BMC Bioinformatics

RESEARCH ARTICLE Open Access

PseUI: Pseudouridine sites identification based on RNA sequence information

Jingjing He[1], Ting Fang[1], Zizheng Zhang[1], Bei Huang[1], Xiaolei Zhu[1*] and Yi Xiong[2*]

Abstract

Background: Pseudouridylation is the most prevalent type of posttranscriptional modification in various stable RNAs of all organisms, which significantly affects many cellular processes that are regulated by RNA. Thus, accurate identification of pseudouridine (Ψ) sites in RNA will be of great benefit for understanding these cellular processes. Due to the low efficiency and high cost of current available experimental methods, it is highly desirable to develop computational methods for accurately and efficiently detecting Ψ sites in RNA sequences. However, the predictive accuracy of existing computational methods is not satisfactory and still needs improvement.

Results: In this study, we developed a new model, PseUI, for Ψ sites identification in three species, which are *H. sapiens*, S. cerevisiae, and *M. musculus*. Firstly, five different kinds of features including nucleotide composition (NC), dinucleotide composition (DC), pseudo dinucleotide composition (pseDNC), position-specific nucleotide propensity (PSNP), and position-specific dinucleotide propensity (PSDP) were generated based on RNA segments. Then, a sequential forward feature selection strategy was used to gain an effective feature subset with a compact representation but discriminative prediction power. Based on the selected feature subsets, we built our model by using a support vector machine (SVM). Finally, the generalization of our model was validated by both the jackknife test and independent validation tests on the benchmark datasets. The experimental results showed that our model is more accurate and stable than the previously published models. We have also provided a user-friendly web server for our model at http://zhulab.ahu.edu.cn/PseUI, and a brief instruction for the web server is provided in this paper. By using this instruction, the academic users can conveniently get their desired results without complicated calculations.

Conclusion: In this study, we proposed a new predictor, PseUI, to detect Ψ sites in RNA sequences. It is shown that our model outperformed the existing state-of-art models. It is expected that our model, PseUI, will become a useful tool for accurate identification of RNA Ψ sites.

Keywords: Pseudouridine site, Position specific nucleotide propensity, Nucleotide composition

Background

Pseudouridylation, which occurs at the uridine site and is catalyzed by pseudouridine synthase (PUS), has been observed in various RNAs of all organisms [1–4]. As the most abundant posttranscriptional modification, pseudouridylation plays an important role in the structure, function and metabolism of RNAs [5–9]. Therefore, it is crucial to identify pseudouridylation information for revealing the biological principles.

Although some experimental techniques for identifying Ψ sites have been developed, they are both time-consuming and costly [10–13]. Facing the exponential-increasing of RNA sequences in the post-genomic era, it is urgent to have an accurate, efficient and low-cost method to identify Ψ sites on RNA segments. Former studies suggest that computational methods or statistical learning methods are promising candidates because of their low cost and reasonable efficiency [14, 15].

Unfortunately, to the best of our knowledge, only two computational methods have been developed to predict Ψ sites in RNAs. Li et al. [15] built a model called PPUS to predict the PUS-specific Ψ sites in *H. sapiens* and *S. cerevisiae*. This model employed support vector machine

* Correspondence: xlzhu_mdl@hotmail.com; xiongyi@sjtu.edu.cn
[1]School of Life Sciences, Anhui University, Hefei 230601, Anhui, China
[2]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

(SVM) as the classifier and used the nucleotides around Ψ as features. Besides this PPUS model, Chen et al. [14] developed another model called iRNA-PseU to identify Ψ sites in *H. sapiens*, S. cerevisiae, and *M. musculus*. This model was built by incorporating the chemical properties of nucleotides and their occurrence frequency density distributions into the general form of pseudo nucleotide composition (pseKNC) [14]. Despite the promising results offered by these two computational methods, it is suggested that the performance of computational methods can be further improved by introducing other effective features such as position-specific nucleotide propensity and position-specific dinucleotide propensity [16].

In this study, we have developed a new model, PseUI, for Ψ sites identification from RNA sequences in *H. sapiens*, S. cerevisiae, and *M. musculus*. Based on the RNA sequence segment, we first generated five different kinds of features including nucleotide composition (NC), dinucleotide composition (DC), pseudo dinucleotide composition (pseDNC), position-specific nucleotide propensity (PSNP), and position-specific dinucleotide propensity (PSDP). Then, we selected a relevant feature combination by using a sequential forward feature selection strategy [17, 18]. Based on the selected features, our model was built by using a support vector machine (SVM). Finally, the prediction results provided by our models for the three species, *H. sapiens*, S. cerevisiae, and *M. musculus*, were compared with iRNA-PseU's results by using both jackknife tests and independent validation tests on the benchmark datasets, and it is convincing from the result of comparison that our model PseUI can offer more accurate identification of Ψ sites than iRNA-PseU.

To develop a really useful feature-based analysis method for a biological system as reported in a series of recent studies [19–23], one should observe the 5-step rule [24]: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) develop a powerful algorithm (or engine) to operate the prediction; (iv) perform cross-validation and independent tests properly to objectively evaluate the anticipated accuracy of the predictor; and (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by- one.

Methods

Benchmark datasets

Three benchmark datasets, H_990, S_628, and M_944, were used for training in this study, where H, S, and M represent for *H. sapiens*, S. cerevisiae, and *M. musculus*, respectively, and 990, 628, 944 are the number of examples in each dataset. These three datasets are the same as that were used in Chen et al.'s work [14]. In their work, they downloaded RNA sequences with experimentally validated Ψ sites of *H. sapiens*, M. musculus and *S. cerevisiae* from RMBase [25]. In addition, they collected the RNA segments with uridine at the center but not experimentally conformed as Ψ sites from genomes as negative samples. More details about how to construct these datasets can be found in the reference [14].

The positive subset of H_990, S_628, and M_944 contains 495, 314, and 472 RNA segments, respectively, and each of these RNA segments has a uridine at the center position that can be pseudouridylated. The negative subset is composed of 495, 314, and 472 RNA segments, respectively, and each of these RNA segments has a uridine at the center position that cannot be pseudouridylated.

We can formulate each RNA segment, denoted as $R_\xi(U)$, in these datasets as follow:

$$R_\xi(U) = N_{-\xi}N_{-(\xi-1)}\cdots N_{-1}UN_1\cdots N_{+(\xi-1)}N_\xi \qquad (1)$$

where the center U represents 'uridine,' $N_{-\xi}$ represents the ξ-th upstream nucleotide from the central uridine and $N_{+\xi}$ represents the ξ-th downstream nucleotide.

