences concerned (**Figure 5**), and collected were only those RNA segments that have uridine (U) at the center (see Equation 5). (ii) If the upstream or downstream in an RNA was less than ξ or greater than L–ξ (L is the RNA's length), the lacking nucleotide was filled with its mirror image (**Figure 6**). (iii) The RNA samples thus obtained were deemed as the positive ones if their centers have been experimentally confirmed as the Ψ sites; otherwise, the negative. (iv) Using the CD-HIT software,[62] the aforementioned samples were further subject to a screening procedure to winnow those that had ≥60% pairwise sequence identity to any other in a same class

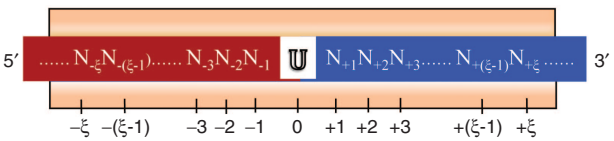

Figure 5 Schematic drawing to show how to use the flexible scaled window along an RNA sequence to collect the potential Ψ-site-containing sequence samples.

a
Mirror image for 5′ terminus
$N_{-1}N_{-2} \cdots N_{-\xi+1}N_{-\xi} \Leftrightarrow N_{-\xi}N_{-\xi+1} \cdots N_{-2}N_{-1}$

b
Mirror image for 3′ terminus
$N_{L-\xi}N_{L-\xi+1} \cdots N_{L-1}N_L \Leftrightarrow N_LN_{L-1} \cdots N_{L-\xi+1}N_{L-\xi}$

Figure 6 Schematic illustration to show the mirror image of (**a**) the 5' RNA terminal segment, and (**b**) the 3' RNA terminal segment. The symbol ⇔ represents a mirror, and the real RNA segment is colored in blue, while its mirror image in red.

because a dataset containing many high similar samples would lack statistical representativeness.[12] (v) The number of negative samples thus obtained would be substantially greater than that of positive ones; to avoid the bias caused by such a skewed dataset,[15] a randomly picking procedure was adopted to make the negative subset have the same size as the positive subset.[25] (vi) The length of samples collected via the above procedures would depend on the value of ξ, h owever, preliminary tests had indicated that best prediction results were achieved when ξ = 10 for the case of *H. sapiens* or *M. musculus*, whereas ξ = 15 for the case of *S. cerevisiae* (see **Figure 7**). Accordingly, hereafter we shall focus on the RNA samples with 21 nucleotides when analyzing the genome from *H. sapiens* or *M. musculus*, while those with 31 nucleotides when analyzing the genome from *S. cerevisiae*.

After going through the above six procedures, we finally obtained three benchmark datasets, as formulated below

$$\begin{cases} \mathbb{S}(1) = \mathbb{S}^+(1) \cup \mathbb{S}^-(1) \\ \mathbb{S}(2) = \mathbb{S}^+(2) \cup \mathbb{S}^-(2) \\ \mathbb{S}(3) = \mathbb{S}^+(3) \cup \mathbb{S}^-(3) \end{cases} \tag{9}$$

where $\mathbb{S}(1)$, $\mathbb{S}(2)$, and $\mathbb{S}(3)$ and denote the benchmark datasets for *H. sapiens*, *S. cerevisiae*, and *M. musculus*, respectively. The RNA samples in $\mathbb{S}(1)$ and $\mathbb{S}(3)$ are each formed by 21 nucleotides, while those in $\mathbb{S}(2)$ are each formed by 31 nucleotides. The subsets $\mathbb{S}^+(1)$, $\mathbb{S}^+(2)$, and $\mathbb{S}^+(3)$ contain 495, 314, and 472 positive samples, while the subsets $\mathbb{S}^-(1)$, $\mathbb{S}^-(2)$, and $\mathbb{S}^-(3)$ contain 495, 314, and 472 negative samples, respectively.

The detailed sequences for the three benchmark datasets are given in **Supplementary Information S1, Supplementary Information S2**, and **Supplementary Information S3**, respectively.

