Molecular Therapy
Nucleic Acids
Original Article

Is There Any Sequence Feature in the RNA Pseudouridine Modification Prediction Problem?

Lijun Dou,[1,2,6] Xiaoling Li,[3,6] Hui Ding,[4] Lei Xu,[5] and Huaikun Xiang[1]

[1]School of Automotive and Transportation Engineering, Shenzhen Polytechnic, Shenzhen, China; [2]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China; [3]Department of Oncology, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China; [4]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China; [5]School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China

Pseudouridine (Ψ) is the most abundant RNA modification and has been found in many kinds of RNAs, including snRNA, rRNA, tRNA, mRNA, and snoRNA. Thus, Ψ sites play a significant role in basic research and drug development. Although some experimental techniques have been developed to identify Ψ sites, they are expensive and time consuming, especially in the post-genomic era with the explosive growth of known RNA sequences. Thus, highly accurate computational methods are urgently required to quickly detect the Ψ sites on uncharacterized RNA sequences. Several predictors have been proposed using multifarious features, but their evaluated performances are still unsatisfactory. In this study, we first identified Ψ sites for *H. sapiens*, *S. cerevisiae*, and *M. musculus* using the sequence features from the bi-profile Bayes (BPB) method based on the random forest (RF) and support vector machine (SVM) algorithms, where the performances were evaluated using 5-fold cross-validation and independent tests. It was found that the SVM-based accuracies were 3.55% and 5.09% lower than the iPseU-CUU predictor for the H_990 and S_628 datasets, respectively. Almost the same-level results were obtained for M_994 and an independent H_200 dataset, even showing a 5.0% improvement for S_200. Then, three different kinds of features, including basic Kmer, general parallel correlation pseudo-dinucleotide composition (PC-PseDNC-General), and nucleotide chemical property (NCP) and nucleotide density (ND) from the iRNA-PseU method, were combined with BPB to show their comprehensive performances, where the effective features are selected by the max-relevance-max-distance (MRMD) method. The best evaluated accuracies of the combined features for the S_628 and M_994 datasets were achieved at 70.54% and 72.45%, which were 2.39% and 0.65% higher than iPseU-CUU. For the S_200 dataset, it was also improved 8% from 69% to 77%. However, there was no obvious improvement for H. sapiens, which was evaluated as approximately 63.23% and 72.0% for the H_990 and H_200 datasets, respectively. The overall performances for Ψ identification using BPB features as well as the combined features were not obviously improved. Although some kinds of feature extraction methods based on the RNA sequence information have been applied to construct the predictors in previous studies, the corresponding accuracies are generally in the range of 60%–70%. Thus, researchers need to reconsider whether there is any sequence feature in the RNA Ψ modification prediction problem.

INTRODUCTION

Pseudouridine (Ψ) is the most prevalent post-transcriptional modification, and it has been widely found in a series of biological and cellular processes.[1,2] Recent studies have demonstrated that Ψ sites exist in many kinds of RNAs, such as small nuclear RNA (snRNA), rRNA, tRNA, mRNA, and small nucleolar RNA (snoRNA).[3–11] Thus, the Ψ site plays a crucial role in biological research and drug development. More specifically, Ψ is an isomer of uridine catalyzed by the Ψ synthase (PUS) that removes the uridine residue's base from its sugar, followed by "rotating" it 180° along the N3-C6 axis, and subsequently reattaches the base's 5-carbon to the 1'-carbon of the sugar.[12]

Although there are several experimental methods based on the high-throughput techniques that have been developed to recognize the Ψ modifications, they are both costly and time consuming.[13–17] In addition, researchers are facing an explosive increase of RNA data in the post-genomic age.[18–30] Therefore, intelligent computational approaches are highly desirable to predict Ψ sites on RNA sequences.

To the best of our knowledge, six predictors have been reported to identify Ψ sites. Specifically, Panwar and Raghava[31] first proposed the tRNAmod model to predict Ψ sites in tRNA. Li et al.[32] then developed the PPUS method based on the support vector machine (SVM) to identify PUS-specific Ψ sites. Later, Chen et al.[33] provided the iRNA-PseU predictor, and He et al.[34] introduced the PseUI predictor, which are both based on the SVM classifier. In addition, Tahir et al.[35]

Received 8 September 2019; accepted 11 November 2019;
https://doi.org/10.1016/j.omtn.2019.11.014.

[6]These authors contributed equally to this work.

Correspondence: Lei Xu, School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China.
E-mail: csleixu@szpt.edu.cn
Correspondence: Huaikun Xiang, School of Automotive and Transportation Engineering, Shenzhen Polytechnic, Shenzhen, China.
E-mail: xianghuaikun@szpt.edu.cn

Table 1. Results of the Proposed iRNA-PseU, PseUI, iPseU-CUU, and XG-PseU Predictors for Training Datasets H_990, S_628, and M_944 and Testing Datasets H_200 and S_200

Predictors	Training Datasets	Acc (%)	MCC	Sn (%)	Sp (%)	Testing Datasets	Acc (%)	MCC	Sn (%)	Sp (%)
iRNA-PseU[a]	H_990	60.4	0.21	61.01	59.8	H_200	65.00	0.30	60.00	70.00
PseUI[b]		64.24	0.28	64.85	63.64		65.50	0.31	63.00	68.00
iPseU-CUU[c]		66.68	0.34	65.00	68.78		69.00	0.40	77.72	60.81
XG-PseU[d]		65.44	0.31	63.64	67.24		67.00	0.34	67.00	67.00
iRNA-PseU[a]	S_628	64.49	0.29	64.65	64.33	S_200	73.00	0.46	81.00	65.00
PseUI[b]		66.56	0.33	62.1	71.02		68.50	0.37	72.00	65.00
iPseU-CUU[c]		68.15	0.37	66.36	70.45		73.50	0.47	68.76	77.82
XG-PseU[d]		68.15	0.37	66.84	69.45		71.00	0.42	75.00	67.00
iRNA-PseU[a]	M_944	69.07	0.38	73.31	64.83					
PseUI[b]		70.44	0.41	74.58	66.31					
iPseU-CUU[c]		71.81	0.44	74.49	69.11					
XG-PseU[d]		72.03	0.45	76.48	67.57					

[a]The predictor developed by Chen et al.[33]
[b]The predictor proposed by He et al.[34]
[c]The predictor constructed by Tahir et al.[35]
[d]The predictor constructed by Liu et al.[36]

built the iPseU-CUU model based on the convolution neural network (CNN). Most recently, Chen et al.[36] proposed an eXtreme Gradient Boosting (xgboost)-based method (XG-PseU). It should be noted that the same datasets, built by Chen et al.,[33] were applied in the three studies (iRNA-PseU, PseUI, and iPseU-CUU) to build the predictors, including the benchmark training datasets (H_990, S_628, and M_944) and the independent testing datasets (H_200 and S_200). Here, H, S, and M represent the RNA samples for *H. sapiens*, *S. cerevisiae*, and *M. musculus*, while 990, 628, 944, and 200 indicate the corresponding sample numbers in each dataset. Thus, we used the datasets mentioned earlier in this article for convenient comparisons. The performances of the four predictors (iRNA-PseU, PseUI, iPseU-CUU, and XG-PseU) are listed in Table 1, where the XG-PseU results for independent datasets were obtained by the web server at http://www.bioml.cn. The jackknife test, 5-fold cross-validation, and 10-fold cross-validation are used for the iRNA-PseU, PseUI/iPseU-CUU, and XG-PseU models, respectively. It can be seen that their overall performances are gradually improved through the scientists' efforts. Taking H_990 as an example, the accuracies have been improved by 6.28% from 60.40% (iRNA-PseU) to 61.24% (PseUI) and to 66.68% (iPseU-CUU). However, it must be noted that these predictive accuracies are still unsatisfactory.

