# *XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites*

**Kewei Liu, Wei Chen & Hao Lin**

**Molecular Genetics and Genomics**

ISSN 1617-4615
Volume 295
Number 1

Mol Genet Genomics (2020) 295:13-21
DOI 10.1007/s00438-019-01600-9

Springer

Molecular Genetics and Genomics (2020) 295:13–21
https://doi.org/10.1007/s00438-019-01600-9

ORIGINAL ARTICLE

# XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites

**Kewei Liu**[1] · **Wei Chen**[1,2] · **Hao Lin**[3]

Received: 16 June 2019 / Accepted: 29 July 2019 / Published online: 7 August 2019


## Abstract

As one of the most popular post-transcriptional modifications, pseudouridine (Ψ) participates in a series of biological processes. Therefore, the efficient detection of pseudouridine sites is very important in revealing its functions in biological processes. Although experimental techniques have been proposed for identifying Ψ sites at single-base resolution, they are still labor intensive and expensive. Recently, to fill the experimental method's gap, computational methods have been proposed for identifying Ψ sites. However, their performances are still unsatisfactory. In this paper, we proposed an eXtreme Gradient Boosting (xgboost)-based method, called XG-PseU, to identify Ψ sites based on the optimal features obtained using the forward feature selection together with increment feature selection method. Our results demonstrated that XG-PseU is superior or at least complementary to existing methods for identifying pseudouridine sites. Finally, a freely available online web server for XG-PseU was established at http://www.bioml.cn/. We wish that XG-PseU will become a useful tool for computationally identifying Ψ sites.

**Keywords** Pseudouridine · eXtreme Gradient Boosting · Feature selection · Web server

## Introduction

Pseudouridine (Ψ) is one kind of RNA modifications (Boccaletto et al. 2018) and has been found in various RNAs from all kingdoms of life (Ge and Yu 2013; Hudson et al. 2013). There are two ways to catalyze pseudouridine modification. One way is through pseudouridine synthases (PUS) which isomerize uridines at specific position in RNA (Ferre-D'Amare 2003; Hamma and Ferre-D'Amare 2006), and the other one is dependent on the H/ACA nucleic acid protein complex (Kiss et al. 2006; Ye 2007). The formation of Ψ is through the isomerization of uracil in which the uracil binds to ribose via C5 instead of N1.

Recently, with the deepening of epigenetics research, more and more researches about Ψ have been done. For example, changes of pseudouridine in rRNA can affect the susceptibility of bacteria to antibiotics (Toh and Mankin 2008). Elimination of rRNA by CBF5 deletion in *S. cerevisiae* is fatal (Jiang et al. 1993; Zebarjadian et al. 1999). The absence of PUS1 leads to growth defects in *S. cerevisiae* and mutations in human PUS1 lead to mitochondrial myopathy and sideroblastic anemia. (Fujiwara and Harigae 2013, 2019). In addition, PUS7 inactivation in embryonic stem cells impairs tRNA-derived small fragments-mediated translation regulation, leading to increased protein biosynthesis and defective germ layer specification (Guzzi et al. 2018). Thus, it is necessary to reveal the biological functions of pseudouridine. The key step for this aim is to accurately find the pseudouridine site in the transcriptome.

There are two main ways to detect Ψ sites. One way is to use experimental methods such as Ψ-seq, Pseudo-seq,

Communicated by Stefan Hohmann.

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s00438-019-01600-9) contains supplementary material, which is available to authorized users.

✉ Wei Chen
chenweiimu@gmail.com

✉ Hao Lin
hlin@uestc.edu.cn

[1] School of Life Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China

[2] Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China

[3] Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China

and CeU-Seq (Basak and Query 2014; Carlile et al. 2014; Schwartz et al. 2014; Li et al. 2015a). Although these methods are labor intensive and expensive for transcriptome-wide detection of Ψ sites, they play important roles and provide key clues for the researches on Ψ modifications. The other one is to use computational methods to predict Ψ sites. In 2015, Li et al. (2015b) built the first computational model called PPUS to predict the PUS-specific Ψ sites in *H. sapiens* and *S. cerevisiae*. Later on, Chen et al. (2016b) developed another model called iRNA-PseU to identify Ψ sites in *H. sapiens*, *S. cerevisiae*, and *M. musculus*. The model improves the accuracy of the prediction of Ψ sites by the method of nucleotide density and their chemical properties. Inspired by these works, to further improve the accuracy for identifying Ψ sites, He et al. proposed the PseUI (He et al. 2018), in which hybrid features including nucleotide composition, pseDNC position-specific nucleotide propensity were used to encode the RNA sequences. More recently, iPseU-CNN model was developed by using the method of convolutional neural network (Tahir et al. 2019). Although iPseU-CNN improves the accuracy of computationally identifying Ψ sites, it is not convenient to use since no accessible web server or open-source code was provided.

In this article, we proposed a new model called XG-PseU to identify Ψ sites. According to the latest RMBase2.0 database, we first built the high-quality benchmark datasets of *H. sapiens*, *M. musculus*, and *S. cerevisiae*. The nucleotide composition, dinucleotide composition, trinucleotide composition, nucleotide chemical property, nucleotide density (ND), one-hot and their combinations were used to represent samples in the datasets. The forward feature selection and increment feature selection techniques were used to find the optimal features which were used as the input of eXtreme Gradient Boosting (XGboost) to perform the prediction. The flowchart for building XG-PseU is shown in Fig. 1.

