M6AMRFS: Robust Prediction of N6-Methyladenosine Sites With Sequence-Based Features in Multiple Species

Xiaoli Qiang[1], Huangrong Chen[2], Xiucai Ye[3], Ran Su[4]* and Leyi Wei[2]*

[1] Institute of Computing Science and Technology, Guangzhou University, Guangzhou, China, [2] School of Computer Science and Technology, Tianjin University, Tianjin, China, [3] Department of Computer Science, University of Tsukuba, Tsukuba, Japan, [4] School of Software, Tianjin University, Tianjin, China

As one of the well-studied RNA methylation modifications, N6-methyladenosine (m^6A) plays important roles in various biological progresses, such as RNA splicing and degradation, etc. Identification of m^6A sites is fundamentally important for better understanding of their functional mechanisms. Recently, machine learning based prediction methods have emerged as an effective approach for fast and accurate identification of m^6A sites. In this paper, we proposed "M6AMRFS", a new machine learning based predictor for the identification of m^6A sites. In this predictor, we exploited a new feature representation algorithm to encode RNA sequences with two feature descriptors (dinucleotide binary encoding and Local position-specific dinucleotide frequency), and used the F-score algorithm combined with SFS (Sequential Forward Search) to enhance the feature representation ability. To predict m^6A sites, we employed the eXtreme Gradient Boosting (XGBoost) algorithm to build a predictive model. Benchmarking results showed that the proposed predictor is competitive with the state-of-the art predictors. Importantly, robust predictions for multiple species by our predictor demonstrate that our predictive models have strong generalization ability. To the best of our knowledge, M6AMRFS is the first tool that can be used for the identification of m^6A sites in multiple species. To facilitate the use of our predictor, we have established a user-friendly webserver with the implementation of M6AMRFS, which is currently available in http://server.malab.cn/M6AMRFS/. We anticipate that it will be a useful tool for the relevant research of m^6A sites.

Keywords: N6-methyladenosine site, eXtreme Gradient Boosting, machine learning, feature representation, RNA methylation, feature selection

INTRODUCTION

To date, more than 150 types of RNA modifications have been discovered (Maden, 1990; Wang X. et al., 2014). Of these modifications, N6-methyladenosine (m^6A) is the most common and abundant one and exists in various species. It is found to be closely associated with diverse biological processes, such as RNA localization and degradation (Wang X. et al., 2014), RNA structural dynamics (Roost et al., 2015), alternative splicing (Liu N. et al., 2015), primary microRNA

processing (Alarcón et al., 2015), cell differentiation, and reprogramming (Chen et al., 2015), and regulation of circadian clock (Geula et al., 2015). Thus, identification of m⁶A sites is of great importance for better understanding of their functional mechanisms. In the past few years, high-throughput experimental methods, such as MERIP (Meyer et al., 2012) and m⁶A-seq (Dominissini et al., 2012), have been utilized to identify m⁶A modifications, and more and more m⁶A peaks have been characterized. However, they have the following limitations: (1) they cannot accurately locate the positions of m⁶A sites; (2) they are highly cost; and (3) they are not applicable for the large-scale identification of m⁶A sites. Hence, it is highly desirable to develop fast and accurate computational methods for the identification of m⁶A sites (Chen et al., 2015b, 2016).

In recent years, machine learning based prediction methods have emerged as effective approach for predicting m⁶A sites. For example, Chen et al. (2015a) developed the first machine learning based predictor, called "iRNA-Methyl", for m⁶A site identification. They exploited physicochemical properties and sequence-order information embedded in PseDNC (pseudo dinucleotide composition) (Liu B. et al., 2015), and used support vector machine for model construction. Later, Liu Z. et al. (2016) proposed to incorporate more additional physicochemical properties coupled with a scalable transformation algorithm into their feature extraction model. To improve the predictive performance, Jia et al. proposed to fuse three types of feature descriptors, such as bi-profile Bayes, dinucleotide composition and KNN scores. Their results showed that this fusion strategy is able to achieve better performance than single one feature descriptor (Jia et al., 2016). Similarly, Xiang et al. (2016b) found that combining binary encoding scheme together with k-mer frequency could contribute to the improved performance. Recently, Zhou et al. (2016) developed "SRAMP", a powerful prediction tool using multiple types of feature descriptors, including positional binary encoding of nucleotide sequence, k-nearest neighbor encoding, nucleotide pair spectrum encoding, and secondary structure pattern, to train an ensemble predictive model with random forest for the identification of m⁶A sites. SRAMP is reported to achieve relatively good performance as compared to other predictors. More recently, Xiang et al. (2016a) proposed a new predictor called "RNAMethyPre", using compositional information and position-specific information to build predictive models for the prediction of m⁶A sites on both human and mouse. Additionally, in our previous study, we proposed to use deep learning algorithm to generate high-latent features to improve the predictive performance (Wei et al., 2018d). However, we found that most of existing predictors are species-specific. Currently, there is not any predictor that is capable of predicting m⁶A sites for multiple species.

For this purpose, we proposed a novel sequence-based predictor, namely "M6AMRFS" for detecting m⁶A sites in RNA sequences. For feature extraction (Mrozek et al., 2007, 2013), we proposed a feature representation algorithm to encode sequences with dinucleotide binary encoding and local position-specific dinucleotide frequency. To optimize the feature space, we combined the F-score algorithm with SFS (Sequential Forward Search) (Wei et al., 2018a,c,e) to improve the representation

ability of our features. For model training, we trained the optimal feature representations under XGBoost algorithm. Our experimental results showed that the proposed M6AMRFS is able to achieve competitive and robust performance as compared to state-of-the-art predictors for four different species. To the best of our knowledge, this is the first predictor that is applicable for multiple species. Furthermore, we have established a user-friendly webserver that implements the proposed M6AMRFS, which is currently available in http://server.malab.cn/M6AMRFS/. We anticipate that it will be a useful tool complementary for existing tools, facilitating to further reveal the functional mechanisms of m⁶A sites.