The RNA samples in both of H_990 and M_944 are all composed of 21 nucleotides, while those in S_628 are composed of 31 nucleotides. Namely, the value of ξ is 10 and the RNA segment length is 2 × 10 + 1 for the datasets H_900 and M_944. The value of ξ is 15 and the RNA segment length is 2 × 15 + 1 for the dataset S_628.

Corresponding to the training datasets, Chen et al. [14] provided two independent testing datasets for *H. sapiens* and *S. cerevisiae*, i.e. H_200 and S_200, but not for *M. musculus*. The detailed sequence information for all the aforementioned datasets is given in Table 1; and the sequences of the five datasets can be found in Additional files 1, 2, 3, 4 and 5.

Feature representation of the RNA samples

One of the key problems in designing a predictor based on machine learning is how to encode an RNA sequence as a feature vector containing highly discriminative information. With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to represent a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms can only handle vectors with equal lengths for all sequence samples, as elucidated in a comprehensive

Table 1 The information of training datasets and independent testing datasets

Species	The name of training/ testing datasets[a]	The length of the RNA sequences (bp)	The number of positive samples	The number of the negative samples
H. sapiens	H_990	21	495	495
	H_200	21	100	100
S. cerevisiae	S_628	31	314	314
	S_200	31	100	100
M. musculus	M_944	21	472	472
	–	–	–	–

[a] H_900, S_628, M_944 are the training datasets for H. sapiens, S. cerevisiae, M. musculus, respectively; H_200 and S_200 are the independent testing datasets for H. sapiens and S. cerevisiae, respectively

review [26]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition [27] or PseAAC [28] was proposed. Encouraged by the success of using PseAAC to represent protein/peptide sequences, the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) [29] was developed for generating various feature vectors to represent DNA/RNA sequences. Particularly, recently a very powerful web-server called Pse-in-One [30] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies. In the current study, five types of features, nucleotide composition (NC) feature, dinucleotide composition (DC) feature, pseudo dinucleotide composition (pseDNC) feature, position-specific nucleotide propensity (PSNP) feature, and position-specific dinucleotide propensity (PSDP) feature, were proposed to encode the RNA segments for identifying pseudouridine sites in RNA. Three of them, NC, DC, and pseDNC, can also be generated by Pse-in-One server [30].

Nucleotide composition (NC) and dinucleotide composition (DC) feature

Nucleotide composition, a classic method for the characterization of nucleotide sequences, is widely used in previous studies [31–33]. Theoretically, a k-mer nucleotide composition for an RNA sequence is a 4^k-dimensional vector which is consisted of the frequency of each k-mer types. Thus, we can obtain 4 types of nucleotide frequencies and 16 types of dinucleotide frequencies when k is equal to 1 and 2, respectively. We called these two features as NC and DC, respectively, and a 4-dimensional NC feature vector and a 16-dimensional DC feature vector were generated for an RNA segment.

Pseudo dinucleotide composition (pseDNC) feature

The pseudo oligonucleotide composition, or pseudo K-tuple nucleotide composition (PseKNC) [34–37], can be used to represent an RNA sequence with a discrete model or vector. This type of pseudo composition can still keep considerable sequence order information, particularly the global or long-range sequence order information, via the physicochemical properties of its constituent oligonucleotides [38]. In this study, we choose the value of K to be 2, namely, using pseudo dinucleotide composition (pseDNC) feature to represent the information of RNA sequences. Three physicochemical properties, free energy, hydrophilicity, and stacking energy, were used to generate features of pseudo dinucleotide composition (pseDNC), which are listed in Table 2.

Position-specific nucleotide propensity (PSNP) and position-specific dinucleotide propensity (PSDP) feature

While position-specific amino acid preferences have been widely used in bioinformatics to predict functional site in biological sequences [39–42], the position-specific nucleotide preferences were first introduced in Li et al.'s paper [16], which were obtained by calculating the differences of the frequency of nucleotides in specific locations between positive and negative RNA segments.

For position-specific nucleotide propensity (PSNP) feature, according to the equation (1), the RNA segment can be reformulated as:

$$R_\xi = N_1 N_2 \dots N_{2\xi+1} \tag{2}$$

Table 2 Three types of physicochemical properties of dinucleotides in RNA

Dinucleotide	Free energy	Hydrophilicity	Stacking energy
GG	−3.260	0.170	−11.100
GA	− 2.350	0.100	− 14.200
GC	−3.420	0.260	−16.900
GU	−2.240	0.270	−13.800
AG	−2.080	0.080	−14.000
AA	−0.930	0.040	−13.700
AC	−2.240	0.140	−13.800
AU	−1.100	0.140	−15.400
CG	−2.360	0.350	−15.600
CA	−2.110	0.210	−14.400
CC	−3.260	0.490	−11.100
CU	−2.080	0.520	−14.000
UG	−2.110	0.340	−14.400
UA	−1.330	0.210	−16.000
UC	−2.350	0.480	−14.200
UU	−0.930	0.440	−13.700

More details about the pseudo dinucleotide composition (pseDNC) feature refer to [38] where N_j(j=1,2,...,2ξ+1) represents the nucleotide at the j-th position of the RNA segment, and can be any one of the 4 nucleotides, i.e., $N_j \in \{A, C, G, U\}$.

First, we calculated the frequency of occurrence at the j-th position for the 4 types of nucleotides from both the positive and negative samples, respectively. Then, we combined the 4-dimensional positive vectors and the 4-dimendional negative vectors individually. In this way, we obtained two 4× $(2\xi + 1)$ position-specific occurrence frequency matrixes, i.e., Z^+ and Z^-, where Z^+ was obtained from all the positive samples, and Z^- was obtained from all the negative samples. Next, we defined the position-specific nucleotide propensity (PSNP) matrixes, denoted as Z_{PSNP}, as below:

$$Z_{PSNP} = Z^+ - Z^- \tag{3}$$

As for position-specific dinucleotide propensity (PSDP) feature, according to equation (2), the RNA segment can be rewritten in a dinucleotide form:

$$R_\xi = N_1N_2 \dots N_{2\xi+1} = D_1D_2 \dots D_{2\xi} \tag{4}$$

where $D_j = N_jN_{j+1}(j = 1, 2, \dots, 2\xi)$ represents the dinucleotide at the j-th position of the RNA segment, and can be any of 16 types of dinucleotides, i.e., $D_j \in$ {AA, AC, AG, …, UU}.

Similarly, following the principle we used to generate the Z_{PSNP} matrix, we can get the 16 × 2ξ position-specific dinucleotide propensity (PSDP) matrix. Both of the PSNP matrix and PSDP matrix can then be used to encode the new samples.

For the features encoded by PSNP and PSDP, we should pay particular attention to the fact that the propensity matrices (Z_{PSNP}/Z_{PSDP}) were only generated from the training samples without the one validation sample when evaluating the model using the jackknife test.