## Benchmark datasets

In 2016, Chen et al. built the training datasets for computationally identifying Ψ sites in *H. sapiens*, *M. musculus*, and *S. cerevisiae*. In the current study, since the RMBase was updated, we updated the training datasets based on the RMBase v2.0 (Xuan et al. 2018) and the datasets built by Chen et al. (2016b), and obtained three new training datasets based namely NH_990, NM_944, and NS_627, which have 26, 10, and 1 more samples for *H. sapiens*, *M. musculus*, and *S. cerevisiae* than those in the original dataset. The new datasets are available at http://www.bioml.cn/data.html. According to our previous experience (Chen et al. 2016b), the sequences in NH_990 and NM_944 are 21 nt, while those in NS_627 are 31 nt. The detail information of these datasets is shown in Table 1.

The independent datasets (H_200 and S_200) built by Chen et al. were also used to evaluate the proposed method, and the detail information of the independent datasets is introduced by Chen et al. (2016b).

## Feature extraction

Feature extraction is an important step in model construction. The commonly used six features namely nucleotide composition (NC), dinucleotide composition (DNC),

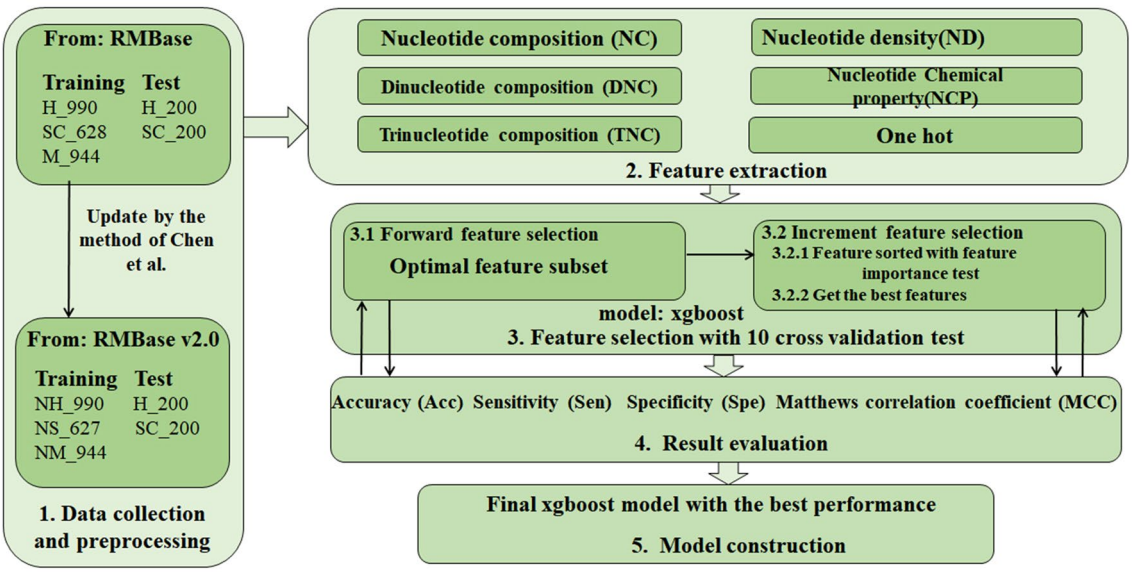


**Fig. 1** The framework of building XG-PseU

**Table 1** The information of training datasets and independent datasets

| Species | *H. sapiens* | *M. musculus* | *S. cerevisiae* | *H. sapiens* | *S. cerevisiae* |
|---|---|---|---|---|---|
| Positive | 495 | 472 | 314 | 100 | 100 |
| Negative | 495 | 472 | 313 | 100 | 100 |
| Length (bp) | 21 | 21 | 31 | 21 | 31 |

NH_990, NS_627, NM_944 are the training datasets for *H. sapiens*, *S. cerevisiae*, and *M. musculus*, respectively

H_200, S_200 are the independent test datasets for *H. sapiens*, *S. cerevisiae*

trinucleotide composition (TNC), nucleotide chemical property (NCP), nucleotide density (ND), and one-hot encode (one hot) were employed to represent the samples in the dataset.

### NC, DNC, TNC

NC, DNC, and TNC are the most common feature extraction methods (Zhang et al. 2011; Brayet et al. 2014). NC is a four-dimensional vector that includes the occurrence frequency of the four nucleotides. Similarly, DNC and TNC are 16-dimensional and 64-dimensional vectors that include the frequency of the 16 dinucleotides and 64 trinucleotides, respectively. They can be written as following vector,

$$D_k = \left[f_1^k, f_2^k, \ldots f_i^k, \ldots, f_{4^k}^k\right] \tag{1}$$

where $f_i^k$ is the frequency of the $i$th $k$-tuple nucleotide in the RNA sequence. $k = 1, 2$, and 3 corresponds to NC, DNC, and TNC, respectively.