MATERIALS AND METHODS

Benchmark Datasets
To predict the m⁶A sites in multiple species, we employed four benchmark datasets from four species, including *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Musculus*, and *Homo sapiens*. The detail of the four benchmark datasets is listed in **Table 1**. For the four benchmark datasets, the positives are the sequences centered with true m⁶A sites, while the negatives are usually the sequences centered with adenines but without any m⁶A peaks detected. The datasets can be found in the following website: http://server.malab.cn/M6AMRFS/.

Prediction Framework of the Proposed Predictor
Figure 1 illustrates the overall procedure of the proposed predictor. As we can see from **Figure 1**, there are two steps in the predictor. The first step is data pre-processing, including data clean and feature extraction. It filters out those irrelevant sequences from input sequences. Then, the resulting sequences are submitted into the feature representation algorithm, in which the sequences are encoded with feature vectors. The second step is feature optimization and model training. For feature space optimization, we used the F-score algorithm combined with SFS (Sequential Forward Search) to search for the optimal features. Afterward, the resulting optimal feature representations are fed into a well-trained XGBoost model to predict whether the sequences are true m⁶A sites or not. In our predictor, the predicted outcome for each sequence is 0 or 1, where 0 denotes non-m⁶A site and 1 denotes true m⁶A site.

Feature Representation
In this work, we present a new feature representation algorithm that combines two feature descriptors. One is named "Dinucleotide binary encoding" and the other is "Local position-specific dinucleotide frequency", which are described as follows,

Dinucleotide Binary Encoding
The feature descriptor encapsulates the positional information of the dinucleotide at each position in the sequence. Obviously, there are a total of 16 possible dinucleotides. In this descriptor, each dinucleotide can be encoded into a 4-dimensional 0/1 vector. For example, AA is encoded as (0,0,0,0); AT is encoded as

Datasets	Species	Positives	Negatives	Total	Sequence length	Reference
Dataset-S51	*Saccharomyces cerevisiae*	1307	1307	2614	51 nt	Chen et al., 2015a
Dataset-H41	*Homo sapiens*	1130	1130	2260	41 nt	Chen et al., 2017
Dataset-M41	*Musculus*	725	725	1450	41 nt	Dominissini et al., 2012
Dataset-A101	*Arabidopsis thaliana*	1000	1000	2000	101 nt	Wang and Yan, 2018
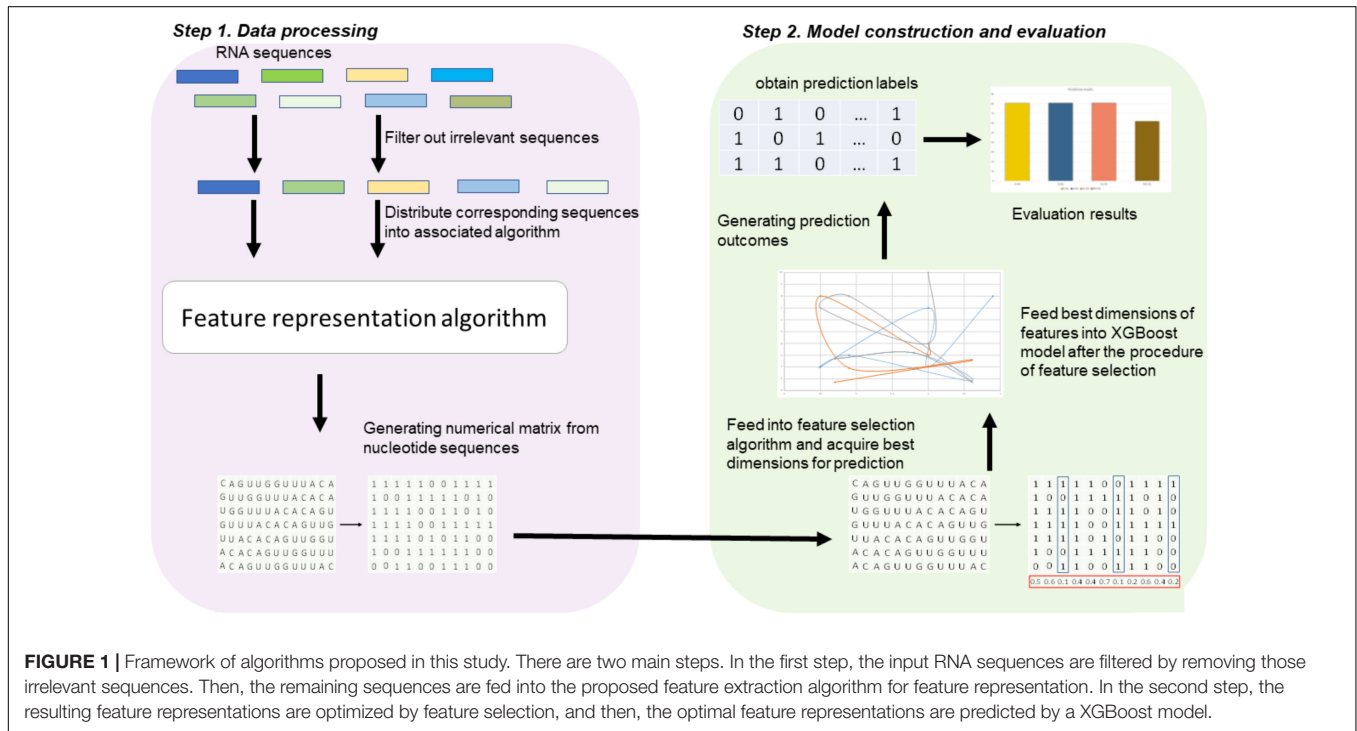
FIGURE 1 | Framework of algorithms proposed in this study. There are two main steps. In the first step, the input RNA sequences are filtered by removing those irrelevant sequences. Then, the remaining sequences are fed into the proposed feature extraction algorithm for feature representation. In the second step, the resulting feature representations are optimized by feature selection, and then, the optimal feature representations are predicted by a XGBoost model.