Figure 1 clearly described the jackknife cross validation for features encoded by PSNP/PSDP. The validation process has four steps: (1) Input the dataset (R), e.g., H_990, S_628, or M_944, which is assumed to have n samples. (2) Divide the dataset (R) into n subsets and each subset will contain only one sample. (3) One subset is selected as the validation set, and the rest are used as the training set. The samples of the training set will be used to calculate the frequency of nucleotides at specific locations, and the position specific propensity matrices (Z_{PSNP}/Z_{PSDP}) will be obtained and then used to encode the RNA segments in the training set and the validation set. In such way, the feature matrices R^T(PSNP/PSDP) and R^V(PSNP/PSDP) can be obtained to represent the statistical information extracted from the training set and the validation set, respectively. A model will be then built by SVM based on the training set, and evaluated on the validation set. The whole process will be repeated for n times and each time a different sample will be selected as the validation set. (4) Count the results from the previous steps and calculate the evaluation parameter, i.e., Sen, Spe, Acc, and MCC, which are described in “Evaluation parameter” section.

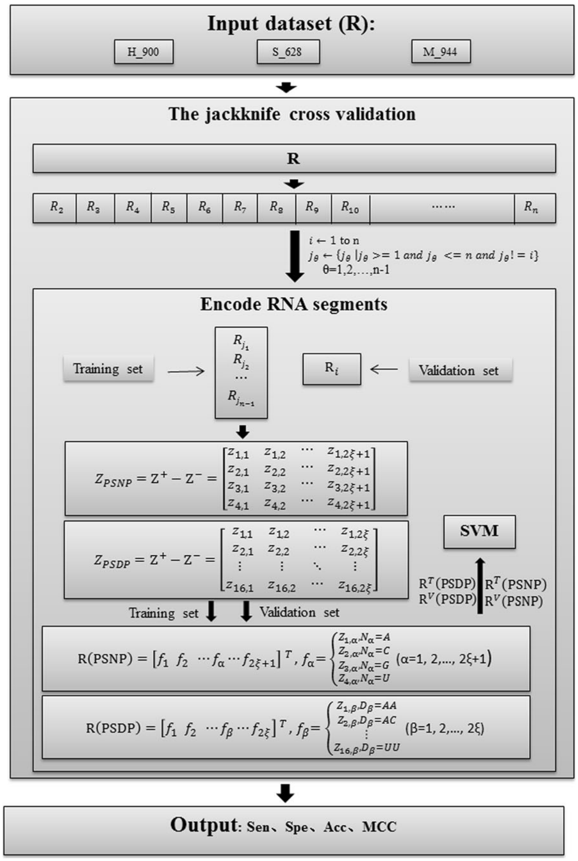
Input dataset (R):
H_900
S_628
M_944
The jackknife cross validation
R
R_2 R_3 R_4 R_5 R_6 R_7 R_8 R_9 R_{10} …… R_n
$i \leftarrow 1$ to n
$j_\theta \leftarrow \{j_\theta \mid j_\theta >= 1 \; and \; j_\theta <= n \; and \; j_\theta != i\}$
θ=1,2,…,n-1
Encode RNA segments
Training set
R_{j_1} R_{j_2} … $R_{j_{n-1}}$
R_i
Validation set
$Z_{PSNP} = Z^+ - Z^- = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,2\xi+1} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,2\xi+1} \\ z_{3,1} & z_{3,2} & \cdots & z_{3,2\xi+1} \\ z_{4,1} & z_{4,2} & \cdots & z_{4,2\xi+1} \end{bmatrix}$
$Z_{PSDP} = Z^+ - Z^- = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,2\xi} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,2\xi} \\ \vdots & \vdots & \ddots & \vdots \\ z_{16,1} & z_{16,2} & \cdots & z_{16,2\xi} \end{bmatrix}$
SVM
R^T(PSDP) R^T(PSNP)
R^V(PSDP) R^V(PSNP)
Training set
Validation set
$R(PSNP) = [f_1 \; f_2 \; \cdots f_\alpha \cdots f_{2\xi+1}]^T, f_\alpha = \begin{cases} Z_{1,\alpha}, N_\alpha = A \\ Z_{2,\alpha}, N_\alpha = C \\ Z_{3,\alpha}, N_\alpha = G \\ Z_{4,\alpha}, N_\alpha = U \end{cases}$ (α=1, 2,..., 2ξ+1)
$R(PSDP) = [f_1 \; f_2 \; \cdots f_\beta \cdots f_{2\xi}]^T, f_\beta = \begin{cases} Z_{1,\beta}, D_\beta = AA \\ Z_{2,\beta}, D_\beta = AC \\ \vdots \\ Z_{16,\beta}, D_\beta = UU \end{cases}$ (β=1, 2,..., 2ξ)
Output: Sen、Spe、Acc、MCC

Fig. 1 Flow charts of the jackknife cross validation for features encoded by PSNP or PSDP

Model construction

Support vector machine

As a popular statistical learning method, SVM has been extensively used to build bioinformatics models [43–52]. Both of the PPUS and iRNA-PseU models [14, 15] mentioned in the background section were built by using SVM due to its high efficiency and robust output. In this study, we used the Matlab function FITCSVM to build our models. Different kernel functions can be used in SVM training, and we selected the radial basis function in this study. Two parameters c and g were referred for the radial basis function, which were called box constraint and kernel scale in FITCSVM, respectively. Here, we optimized these two parameters based on the jackknife test using a grid search.

In statistical analysis fields, three different validation methods have mostly been used to evaluate the performance of a machine learning model: independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test [53]. The jackknife test has already proved its effectiveness in many aspects [54, 55]. It is not affected by the random partition of the samples, and the final result is unique. In addition, the training set used by the jackknife test is only one sample less than the initial training set. Therefore, in most cases, the actual model evaluated by the jackknife test is very close to the expected model, which will offer more accurate results. Based on all these advantages, the jackknife test was used to evaluate the performance of our models.