Representation of RNA sequence samples. With the explosive growth of biological sequences generated in the postgenomic age, one of the most challenging problems in computational

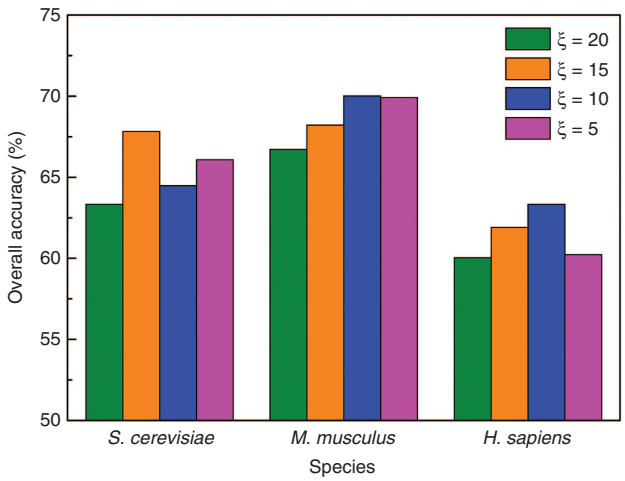

Figure 7 A histogram to show the overall accuracy obtained by the proposed predictor in identifying Ψ site with different ξ values. The accuracy for *H.sapiens* or *M. musculus* reaches a peak when ξ = 10, while that for *S. cerevisiae* reaches a peak when ξ = 15.

biology is how to formulate a biological sequence with a discrete model or vector, yet still considerably keep its key pattern or sequence order information. This is because almost all the existing machine-learning algorithms were developed to handle vector but not sequence samples, as elaborated in a recent review.[63] Unfortunately, a vector defined in a discrete model may completely lose all the sequence-order information or sequence pattern characteristics. To overcome such a problem for protein/peptide and DNA/RNA sequences, the pseudo amino acid composition (PseAAC)[64–69] and pseudo nucleotide composition (PseKNC)[70–73] were introduced, respectively. Ever since they were introduced, PseAAC has been widely used in computational proteomics (see a long list or references cited[12,74]) and PseKNC has been increasingly used in computational genomics.[75] Recently, a web-server called "Pse-in-One" was established for generating various modes of pseudo components for DNA/RNA and protein/peptide sequences.[76]

According to a recent research,[75] the general form of PseKNC for an RNA sequence sample can be formulated as

$$\mathbf{R} = [\phi_1 \quad \phi_2 \quad \cdots \quad \phi_u \quad \cdots \quad \phi_Z]^{\mathbf{T}} \tag{10}$$

where **T** is a transpose operator, while the subscript Z is an integer and its value as well as the components ϕ_u ($u = 1, 2, ..., Z$) will depend on how to extract the desired information from the RNA sequence sample. In order to make Equation 10 able to cover the RNA sample's local site information as well as its global sequence pattern characteristics, below let us use the nucleotide chemical property and nucleotide density to define the components therein.

Nucleotide chemical property. RNA is comprised of four kinds of nucleotides: adenosine (A), guanosine (G), cytidine (C), and uridine (U). Each nucleotide has its own chemical structure and internal binding feature. A and G have two rings, while C and U have only one ring (**Figure 8**). When forming the secondary or tertiary structures, the hydrogen bonding between G and C is stronger than that between A and U (**Figure 8**). Furthermore, according to the chemical functionality, A and C can be classified as amino group, while G and U as keto group. Therefore, the four types of nucleotides can be classified into three different groups as shown in **Table 2**.