(0,0,0,1); AC is encoded as (0,0,1,0); and so forth, GG is encoded as (1,1,1,1). Therefore, using the dinucleotide binary encoding, we yielded a 160 (=40*4)-dimensional 0/1 vector for the given sequence.

Local Position-Specific Dinucleotide Frequency

For a given sequence, the feature vector of this descriptor can be denoted as $(f_2, f_3, \ldots, f_l)$, where f_i is calculated as follows,

$$f = \frac{1}{|N_i|} C\left(X_{i-1}X_i\right), 2 \leq i \leq l$$

where l is the length of the given sequence, $|N_i|$ is the length of the ith prefix string $\{X_1 X_2 \ldots X_i\}$ in the sequence, and $C\left(X_{i-1}X_i\right)$ is the occurrence number of the dinucleotide $X_{i-1}X_i$ in position i of the ith prefix string.

Feature Selection

Feature selection is an important process to improve the classification performance (Mrozek et al., 2009; Mrozek et al., 2014; Zeng et al., 2016; Zou et al., 2016a,b; Liu, 2017). Here, we used the F-score algorithm together with the SFS strategy to search the most discriminative features (Peng et al., 2005). **Figure 2** illustrates the procedure of the feature selection strategy,

which is described as follows. Firstly, the F-score algorithm is utilized to rank all the features from the highest scores to the lowest scores, generating a ranked feature list. Secondly, we added the features one by one from the ranked list, and respectively trained the predictive models. Lastly, the feature subset corresponding to the highest accuracy of the predictive model is used as the optimal features. The results of feature selection were discussed in section of "Results and Discussion".

XGBoost (eXtreme Gradient Boosting)

eXtreme Gradient Boosting, which was proposed by Chen and Guestrin (2016), has been shown to be a powerful classification algorithm. The general idea of XGBoost is to enumerate several candidates that may be the segmentation points according to the percentile method, and then to find the best segmentation point from the candidates for calculating the segmentation points. The main advantage of XGBoost is to combine multithreading, data compression, and fragmentation methods to improve the efficiency of the algorithm as much as possible. Moreover, the regularization terms added by XGBoost in the loss function can be used to control the complexity of the model and avoid overfitting. Parameters, such as subsamples, max depth, and estimators, are utilized to optimize evaluation performance
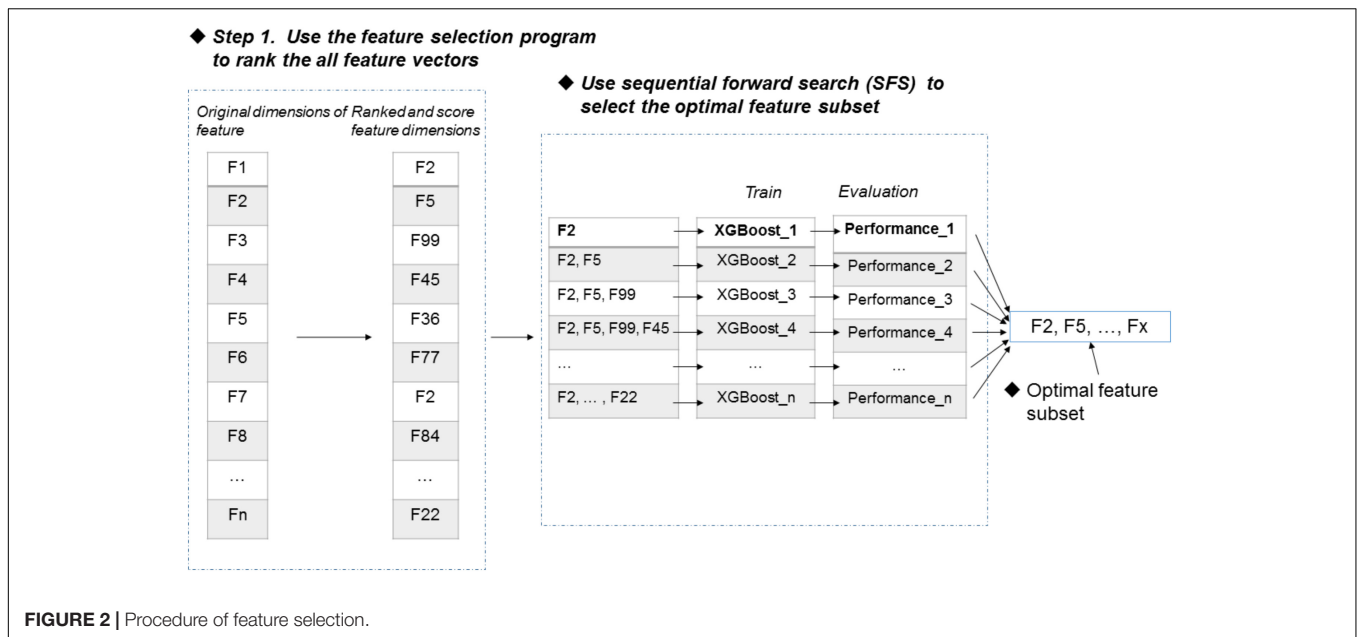
FIGURE 2 | Procedure of feature selection.

via parallelization program namely "Grid Search". For the implementation of XGBoost in our predictor, the range of max depth is set from 2 to 10; learning rate is ranged from 0.1 to 0.8; and estimators are ranged from 1 to 10.