Evaluation parameters

In recent studies, four evaluation parameters, Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), and the Matthews correlation coefficient (MCC) have been frequently used to measure the predictor's quality [46, 56]. The original formulas of the four parameters, particularly the MCC, are lacking intuitiveness and not easy to understand for most biologists. To make the most readers easy to understand, we here introduced the Chou's intuitive formulas of the four parameters, as elaborated by the four sub-equations in Eq. 19 of [57], or the four sub-equations in Eq. 14 of [58]. Particularly, the advantages of Chou's intuitive metrics have been analyzed and concurred by a series of studies published very recently [19, 20, 22, 59, 60]. The Chou's intuitive metrics are formulated as below:

$$\begin{cases} \mathrm{Sen} = 1 - \dfrac{N_-^+}{N^+}, 0 << Sen << 1 \\ \mathrm{Spe} = 1 - \dfrac{N_+^-}{N^-}, 0 << Spe << 1 \\ \mathrm{Acc} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-}, 0 << Acc << 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}}, -1 << MCC << 1 \end{cases} \tag{5}$$

Where N^+ represents the total number of positive RNA samples; N^- represents the total number of negative RNA samples; N_-^+ represents the number of positive RNA samples that are incorrectly predicted as negative RNA samples; N_+^- represents the number of negative RNA samples that are incorrectly predicted as positive RNA samples. In addition, it should be noted that the set of metrics in eq. (5) is only valid for the single-label systems (in which each sample only belongs to one class). For the multi-label systems (in which a sample might belong to several classes), whose existence has become more frequent in system biology [61] and system medicine [20] and biomedicine [60], a completely different set of metrics as defined in [62] is needed.

Feature selection

In this study, we generated five types of features which composed a high dimensional feature vector for each sample. In order to obtain a more compact and effective feature subset, we conducted a sequential forward feature selection (SFS) [17, 18] process on the original features, which is described as follows:

In the first round, the performance metrics of each of the five types of features were calculated based on the jackknife test using a specific prediction engine, respectively. According to Acc or MCC, the best type of feature was selected to enter the next round of calculation. In the second round, the remaining four types of features were added to the type of feature selected by the first round. Similarly, according to Acc or MCC, the best combination of features was selected to enter the next round of calculation. This process continued to run until the Acc or MCC converged. The subset obtained with the highest Acc or MCC value will be regarded as the optimal feature subset.

Results and discussion

Performance of single type of feature

In this section, we evaluated the performance of each type of features using SVM over the rigorous jackknife test, and the feature PSNP was found to be particularly excellent for identifying Ψ sites. The performance of each evaluation index for the three species, i.e., *H. sapiens*, S. cerevisiae, and *M. musculus*, were listed in Tables 3, 4, and 5, respectively.

In addition, the receiver operating characteristic (ROC) curves [63] were employed to show the results more clearly. On the ROC curve, the diagonal line from point (0, 0) to (1, 1) corresponds to the random guessing model, and the point (0, 1) corresponds to the ideal model with no positive example wrongly predicted. When comparing models, if the ROC curve of one model is completely enveloped by the curve of the other model, it can be asserted that the latter model is superior to the former in performance. However, it is difficult to judge when the ROC curves of two models cross. In this situation, the area under the ROC curve (AUC) will be used as the more reasonable criteria for comparing model performance, and the lager AUC indicates better performance. The ROC curves of the five types of feature for each species were plotted in Fig. 2, together with the AUC values.

Table 3 The results of feature selection for H_990

Feature subset	Sen (%)	Spe (%)	Acc (%)	MCC	Kernel scale	Box constraint
NC	62.83	51.31	57.07	0.1424	0.5	4
DC	46.87	74.95	60.91	0.2273	2	256
pseDNC	44.24	76.57	60.40	0.2199	4	1024
PSNP	66.06	60.61	63.33	0.2671	8	512
PSDP	55.15	57.17	56.16	0.1233	0.5	1024
PSNP+NC	65.05	61.21	63.13	0.2628	1	4
PSNP + DC	*64.85*	*63.64*	*64.24*	*0.2849*	*2*	*8*
PSNP+pseDNC	64.44	62.42	63.43	0.2687	1	8
PSNP+PSDP	66.26	59.39	62.83	0.2572	8	1024
PSNP+DC + NC	64.85	63.43	64.14	0.2829	8	128
PSNP+DC + pseDNC	63.03	63.23	63.13	0.2626	4	32
PSNP+DC + PSDP	64.24	63.43	63.84	0.2768	1	2

The feature combination with the maximum MCC was italicized in the table

As shown in Fig. 2, the AUC values of PSNP are 0.6569, 0.6441, and 0.7443 for *H. sapiens*, S. cerevisiae, and *M. musculus*, respectively. For *H. sapiens* and *M. musculus*, the AUC values of PSNP are much higher than those of the other four types of features. For *S. cerevisiae*, the AUC value of PSNP is only 0.0077 lower than the highest AUC value 0.6518 given by DC. Moreover, the accuracy was improved from 62.10 to 64.49% when PSNP was added in the second round of SFS for *S. cerevisiae*, which was shown in Table 4. These results all indicate that PSNP offered the best performance among these five types of features and the addition of PSNP provided a great possibility of improving the model performance, which may lay the foundation for our future works.

Feature subsets selected by SFS

For the selection of feature subset with SFS described in the "Feature selection" section, we run three rounds of calculation for the datasets H_990 and M_944, respectively. Finally, the subset that made up of DC and PSNP features was chosen as the optimal feature subset. The results of each round for *H. sapiens* and *M. musculus* are shown in Tables 3 and 5, respectively. For both *H. sapiens* and *M. musculus*, the best models were built based on the feature subset PSNP+DC.

For the dataset S_628, four rounds of calculation were conducted, and the subset with a combination of DC, pseDNC, and PSNP, was selected as the optimal feature subset. The results of each round are listed in Table 4. The best model of *S. cerevisiae* is built based on the feature subset DC + PSNP+pseDNC.

Table 4 The results of feature selection for S_628

Feature subset	Sen (%)	Spe (%)	Acc (%)	MCC	Kernel scale	Box constraint
NC	71.97	45.22	58.60	0.1785	1	8
DC	64.33	59.87	62.10	0.2423	0.25	1
pseDNC	58.92	62.42	60.67	0.2135	0.25	0.5
PSNP	50.96	72.93	61.94	0.2448	1	0.125
PSDP	49.36	73.57	61.46	0.2363	0.25	0.03125
DC + NC	59.55	61.78	60.67	0.2134	4	512
DC + pseDNC	62.42	60.51	61.46	0.2293	1	1024
DC + PSNP	63.69	65.29	64.49	0.2898	0.5	16
DC + PSDP	60.51	66.88	63.69	0.2744	0.125	2
DC + PSNP+NC	61.78	65.61	63.69	0.2741	0.25	1
DC + PSNP + pseDNC	*64.97*	*66.88*	*65.92*	*0.3185*	*0.25*	*2*
DC + PSNP+PSDP	63.38	67.20	65.29	0.3060	0.25	2
DC + PSNP+pseDNC+NC	61.78	65.92	63.85	0.2773	0.25	2
DC + PSNP+pseDNC+PSDP	62.74	67.52	65.13	0.3029	0.25	4