In order to incorporate these chemical property features into the representation for a RNA sample, similar to the approach in studying the codon usage in HIV proteins[77] and *E. Coli* proteins,[78] let us formulate the i-th nucleotide in Equation 5 by

$$\mathrm{N}_i = (x_i, y_i, z_i) \tag{11}$$

where[79]

$$x_i = \begin{cases} 1, \text{if } \mathrm{N}_i \in \{\mathrm{A}, \mathrm{G}\} \\ 0, \text{if } \mathrm{N}_i \in \{\mathrm{C}, \mathrm{U}\} \end{cases}; y_i = \begin{cases} 1, \text{if } \mathrm{N}_i \in \{\mathrm{A}, \mathrm{C}\} \\ 0, \text{if } \mathrm{N}_i \in \{\mathrm{G}, \mathrm{U}\} \end{cases}; z_i = \begin{cases} 1, \text{if } \mathrm{N}_i \in \{\mathrm{A}, \mathrm{U}\} \\ 0, \text{if } \mathrm{N}_i \in \{\mathrm{C}, \mathrm{G}\} \end{cases} \tag{12}$$

Thus, according to **Table 2**, the nucleotide A can be formulated as (1, 1, 1), C as (0, 1, 0), G as (1, 0, 0), and U as (0, 0, 1).

Nucleotide density. In order to incorporate the local occurrence frequency of a nucleotide and its distribution in a RNA sequence, let us introduce the following equations

$$d_i = \frac{1}{\| \mathrm{S}_i \|} \sum_{j=1}^{\ell} f(\mathrm{N}_j) \tag{13}$$

where d_i is the density of the nucleotide N_i at position i of a RNA sequence, $\| \mathrm{S}_i \|$ is the length of the sliding substring concerned, ℓ the corresponding locator's sequence position, and

$$\mathrm{f}(\mathrm{N}_j) = \begin{cases} 1, & \text{if } \mathrm{N}_j = \text{the nucleotide concerned} \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

For example, suppose a RNA sequence "AGCGUAAC". The density of "A" is 1 (1/1), 0.33 (2/6), 0.43 (3/7) at positions 1, 6, and 7, respectively. The density of "C" is 0.33 (1/3), 0.25 (2/8) at positions 3 and 8, respectively. The density of "G" is 0.5 (1/2), 0.5 (2/4) at positions 2 and 4, respectively. The density of "U" is 0.2 (1/5) at position 5.

Pseudo nucleotide composition (PseKNC). By integrating both the nucleotide chemical property (Equation 11) and nucleotide frequency information (Equation 13), we have

$$\mathrm{N}_i = (x_i, y_i, z_i, d_i) \tag{15}$$

Thus, the nucleotides in the RNA sequence "AGCGUAAC" can be consecutively denoted by the following eight groups of digits: (1, 1, 1, 1), (1, 0, 0, 0.5), (0, 1, 0, 0.33), (1, 0, 0, 0.5), (0, 0, 1, 0.2), (1, 1, 1, 0.33), (1, 1, 1, 0.43), and (0, 1, 0, 0.25).

Or, according to the formulation of PseKNC (see Equation 10), we have

$$\mathbf{R}(\mathbf{AGCGUAAC}) = [1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0.5 \ \cdots \ 0 \ 1 \ 0 \ 0.25]^{\mathbf{T}} \tag{16}$$

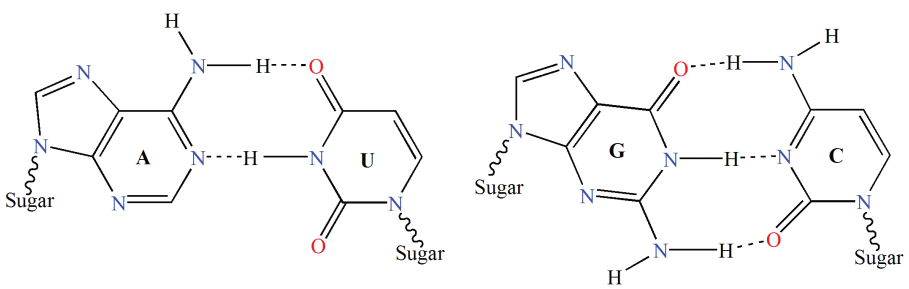

Figure 8 Illustration to show the structure of paired nucleic acid residues. The left panel is the A-U pair bonded to each other with two hydrogen bonds; the right panel is the G-C pair with three hydrogen bonds.