### NCP

According to their chemical structure and functionality, A, C, G, and U can be classified into three different groups and defined as follows (Chen et al. 2016b). Therefore,

$$N_i = \left(x_i, y_i, z_i\right) \tag{2}$$

where

$$x_i = \begin{cases} 1, & \text{if } N_i \in \{\mathrm{A}, \mathrm{G}\} \\ 0, & \text{if } N_i \in \{\mathrm{C}, \mathrm{U}\} \end{cases}; \quad y_i = \begin{cases} 1, & \text{if } N_i \in \{\mathrm{A}, \mathrm{C}\} \\ 0, & \text{if } N_i \in \{\mathrm{G}, \mathrm{U}\} \end{cases}; \quad z_i = \begin{cases} 1, & \text{if } N_i \in \{\mathrm{A}, \mathrm{U}\} \\ 0, & \text{if } N_i \in \{\mathrm{C}, \mathrm{G}\} \end{cases}. \tag{3}$$

Thus, A, U, C, and G can be represented by (1, 1, 1), (0, 0, 1), (0, 1, 0), (1, 0, 0), respectively.

### ND

ND considers nucleotide position and frequency (Chen et al. 2016b), and is defined as follows:

$$d_i = \frac{1}{||S_i||} \sum_{j=1}^{\xi} f\left(N_j\right) \tag{4}$$

where $d_i$ is the density of the nucleotide $N_j$ at position $i$ of a RNA sequence, $||S_i||$ is the length of the sliding substring, $\xi$ the corresponding locator's sequence position.

$$f\left(N_j\right) = \begin{cases} 1, & \text{if } N_j \text{ is the nucleotide concerned} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

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

### One-hot

According to one-hot encoding scheme, the four bases in RNA can be encoded by 0 and 1, and thus A, U, C, G can be converted to (1, 0, 0, 0), (0, 1, 0, 0) (0, 0, 1, 0) (0, 0, 0, 1), respectively. Therefore, a ξ-nt long RNA sequence can be represented by a 4ξ-dimensional vector (Tahir et al. 2019).

## Classification engine

eXtreme Gradient Boosting (XGboost) (Chen and Guestrin 2016) is a machine learning algorithm based on tree model. It has been widely used in the recent papers (Wang et al. 2019; Yao et al. 2019). By adding regularization items to the cost function, the complexity of the model was controlled. This procedure prevents overfitting of the model. What's more, the XGboost supports parallel computing, which makes it faster and more flexible to use. Therefore, the XGboost was used to perform the classification in the present work. The range of its parameter selection is shown in Table 2.

**Table 2** Parameter selection range and step size of the Xgboost

| Parameter | Range | Step |
|---|---|---|
| Learning_rate | 0.1 | – |
| N_estimators | (30, 50) | 10 |
| Max_depth | (3, 7) | 2 |
| Min_child_weight | (3, 7) | 2 |
| Subsample | (0.6, 0.9) | 0.1 |
| Colsample_bytree | (0.6, 0.9) | 0.1 |
| Reg_alpha | 0.1 | – |
| Reg_lambda | 0.1 | – |

## Feature selection

To effectively represent RNA sequences, as described above, six kinds of sequence encoding scheme will be employed to encode RNA sequences. To avoid the overfitting problem, a two-step feature selection technique was used to determine the optimal features.

At first, the forward feature selection (Wang et al. 2011; Liu et al. 2015) method was used to select optimal features from the six kinds of feature extraction methods. For this aim, we initially calculated the prediction accuracy of the six feature extraction methods and selected out the feature with the highest accuracy. The process was repeated until the selected feature combination has the highest accuracy. Subsequently, the increment feature selection (IFS) (Chen et al. 2016c; Yang et al. 2016; Tang et al. 2018) was used to sort the importance of the features obtained by forward feature selection. To do so, we built the XGboost model by using the feature that ranks the first in the forward feature selection step. Then the second feature was added to build a new model. This process was repeated until all the features obtained in the forward feature selection step were added. For each iteration, an XGboost model will be built and an accuracy will be obtained. Accordingly, a IFS can be plotted with the abscissa indicating the number of features and the ordinate indicating the accuracy. When the IFS curve reaches its peak, the optimal features were obtained.

## Cross validation

Since the $K$-fold ($K = 5$ or 10) cross validation is widely used to evaluate a computational model (Li et al. 2016; Vuckovic et al. 2016; Dezman et al. 2017), the 10-cross-validation test was used to measure the performance of the proposed method. To objectively validate its stability, the proposed method was also tested on the independent dataset.