The feature combination with the maximum MCC was italicized in the table

Table 5 The results of feature selection for M_944

Feature subset	Sen (%)	Spe (%)	Acc (%)	MCC	Kernel scale	Box constraint
NC	56.99	53.18	55.08	0.2233	2	2
DC	61.86	52.75	57.31	0.1468	4	1024
pseDNC	72.46	44.28	58.37	0.1744	4	128
PSNP	73.31	66.31	69.81	0.3972	0.5	1
PSDP	68.22	60.38	64.30	0.2869	1	256
PSNP+NC	69.70	70.34	70.02	0.4004	0.25	0.125
PSNP + DC	*74.58*	*66.31*	*70.44*	*0.4103*	*1*	*2*
PSNP+pseDNC	74.15	66.53	70.34	0.4080	0.5	1
PSNP+PSDP	68.64	70.97	69.81	0.3963	0.125	0.5
PSNP+DC + NC	74.15	66.10	70.13	0.4039	0.5	0.25
PSNP+DC + pseDNC	73.09	67.80	70.44	0.4095	0.5	0.5
PSNP+DC + PSDP	74.58	66.31	70.44	0.4103	0.5	0.25

The feature combination with the maximum MCC was italicized in the table

Comparison with existing methods

In this section, we compared our model PseUI with the latest model iRNA-PseU [14] by using two validation methods (i.e., the jackknife cross validation and independent tests) to confirm the predictability of our model.

Unfortunately, after a careful study of Chen et al.'s article [14], we found that some of the results reported by the authors were not reasonable. For example, the values of Sen (Sensitivity) and Spe (Specificity) for *S. cerevisiae* using the jackknife cross validation were 64.65 and 64.33% (see Table 6). However, according to the ROC curve in Chen et al.'s paper [14], the value of "1-Specificity" is estimated to be approximately 0.24, thus the "Specificity" value should be approximately 0.76, when "Sensitivity" is 0.6465. This "specificity" value (0.76) is significantly different from the aforementioned "specificity" value (64.33%). Besides this big discrepancy in "specificity" values, the optimized parameters g and c were not reported in the paper.

To have a more accurate comparison with Chen et al.'s method, we wrote our programs in strict accordance with the description of their paper to re-implement iRNA-PseU. The software LIBSVM-3.22 was used to train the SVM models. To obtain the best performance of the jackknife cross validation, we used a grid search to optimize the SVM parameter g from 2^{-15} to 2^{-5} and parameter c from 2^{-5} to 2^{15} with a step of 2. Finally, the parameters g and c were set at 0.01562 and 2 for *H. sapiens*, 0.0003 and 32,768 for S. cerevisiae, and 0.00098 and 4 for *M. musculus*, respectively.

Then, we compared the proposed PseUI with the re-implemented iRNA-PseU (named re-iRNA-PseU) by using the jackknife cross validation. The comparison results for the three training datasets, i.e., H_990, S_628,

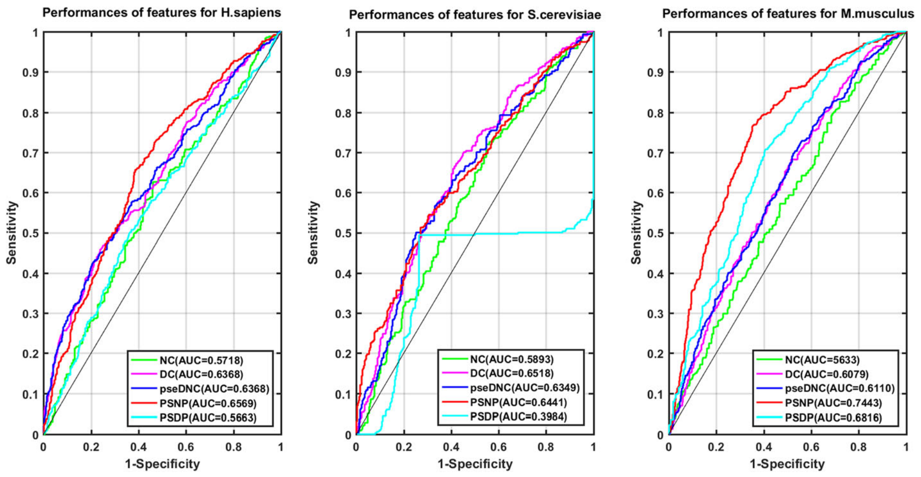

Fig. 2 The ROC curves that show the performances of the five type of features for H.sapiens, S.cerevisiae, and M.musculus, respectively

Table 6 A comparison of PseUI with iRNA-PseU and re-iRNA-PseU on three training datasets

Training datasets	Predictor	Sen (%)	Spe (%)	Acc (%)	MCC	AUC
H_990	iRNA-PseU[a]	61.01	59.80	60.40	0.21	0.64
	re-iRNA-PseU[b]	65.05	58.79	61.92	0.24	0.65
	PseUI[c]	64.85	63.64	64.24	0.28	0.68
S_628	iRNA-PseU[a]	64.65	64.33	64.49	0.29	0.81
	re-iRNA-PseU[b]	66.88	64.33	65.61	0.31	0.69
	PseUI[c]	62.10	71.02	66.56	0.33	0.69
M_944	iRNA-PseU[a]	73.31	64.83	69.07	0.38	0.75
	re-iRNA-PseU[b]	79.87	60.81	70.34	0.41	0.75
	PseUI[c]	74.58	66.31	70.44	0.41	0.77

[a] The predictor developed by Chen et al. [14]
[b] The predictor we re-implemented by the method proposed by Chen et al. [14]
[c] The predictor proposed in this paper

and M_944, were listed in Table 6, and the ROC curves of PseUI were shown in Fig. 3. As shown in Table 6, both Acc and MCC obtained by PseUI are higher than those obtained by re-iRNA-PseU. For Acc, improvements of 2.32%, 0.95%, and 0.10% were observed for H_990, S_628, and M_944, respectively, and for MCC, improvements of 4 and 2% were observed for H_990 and S_628. In addition, as shown in Fig. 3, the AUC values of PseUI are 0.68 and 0.77, which are 0.03 and 0.02 higher than the corresponding AUC values of re-iRNA-PseU for *H. sapiens* and *M. musculus*, respectively. These findings confirmed that the PseUI outperformed the re-iRNA-PseU in both accuracy and stability for identifying Ψ sites. Note that the re-iRNA-PseU is superior to iRNA-PseU according to the evaluation metrics shown in Table 6.

Next, we compared our models PseUI with the re-iRNA-PseU on the independent datasets. In this study, independent datasets are only available for the species of H. sapiens and *S. cerevisiae* (i.e., H_200 and S_200), so the comparison was only conducted on these two datasets. The results were listed in Table 7.