Table 2 Nucleotide chemical property[a]

Chemical property	Class	Nucleotides
Ring structure	Purine	A, G
	Pyrimidine	C, U
Functional group	Amino	A, C
	Keto	G, U
Hydrogen bond	Strong	C, G
	Weak	A, U

[a]See the text in section "Nucleotide chemical property" for further explanation.

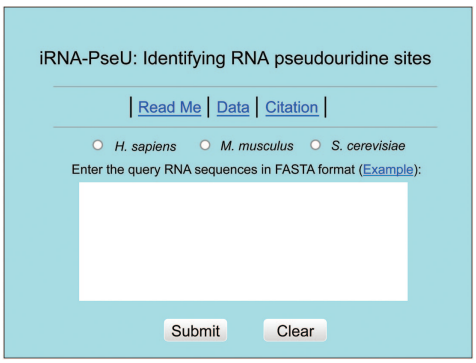

Figure 9 A semi-screenshot for the top-page of the **iRNA-PseU** web-server at *http://lin.uestc.edu.cn/server/ iRNA-PseU*.

meaning that the 8-tuple nucleotide example can be denoted by an 8×4 = 32-D (dimensional) PseKNC vector. Accordingly, a sample in $\mathbb{S}(1)$ or $\mathbb{S}(3)$ can be formulated by a 21×4 = 84-D vector, and that in $\mathbb{S}(2)$ by a 31×4 = 124-D vector (see Equation 9 and the follow-up text).

Support vector machine (SVM). Being a machine learning algorithm based on statistical learning theory, SVM has been widely and successfully used in the realm of bioinformatics[16,80,81] and computational biology.[13–15,26,82] The basic idea of SVM is to transform the input data into a high dimensional feature space and then determine the optimal separating hyperplane.

For a brief formulation of SVM and how it works, see the papers[83,84]; for more details about SVM, see a monograph.[85]

In the current study, the LibSVM package 3.18 was used to implement SVM, which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Because of its effectiveness and speed in training process, the radial basis kernel function (RBF) was used to obtain the best classification hyperplane here. In the SVM operation engine, the regularization parameter C and the kernel width parameter γ were optimized via an optimization procedure using the grid search approach as defined by

$$\begin{cases} 2^{-5} \le C \le 2^{15} \text{ with step of } 2 \\ 2^{-15} \le \gamma \le 2^{-5} \text{ with step of} 2^{-1} \end{cases} \quad (17)$$

The predictor obtained via the above procedures is called **iRNA-PseU**, where "i" stands for "identify", "Pse" for "pseudo", and "U" for "uridine".

Web-server and user guide. As demanded by most experimental scientists, a publicly accessible web-server for **iRNA-PseU** has been established. Moreover, to maximize their convenience, below we are to give a step-by-step guide of the web-server, by which users can easily get their desired results without the need to go through the detailed mathematical equations involved.

Step 1. Open the web server at http://lin.uestc.edu.cn/server/iRNA-PseU and you will see the top page of the **iRNA-PseU** predictor on your computer screen, as shown in **Figure 9**. Click on the *Read Me* button to see a brief introduction about the predictor and the caveat when using it.

Step 2. Select the organism or species by checking on the corresponding open circle. Either type or copy/paste the query RNA sequences into the input box at the center of **Figure 9**. The input sequence should be in FASTA format. For the examples of RNA sequences in FASTA format, click the *Example* button right above the input box.

Step 3. Click on the *Submit* button to see the predicted result. For example, if using the three query RNA sequences from the *H. sapiens* species in the *Example* window as the input and checking on the *H. sapiens* button, after clicking the *Submit* button, you will see the following shown on the screen of your computer. (i) The first query sequence includes 5 U (uridine) residues, of which the one at position 11 can be modified to be of pseudouridine (Ψ site). (ii) The second query sequence includes 3 U residues, of which none can be modified to be of pseudouridine. (iii) The third query sequence includes 7 U residues, of which the one at position 21 can be modified to be of pseudouridine. All these results are fully consistent with the experimental observations. **Note**: to get the anticipated prediction accuracy, the species button must be consistent with the source of query sequences: if the query sequences are from *H. sapiens*, check on the *H. sapiens* button; from *M. musculus*, check on the *M. musculus* button; from *S. cerevisiae*, check on the *S. cerevisiae* button.