**Table 3** Optimal subset of features after forward feature selection on training datasets

| Specie | Accuracy (%) | Feature subset |
|---|---|---|
| *H. sapiens* | 64.35 | NCP + DNC + TNC |
| *M. musculus* | 72.02 | One-hot + TNC + NCP + ND + DNC |
| *S. cerevisiae* | 67.29 | NCP + TNC + DNC + One-hot |

## Evaluation metrics

The parameters used to evaluate the quality of the classifier are sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (Mcc), which have been widely used to evaluate the quality of the classifier (Chou 2001; Feng et al. 2013; Chen et al. 2016a, 2019; Le 2019; Le et al. 2019a, b). Their definitions are as follows:

$$\begin{cases} \mathrm{Sn} = 1 - \frac{N_-^+}{N^+} \\ \mathrm{Sp} = 1 - \frac{N_+^-}{N^-} \\ \mathrm{Acc} = 1 - \frac{N_-^+ + N_+^-}{N^+ N^-} \\ \mathrm{Mcc} = \frac{1 - \frac{N_-^+ + N_+^-}{N^+ N^-}}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \end{cases} \quad (6)$$

where $N^+$ is the number of the Ψ site containing sequences; $N^-$ the number of non-Ψ site containing sequences. $N_-^+$ is the number of Ψ site containing sequences incorrectly predicted as non-Ψ site containing sequences. $N_+^-$ represents the number of non-Ψ site containing sequences incorrectly predicted as Ψ site containing sequences.

## Results

### Parameter optimization

To determine the optimal features, we first compared the contributions of the six kinds of features for the identification of Ψ sites using forward feature selection. The predictive accuracy for identifying Ψ sites using the six kinds of features and their combinations is provided in Supplementary Tables S1–S3. The optimal features and their corresponding accuracies for identifying Ψ sites in *H. sapiens*, *M. musculus*, and *S. cerevisiae* are reported in Table 3.

It was found that, for different species, the optimal feature subsets are different. The optimal feature subsets for *H. sapiens* is the combination of NCP, DNC, and TNC, that for *M. musculus* is one-hot, TNC, NCP, ND, and DNC, and

the combination of NCP, TNC, DNC, and one-hot is the optimal features of *S. cerevisiae*. This may be due to the different nucleotide compositions surrounding Ψ sites in these species.

To test this hypothesis, we compared the nucleotide composition between the positive and negative samples of these species using the Two Sample Logo (Vacic et al. 2006) software. It was found that the nucleic acid sequences of the positive and negative samples near the central site (Ψ/U) have a strong species positional specificity, Fig. 2. For *H. sapiens*, the enrichment of G, U was observed at the 6th, 9th, and 10th positions of the upstream sequences, and the enrichment of A/U, C, G was observed at the 12th, 13th, 14th, 18th, and 20th positions of the downstream sequences. For *M. musculus*, the enrichment of C, G, U was observed at the 2nd, 9th, and 10th positions of the upstream sequences, and the enrichment of A/U, C/U, U was observed at the 12th, 13th, and 18th positions of the downstream sequences. For *S. cerevisiae*, the enrichment of U was observed at the 14th position of the upstream sequences, and the enrichment of G was observed in both upstream and downstream sequences. This result explained why the different optimal feature combinations were obtained for different species.

To improve the performance, the IFS curves for *H. sapiens*, *M. musculus*, and *S. cerevisiae* are plotted in Fig. 3. As shown in Fig. 3, when the 83, 52, and 35 optimal features were used, the best predictive accuracies of 66.1%, 73.4%, and 71.1% were obtained for identifying Ψ sites in *H. sapiens*, *M. musculus*, and *S. cerevisiae*, respectively. It was found that the accuracies are 1.75%, 1.38%, and 3.81% higher than that by only using forward feature selection. Therefore, in the following analysis, the optimal features thus obtained were used to build computational models in *H. sapiens*, *M. musculus*, and *S. cerevisiae*, respectively. The corresponding parameters of the XGboost for each species