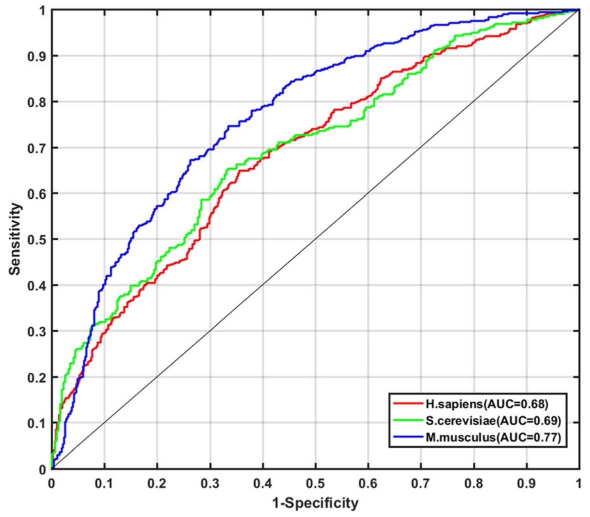

Fig. 3 The ROC curves of the best models for H.sapiens, S.cerevisiae, and M.musculus, respectively

As shown in Table 7, the predictive Accs of H_200 and S_200 are 65.50 and 68.50%, which are similar to the corresponding cross validation Accs on the training datasets. This means that our model is stable and has good generalization ability for predicting Ψ sites. When compared with re-iRNA-PseU, the proposed PseUI model showed improvements of 4 and 8.5% of the Accs values on the two independent test sets, respectively. As for MCC, PseUI outperformed re-iRNA-PseU with improvements of 0.08 and 0.17 for H_200 and S_200, respectively. All these results confirmed that our proposed model PseUI is superior to re-iRNA-PseU.

Web implementation

As demonstrated in a series of recent publications [58, 61, 64–75], user-friendly and publicly accessible web-servers or source codes represent the future direction for developing practically more useful analysis methods and computational tools. Actually, many practically useful web-servers have significant impacts on medical science [26], driving medicinal chemistry into an unprecedented revolution [76]. For the convenience of academic users, we did the same and established a user-friendly and publicly accessible web server for PseUI, which is freely accessible at http://zhulab.ahu.edu.cn/PseUI. Users can

Table 7 A comparison of PseUI with the re-iRNA-PseU on two independent datasets

Datasets	Predictor	Sen (%)	Spe (%)	Acc (%)	MCC
H_200	re-iRNA-PseU[a]	58.00	65.00	61.50	0.23
	PseUI[b]	63.00	68.00	65.50	0.31
S_200	re-iRNA-PseU[a]	63.00	57.00	60.00	0.20
	PseUI[b]	72.00	65.00	68.50	0.37

[a] The predictor we re-implemented by the method proposed by Chen et al. [14]
[b] The predictor proposed in this paper

easily get their desired results without complicated mathematic calculations. The final online PseUI method was trained on H_990, S_628, and M_944, which are composed of 21, 31, and 21 nucleotides, respectively. The detailed procedure to predict Ψ sites by using PseUI method is as follows:

Firstly, a query RNA sequence is submitted and the RNA sequence should be longer than 21 bp for *H.sapiens* and *M.musculus* or longer than 31 bp for S.cerevisiae in FASTA format. Secondly, PseUI identifies each uridine site in the query RNA sequence, and a corresponding 21-nt RNA segment for *H.sapiens* and *M.musculus* or 31-nt RNA segment for S.cerevisiae is constructed by placing a sliding window centered on the uridine site. Thirdly, according to the reconstructed RNA segment, the vector for the statistical information of the sequence is extracted by the features, and then submitted to the SVM classification engine for prediction. Finally, the users can get the result they desired. Please notice that the reconstructed RNA segment for unequal number of nucleotides around the target uridine is filled with its mirror image [47].

Conclusion

In this study, we proposed a model, PseUI, for accurate and efficient identification of Ψ sites in RNA sequences. We compared our model PseUI with the latest Ψ site identification model iRNA-PseU [14] by using two different methods, jackknife cross validation and independent tests. The results showed that our model is more accurate and stable than iRNA-PseU. In addition, the performances of the five types of features used in this study were systematically evaluated and compared, and the feature of PSNP was found to show the best performance. To facilitate the use of our model, a web server was built at http://zhulab.ahu.edu.cn/PseUI, which allows the academic users to easily use our model to predict the Ψ sites in RNA sequences.

Additional files

Additional file 1: The benchmark dataset H_990 for H.sapiens. The benchmark dataset H_990, S_628, and M_944 is formed by 495, 314 and 472 Ψ-site-containing sequences and 495, 314 and 472 false Ψ-site-containing sequences, respectively. Both H_200 and S_200 are formed by 100 Ψ-site-containing sequences and 100 false Ψ-site-containing sequences, and none of the samples included here occur in the corresponding benchmark datasets. Each of these samples for *H.sapiens* and *M.musculus* is 21-bp long with the uridine located at the center, and each of these samples for S.cerevisiae is 31-bp long with the uridine located at the center. None of the sequences included here has ≥60% pairwise sequence identity to any other in a same subset. (DOCX 56 kb)

Additional file 2: The benchmark dataset S_628 for S.cerevisiae. The benchmark dataset H_990, S_628, and M_944 is formed by 495, 314 and 472 Ψ-site-containing sequences and 495, 314 and 472 false Ψ-site-containing sequences, respectively. Both H_200 and S_200 are formed by 100 Ψ-site-containing sequences and 100 false Ψ-site-containing sequences, and none of the samples included here occur in the corresponding benchmark datasets. Each of these samples for *H.sapiens* and *M.musculus* is 21-bp long with the uridine located at the center, and each of these samples for S.cerevisiae is 31-bp long with the uridine located at the center. None of the sequences included here has ≥60% pairwise sequence identity to any other in a same subset. (DOCX 45 kb)

Additional file 3: The benchmark dataset M_944 for M.musculus. The benchmark dataset H_990, S_628, and M_944 is formed by 495, 314 and 472 Ψ-site-containing sequences and 495, 314 and 472 false Ψ-site-containing sequences, respectively. Both H_200 and S_200 are formed by 100 Ψ-site-containing sequences and 100 false Ψ-site-containing sequences, and none of the samples included here occur in the corresponding benchmark datasets. Each of these samples for *H.sapiens* and *M.musculus* is 21-bp long with the uridine located at the center, and each of these samples for S.cerevisiae is 31-bp long with the uridine located at the center. None of the sequences included here has ≥60% pairwise sequence identity to any other in a same subset. (DOCX 54 kb)