As a crucial step toward building a machine-learning-based predictor, feature extraction becomes a particularly important process. Several sequence representation methods have been used in previous works to obtain feature vectors. For example, a hybrid approach of the binary profile of patterns (BPP) and structural information is applied in the tRNAmod.[31] In addition, the PPUS model uses the nucleotides around Ψ as the features to identify.[32] For the successful iRNA-PseU method, dinucleotide chemical properties (DCP) and nucleotide density (ND) are incorporated for identification.[33] For the PseUI, the effective features are selected from five different feature extraction techniques using the sequential forward-feature-selection method, including nucleotide composition (NC), dinucleotide composition (DNC), pseudo-dinucleotide composition (PseDNC), position-specific nucleotide propensity (PSNP), and position-specific dinucleotide propensity (PSDP).[37,38] For the iPseU-CUU method, the features are obtained automatically by a CNN model based on a deep learning machine, which is widely used in bioinformatics.[39–42] Furthermore, two additional feature extraction techniques, n-gram and multivariate mutual information (MMI), are also applied for the machine learning approach by the SVM method, where they still give a low accuracy (Acc).[35] For the newly reported XG-PseU predictor, six feature extraction techniques are used, namely, NC, DNC, trinucleotide composition (TNC), nucleotide chemical property (NCP), ND, and one-hot encode (one hot).

At the same time, the identification of many types of RNA modifications using the machine-learning-based computational approaches shows the excellent performance, including for N6-methyladenosine (m6A),[43–45] 5-methylcytosine (m5C),[46–53] N1-methyladenosine (m1A),[54–56] and so forth. The related kinds of computational models used for these purposes have been summarized in a review,[57] in which the recently reported overall accuracies are basically above 90%. In particular, the SVM-based iRNA(m6A)-PseDNC model demonstrates an Acc of 91.24% of 10-fold cross-validation for m6A identification for *S. cerevisiae*.[43] For the m5C site, the recently developed iRNA-m5C predictor by the Random Forest (RF) algorithm shows a jackknife test Acc up to 92.9% for *H. sapiens*.[52] For m1A, the SVM-based iRNA-3typeA method obtains a jackknife validation Acc of 99.13% on *H. sapiens* and 98.73% for *M. musculus*.[56] However, as mentioned earlier, the evaluated accuracies of Ψ site identification of different models are basically only 60%–70%, where there is still a large amount of improvement possible.

Table 2. Comparison of Our Results based on the RF and SVM Methods Using the BPB Features with the iPseU-CUU Predictor

Predictors	Training Datasets	Acc (%)	MCC	Sn (%)	Sp (%)	Testing Datasets	Acc (%)	MCC	Sn (%)	Sp (%)
iPseU-CNN[a]	H_990	66.68	0.34	65.00	68.78	H_200	69.00	0.40	77.72	60.81
RF[b]		58.28	0.17	60.00	56.57		59.00	0.18	61.00	57.00
SVM[c]		63.13	0.26	64.04	62.22		74.00	0.48	78.00	70.00
iPseU-CNN[a]	S_628	68.15	0.37	66.36	70.45	S_200	73.50	0.47	68.76	77.82
RF[b]		62.58	0.25	63.69	61.46		74.00	0.48	70.00	78.00
SVM[c]		63.06	0.27	52.87	73.25		73.00	0.49	60.00	86.00
iPseU-CNN[a]	M_944	71.81	0.44	74.49	69.11					
RF[b]		67.27	0.35	69.28	65.25					
SVM[c]		71.40	0.43	75.00	67.80					

[a]The predictor proposed by Tahir et al.[35]
[b]The RF-based predictor using BPB features.
[c]The SVM-based predictor using BPB features.

We noticed that a predictor called "KELMPSP" reported a better performance, where the accuracies for the H_990, S_628, M_949, H_200, and S_200 datasets are up to 74.55%, 85.53%, 79.45%, 72.5%, and 76.00%, respectively.[58] In this method, the kernel extreme learning machine (KELM) algorithm is applied, where the final features are obtained by combining NCP, nucleotide concentrations, and position-specific mononucleotide, dinucleotide, and trinucleotide propensity characteristics. However, the related web server at http://39.105.77.161:8890/KELMPSP is no longer available.

In this paper, we first applied the bi-profile Bayes method (BPB)[59] to extract the RNA sequence features to identify the Ψ sites. Two algorithms, RF and SVM, were both used to construct the models, where the performances were evaluated by 5-fold cross-validation and independent tests. Then, we incorporated three different features with BPB to show their comprehensive performance, including basic Kmer (Kmer),[60] general parallel correlation pseudo-dinucleotide composition (PC-PseDNC-General) generated from the web server Pse-in-One,[61] and NCP with ND (NCP+ND). Also, high-quality features were selected using the MRMD[62] method to predict the Ψ sites.

RESULTS AND DISCUSSION

Performance of the BPB Features

First, we extracted the RNA features using the BPB method for Ψ site prediction. The performances were evaluated over the 5-fold cross-validation for the benchmark datasets H_990, S_628, and M_944 and independent dataset for H_200 and S_200. Table 2 gives a comparison of our results using BPB features with the iPseU-CUU predictor, where RF and SVM indicate the results from the RF and SVM classifiers, respectively. It is obvious that the SVM generally performed better than the RF. Specifically, the accuracies of the SVM method were improved 4.85% and 4.13% for the training datasets H_990 and M_994, respectively. For the independent dataset test, the Acc and MCC were obviously increased by 15.0% and 0.3 for the H_200 datasets. However, for the S_628 datasets, the Acc was only increased from 62.58% to 63.06%. Here, we found that, although the specificity (Sp) increased from 61.46% to 73.25%, the sensitivity (Sn) actually declined from 63.69% to 52.87%, which means that one half of the positive samples were incorrectly predicted to be the false one. Similar results can also be observed in S_200. From the comparison between RF and SVM, it can be concluded that the SVM algorithm is more efficient than RF for the Ψ prediction of RNA sequences for *H. sapiens* and *M. musculus*.

Compared with the iPseU-CUU model, the SVM method showed accuracies reduced by 3.55% and 5.09% for the first two training datasets H_990 and S_628. Almost the same results could be found for the training dataset M_994 and independent dataset H_200, where our results are only 0.5% lower than that for iPseU-CUU. Additionally, the SVM model performed better for S_200, where the Acc and MCC were both improved approximately 5.0% and 0.08, respectively. In general, the SVM algorithm appears to be a better choice than RF for the Ψ modification prediction using BPB features alone, which can be clearly found in Figure 2. However, it must be noted that the overall performance of the SVM method here is unsatisfactory, even lower than that of the latest predictor, iPseU-CUU, for the two datasets H_990 and S_628.