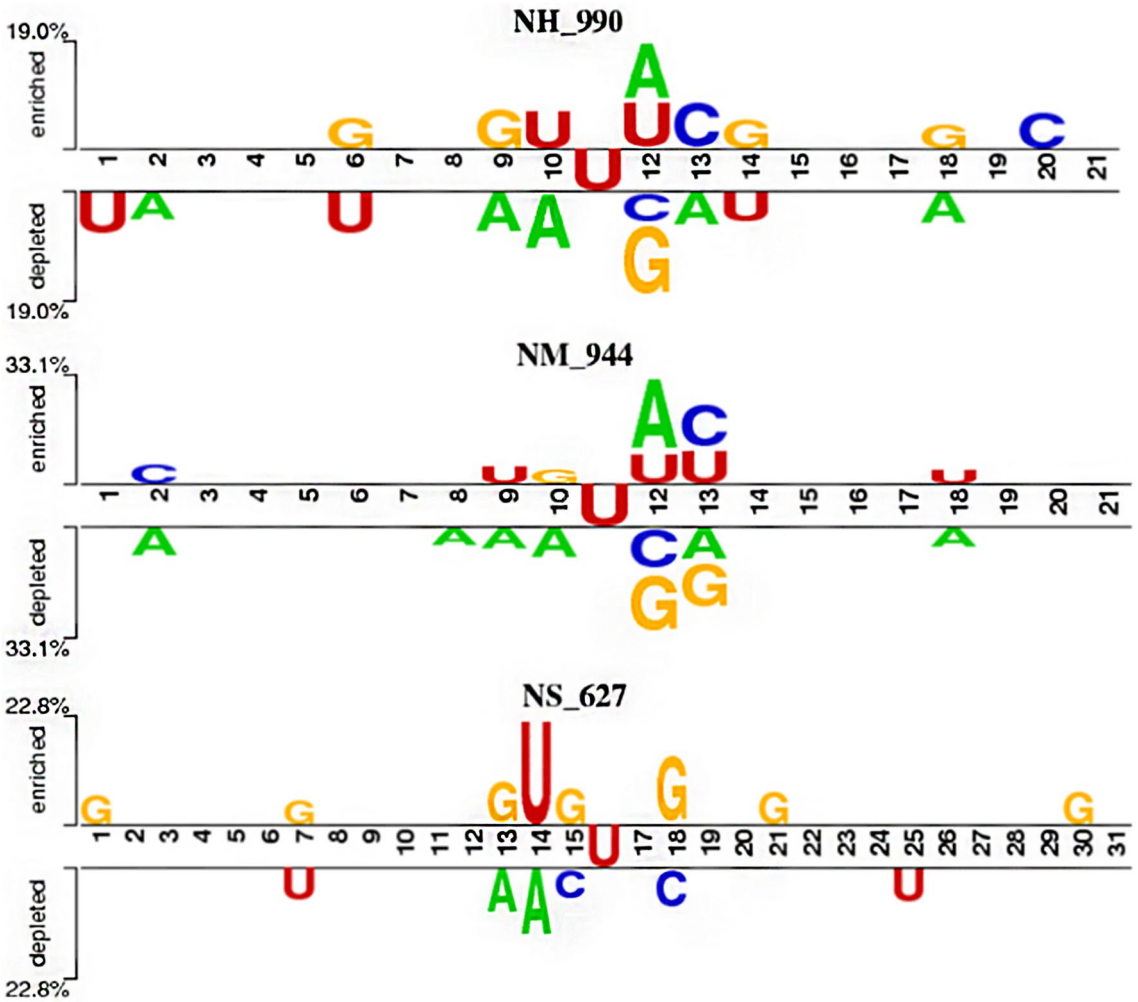


**Fig. 2** The nucleotide composition preferences of Ψ site and non-Ψ site containing sequences. The three figures from top to down are based on the sequences from NH_990 (*H. sapiens*), NM_944 (*M. musculus*), and NS_627 (*S. cerevisiae*), respectively. In each figure, the top panel is for the Ψ site containing sequences and the down panel is for non-Ψ site containing sequences

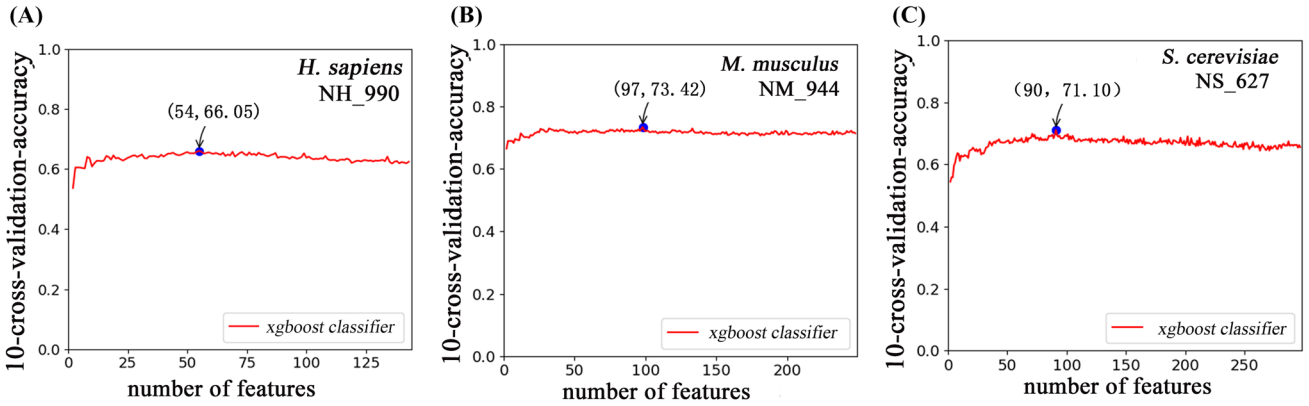


**Fig. 3** The IFS curves used to sort the importance of the features obtained by forward feature selection, **a** *H. sapiens*, **b** *M. musculus*, and **c** *S. cerevisiae*. The abscissa is the number of features and the ordinate is the accuracy for Ψ sites identifying in tenfold cross-validation test. The optimal number of features was obtained when the IFS curve reached the peak

**Table 4** The parameters of the XGboost for the model of each species

| Species | Estimators | Depth | Min_child_weight | Subsample | Colsample |
|---|---|---|---|---|---|
| *H. sapiens* | 40 | 3 | 3 | 0.7 | 0.7 |
| *M. musculus* | 30 | 3 | 5 | 0.6 | 0.6 |
| *S. cerevisiae* | 30 | 5 | 3 | 0.7 | 0.7 |

**Table 5** The performance of the model for identifying Ψ sites in each species

| Species | Acc (%) | Sn (%) | Sp (%) | Mcc |
|---|---|---|---|---|
| *H. sapiens* | 66.05 | 63.45 | 68.65 | 0.32 |
| *M. musculus* | 71.10 | 65.92 | 76.30 | 0.43 |
| *S. cerevisiae* | 73.42 | 77.35 | 69.48 | 0.47 |