Performance Evaluation

In this work, four commonly used performance metrics are used for performance evaluation, including Acc (accuracy), Sn (sensitivity), Sp (specificity), and MCC (Mathew's correlation coefficient), respectively (Zeng et al., 2015; Lai et al., 2017; Zhang et al., 2017; Cheng et al., 2018; Su et al., 2018; Tang et al., 2018; Wei et al., 2018b; Yang et al., 2018). They are formulated as follows

$$
\begin{cases}
Sn = \frac{TP}{TP+FN} \times 100\% \\
Sp = \frac{TN}{TN+FP} \times 100\% \\
Acc = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \\
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}
\end{cases}
$$

where TP denotes true positive; TN denotes true negative; FP denotes false positive; and FN denotes false negative. Sn measures the predictive ability of a predictor for positive samples while Sp measures the predictive ability of a predictor for negative samples. Acc and MCC are two metrics measuring the overall performance of a predictor.

Besides, we used Receiver Operating Characteristic (ROC) curve to intuitively evaluate the overall performance (Liu et al., 2013, 2016b). It is plotted with true positive rate (TPR) against false positive rate (FPR) under different classification thresholds. The TPR is the same with sensitivity as described above, while FPR is calculated as 1-specificity. Area under ROC curve (AUC) is usually used as an evaluation metric (Liu et al., 2016a, 2017). The value of AUC ranges from 0.5 to 1. If the AUC is close to 1, it indicates that the predictor has excellent performance. If the

AUC approaches to 0.5, the predictor does not perform well for prediction.

Additionally, we used 10-fold cross validation method and jackknife test to evaluate the predictive performance (Wei et al., 2017a; Zeng et al., 2017a,b; Liao et al., 2018; Zou et al., 2018). The two evaluation methods were chosen since existing methods in the literature used them for performance evaluation.

RESULTS AND DISCUSSION

Comparison of XGBoost and Other Classifiers

To evaluate the effectiveness of the XGBoost classifier, we compared it with five commonly used machine learning algorithms, including Random Forest (RF) (Liu B. et al., 2015; Li et al., 2016; Wei et al., 2017b), Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN)(Huang and Li, 2018), Support Vector Machine (SVM) (Song et al., 2010, 2012, 2018; Wang M. et al., 2014; Wei et al., 2017), and Gradient Boosting Decision Tree (GBDT) (Liao et al., 2018), respectively. For fair comparison, the machine learning algorithms were trained and evaluated with 10-fold cross validation on the benchmark datasets, respectively. The performance of different classifiers is illustrated in **Figure 3**. The detailed results are presented in **Table 2**.

As shown in **Table 2** and **Figure 3**, XGBoost outperforms the other classifiers on three out of the four datasets, with the exception of Dataset-A101, for which the SVM classifier is slightly better than the XGBoost, which is the second best among the compared classifiers. For those datasets that the XGBoost outperforms other classifiers, the XGBoost is able to achieve higher Acc and MCC. To be specific, our Acc and MCC are 0.7314 and 0.4629 in the Dataset-S51, 0.6 and 1.1% higher
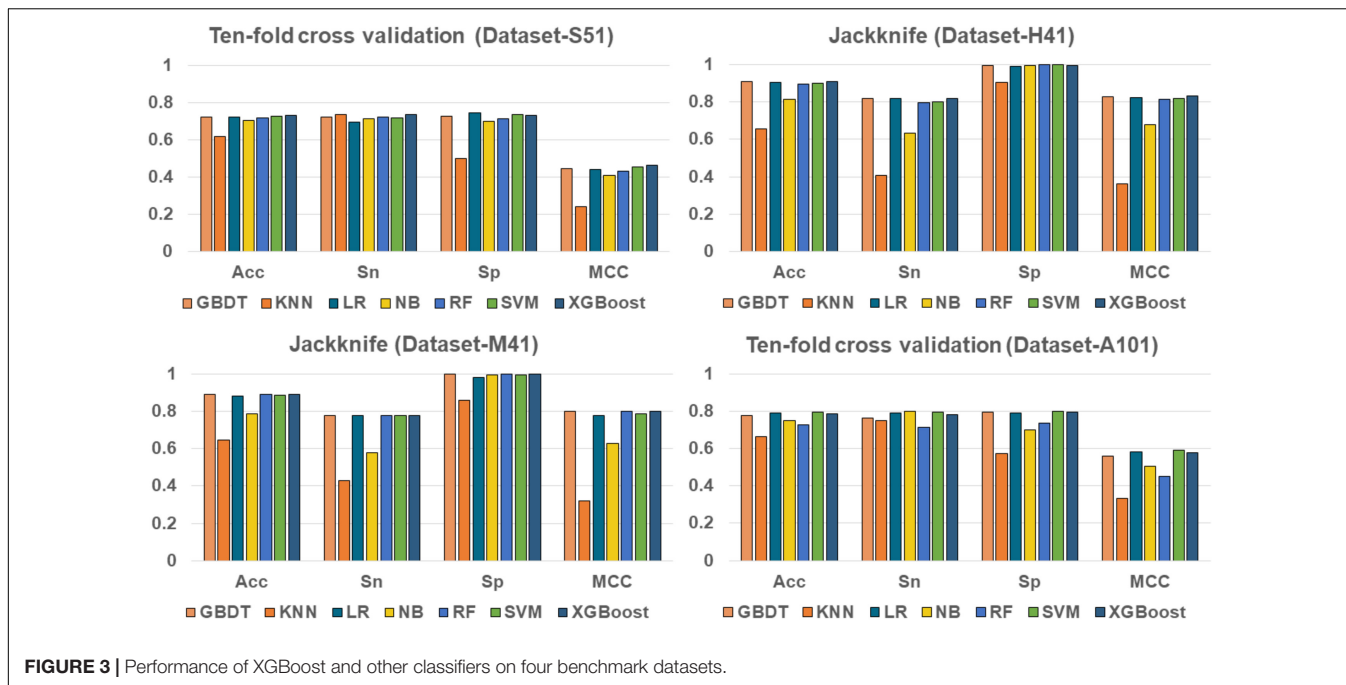
FIGURE 3 | Performance of XGBoost and other classifiers on four benchmark datasets.