Performance of the BPB Features Combining Other Features

For a better performance, three different kinds of features were also investigated: Kmer,[60] PC-PseDNC-General, and NCP+ND from the iRNA-PseU method.[33] At the same time, those features were further combined with BPB to achieve a better result, where the MRMD method was applied to select the important features for experiments.[62] Table 3 lists the results of different feature selection for H_990 datasets using the RF method (left) and SVM method (right).

The first six rows give the performance of each type of feature, including BPB, Kmer (k = 2, 3, 4), PC-PseDNC-General (λ = 6,w = 0.99), and NCP+ND. For the Kmer method,[60] three results with k = 2, 3, and 4 are listed. It can be found that the Kmer(3) shows consistent results, where the accuracies are 58.79% and 59.70% for the RF and SVM classifiers, respectively. In the PC-PseDNC-General method,[63,64] several parameters have been tested, and better results

Table 3. Results of Feature Selection for the H_990 Dataset Using the RF and SVM Methods

	RF				SVM			
Feature Subset	Acc (%)	MCC	Sn (%)	Sp (%)	Acc (%)	MCC	Sn (%)	Sp (%)
BPB	58.28	0.17	60.00	56.57	63.13	0.26	64.04	62.22
Kmer(2)	55.76	0.12	53.13	58.38	60.00	0.23	41.82	78.18
Kmer(3)	58.79	0.18	58.59	58.99	59.70	0.20	53.94	65.45
Kmer(4)	58.59	0.17	59.39	57.78	57.27	0.15	56.57	57.98
PC-PseDNC-General (6,0.99)	58.59	0.17	56.57	60.61	57.78	0.16	49.49	66.06
NCP+ND	56.87	0.14	57.37	56.36	60.34	0.21	60.40	60.28
BPB+Kmer(3)	60.40	0.21	60.61	60.20	63.23[a]	0.27[a]	61.01[a]	65.45[a]
BPB+PC-PseDNC-General (6,0.99)	61.72	0.23	59.39	64.04	62.93	0.26	61.62	64.24
BPB+NCP+NP	61.11	0.22	62.83	59.39	61.11	0.22	58.79	63.43
BPB+PC-PseDNC-General (6,0.99) + Kmer(3)	61.01	0.22	59.39	62.63	62.73	0.25	61.82	63.64

[a]Performance with maximum accuracy.

are obtained with the parameters λ = 6 and w = 0.99. The corresponding SVM-based Acc (57.78%) is slightly lower than the RF-based Acc (58.59%). We also repeated the work by Chen et al.[54] (NCP+ND) with the 5-fold cross-validation, which obtained an Acc of 60.34% compared to the reported jackknife results (60.40%). From the discussion earlier, the performances of the single features are all lower than that of the latest iPseU-CUU predictor (66.68%), among which the BPB features give the best Acc (63.13%) by the SVM method.

Further, we combined the Kmer, PC-PseDNC-General, and NCP+ND features with the BPB, and the final useful features for model constructing were selected using the MRMD method.[62] There were four results for the combined features listed in Table 3 for the H_990 datasets, including BPB+Kmer(3), BPB+PC-PseDNC-General(6,0.99), BPB+NCP+ND, and BPB+PC-PseDNC-General(6,0.99)+Kmer(3). It can be found that the combined results are generally improved 2%–3% over the single BPB results by the RF method, where the best combination with a maximum Acc 61.72% is BPB+PC-PseDNC-General(6,0.99). However, there is no obvious improvement for the SVM-based method and even a 2.02% decrease for the BPB+NCP+ND combination. The feature combination BPB+Kmer(3) showed the best performance by the SVM method, which gave 63.23% Acc, 0.27 MCC, 61.01% Sn, and 65.45% Sp. Applying this model to an independent test for H_200, the obtained Acc, MCC, Sp, and Sn were 72.00%, 0.46, 82%, and 62%, respectively. Compared to the iPseU-CUU predictor, 3% and 0.06 improvement for the Acc and MCC were found.

Tables 4 and 5 list the same results as in Table 3 but for the datasets S_628 and M_944, respectively. For S_628, the feature combination BPB+PC-PseDNC-General(2,0.1)+Kmer(4) gave the best performance, where the ACc and MCC were obviously improved by 7.48% and 0.14, respectively. When compared with the iPseU-CNN model, the evaluated Acc shows 2.39% improvement. Finally, the combined model was tested using the independent dataset S_200, where the Acc, MCC, Sn, and Sp are 77.00%, 0.54, 75%, and 79%, respectively. It can be seen that there were 3.5% and 0.07 improvement for the Acc and MCC compared to those for the iPseU-CUU model. For M_994, the best performance was given by feature combination BPB+Kmer(3), for which the Acc was 72.46%, MCC was 0.45, Sn was 75.85%, and Sp was 69.07%. Compared with the Acc of the iPseU-CUU method, there was only 0.65% improvement obtained. Figure 3 shows an intuitive comparison of the evaluated performance of the iPseU-CUU (orange bars), XG-PseU (green bars), and the constructed SVM-based model using the combined features in this work (blue bars).

Finally, we investigated several kinds of features from the two state-of-the-art tools iLearn[65] and BioSeq-Analysis2.0[66] for *H. sapiens*, including Mismatch (k = 2,3,4), subsequence (k = 2,3,4), the enhanced nucleic acid composition (ENAC) with the sequence window 5, electron-ion interaction pseudopotentials of trinucleotide (EIIP), electron-ion interaction pseudopotentials of trinucleotide (PseEIIP), binary encoding (BE), dinucleotide-based auto covariance (DAC), dinucleotide-based cross covariance (DCC), and dinucleotide-based auto-cross covariance (DACC). It was found that the average Acc of subsequence, ENAC, and autocorrelation features using the SVM algorithm is approximately 55%. The evaluated performances of other features as well as the combined features with the best performances BPB+Kmer(3) are listed in Table 6. It can be seen that the feature combination BPB+Kmer(3)+EIIP gives the accuracies 63.33% and 75% on the H_990 and H_200 datasets, which are improved by 0.1% and 3% compared with our original feature combination BPB+Kmer(3), respectively.