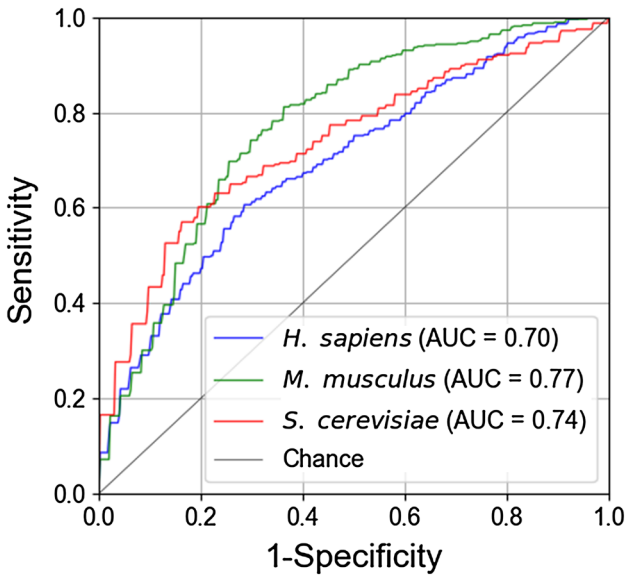


**Fig. 4** A graphical illustration to show the performance of the model by means of the ROC curves obtained from the tenfold cross-validation test. The vertical coordinate is the true positive rate (sensitivity) while the horizontal coordinate is the false positive rate (1-specificity)

are listed in Table 4, which is obtained by a grid search under the 10-cross-validation test.

Based on the optimal features and parameters obtained above, the tenfold cross-validation test results of the proposed methods for identifying Ψ sites in each species are listed in Table 5. Moreover, we also plotted the ROC curves to objectively evaluate the models for identifying Ψ sites in *H. sapiens*, *M. musculus,* and *S. cerevisiae*, as shown in Fig. 4. In the tenfold cross-validation test, the area under the ROC curve (AUC) of *H. sapiens*, *S. cerevisiae*, and *M. musculus* was 0.70, 0.77, and 0.74, respectively.

## Comparison with other methods

To further demonstrate its performance, we compared XG-PseU with other existing methods (iRNA-PseU, PseUI, iPseU-CNN) for identifying Ψ sites. Since iRNA-PseU, PseUI, iPseU-CNN are all trained based on the dataset built by Chen et al., for a fair comparison, we retrained the XGboost on the same dataset. The detailed information about the optimal features and parameters of the XGboost-based model is provided in Supplementary Tables S4–S6 and Supplementary Figure S1. The tenfold cross-validation test results are reported in Table 6.

Although the accuracy of our proposed XGboost-based method for identifying Ψ sites in *H. sapiens* is a little lower than that of iPseU-CNN which is the best predictor

**Table 6** Comparative results of different methods for identifying Ψ sites in each species

| Species | Methods | Acc (%) | Sn (%) | Sp (%) | Mcc |
|---|---|---|---|---|---|
| *H. sapiens* | iRNA-PseU | 60.40 | 61.01 | 59.80 | 0.21 |
| | PseUI | 64.24 | 64.85 | 63.64 | 0.28 |
| | iPseU-CNN | 66.68 | 65.00 | 68.78 | 0.34 |
| | XGboost | 65.44 | 63.64 | 67.24 | 0.31 |
| *S. cerevisiae* | iRNA-PseU | 64.49 | 64.65 | 64.33 | 0.29 |
| | PseUI | 65.13 | 62.74 | 67.52 | 0.30 |
| | iPseU-CNN | 68.15 | 66.36 | 70.45 | 0.37 |
| | XGboost | 68.15 | 66.84 | 69.45 | 0.37 |
| *M. musculus* | iRNA-PseU | 69.07 | 73.31 | 64.83 | 0.38 |
| | PseUI | 70.44 | 74.58 | 66.31 | 0.41 |
| | iPseU-CNN | 71.81 | 74.79 | 69.11 | 0.44 |
| | XGboost | 72.03 | 76.48 | 67.57 | 0.45 |

iRNA-PseU is the predictor developed by Chen et al.; PseUI is the predictor developed by He et al.; and iPseU-CNN is the predictor developed by Tahir et al. in 2019

at present, the proposed method obtained an accuracy of 68.15% which is the same as that of iPseU-CNN for identifying Ψ sites in *S. cerevisiae*. Moreover, our proposed XGboost-based method obtained the best accuracy of 72.03% for identifying Ψ sites in *M. musculus*. These results demonstrate the reliability of XGboost-based method in identifying Ψ sites.

### Validation of the methods on independent dataset

To further demonstrate the generalization ability of the existing methods for identifying Ψ sites, we validated them on the independent datasets built by Chen et al. (2016b). It should be pointed out that since no web server or source code was provided by iPseU-CNN, the comparisons were performed between the proposed method with the other remaining methods (i.e., PPUS, iRNA-PseU and PseUI). To provide an intuitive view of the performance of different methods, their predictive results in the independent datasets are shown in Fig. 5. As indicated in Fig. 5, our proposed XGboost-based method with the accuracy of 67.5% ranks the first for identifying Ψ sites in *H. sapiens*, and is comparable with iRNA-PseU and PPUS for identifying Ψ sites in *S. cerevisiae*. These results demonstrate that the proposed XGboost-based method, PPUS, iRNA-PseU, and PseUI are complementary tools for identifying Ψ sites.