Step 4. Click on the *Data* button to download the datasets used to train and test the **iRNA-PseU** predictor.

Step 5. Click on the *Citation* button to find the relevant papers that document the detailed development and algorithm of **iRNA-PseU.**

Supplementary material

Information S1. The benchmark dataset $\mathbb{S}(1)$ for *H. sapiens*.
Information S2. The benchmark dataset $\mathbb{S}(2)$ for *S. cerevisiae*.
Information S3. The benchmark dataset $\mathbb{S}(3)$ for *M. musculus*.
Information S4. The independent dataset $\mathbb{S}(4)$ for *H. sapiens*.
Information S5. The independent dataset $\mathbb{S}(5)$ for *S. cerevisiae*.

Acknowledgments The authors wish to thank the three anonymous reviewers for their constructive comments, which were very helpful for strengthening the presentation

of this paper. This work was supported by Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028), the Outstanding Youth Foundation of North China University of Science and Technology (No. JP201502), the Applied Basic Research Program of Sichuan Province (No. 2015JY0100 and LZ-LY-45), the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122), and the Fundamental Research Funds for the Central Universities, China (Nos. ZYGX2015J144, ZYGX2015Z006).

1. Hudson, GA, Bloomingdale, RJ and Znosko, BM (2013). Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. *RNA* **19**: 1474–1482.
2. Ge, J and Yu, YT (2013). RNA pseudouridylation: new insights into an old modification. *Trends Biochem Sci* **38**: 210–218.
3. Charette, M and Gray, MW (2000). Pseudouridine in RNA: what, where, how, and why. *IUBMB Life* **49**: 341–351.
4. Davis, DR, Veltri, CA and Nielsen, L (1998). An RNA model system for investigation of pseudouridine stabilization of the codon-anticodon interaction in tRNALys, tRNAHis and tRNATyr. *J Biomol Struct Dyn* **15**: 1121–1132.
5. Basak, A and Query, CC (2014). A pseudouridine residue in the spliceosome core is part of the filamentous growth program in yeast. *Cell Rep* **8**: 966–973.
6. Karijolich, J and Yu, YT (2015). The new era of RNA modification. *RNA* **21**: 659–660.
7. Carlile, TM, Rojas-Duran, MF, Zinshteyn, B, Shin, H, Bartoli, KM and Gilbert, WV (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* **515**: 143–146.
8. Lovejoy, AF, Riordan, DP and Brown, PO (2014). Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. *PLoS One* **9**: e110799.
9. Schwartz, S, Bernstein, DA, Mumbach, MR, Jovanovic, M, Herbst, RH, León-Ricardo, BX *et al.* (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* **159**: 148–162.
10. Li, X, Zhu, P, Ma, S, Song, J, Bai, J, Sun, F *et al.* (2015). Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nat Chem Biol* **11**: 592–597.
11. Li, YH, Zhang, G and Cui, Q (2015). PPUS: a web server to predict PUS-specific pseudouridine sites. *Bioinformatics* **31**: 3362–3364.
12. Chou, KC (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J Theor Biol* **273**: 236–247.
13. Chen, W, Feng, PM, Lin, H and Chou, KC (2013). iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res* **41**: e68.
14. Lin, H, Deng, EZ, Ding, H, Chen, W and Chou, KC (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.
15. Liu, Z, Xiao, X, Qiu, WR and Chou, KC (2015). iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. *Anal Biochem* **474**: 69–77.
16. Chen, W, Feng, P, Ding, H, Lin, H and Chou, KC (2015). iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal Biochem* **490**: 26–33.
17. Jia, J, Liu, Z, Xiao, X, Liu, B and Chou, KC (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.
18. Liu, B, Fang, L, Long, R, Lan, X and Chou, KC (2016). iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* **32**: 362–369.
19. Jia, J, Liu, Z, Xiao, X, Liu, B and Chou, KC (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.
20. Liu, Z, Xiao, X, Yu, DJ, Jia, J, Qiu, WR and Chou, KC (2016). pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. *Anal Biochem* **497**: 60–67.
21. Jia, J, Liu, Z, Xiao, X, Liu, B and Chou, KC (2016). iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. *Molecules* **21**:95.
22. Chen, J, Liu, H, Yang, J and Chou, KC (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* **33**: 423–428.