Conclusions

Ψ identification plays an important role in academic research and drug development. In this study, we first extracted the RNA features using the BPB method[59,67–69] for Ψ site prediction, which gives the RNA sequence information from both positive and negative training samples. The evaluated accuracies using the SVM method are 3.55% and 5.09% lower than the iPseU-CUU[35] for the H_990 and S_628 datasets. Almost the same results and 5.0% improvement were obtained

Table 4. Results of Feature Selection for the S_628 Dataset Using the RF and SVM Methods

	RF				SVM			
Feature Subset	Acc (%)	MCC	Sn (%)	Sp (%)	Acc (%)	MCC	Sn (%)	Sp (%)
BPB	62.58	0.25	63.69	61.46	63.06	0.27	52.87	73.25
Kmer (k = 2)	58.12	0.16	58.28	57.96	61.78	0.24	64.33	59.24
Kmer (k = 3)	60.35	0.21	62.10	58.60	61.78	0.24	66.56	57.01
Kmer (k = 4)	59.71	0.19	62.74	56.69	64.97	0.30	67.52	62.42
PC-PseDNC-General (2, 0.11)	58.76	0.18	61.78	55.73	61.15	0.22	64.01	58.28
NCP+ND	60.83	0.22	62.74	58.92	60.99	0.22	57.01	64.97
BPB+Kmer (k = 4)	64.01	0.28	64.33	63.69	68.15	0.36	66.56	69.75
BPB+PC-PseDNC-General (2, 0.11)	62.90	0.26	63.38	62.42	66.08	0.33	57.64	74.52
BPB+NCP+ND	62.74	0.26	65.61	59.87	61.78	0.24	56.37	67.20
BPB+PC-PseDNC-General (2, 0.11) + Kmer(4)	64.49	0.29	65.92	63.06	70.54[a]	0.41[a]	69.43[a]	71.66[a]

[a]Performance with maximum accuracy.

for M_994, H_200, and S_200, respectively. Then, we combined three kinds of features—Kmer, PC-PseDNC-General, and NCP+ND, where the useful features were further selected by the MRMD method.[62] The final accuracies of the combined features using the SVM classifier were achieved at 70.54% and 72.45% for S_628 and M_994, respectively. The predicted Acc of independent S_200 was also improved from 69.0% (BPB features alone) to 77.0% (combined features).

It can be concluded that there are some improvements for the *S. cerevisiae* and *M. musculus* using the combined features by the SVM classifier. However, including the six existing predictors, the general accuracies are still 60%–70%, which needs to be further improved for biologist usage. It is clearly known that many kinds of feature extraction methods have been applied to encode RNA sequences to identify Ψ modification, including BPB, Kmer, PC-PseDNC-General, NCP, ND, Mismatch, subsequence, ENAC, EIIP, PseEIIP, BE, DAC, DCC, and DACC in this paper, as well as PSNP, PSDP, and so forth. In addition, many machine-learning-based computational methods[70–72] for the identification of many types of RNA methylations have shown excellent performance (with an Acc of approximately 90%), including m6A, m5C, m1A, and so forth. Thus, the researchers need to reconsider whether there is any sequence feature in the RNA Ψ modification prediction problem. There may be other methods to identify Ψ modification sites that have better performance.

MATERIALS AND METHODS

In this study, we use the datasets built by Chen et al.[33] from RMBase,[73] including training datasets H_990, S_628, and M_990, and independent testing datasets H_200 and S_200 for *H. sapiens*, *S. cerevisiae*, and *M. musculus*, respectively. Here, the BPB features alone as well as the combination of three other kinds of features (Kmer, PC-PseDNC-General, and NCP+ND) using the MRMD method are prepared. Then, two classifiers, RF and SVM, are used separately for model construction. The schematic flowchart of this work is shown in Figure 1.

Feature Extraction Methods

BPB

BPB is an effective feature extraction approach that has been successfully applied in bioinformatics with good performance.[59,67–69,74–77] It can obtain comprehensive sequence information from not only positive but also negative RNA samples. Considering the RNA sequence $S = R_1R_2R_3..R_l$, the associated BPB feature vector is written as

$$V = \left(p_1, p_2, \ldots, p_l, n_{l+1}, n_{l+2}, \ldots, n_{2l}\right)^T, \quad \text{(Equation 1)}$$

where $(p_1, p_2, \ldots, p_l)$ and $(n_{l+1}, n_{l+2}, \ldots, n_{2l})$ represent the corresponding nucleotide frequency at each position in positive and negative datasets, respectively. Thus, the BPB features for model training can well reflect the positive and negative position-specific information.

Kmer

Kmer is a common method used to give RNA sequence information, where the feature vector is obtained from the frequencies of *k*-neighboring nucleotides.[60,66] The Kmer features are available at the powerful web server Pse-in-one (http://bioinformatics.hitsz.edu.cn/Pse-in-One/RNA/Kmer/).

PC-PseDNC-General

Similarly, the PC-PseDNC-General features[63,64,78] can also be obtained at Pse-in-one (http://bioinformatics.hitsz.edu.cn/Pse-in-One/RNA/PC-PseDNC-General/), where 22 alternative physicochemical properties are provided to generate the pseudo-dinucleotide composition. The corresponding RNA features can be written as:

$$V = \left(x_1x_2\cdots x_{16}x_{16+1}\cdots x_{16+\lambda}\right)^T \quad \text{(Equation 2)}$$

with

Table 5. Results of Feature Selection for the M_944 Dataset Using the RF and SVM Methods

	RF				SVM			
Feature Subset	Acc (%)	MCC	Sn (%)	Sp (%)	Acc (%)	MCC	Sn (%)	Sp (%)
BPB	68.54	0.37	69.28	67.80	71.40	0.43	75.00	67.80
Kmer(2)	52.22	0.04	54.45	50.00	56.78	0.14	61.65	51.91
Kmer(3)	55.51	0.11	57.42	53.60	59.22	0.18	60.81	57.63
Kmer(4)	56.04	0.12	58.05	54.03	58.37	0.17	59.96	56.78
PC-PseDNC-General (2, 0.1)	53.07	0.06	56.14	50.00	57.84	0.16	64.41	51.27
NCP+ND	67.58	0.35	70.34	64.83	68.01	0.36	69.49	66.53
BPB+Kmer(3)	67.37	0.35	71.61	63.14	72.46[a]	0.45[a]	75.85[a]	69.07[a]
BPB+PC-PseDNC-General (2, 0.1)	67.58	0.35	70.97	64.19	71.40	0.43	73.52	69.28
BPB+NCP+ND	68.43	0.37	71.82	65.04	68.11	0.36	69.70	66.53
BPB+PC-PseDNC-General (2, 0.11) + Kmer(3)	68.33	0.37	72.67	63.98	71.72	0.44	75.00	68.43

[a]Performance with maximum accuracy.

$$x_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} & (1 \le k \le 16) \\ \dfrac{\omega \theta_{u-16}}{\sum_{i=1}^{16} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} & (16+1 \le k \le 16+\lambda) \end{cases} . \quad \text{(Equation 3)}$$

Here, $f_k (k = 1, 2, \ldots, 16)$ indicates the normalized occurrence frequency of the 16 dinucleotides; $\omega (0 \le \omega \le 1)$ is the weight factor; and θ_j is the j-tier correlation factor demonstrating the sequence-order correlations between all of the most contiguous dinucleotides along a given RNA sequence, where parameter λ gives the highest counted rank (or tier). It can be further expressed as:

$$\theta_j = \frac{1}{L-j-1} \sum_{i=1}^{L-j-1} C(D_i, D_{i+j}) (j = 1, 2, \cdots, \lambda; \lambda \le l), \quad \text{(Equation 4)}$$

where $C(D_i, D_{i+j})$ is called the correlation function formulated as

$$C_{i,i+j} = \frac{1}{u} \sum_{g=1}^{u} \left[P_g(D_i) - P_g(D_{i+j}) \right]^2. \quad \text{(Equation 5)}$$

Here, u indicates the number of physicochemical properties investigated and $P_g(D_i)$ and $P_g(D_{i+j})$ are the associated values of the gth property for the dinucleotides D_i at position i and D_{i+j} at $i+j$, respectively.