Additional file 4: The independent dataset H_200 for H.sapiens. The benchmark dataset H_990, S_628, and M_944 is formed by 495, 314 and 472 Ψ-site-containing sequences and 495, 314 and 472 false Ψ-site-containing sequences, respectively. Both H_200 and S_200 are formed by 100 Ψ-site-containing sequences and 100 false Ψ-site-containing sequences, and none of the samples included here occur in the corresponding benchmark datasets. Each of these samples for H.sapiens and M.musculus is 21-bp long with the uridine located at the center, and each of these samples for S.cerevisiae is 31-bp long with the uridine located at the center. None of the sequences included here has ≥60% pairwise sequence identity to any other in a same subset. (DOCX 26 kb)

Additional file 5: The independent dataset S_200 for S.cerevisiae. The benchmark dataset H_990, S_628, and M_944 is formed by 495, 314 and 472 Ψ-site-containing sequences and 495, 314 and 472 false Ψ-site-containing sequences, respectively. Both H_200 and S_200 are formed by 100 Ψ-site-containing sequences and 100 false Ψ-site-containing sequences, and none of the samples included here occur in the corresponding benchmark datasets. Each of these samples for *H.sapiens* and *M.musculus* is 21-bp long with the uridine located at the center, and each of these samples for S.cerevisiae is 31-bp long with the uridine located at the center. None of the sequences included here has ≥60% pairwise sequence identity to any other in a same subset. (DOCX 25 kb)

Abbreviations

DC: Dinucleotide Composition; NC: Nucleotide Composition; PSDP: Position-Specific Dinucleotide Preferences; PSNP: Position-Specific Nucleotide Preferences; PUS: Pseudouridine Synthase; SFS: Sequential Forward Selection; SVM: Support Vector Machine

Acknowledgements

The authors thank the editor and two anonymous reviewers for their valuable comments on significantly improving this paper. They also would like to thank Tianle Wu for assistance on manuscript preparation and revision.

Funding

This work was supported by National Natural Science Foundation of China (21403002, 31601074). The fundings had no role in the design of the study and collection, analysis, and interpretation of data and writing the manuscript.

Availability of data and materials

All data generated or analyzed during this study are included in this published article or the Additional files.

Authors' contributions

Conceived the study: XZ, YX. Designed the study: JH, XZ. Participate designed the study: YX, TF, BH. Analyzed the data: JH, TF, YX. Website building: ZZ, XZ. Wrote the paper: JH, XZ, YX, BH. All authors read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 19 April 2018 Accepted: 21 August 2018
Published online: 29 August 2018

References
1. Cantara WA, Crain PF, Rozenski J, Mccloskey JA, Harris KA, Zhang X, Vendeix FA, Fabris D, Agris PF. The RNA modification database, RNAMDB: 2011 update. Nucleic Acids Res. 2011;39(Database issue):D195.
2. Duninhorkawicz S, Czerwoniec A, Gajda MJ, Feder M, Grosjean H, Bujnicki JM. MODOMICS: a database of RNA modification pathways. Nucleic Acids Res. 2006;34(Database issue):D145.
3. Behmansmant I, Urban A, Ma X, Yu YT, Motorin Y, Branlant C. The Saccharomyces cerevisiae U2 snRNA:pseudouridine-synthase Pus7p is a novel multisite-multisubstrate RNA:psi-synthase also acting on tRNAs. Rna-a Publication of the Rna Society. 2003;9(11):1371.
4. Bousquet-Antonelli C, Henry Y, Gélugne JP, Caizergues-Ferrer M, Kiss T. A small nucleolar RNP protein is required for pseudouridylation of eukaryotic ribosomal RNAs. EMBO J. 1997;16(15):4770–6.
5. Junhui Y, Tao Y. RNA pseudouridylation: new insights into an old modification. Trends Biochem Sci. 2013;38(4):210.
6. Grosjean H. DNA and RNA modification enzymes: Structure, Mechanism, Function and Evolution. Austin: Landes Biosciences; 2009.
7. Ofengand J, Fournier MJ: The pseudouridine residues of rRNA: Number, location, biosynthesis, and function. 1998.
8. Ma X, Zhao X, Yu YT. Pseudouridylation (Ψ) of U2 snRNA in S.Cerevisiae is catalyzed by an RNA-independent mechanism. EMBO J. 2003;22(8):1889.
9. Newby MI, Greenbaum NL. A conserved pseudouridine modification in eukaryotic U2 snRNA induces a change in branch-site architecture. Rna-a Publication of the Rna Society. 2001;7(6):833–45.
10. Carlile TM, Rojasduran MF, Zinshteyn B, Shin H, Bartoli KM, Gilbert WV. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature. 2014;515(7525):143–6.
11. Lovejoy AF, Riordan DP, Brown PO. Transcriptome-wide mapping of Pseudouridines: Pseudouridine synthases modify specific mRNAs in S. Cerevisiae. PLoS One. 2014;9(10):e110799.
12. Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, Leónricardo BX, Engreitz JM, Guttman M, Satija R, Lander ES. Transcriptome-wide mapping reveals widespread dynamic regulated pseudouridylation of ncRNA and mRNA. Cell. 2014;159(1):148.
13. Li X, Zhu P, Ma S, Song J, Bai J, Sun F, Yi C. Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. Nat Chem Biol. 2015;11(8):592.
14. Wei C, Hua T, Jing Y, Hao L, Chou KC. iRNA-PseU: identifying RNA pseudouridine sites. Mol Ther Nucleic Acids. 2016;5(7):e332.
15. Li YH, Zhang G, Cui Q. PPUS: a web server to predict PUS-specific pseudouridine sites. Bioinformatics. 2015;31(20):3362–4.
16. Li GQ, Liu Z, Shen HB, Yu DJ: TargetM6A: identifying N6-methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. IEEE Transactions on Nanobioscience 2016, PP(99):1–1.
17. Ververidis D, Kotropoulos C. Sequential forward feature selection with low computational cost. In: Signal processing conference, 2005 European; 2010. p. 1–4.
18. Wang L, Shen C, Hartley R. On the optimality of sequential forward feature selection using class Separability measure. In: International conference on digital image computing techniques and applications; 2012. p. 203–8.
19. Jia J, Liu Z, Xiao X, Liu B, Chou KC. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol. 2015;377:47–56.
20. Cheng X, Zhao SG, Xiao X, Chou KC. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics. 2017;33(3):341–6.
21. Feng P, Ding H, Yang H, Chen W, Lin H, Chou KC. iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Mol Ther Nucleic Acids. 2017;7:155–63.
22. Liu B, Wang S, Long R, Chou KC. iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics. 2017;33(1):35–41.
23. Xu Q, Xiong Y, Dai H, Kumari KM, Xu Q, Ou HY, Wei DQ. PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm. J Theor Biol. 2017;417:1–7.
24. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol. 2011;273(1):236–47.
25. Sun WJ, Li JH, Liu S, Wu J, Zhou H, Qu LH, Yang JH. RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. Nucleic Acids Res. 2016;44(Database issue):D259–65.
26. Chou KC. Impacts of bioinformatics to medicinal chemistry. Med Chem. 2015;11(3):218–34.
27. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins. 2001;43(3):246–55.
28. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005;21(1):10–9.
29. Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst. 2015; 11(10):2620–34.
30. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-one: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015;43(W1):W65–71.
31. Brayet J, Zehraoui F, Jeansonleh L, Israeli D, Tahi F. Towards a piRNA prediction using multiple kernel fusion and support vector machine. Bioinformatics. 2014;30(17):i364.
32. Kamil E, Hashim M, Rosni A. Rare k-mer DNA: identification of sequence motifs and prediction of CpG Island and promoter. J Theor Biol. 2015;387:88–100.
33. Vinje H, Liland KH, Almøy T, Snipen L. Comparing K-mer based methods for improved classification of 16S sequences. BMC Bioinformatics. 2015;16(1):205.
34. Feng P, Ding H, Chen W, Lin H. Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. Mol BioSyst. 2016;12(11):3307.
35. Feng P, Jiang N, Liu N. Prediction of DNase I hypersensitive sites by using Pseudo nucleotide compositions. Thescientificworldjournal. 2014;2014:11):740506.
36. Liu B, Fang L, Long R, Lan X, Chou KC. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics. 2016;32(3):362.
37. Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou KC. PseKNC-general: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics. 2015;31(1):119–20.
38. Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal Biochem. 2014;456(1):53.
39. Tang YR, Chen YZ, Canchaya CA, Zhang Z. GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. Protein Engineering Design & Selection Peds. 2007;20(8):405–12.
40. Thangakani AM, Kumar S, Nagarajan R, Velmurugan D, Gromiha MM. GAP: towards almost 100 percent prediction for β-strand-mediated aggregating peptides with distinct morphologies. Bioinformatics. 2014;30(14):1983–90.
41. Xu Y, Ding YX, Ding J, Wu LY, Deng NY. Phogly–PseAAC: prediction of lysine phosphoglycerylation in proteins incorporating with position-specific propensity. J Theor Biol. 2015;379:10–5.
42. Chou KC. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem. 1993;268(23):16938–48.
43. Zhu X, Mitchell JC. KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins. 2011;79(9):2671–83.
44. Xiong Y, Liu J, Wei DQ. An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins. 2011;79(2):509–17.
45. Liu Z, Xiao X, Qiu WR, Chou KC. iDNA-methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal Biochem. 2015;474:69.