TABLE 2 | Performances of XGBoost and other machine learning algorithms.

Dataset-S51	Acc	Sn	Sp	MCC	Dataset-H41	Acc	Sn	Sp	MCC
GBDT	0.7234	0.7200	0.7269	0.4468	GBDT	0.9089	0.8204	0.9973	0.8308
KNN	0.6167	0.7337	0.4996	0.2400	KNN	0.6566	0.4062	0.9071	0.3620
LR	0.7192	0.6924	0.7460	0.4390	LR	0.9066	0.8204	0.9929	0.8257
NB	0.7050	0.7100	0.7001	0.4101	NB	0.8155	0.6327	0.9982	0.6779
RF	0.7165	0.7192	0.7138	0.4331	RF	0.8982	0.7965	1.0000	0.8135
SVM	0.7257	0.7169	0.7345	0.4515	SVM	0.9018	0.8035	1.0000	0.8195
XGBoost	0.7314	0.7345	0.7284	0.4629	XGBoost	0.9089	0.8195	0.9982	0.8311
Dataset-M41	**Acc**	**Sn**	**Sp**	**MCC**	**Dataset-A101**	**Acc**	**Sn**	**Sp**	**MCC**
GBDT	0.8890	0.7779	1.0000	0.7979	GBDT	0.7795	0.7624	0.7967	0.5594
KNN	0.6448	0.4303	0.8593	0.3207	KNN	0.6638	0.7524	0.5752	0.3329
LR	0.8807	0.7793	0.9821	0.7775	LR	0.7914	0.7910	0.7919	0.5829
NB	0.7862	0.5793	0.9931	0.6288	NB	0.7517	0.8005	0.7029	0.5057
RF	0.8890	0.7779	1.0000	0.7979	RF	0.7260	0.7152	0.7367	0.4520
SVM	0.8848	0.7766	0.9931	0.7884	SVM	0.7971	0.7957	0.7986	0.5943
XGBoost	0.8890	0.7779	1.0000	0.7979	XGBoost	0.7890	0.7824	0.7957	0.5781

than that of the runner-up SVM. Similar results are observed in the Dataset-H41; XGBoost leads by 0.71 and 1.2% in terms of Acc and MCC, respectively. Moreover, in the Dataset-M41, the performances of our XGBoost are the same with the RF and GBDT in terms of Acc, Sn, Sp, and MCC, respectively. In summary, our results demonstrate that as compared to other commonly used classifiers, the XGBoost shows generally better and more robust performance to classify true m⁶A sites to non-m⁶A sites from different species.
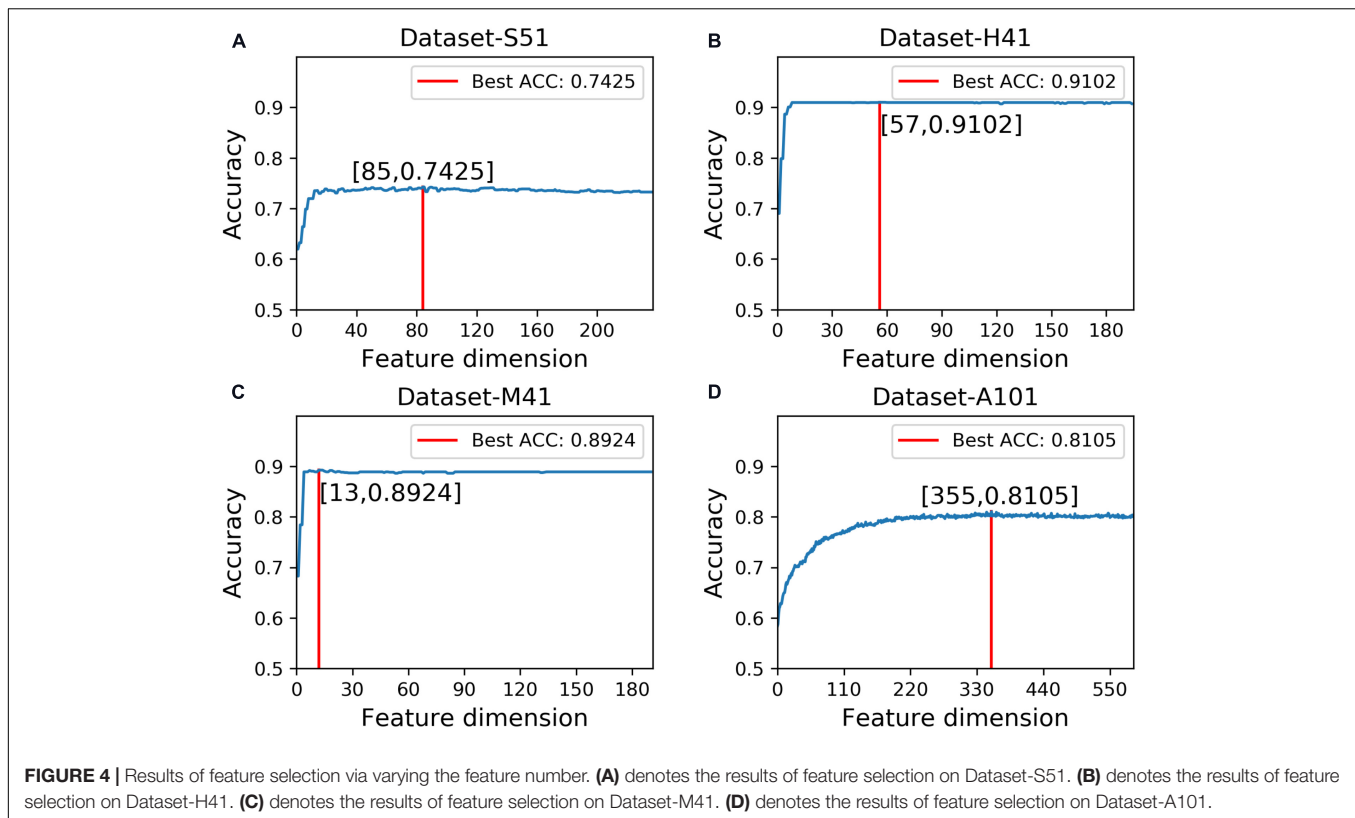
Impact of Feature Selection