NCP+ND

In the iRNA-PseU method, the feature vectors are obtained by incorporating three NCPs (ring structure, hydrogen bond, and functional group) and accumulated occurrence frequency.[33] The related chemical properties are described as follows: A and G purines with two rings encoded as 1; C and U pyrimidines with one ring, as 0; the strong hydrogen bonds formed between C and G, as 1 when constructing secondary structures; the weak hydrogen bonds between A and U, as 0; the amino groups A and C, as 1; and the keto groups G and U, as 0. Then, the four nucleotides A, C, G, and U can be encoded as (1, 1, 1), (0, 0, 1), (1, 0, 0), and (0, 1, 0), respectively. In addition, the nucleotide density d_i is defined as

$$d_i = \frac{1}{|N_i|} \sum_{j=1}^{l} f(n_j), f(n_j) = \begin{cases} 1 & if\ n_j = \text{A, C, G or } U \\ 0 & otherwise \end{cases}, \quad \text{(Equation 6)}$$

where $|N_i|$ is the length of the ith prefix string $n_1, n_2, \ldots, n_i$. Finally, the RNA sequence can be simply represented by a $4l$-dimensional vector according to the formulation of PseKNC.

Classifiers and Cross-Validation

RF

RF is a widely used algorithm in prediction problems that effectively combines ensemble tree-structured classifiers.[79–87] It is usually applied to research with a very large number of feature vectors. This classifier consists of hundreds of decision trees, and the final prediction is obtained by major votes. In this article, we used the RF method implemented on the Weka data mining suite with the default parameters for analysis.[88]

SVM

SVM is a successful machine learning algorithm based on statistical learning theory,[89–96] which has been widely applied in bioinformatics and computational biology.[90,97–108] In this method, the original input data are transformed into a higher dimensional feature space (Hilbert space), and then the optimal separating hyperplanes are determined. Here, the LIBSVM package v.3.21[109] was used to implement the SVM, where the radial basis kernel function (RBF) was chosen to obtain the best classification hyperplane. The related regularization parameter C and kernel width γ were determined

Table 6. Results of Feature Selection for H_990 and H_200 Datasets Using Several Kinds of Features from iLearn and BioSeq-Analysis 2.0

	H_990				H_220			
Feature Subset	Acc	MCC	Sn	Sp	Acc	MCC	Sn	Sp
BE	60.10	0.20	58.79	61.41	66.50	0.33	64.00	69.00
Mismatch (3)	60.81	0.22	57.37	64.24	59.50	0.19	58.00	61.00
EIIP	57.37	0.15	54.55	60.20	58.00	0.16	56.00	60.00
PseEIIP	58.99	0.18	54.75	63.23	58.00	0.16	55.00	61.00
BE	60.10	0.20	58.79	61.41	66.50	0.33	64.00	69.00
BPB+Kmer(3)+EIIP[a]	63.33	0.27	62.63	64.04	75.00	0.51	81.00	69.00
BPB+Kmer(3)+PseEIIP	63.13	0.26	61.01	65.25	70.50	0.43	82.00	59.00
BPB+Kmer(3)+BE	60.91	0.22	58.99	62.83	68.00	0.36	69.00	67.00
BPB+Kmer(3)+mismatch(3)	61.11	0.22	56.77	65.45	60.20	0.20	61.00	59.41
BPB+Kmer(3)+EIIP+mismatch(3)	61.21	0.23	56.97	65.45	60.20	0.20	61.00	59.41

[a]All values in this row indicate performance with maximum accuracy.

through the optimization procedure, using the default grid search approach written as:

$$\begin{cases} 2^{-5} \le C \le 2^{15} & \text{with step of } 2 \\ 2^{-15} \le \gamma \le 2^{3} & \text{with step of } 2 \end{cases}. \qquad \text{(Equation 7)}$$

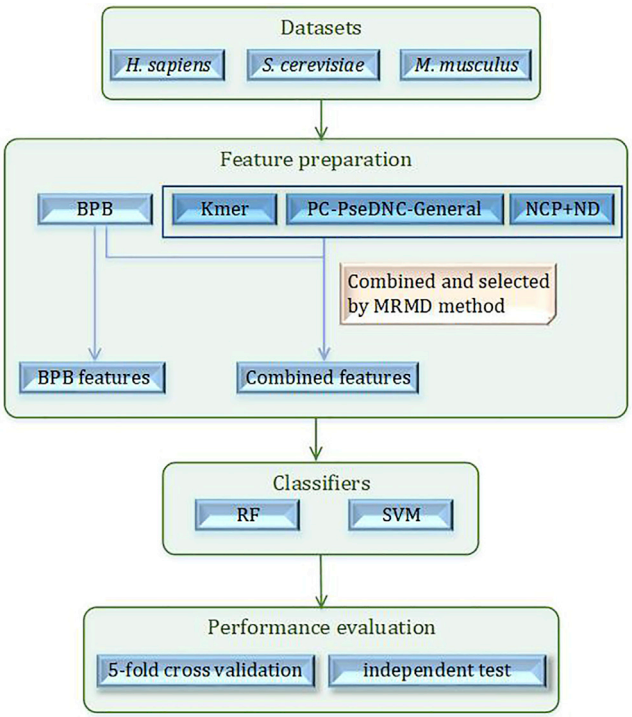

Figure 1. Flowchart of Constructed Predictors for Ψ Identification Using the BPB Features

5-Fold Cross-Validation

Although the jackknife test is effective and stable and has been applied in the iRNA-PseU[33] and PseUI,[34] it is a very time-consuming process. On the other hand, the predictor iPseU-CUU[35] uses 5-fold cross-validation to evaluate performance. Therefore, we chose 5-fold cross-validation on the benchmark datasets for a convenient comparison. Specifically, the benchmark datasets are equally divided into five subsets separately. Then, the four subsets are used to train the model and the remaining one to test. This process is repeated five times when all subsets are applied once for testing. The final performances are an average value of all five testing experiments.[110]

MRMD

Feature selection aims to select a subset of features by removing redundancy and keeping the most discriminative features.[111–114]

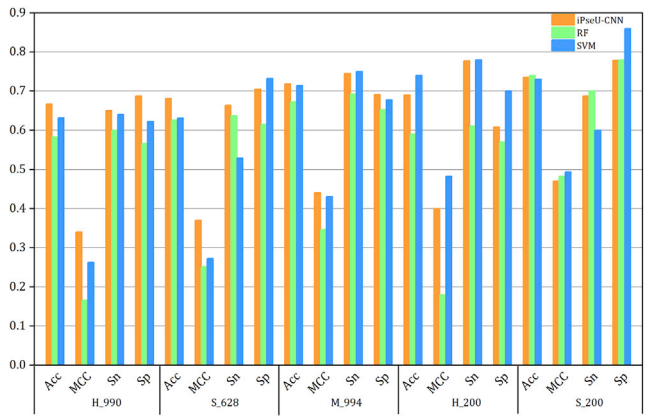

Figure 2. This Histogram Shows the Results of the iPseU-CUU Predictor and the Constructed Model Based on the RF and SVM Classifiers Using the BPB Features