## Discussion

In this work, a new method called XG-PseU was proposed to identify pseudouridine sites. Due to the distinct nucleotide compositions between Ψ sites and non-Ψ sites containing sequences in *H. sapiens*, *S. cerevisiae*, and *M. musculus*, different parameters were set for XGboost to build the computational models in these three species. Considering that the existing methods were trained or tested based on different dataset or different cross-validation methods, to objectively validate its performance, the XG-PseU was compared with the existing methods on an independent dataset. The results demonstrate that XG-PseU is superior or at least complementary to the existing methods of identifying pseudouridine sites. Based on the proposed method, a freely accessible web server was proposed at http://www.bioml.cn/, by which the users can detect the potential pseudouridine sites in *H. sapiens, S. cerevisiae, and M. musculus*. We hope that XG-PseU will become a useful computational tool for identifying pseudouridine sites.

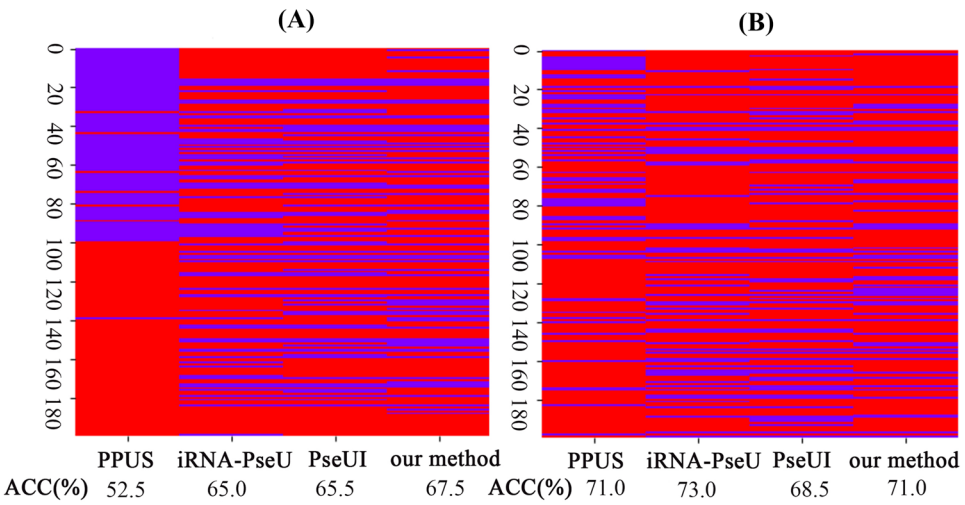


**Fig. 5** The detail predictive results of PPUS, iRNA-PseU, PseUI, and the proposed method based on the independent dataset of **a** *H. sapiens* and **b** *S. cerevisiae*. Each row is a sample in the independent dataset (the first 100 rows indicate the positive samples and the remaining 100 indicate the negative samples). The correctly identified ones are highlighted in red and the counterparts are in blue. The accuracies of each method are provided on the bottom panel (color figure online)

**Acknowledgements** This work was supported by the National Nature Scientific Foundation of China (31771471, 61772119) and the Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244).

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants performed by any of the authors.

## References

Basak A, Query CC (2014) A pseudouridine residue in the spliceosome core is part of the filamentous growth program in yeast. Cell Rep 8:966–973

Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, Wirecki TK, de Crecy-Lagard V, Ross R, Limbach PA, Kotter A, Helm M, Bujnicki JM (2018) MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res 46:D303–D307

Brayet J, Zehraoui F, Jeanson-Leh L, Israeli D, Tahi F (2014) Towards a piRNA prediction using multiple kernel fusion and support vector machine. Bioinformatics 30:I364–I370

Carlile TM, Rojas-Duran MF, Zinshteyn B, Shin H, Bartoli KM, Gilbert WV (2014) Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature 515:143–146

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Acm sigkdd international conference on knowledge discovery & data mining

Chen W, Ding H, Feng PM, Lin H, Chou KC (2016a) IACP: a sequence-based tool for identifying anticancer peptides. Oncotarget 7:16895–16909

Chen W, Tang H, Ye J, Lin H, Chou KC (2016b) iRNA-PseU: identifying RNA pseudouridine sites. Mol Ther Nucleic Acids 5:e332

Chen XX, Tang H, Li WC, Wu H, Chen W, Ding H, Lin H (2016c) Identification of bacterial cell wall lyases via pseudo amino acid composition. Biomed Res Int 2016:1654623

Chen W, Lv H, Nie F, Lin H (2019) i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz015

Chou KC (2001) Using subsite coupling to predict signal peptides. Protein Eng 14:75–79

Dezman ZDW, Gao C, Yang SM, Hu P, Yao L, Li HC, Chang CI, Mackenzie C (2017) Anomaly detection outperforms logistic regression in predicting outcomes in trauma patients. Prehospital Emerg Care 21:174–179