In this study, we employed the F-score with the SFS for feature selection. The results of feature selection are summarized in **Table 3** and illustrated in **Figure 4** as well. As seen from **Table 3**,

TABLE 3 | Performance of features before and after feature selection.

Datasets	Methods	Acc	Sn	Sp	MCC
Dataset-S51	*Before*	*0.7314*	*0.7345*	*0.7284*	*0.4629*
	After	*0.7425*	*0.7521*	*0.7330*	*0.4852*
Dataset-H41	*Before*	*0.9089*	*0.8195*	*0.9982*	*0.8311*
	After	*0.9102*	*0.8204*	*1.0000*	*0.8339*
Dataset-M41	*Before*	*0.8890*	*0.7779*	*1.0000*	*0.7979*
	After	*0.8924*	*0.7890*	*0.9959*	*0.8022*
Dataset-A101	*Before*	*0.7890*	*0.7824*	*0.7957*	*0.5781*
	After	*0.8105*	*0.8067*	*0.8143*	*0.6210*

before feature selection, the performances of the predictive model in the Dataset-S51 are 0.7314, 0.7345, 0.7284, and 0.4629 in

FIGURE 4 | Results of feature selection via varying the feature number. **(A)** denotes the results of feature selection on Dataset-S51. **(B)** denotes the results of feature selection on Dataset-H41. **(C)** denotes the results of feature selection on Dataset-M41. **(D)** denotes the results of feature selection on Dataset-A101.

terms of Acc, Sn, Sp, and MCC, respectively. After applying the feature selection, we observed that the performances in terms of all the metrics were improved. To be specific, the Acc and MCC were improved to 0.7425 and 0.4852, respectively. This indicates that the feature selection strategy to yield more informative features to distinguish true m^6A sites from non-m^6A sites. For the other datasets from different species, similar results were observed. We can see from **Table 3** that almost all the performances were improved by using feature selection, demonstrating that feature selection is an effective way to enhance the predictive performance of the predictor. Moreover, **Figure 4** illustrates the Acc of the features by varying the feature number when conducting feature selection. As seen in **Figure 4**, we pointed out the optimal feature number and their corresponding highest Acc for each dataset. The optimal feature number for the four datasets are 85, 57, 13, and 355, giving the highest Acc of 0.7425, 0.9102, 0.8924, and 0.8105, respectively.

Comparison With Other Feature Representation Algorithms

To examine the performance of the proposed feature algorithm, we evaluated and compared it with existing feature representation algorithms, including RFH, PseDNC, PCP (physical and chemical properties), KNN (K-Nearest Neighbors), and AthMethPre, respectively. These algorithms were reported to have relatively strong power for the identification of m^6A sites. Thus, they were chosen for comparison. The results of

the above algorithms were presented in **Table 4**. As we can see from **Table 4**, the proposed features are competitive with the best-performing AthMethPre other feature representation methods and remarkably outperform the other existing features in all the four datasets. Note that for the Dataset-S51 and the Dataset-A101, our method performs slightly worse than the best-performing AthMethPre; while for the other two datasets, our method is slightly better. As well known, for the genome-wide identification, the running time for a predictor is important as well. Therefore, we further compared the feature number of AthMethPre and our feature representation method. We found that the feature number of the AthMethPre method for each dataset are 540, 500, 500, and 740, while ours are 85, 57, 13, and 355, respectively. As can be seen, our feature numbers for all the four datasets are averagely much fewer than the AthMethPre method. This indicates that the computation time by our predictive models costs less. In general, it can be concluded that our features are at least effective for the representatives of m^6A sites in multiple species with different sequence lengths.