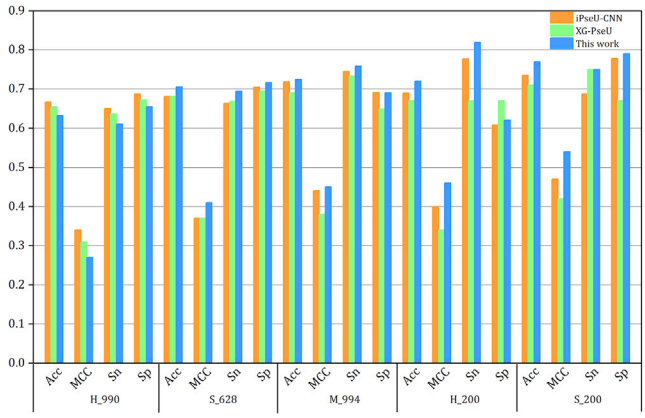

Figure 3. Comparisons of the Evaluated Performance of Predictors iPseU-CUU ,XG-PseU and the Constructed Model Using the Combined Features in This Work

MRMD[62] is an effective feature selection method to reduce dimensionalities of feature vectors, where the Acc and stability of feature ranking and prediction tasks are both considered. As Xu et al.'s[115] related work shows, the performances are improved based on the selected features using the MRMD method. In this method, the features with the maximum relevance and distance are selected as the ultimate sub-feature set for experiments.

Evaluation Parameters

The performance of the constructed models is frequently evaluated using Sn, Sp, Acc, and Matthews correlation coefficient (MCC), which are expressed as:[116–121]

$$\begin{cases} Sn = 1 - \dfrac{N_-^+}{N^+} & 0 \le Sn \le 1 \\ Sp = 1 - \dfrac{N_+^-}{N^-} & 0 \le Sp \le 1 \\ Acc = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le Acc \le 1 \\ MCC = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le MCC \le 1 \end{cases}, \quad \text{(Equation 8)}$$

where N^+ and N^- represent the total number of positive and negative RNA samples considered, in which the incorrectly predicted samples are indicated by N_-^+ and N_+^-, respectively.

AUTHOR CONTRIBUTIONS

L.X. and H.X. conceived the idea and designed the overall research. L.D. constructed the predictors, evaluated the performance, and drafted the manuscript. X.L. and H.D. helped to revise the paper; Both authors read, critically revised and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China (no. 61902259), the Natural Science Foundation of Guangdong Province (grant no. 2018A0303130084), and the Scientific Research Foundation in Shenzhen (JCYJ20170818100431895, JCYJ20180305163701198, and JCYJ20180306172207178).

REFERENCES

1. Hudson, G.A., Bloomingdale, R.J., and Znosko, B.M. (2013). Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. RNA *19*, 1474–1482.
2. Sloan, K.E., Warda, A.S., Sharma, S., Entian, K.D., Lafontaine, D.L.J., and Bohnsack, M.T. (2017). Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. RNA Biol. *14*, 1138–1152.
3. Ge, J., and Yu, Y.T. (2013). RNA pseudouridylation: new insights into an old modification. Trends Biochem. Sci. *38*, 210–218.
4. Han, S., Liang, Y., Ma, Q., Xu, Y., Zhang, Y., Du, W., Wang, C., and Li, Y. (2018). LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. Brief. Bioinform. Published online July 31, 2018. https://doi.org/10.1093/bib/bby065.
5. Lu, S.J., Xie, J., Li, Y., Yu, B., Ma, Q., and Liu, B.Q. (2019). Identification of lncRNAs-gene interactions in transcription regulation based on co-expression analysis of RNA-seq data. Math. Biosci. Eng. *16*, 7112–7125.
6. Cantara, W.A., Crain, P.F., Rozenski, J., McCloskey, J.A., Harris, K.A., Zhang, X., Vendeix, F.A., Fabris, D., and Agris, P.F. (2011). The RNA modification database, RNAMDB: 2011 update. Nucleic Acids Res. *39*, D195–D201.
7. Boccaletto, P., Machnicka, M.A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T.K., de Crécy-Lagard, V., Ross, R., Limbach, P.A., Kotter, A., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res. *46* (D1), D303–D307.
8. Tang, J., Fu, J., Wang, Y., Luo, Y., Yang, Q., Li, B., Tu, G., Hong, J., Cui, X., Chen, Y., et al. (2019). Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. Mol. Cell. Proteomics *18*, 1683–1699.
9. Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. *47* (D1), D140–D144.
10. Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). *OAHG*: an integrated resource for annotating human genes with multi-level ontologies. Sci. Rep. *6*, 34820.
11. Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2019). HOGMMNC: a higher order graph matching with multiple network constraints model for gene-drug regulatory modules identification. Bioinformatics *35*, 602–610.
12. Charette, M., and Gray, M.W. (2000). Pseudouridine in RNA: what, where, how, and why. IUBMB Life *49*, 341–351.
13. Carlile, T.M., Rojas-Duran, M.F., Zinshteyn, B., Shin, H., Bartoli, K.M., and Gilbert, W.V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature *515*, 143–146.
14. Lovejoy, A.F., Riordan, D.P., and Brown, P.O. (2014). Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. PLoS ONE *9*, e110799.

15. Schwartz, S., Bernstein, D.A., Mumbach, M.R., Jovanovic, M., Herbst, R.H., León-Ricardo, B.X., Engreitz, J.M., Guttman, M., Satija, R., Lander, E.S., et al. (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell *159*, 148–162.

16. Li, X., Zhu, P., Ma, S., Song, J., Bai, J., Sun, F., and Yi, C. (2015). Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. Nat. Chem. Biol. *11*, 592–597.

17. Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., Yang, Q., Cui, X., Hong, J., Li, X., Chen, Y., et al. (2019). ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. Brief. Bioinform. Published online January 15, 2019. https://doi.org/10.1093/bib/bby127.

18. Zhou, M., Zhao, H., Wang, X., Sun, J., and Su, J. (2019). Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. Brief. Bioinform. *20*, 598–608.

19. Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018). An immune-related six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. Mol. Neurobiol. *55*, 3684–3697.

20. Zhou, M., Hu, L., Zhang, Z., Wu, N., Sun, J., and Su, J. (2018). Recurrence-associated long non-coding RNA signature for determining the risk of recurrence in patients with colon cancer. Mol. Ther. Nucleic Acids *12*, 518–529.

21. Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. Mol. Cancer *16*, Article 16.

22. Zhou, M., Zhao, H., Wang, Z., Cheng, L., Yang, L., Shi, H., Yang, H., and Sun, J. (2015). Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. J. Exp. Clin. Cancer Res. *34*, 102.

23. Yu, L., Zhao, J., and Gao, L. (2018). Predicting potential drugs for breast cancer based on miRNA and tissue specificity. Int. J. Biol. Sci. *14*, 971–982.

24. Tang, G., Shi, J., Wu, W., Yue, X., and Zhang, W. (2018). Sequence-based bacterial small RNAs prediction using ensemble learning strategies. BMC Bioinformatics *19* (*Suppl. 20*), 503.

25. Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. Neurocomputing *273*, 526–534.

26. Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018). SFPEL-LPI: sequence-based feature projection ensemble learning for predicting lncRNA-protein interactions. PLoS Comput. Biol. *14*, e1006616.

27. Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019). A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. IEEE/ACM Trans. Comput. Biol. Bioinform. Published online July 29, 2019. https://doi.org/10.1109/TCBB.2019.2931546.31369383.

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

29. Liao, Z., Li, D., Wang, X., Li, L., and Zou, Q. (2018). Cancer diagnosis from isomiR expression with machine learning method. Curr. Bioinform. *13*, 57–63.

30. Xu, A., Chen, J., Peng, H., Han, G., and Cai, H. (2019). Simultaneous interrogation of cancer omics to identify subtypes with significant clinical differences. Front. Genet. *10*, 236.

31. Panwar, B., and Raghava, G.P.S. (2014). Prediction of uridine modifications in tRNA sequences. BMC Bioinformatics *15*, 326.

32. Li, Y.H., Zhang, G., and Cui, Q. (2015). PPUS: a web server to predict PUS-specific pseudouridine sites. Bioinformatics *31*, 3362–3364.

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

34. He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: pseudouridine sites identification based on RNA sequence information. BMC Bioinformatics *19*, 306.

35. Tahir, M., Tayara, H., and Chong, K.T. (2019). iPseU-CNN: identifying RNA pseudouridine sites using convolutional neural networks. Mol. Ther. Nucleic Acids *16*, 463–470.

36. Liu, K., Chen, W., and Lin, H. (2019). XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites. Mol. Genet. Genomics. Published online August 7, 2019. https://doi.org/10.1007/s00438-019-01600-9.

37. Chen, C.Y., and Chuang, T.J. (2019). Comment on "A comprehensive overview and evaluation of circular RNA detection tools". PLoS Comput. Biol. *15*, e1006158.

38. Xin, Z., Ma, Q., Ren, S., Wang, G., and Li, F. (2017). The understanding of circular RNAs as special triggers in carcinogenesis. Brief. Funct. Genomics *16*, 80–86.

39. Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., and Peng, S. (2019). Deep learning in omics: a survey and guideline. Brief. Funct. Genomics *18*, 41–57.

40. Wei, L., Su, R., Wang, B., Li, X., Zou, Q., and Gao, X. (2019). Integration of deep feature representations and handcrafted features to improve the prediction of N 6-methyladenosine sites. Neurocomputing *324*, 3–9.

41. Lv, Z., Ao, C., and Zou, Q. (2019). Protein function prediction: from traditional classifier to deep learning. Proteomics *19*, e1900119.

42. Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018). Prediction of human protein subcellular localization using deep learning. J. Parallel Distrib. Comput. *117*, 212–217.

43. Chen, W., Ding, H., Zhou, X., Lin, H., and Chou, K.C. (2018). iRNA(m6A)-PseDNC: identifying N^6-methyladenosine sites using pseudo dinucleotide composition. Anal. Biochem. *561-562*, 59–65.

44. Chen, K., Wei, Z., Zhang, Q., Wu, X., Rong, R., Lu, Z., Su, J., de Magalhães, J.P., Rigden, D.J., and Meng, J. (2019). WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res. *47*, e41.

45. Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA. RNA *25*, 205–218.

46. Feng, P., Ding, H., Chen, W., and Lin, H. (2016). Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. Mol. Biosyst. *12*, 3307–3311.

47. Qiu, W.R., Jiang, S.Y., Xu, Z.C., Xiao, X., and Chou, K.C. (2017). iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget *8*, 41178–41188.

48. Zhang, M., Xu, Y., Li, L., Liu, Z., Yang, X., and Yu, D.J. (2018). Accurate RNA 5-methylcytosine site prediction based on heuristic physical-chemical properties reduction and classifier ensemble. Anal. Biochem. *550*, 41–48.

49. Sabooh, M.F., Iqbal, N., Khan, M., Khan, M., and Maqbool, H.F. (2018). Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. J. Theor. Biol. *452*, 1–9.

50. Li, J., Huang, Y., Yang, X., Zhou, Y., and Zhou, Y. (2018). RNAm5Cfinder: a web-server for predicting RNA 5-methylcytosine (m5C) sites based on random forest. Sci. Rep. *8*, 17299.

51. Song, J., Zhai, J., Bian, E., Song, Y., Yu, J., and Ma, C. (2018). Transcriptome-wide annotation of m^5C RNA modifications using machine learning. Front. Plant Sci. *9*, 519.

52. Lv, H., Zhang, Z.M., Li, S.H., Tan, J.X., Chen, W., and Lin, H. (2019). Evaluation of different computational methods on 5-methylcytosine sites identification. Brief. Bioinform. bbz048.

53. Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., and Zhu, F. (2018). What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. ACS Chem. Neurosci. *9*, 1128–1140.

54. Chen, W., Feng, P., Tang, H., Ding, H., and Lin, H. (2016). RAMPred: identifying the N(1)-methyladenosine sites in eukaryotic transcriptomes. Sci. Rep. *6*, 31080.

55. Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., and Chou, K.-C. (2017). iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Mol. Ther. Nucleic Acids *7*, 155–163.

56. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.C. (2018). iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. Mol. Ther. Nucleic Acids *11*, 468–474.

57. Chen, X., Sun, Y.Z., Liu, H., Zhang, L., Li, J.Q., and Meng, J. (2019). RNA methylation and diseases: experimental results, databases, Web servers and computational models. Brief. Bioinform. *20*, 896–917.

58. Li, Y.Z., FY, X., and FY, X. (2018). KELMPSP: pseudouridine sites identification based on kernel extreme learning machine. Chin. J. Biochem. Mol. Biol. *34*, 785–793.

59. Shao, J., Xu, D., Tsai, S.-N., Wang, Y., and Ngai, S.-M. (2009). Computational identification of protein methylation sites through bi-profile Bayes feature extraction. PLoS ONE *4*, e4920.

60. Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and promising identification of human microRNAs by incorporating a high-quality negative set. IEEE/ACM Trans. Comput. Biol. Bioinformatics *11*, 192–201.

61. Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. *43* (W1), W65–W71.

62. Zou, Q., Zeng, J.C., Cao, L.J., and Ji, R.R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing *173*, 346–354.

63. Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.C. (2015). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics *31*, 119–120.

64. Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018). iRNA-2OM: a sequence-based predictor for identifying 2′-O-methylation sites in *Homo sapiens*. J. Comput. Biol. *25*, 1266–1277.

65. Chen, Z., Zhao, P., Li, F., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I., et al. (2019). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief. Bioinform. bbz041.

66. Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res. *47*, e127.

67. Jia, C., Liu, T., Chang, A.K., and Zhai, Y. (2011). Prediction of mitochondrial proteins of malaria parasite using bi-profile Bayes feature extraction. Biochimie *93*, 778–782.

68. Zhao, X., Zhang, J., Ning, Q., Sun, P., Ma, Z., and Yin, M. (2013). Identification of protein pupylation sites using bi-profile Bayes feature extraction and ensemble learning. Math. Probl. Eng. *2013*, 283129.

69. Jia, C.Z., He, W.Y., and Yao, Y.H. (2017). OH-PRED: prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids. J. Biomol. Struct. Dyn. *35*, 829–835.