46. Wei C, Hui D, Feng P, Hao L, Chou KC. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7(13):16895.
47. Chen W, Feng P, Ding H, Lin H, Chou KC. iRNA-methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal Biochem. 2015;490:26.
48. Liu Z, Xiao X, Yu DJ, Jia J, Qiu WR, Chou KC. pRNAm-PC: predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. Anal Biochem. 2015;497:60–7.
49. Shao J, Dong X, Sau-Na T, Wang Y, Sai-Ming N. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. PLoS One. 2009;4(3):e4920.
50. Song J, Tan H, Shen H, Mahmood K, Boyd SE, Webb GI, Akutsu T, Whisstock JC. Cascleave: towards more accurate prediction of caspase substrate cleavage sites. Bioinformatics. 2010;26(6):752–60.
51. Jia C, Liu T, Chang AK, Zhai Y. Prediction of mitochondrial proteins of malaria parasite using bi-profile Bayes feature extraction. Biochimie. 2011;93(4):778.
52. Wang Y, Zhang Q, Sun MA, Guo D. High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. Bioinformatics. 2011;27(6):777.
53. Chou KC, Zhang CT. Prediction of protein structural classes. Crc Critical Reviews in Biochemistry. 1995;30(4):275–349.
54. Rodgers JL. The bootstrap, the jackknife, and the randomization test: a sampling taxonomy. Multivar Behav Res. 1999;34(4):441.
55. Dalgleish LI. Discriminant analysis: statistical inference using the jackknife and bootstrap procedures. Psychol Bull. 1994;116(3):498–508.
56. Chou KC. Using subsite coupling to predict signal peptides. Protein Eng. 2001;14(2):75.
57. Xu Y, Shao XJ, Wu LY, Deng NY, Chou KC. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ. 2013;1:e171.
58. Chen W, Feng PM, Lin H, Chou KC. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 2013;41(6):e68.
59. Liu B, Long R, Chou KC. iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics. 2016;32(16):2411–8.
60. Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics. 2016;32(20):3116–23.
61. Cheng X, Zhao SG, Lin WZ, Xiao X, Chou KC. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics. 2017;33(22):3524.
62. Chou KC. Some remarks on predicting multi-label attributes in molecular biosystems. Mol BioSyst. 2013;9(6):1092–100.
63. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74.
64. Lin H, Deng EZ, Ding H, Chen W, Chou KC. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 2014;42(21): 12961–72.
65. Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, Song J, Chou KC, Lithgow T. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics. 2017;33(17):2756–8.
66. Song J, Li F, Leier A, Marquez-Lago TT, Akutsu T, Haffari G, Chou KC, Webb GI, Pike RN, Hancock J. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Bioinformatics. 2018;34(4):684–7.
67. Cheng X, Xiao X, Chou KC. pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics. 2018;34(9):1448–56.
68. Noutahi E, Calderon V, Blanchette M, Lang FB, El-Mabrouk N. CoreTracker: accurate codon reassignment prediction, applied to mitochondrial genomes. Bioinformatics. 2017;33(21):3331–9.
69. Leclercq M, Diallo AB, Blanchette M. Prediction of human miRNA target genes using computationally reconstructed ancestral mammalian sequences. Nucleic Acids Res. 2017;45(2):556–66.
70. Cingolani P, Sladek R, Blanchette M. BigDataScript: a scripting language for data pipelines. Bioinformatics. 2015;31(1):10–6.
71. Qiao Y, Xiong Y, Gao H, Zhu X, Chen P. Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. BMC Bioinformatics. 2018;19(1):14.
72. Yuan Q, Gao J, Wu D, Zhang S, Mamitsuka H, Zhu S. DrugE-rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. Bioinformatics. 2016;32(12):i18–27.
73. Sukumar S, Zhu X, Ericksen SS, Mitchell JC. DBSI server: DNA binding site identifier. Bioinformatics. 2016;32(18):2853–5.
74. Zhu X, Xiong Y, Kihara D. Large-scale binding ligand prediction by improved patch-based method patch-Surfer2.0. Bioinformatics. 2015;31(5):707–13.
75. Zhu X, Ericksen SS, Mitchell JC. DBSI: DNA-binding site identifier. Nucleic Acids Res. 2013;41(16):e160.
76. Chou KC. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Curr Top Med Chem. 2017;17(21):2337–58.