Comparison With State-of-the-Art Predictors

To assess the effectiveness of our predictor, we compared it with existing predictors including pRNAm-PC (Liu Z. et al., 2016), MehtyRNA (Chen et al., 2017), and RFAthM6A (Wang and Yan, 2018), respectively. There were chosen since they were reported to have the best performance on the four benchmark

TABLE 4 | Comparison with other feature representation algorithms.

Dataset-S51	Acc	Sn	Sp	MCC	Dataset-H41	Acc	Sn	Sp	MCC
RFH	0.7295	0.7582	0.7008	0.4598	RFH	0.9097	0.8195	1	0.8332
PseDNC	0.64	0.6993	0.5807	0.282	PseDNC	0.6956	0.5973	0.7938	0.3989
PCP	0.627	0.6389	0.6151	0.2541	PCP	0.6447	0.6177	0.6717	0.2898
KNN	0.7131	0.6917	0.7345	0.4266	KNN	0.8235	0.7363	0.9106	0.657
AthMethPre	0.7536	0.7605	0.7467	0.5073	AthMethPre	0.9071	0.8142	1	0.8286
Our features	0.7425	0.7521	0.733	0.4852	Our features	0.9102	0.8204	1	0.8339
Dataset-M41	**Acc**	**Sn**	**Sp**	**MCC**	**Dataset-A101**	**Acc**	**Sn**	**Sp**	**MCC**
RFH	0.8903	0.7848	0.9959	0.7987	RFH	0.7993	0.7705	0.8281	0.5996
PseDNC	0.6228	0.6386	0.6069	0.2456	PseDNC	0.8138	0.8057	0.8219	0.6277
PCP	0.6166	0.5669	0.6662	0.2343	PCP	0.8257	0.8281	0.8233	0.6514
KNN	0.8283	0.7448	0.9117	0.6659	KNN	0.8238	0.8462	0.8014	0.6483
AthMethPre	0.8897	0.7793	1	0.799	AthMethPre	0.85	0.85	0.85	0.7
Our features	0.8924	0.789	0.9959	0.8022	Our features	0.8105	0.8067	0.8143	0.6210

TABLE 5 | Results of the proposed predictor and the state-of-the-art predictors on benchmark datasets from different species.

Dataset-S51	Acc	Sn	Sp	MCC	Dataset-H41	Acc	Sn	Sp	MCC
pRNAm-PC	0.6974	0.6972	0.6975	0.4000	MethyRNA	0.9038	0.8168	0.9911	N.A.
M6AMRFS	0.7425	0.7521	0.7330	0.4852	M6AMRFS	0.9102	0.8204	1.0000	0.8339
Dataset-M41	**Acc**	**Sn**	**Sp**	**MCC**	**Dataset-A101**	**Acc**	**Sn**	**Sp**	**MCC**
MethyRNA	0.8839	0.7779	1.0000	N.A.	RFAthM6A	0.8545	0.8738	0.8352	0.7095
M6AMRFS	0.7933	0.8281	0.7584	0.588	M6AMRFS	0.8105	0.8067	0.8143	0.6210

N.A., denotes not available.

datasets used in this work. The results were presented in **Table 5**.

As shown in **Table 5**, M6AMRFS outperforms pRNAm-PC on the Dataset-S51. The Acc, Sn, Sp, and MCC by our predictor are 0.7425, 0.7521, 0.7339, and 0.4852, respectively. The performances are higher than that of the second best pRNAm-PC on this dataset. To be specific, our overall performances are 0.0451 and 0.0852 higher in terms of Acc and MCC, respectively. As for the other datasets (Dataset-H41 and Dataset-M41), we observed similar results that our overall performance outperforms the existing predictors. Only on Dataset-A101, our predictor performs slightly worse than RFAthM6A. To be concluded, our results demonstrate that the proposed predictor is better than existing predictors or at least competitive with existing predictors on multiple benchmark datasets from different species. Importantly, our predictor exhibits robust performance for multiple species, demonstrating that our predictor is able of capturing the characteristics of m6A sites in different species. This also implies that the m6A sites from different species might share the common patterns.

CONCLUSION

In this study, we have developed a machine learning based predictor, namely M6AMRFS, for the identification of m6A sites in multiple species. We have conducted a series of comparative study, and our experimental results indicate that our predictor is at least competitive as compared to previously published predictors. Importantly, we found that our predictor is able to achieve robust performance in several species. To the best of our knowledge, it is the first predictor that can provide predictions in multiple species. For further analysis, we found that the robust performance contributes to the following two possible reasons. One reason is the XGBoost classifier we used for model training. We have compared XGBoost with other machine learning algorithms. XGBoost is shown to perform better than other classification algorithms. The other reason is that our feature selection strategy helps to adaptively select the optimal features for specific species. We anticipate that the tool and webserver we have established will be useful for facilitating to reveal the functional mechanisms of m6A sites.

AUTHOR CONTRIBUTIONS

XQ and HC wrote the manuscript. HC developed the webserver and analyzed the results. XY analyzed the results. RS and LW designed the experiments. All authors read and approved the manuscript.