70. Song, T., Rodríguez-Patón, A., Zheng, P., and Zeng, X. (2018). Spiking neural P systems with colored spikes. IEEE Trans. Cogn. Dev. Syst. *10*, 1106–1115.

71. Xu, H., Zeng, W., Zhang, D., and Zeng, X. (2019). MOEA/HD: a multiobjective evolutionary algorithm based on hierarchical decomposition. IEEE Trans. Cybern. *49*, 517–526.

72. Cabarle, F.G.C., Adorna, H.N., Jiang, M., and Zeng, X. (2017). Spiking neural P systems with scheduled synapses. IEEE Trans. Nanobioscience *16*, 792–801.

73. Sun, W.J., Li, J.H., Liu, S., Wu, J., Zhou, H., Qu, L.H., and Yang, J.H. (2016). RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. Nucleic Acids Res. *44* (D1), D259–D265.

74. He, W., Jia, C., and Zou, Q. (2019). 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics *35*, 593–601.

75. Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., Chen, Y., Xue, W., Li, X., and Zhu, F. (2017). NOREVA: normalization and evaluation of MS-based metabolomics data. Nucleic Acids Res. *45* (W1), W162–W170.

76. Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics *34*, 1953–1956.

77. Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., and Hu, Y. (2018). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. BMC Genomics *19* (*Suppl. 1*), 919.

78. Liu, B. (2017). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Brief. Bioinform. *20*, 1280–1294.

79. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

80. Li, Y., Shi, X., Liang, Y., Xie, J., Zhang, Y., and Ma, Q. (2017). RNA-TVcurve: a web server for RNA secondary structure comparison based on a multi-scale similarity of its triple vector curve representation. BMC Bioinformatics *18*, 51.

81. Liu, B., Yang, F., Huang, D.S., and Chou, K.C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics *34*, 33–40.

82. Ding, Y., Tang, J., and Guo, F. (2016). Identification of protein-protein interactions via a novel matrix-based sequence representation model with amino acid contact information. Int. J. Mol. Sci. *17*, E1623.

83. Ding, Y., Tang, J., and Guo, F. (2016). Predicting protein-protein interactions via multivariate mutual information of protein sequences. BMC Bioinformatics *17*, 398.

84. Yu, L., Su, R., Wang, B., Zhang, L., Zou, Y., Zhang, J., and Gao, L. (2017). Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. IEEE/ACM Trans. Comput. Biol. Bioinformatics *14*, 966–977.

85. Ru, X., Li, L., and Zou, Q. (2019). Incorporating distance-based Top-n-gram and random forest to identify electron transport proteins. J. Proteome Res. *18*, 2931–2939.

86. Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. Methods *166*, 91–102.

87. Xu, L., Liang, G., Liao, C., Chen, G.D., and Chang, C.C. (2019). k-skip-n-gram-RF: a random forest based method for Alzheimer's Disease protein identification. Front. Genet. *10*, 33.

88. Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I.H. (2004). Data mining in bioinformatics using Weka. Bioinformatics *20*, 2479–2481.

89. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.

90. Nello Cristianini, J.S.-T. (2000). An Introduction of Support Vector Machines and Other Kernel-Based Learning Methods (Cambridge University Press).

91. Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. (2019). Meta-path methods for prioritizing candidate disease miRNAs. IEEE/ACM Trans. Comput. Biol. Bioinformatics *16*, 283–291.

92. Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016). Similarity computation strategies in the microRNA-disease network: a survey. Brief. Funct. Genomics *15*, 55–64.

93. Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and validation of disease genes using HeteSim scores. IEEE/ACM Trans. Comput. Biol. Bioinformatics *14*, 687–695.

94. Sun, Y., Xiong, Y., Xu, Q., and Wei, D. (2014). A hadoop-based method to predict potential effective drug combination. BioMed Res. Int. *2014*, 196858.

95. Xu, L., Liang, G., Liao, C., Chen, G.-D., and Chang, C.-C. (2018). An efficient classifier for Alzheimer's Disease genes identification. Molecules *23*, 3140.

96. Xu, L., Liang, G., Shi, S., and Liao, C. (2018). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. Int. J. Mol. Sci. *19*, E17773.

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

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

99. Zhu, X.J., Feng, C.Q., Lai, H.Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowl. Based Syst. *163*, 787–793.

100. Chen, W., Feng, P., Liu, T., and Jin, D. (2019). Recent advances in machine learning methods for predicting heat shock proteins. Curr. Drug Metab. *20*, 224–228.

101. Li, Y.H., Li, X.X., Hong, J.J., Wang, Y.X., Fu, J.B., Yang, H., Yu, C.Y., Li, F.C., Hu, J., Xue, W.W., et al. (2019). Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. Brief. Bioinform. Published January 23, 2019. https://doi.org/10.1093/bib/bby130.

102. Liu, B., and Li, K. (2019). iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. Mol. Ther. Nucleic Acids *18*, 80–87.

103. Liu, B., Li, C.C., and Yan, K. (2019). DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. Brief. Bioinform. Published online October 28, 2019. https://doi.org/10.1093/bib/bbz098.

104. Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. Inf. Sci. *418-419*, 546–560.

105. Xiong, Y., Qiao, Y., Kihara, D., Zhang, H.Y., Zhu, X., and Wei, D.Q. (2019). Survey of machine learning techniques for prediction of the isoform specificity of cytochrome P450 substrates. Curr. Drug Metab. *20*, 229–235.

106. Xiong, Y., Liu, J., and Wei, D.Q. (2011). An accurate feature-based method for identifying DNA-binding residues on protein surfaces. Proteins *79*, 509–517.

107. Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif. Intell. Med. *83*, 67–74.

108. Wei, L., Wan, S., Guo, J., and Wong, K.K. (2017). A novel hierarchical selective ensemble classifier with bioinformatics application. Artif. Intell. Med. *83*, 82–90.

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

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

111. Zhu, P.F., Xu, Q., Hu, Q.H., and Zhang, C.Q. (2018). Co-regularized unsupervised feature selection. Neurocomputing *275*, 2855–2863.

112. Zhu, P.F., Xu, Q., Hu, Q.H., Zhang, C.Q., and Zhao, H. (2018). Multi-label feature selection with missing labels. Pattern Recognit. *74*, 488–502.

113. Zhu, P.F., Zhu, W.C., Hu, Q.H., Zhang, C.Q., and Zuo, W.M. (2017). Subspace clustering guided unsupervised feature selection. Pattern Recognit. *66*, 364–374.

114. Yu, L., Yao, S., Gao, L., and Zha, Y. (2019). Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments. Front. Genet. *9*, 745.

115. Xu, L., Liang, G., Wang, L., and Liao, C. (2018). A novel hybrid sequence-based model for identifying anticancer peptides. Genes (Basel) *9*, 158.

116. Chen, W., Feng, P.M., Lin, H., and Chou, K.C. (2013). iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. *41*, e68.

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

118. Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics *35*, 2796–2800.

119. Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via semi-supervised model and multiple kernel learning. IEEE J. Biomed. Health Inform. *23*, 2619–2632.

120. Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. Neurocomputing *325*, 211–224.

121. Shen, Y., Tang, J., and Guo, F. (2019). Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. J. Theor. Biol. *462*, 230–239.