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

Ferre-D'Amare AR (2003) RNA-modifying enzymes. Curr Opin Struct Biol 13:49–55

Fujiwara T, Harigae H (2013) Pathophysiology and genetic mutations in congenital sideroblastic anemia. Pediatr Int 55:675–679

Fujiwara T, Harigae H (2019) Molecular pathophysiology and genetic mutations in congenital sideroblastic anemia. Free Radic Biol Med 133:179–185

Ge J, Yu YT (2013) RNA pseudouridylation: new insights into an old modification. Trends Biochem Sci 38:210–218

Guzzi N, Ciesla M, Ngoc PCT, Lang S, Arora S, Dimitriou M, Pimkova K, Sommarin MNE, Munita R, Lubas M, Lim Y, Okuyama K, Soneji S, Karlsson G, Hansson J, Jonsson G, Lund AH, Sigvardsson M, Hellstrom-Lindberg E, Hsieh AC, Bellodi C (2018) Pseudouridylation of tRNA-derived fragments steers translational control in stem cells. Cell 173(1204–1216):e1226

Hamma T, Ferre-D'Amare AR (2006) Pseudouridine synthases. Chem Biol 13:1125–1135

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

Hudson GA, Bloomingdale RJ, Znosko BM (2013) Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. RNA 19:1474–1482

Jiang W, Middleton K, Yoon HJ, Fouquet C, Carbon J (1993) An essential yeast protein, CBF5p, binds in vitro to centromeres and microtubules. Mol Cell Biol 13:4884–4893

Kiss T, Fayet E, Jady BE, Richard P, Weber M (2006) Biogenesis and intranuclear trafficking of human box C/D and H/ACA RNPs. Cold Spring Harb Symp Quant Biol 71:407–417

Le NQK (2019) iN6-methylat (5-step): identifying DNA N6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule. Mol Genet Genomics. https://doi.org/10.1007/s00438-019-01570-y

Le NQ, Yapp EK, Ho QT, Nagasundaram N, Ou YY, Yeh HY (2019a) iEnhancer-5Step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. Anal Biochem 571:53–61

Le NQ, Yapp EK, Ou YY, Yeh HY (2019b) iMotor-CNN: identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule. Anal Biochem 575:17–26

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

Li YH, Zhang G, Cui Q (2015b) PPUS: a web server to predict PUS-specific pseudouridine sites. Bioinformatics 31:3362–3364

Li GQ, Liu Z, Shen HB, Yu DJ (2016) Target M6A: identifying N-6-methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. IEEE Trans Nanobiosci 15:674–682

Liu Y, Gu W, Zhang W, Wang J (2015) Predict and analyze protein glycation sites with the mRMR and IFS methods. Biomed Res Int 2015:561547

Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, Leon-Ricardo BX, Engreitz JM, Guttman M, Satija R, Lander ES, Fink G, Regev A (2014) Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell 159:148–162

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

Tang H, Zhao YW, Zou P, Zhang CM, Chen R, Huang P, Lin H (2018) HBPred: a tool to identify growth hormone-binding proteins. Int J Biol Sci 14:957–964

Toh SM, Mankin AS (2008) An indigenous posttranscriptional modification in the ribosomal peptidyl transferase center confers resistance to an array of protein synthesis inhibitors. J Mol Biol 380:593–597

Vacic V, Iakoucheva LM, Radivojac P (2006) Two sample logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 22:1536–1537

Vuckovic F, Theodoratou E, Thaci K, Timofeeva M, Vojta A, Stambuk J, Pucic-Bakovic M, Rudd PM, Derek L, Servis D, Wennerstrom A, Farrington SM, Perola M, Aulchenko Y, Dunlop MG,

Campbell H, Lauc G (2016) IgG glycome in colorectal cancer. Clin Cancer Res 22:3078–3086

Wang L, Shen C, Hartley R (2011) On the optimality of sequential forward feature selection using class separability measure. In: International conference on digital image computing: techniques & applications

Wang Q, Zhao D, Wang Y, Hou X (2019) Ensemble learning algorithm based on multi-parameters for sleep staging. Med Biol Eng Comput 57(8):1693–1707. https://doi.org/10.1007/s11517-019-01978-z

Xuan JJ, Sun WJ, Lin PH, Zhou KR, Liu S, Zheng LL, Qu LH, Yang JH (2018) RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res 46:D327–D334

Yang H, Tang H, Chen XX, Zhang CJ, Zhu PP, Ding H, Chen W, Lin H (2016) Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition. Biomed Res Int 2016:5413903

Yao L, Cai M, Chen Y, Shen C, Shi L, Guo Y (2019) Prediction of antiepileptic drug treatment outcomes of patients with newly diagnosed epilepsy by machine learning. Epilepsy Behav 96:92–97

Ye K (2007) H/ACA guide RNAs, proteins and complexes. Curr Opin Struct Biol 17:287–292

Zebarjadian Y, King T, Fournier MJ, Clarke L, Carbon J (1999) Point mutations in yeast CBF5 can abolish in vivo pseudouridylation of rRNA. Mol Cell Biol 19:7461–7472

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