23. Chou, KC (2001). Prediction of protein signal sequences and their cleavage sites. *Proteins* **42**: 136–139.
24. Xu, Y, Ding, J, Wu, LY and 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.
25. Xiao, X, Min, JL, Lin, WZ, Liu, Z, Cheng, X and Chou, KC (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.
26. Chen, W, Ding, H, Feng, P, *et al.* (2016). iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* DOI: 10.18632/oncotarget.17815. **7**: 16895–16909.
27. Chen, W, Feng, PM, Deng, EZ, Lin, H and Chou, KC (2014). iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal Biochem* **462**: 76–83.
28. Chen, W, Feng, PM, Lin, H and Chou, KC (2014). iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed Res Int* **2014**: 623149.
29. Ding, H, Deng, EZ, Yuan, LF, Liu, L, Lin, H, Chen, W *et al.* (2014). iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *Biomed Res Int* **2014**: 286419.
30. Liu, B, Fang, L, Liu, F, Wang, X, Chen, J and Chou, KC (2015). Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS One* **10**: e0121501.
31. Chou, KC (1993). A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J Biol Chem* **268**: 16938–16948.
32. Chou, KC, Wu, ZC 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.
33. Lin, WZ, Fang, JA, Xiao, X and Chou, KC (2013). iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol Biosyst* **9**: 634–644.
34. Xiao, X, Wu, ZC and Chou, KC (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.
35. Xiao, X, Wang, P, Lin, WZ, Jia, JH and Chou, KC (2013). iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal Biochem* **436**: 168–177.
36. Chou, KC (2013). Some remarks on predicting multi-label attributes in molecular biosystems. *Mol Biosyst* **9**: 1092–1100.
37. Chou, KC and Zhang, CT (1995). Prediction of protein structural classes. *Crit Rev Biochem Mol Biol* **30**: 275–349.
38. Shen, HB, Yang, J and Chou, KC (2007). Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids* **33**: 57–67.
39. Chou, KC and Cai, YD (2003). Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J Cell Biochem* **90**: 1250–1260.
40. Chou, KC and Cai, YD (2005). Prediction of membrane protein types by incorporating amphipathic effects. *J Chem Inf Model* **45**: 407–413.
41. Mondal, S and Pai, PP (2014). Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J Theor Biol* **356**: 30–35.
42. 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.
43. Fan, GL, Zhang, XY, Liu, YL, Nang, Y and Wang, H (2015). DSPMP: Discriminating secretory proteins of malaria parasite by hybridizing different descriptors of Chou's pseudo amino acid patterns. *J Comput Chem* **36**: 2317–2327.
44. 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.
45. 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.
46. Chen, W, Feng, P, Ding, H, Lin, H and Chou, KC (2016). Using deformation energy to analyze nucleosome positioning in genomes. *Genomics* **107**: 69–75.
47. Chou, KC, Jiang, SP, Liu, WM, Fee, CH (1979). Graph theory of enzyme kinetics: 1. Steady-state reaction system. *Scientia Sinica* **22**: 341–358.
48. Chou, KC and Forsén, S (1980). Graphical rules for enzyme-catalysed rate laws. *Biochem J* **187**: 829–835.
49. Zhou, GP and Deng, MH (1984). An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. *Biochem J* **222**: 169–176.
50. Chou, KC (1990). Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady-state systems. *Biophys Chem* **35**: 1–24.
51. Althaus, IW, Gonzales, AJ, Chou, JJ, Romero, DL, Deibel, MR, Chou, KC *et al.* (1993). The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. *J Biol Chem* **268**: 14875–14880.
52. Althaus, IW, Chou, JJ, Gonzales, AJ, Deibel, MR, Chou, KC, Kezdy, FJ *et al.* (1993). Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry* **32**: 6548–6554.
53. Chou, KC (2010). Graphic rule for drug metabolism systems. *Curr Drug Metab* **11**: 369–378.
54. Wu, ZC, Xiao, X and Chou, KC (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.
55. Chou, KC, Lin, WZ and Xiao, X (2011). Wenxiang: a web-server for drawing wenxiang diagrams *Natural Science* **3**: 862–865.
56. Zhou, GP (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.

57. Fawcett, JA (2005). An introduction to ROC Analysis. *Pattern Recognition Letters* **27**: 861–874.
58. Davis, J, and Goadrich, M (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. *ACM.* pp. 233–240.
59. Chou, KC and Shen, HB (2007). Recent progress in protein subcellular location prediction. *Anal Biochem* **370**: 1–16.
60. Sun, WJ, Li, JH, Liu, S, Wu, J, Zhou, H, Qu, LH *et al.* (2016). RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res* **44**(D1): D259–D265.
61. Chou, KC and Shen, HB (2007). Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem Biophys Res Commun* **357**: 633–640.
62. Fu, L, Niu, B, Zhu, Z, Wu, S and Li, W (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.
63. Chou, KC (2015). Impacts of bioinformatics to medicinal chemistry. *Med Chem* **11**: 218–234.
64. Chou, KC (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **43**: 246–255.
65. Chou, KC (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **21**: 10–19.
66. 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.
67. Cao, DS, Xu, QS and Liang, YZ (2013). propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* **29**: 960–962.
68. Lin, SX, 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 Biomedical Science and Engineering (JBiSE)* **6**: 435–442.
69. 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.
70. Chen, W, Lei, TY, Jin, DC, Lin, H and Chou, KC (2014). PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal Biochem* **456**: 53–60.
71. Chen, W, Zhang, X, Brooker, J, Lin, H, Zhang, L and Chou, KC (2015). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* **31**: 119–120.
72. Liu, B, Liu, F, Fang, L, Wang, X and Chou, KC (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.
73. Liu, B, Liu, F, Fang, L, Wang, X and Chou, KC (2016). repRNA: a web server for generating various feature vectors of RNA sequences. *Mol Genet Genomics* **291**: 473–481.
74. Chou, KC (2009). Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics* **6**: 262–274.
75. Chen, W, Lin, H and Chou, KC (2015). Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol Biosyst* **11**: 2620–2634.
76. Liu, B, Liu, F, Wang, X, Chen, J, Fang, L and Chou, KC (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.
77. Chou, KC and Zhang, CT (1992). Diagrammatization of codon usage in 339 human immunodeficiency virus proteins and its biological implication. *AIDS Res Hum Retroviruses* **8**: 1967–1976.
78. Zhang, CT and Chou, KC (1994). Analysis of codon usage in 1562 E. Coli protein coding sequences. *J Mol Biol* **238**: 1–8.
79. Golam Bari, ATM, Rokeya Reaz, M, and Jeong, BS (2014). DNA Encoding for Splice Site Prediction in Large DNA Sequence. *MATCH Communications in Mathematical and in Computer Chemistry* **71**: 241–258.
80. Guo, SH, Deng, EZ, Xu, LQ, Ding, H, Lin, H, Chen, W *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.
81. Liu, B, Fang, L, Wang, S, Wang, X, Li, H and Chou, KC (2015). Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *J Theor Biol* **385**: 153–159.
82. Qiu, WR, Xiao, X and Chou, KC (2014). iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int J Mol Sci* **15**: 1746–1766.
83. Chou, KC and Cai, YD (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. *J Biol Chem* **277**: 45765–45769.
84. Cai, YD, Zhou, GP and Chou, KC (2003). Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys J* **84**: 3257–3263.
85. Cristianini, N and Shawe-Taylor, J. *An Introduction of Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press: Cambridge, UK; 2000.