FUNDING

REFERENCES

Alarcón, C. R., Lee, H., Goodarzi, H., Halberg, N., and Tavazoie, S. F. (2015). N 6-methyladenosine marks primary microRNAs for processing. *Nature* 519, 482–485. doi: 10.1038/nature14281

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, (New York, NY: ACM), 785–794. doi: 10.1145/2939672.2939785

Chen, T., Hao, Y.-J., Zhang, Y., Li, M.-M., Wang, M., Han, W., et al. (2015). m 6 A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell* 16, 289–301. doi: 10.1016/j.stem.2015.01.016

Chen, W., Feng, P., Ding, H., and Lin, H. (2016). Identifying N6-methyladenosine sites in the *Arabidopsis thaliana* transcriptome. *Mol. Genet. Genomics* 291, 2225–2229.

Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.-C. (2015a). iRNA-methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* 490, 26–33. doi: 10.1016/j.ab.2015.08.021

Chen, W., Tang, H., and Lin, H. (2017). MethyRNA: a web server for identification of N6-methyladenosine sites. *J. Biomol. Struct. Dyn.* 35, 683–687. doi: 10.1080/07391102.2016.1157761

Chen, W., Tran, H., Liang, Z., Lin, H., and Zhang, L. (2015b). Identification and analysis of the N 6-methyladenosine in the *Saccharomyces cerevisiae* transcriptome. *Sci. Rep.* 5:13859. doi: 10.1038/srep13859

Cheng, J.-H., Yang, H., Liu, M.-L., Su, W., Feng, P.-M., Ding, H., et al. (2018). Prediction of bacteriophage proteins located in the host cell using hybrid features. *Chemometr. Intell. Lab. Syst.* 180, 64–69.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Geula, S., Moshitch-Moshkovitz, S., Dominissini, D., Mansour, A. A., Kol, N., Salmon-Divon, M., et al. (2015). m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* 347, 1002–1006. doi: 10.1126/science.1261417

Huang, G. H., and Li, J. C. (2018). Feature extractions for computationally predicting protein post-translational modifications. *Curr. Bioinform.* 13, 387–395. doi: 10.2174/1574893612666170707094916

Jia, C. Z., Zhang, J. J., and Gu, W. Z. (2016). RNA-MethylPred: a high-accuracy predictor to identify N6-methyladenosine in RNA. *Anal. Biochem.* 510, 72–75. doi: 10.1016/j.ab.2016.06.012

Lai, H.-Y., Chen, X.-X., Chen, W., Tang, H., and Lin, H. (2017). Sequence-based predictive modeling to identify cancerlectins. *Oncotarget* 8, 28169–28175. doi: 10.18632/oncotarget.15963

Li, F., Li, C., Revote, J., Zhang, Y., Webb, G. I., Li, J., et al. (2016). GlycoMine struct: a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features. *Sci. Rep.* 6:34595. doi: 10.1038/srep34595

Liao, Z. J., Wan, S. X., He, Y., and Zou, Q. (2018). Classification of small GTPases with hybrid protein features and advanced machine learning techniques. *Curr. Bioinform.* 13, 492–500. doi: 10.2174/1574893612666171121162552

Liu, B. (2017). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* doi: 10.1093/bib/bbx165 [Epub ahead of print].

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, W65–W71. doi: 10.1093/nar/gkv458

Liu, B., Long, R., and Chou, K.-C. (2016a). iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. *Bioinformatics* 32, 2411–2418. doi: 10.1093/bioinformatics/btw186

Liu, B., Wang, S., Long, R., and Chou, K.-C. (2016b). iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics* 33, 35–41. doi: 10.1093/bioinformatics/btw539

Liu, B., Yang, F., Huang, D.-S., and Chou, K.-C. (2017). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 34, 33–40. doi: 10.1093/bioinformatics/btx579

Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., et al. (2013). Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* 30, 472–479. doi: 10.1093/bioinformatics/btt709

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N 6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* 518, 560–564. doi: 10.1038/nature14234

Liu, Z., Xiao, X., Yu, D. J., Jia, J., Qiu, W. R., and Chou, K. C. (2016). pRNAm-PC: predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. *Anal. Biochem.* 497, 60–67. doi: 10.1016/j.ab.2015.12.017

Maden, B. (1990). "The numerous modified nucleotides in eukaryotic ribosomal RNA," in *Progress in Nucleic Acid Research and Molecular Biology*, Vol. 39, ed. K. Moldave (Amsterdam: Elsevier), 241–303.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* 149, 1635–1646. doi: 10.1016/j.cell.2012.05.003

Mrozek, D., Malysiak, B., and Kozielski, S. (2007). "An optimal alignment of proteins energy characteristics with crisp and fuzzy similarity awards," in *Proceedings of the 2007 IEEE International Fuzzy Systems Conference*, (London: IEEE), 1–6. doi: 10.1109/FUZZY.2007.4295590

Mrozek, D., Małysiak-Mrozek, B., and Kłapciński, A. (2014). Cloud4Psi: cloud computing for 3D protein structure similarity searching. *Bioinformatics* 30, 2822–2825. doi: 10.1093/bioinformatics/btu389

Mrozek, D., Malysiak-Mrozek, B., and Kozielski, S. (2009). "Alignment of protein structure energy patterns represented as sequences of Fuzzy Numbers," in *NAFIPS 2009 Annual Meeting of the North American Fuzzy Information Processing Society*, (Cincinnati, OH: IEEE), 1–6. doi: 10.1109/NAFIPS.2009.5156391

Mrozek, D., Małysiak-Mrozek, B., and Siążnik, A. (2013). search GenBank: interactive orchestration and ad-hoc choreography of Web services in the exploration of the biomedical resources of the National Center For Biotechnology Information. *BMC Bioinformatics* 14:73. doi: 10.1186/1471-2105-14-73

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Roost, C., Lynch, S. R., Batista, P. J., Qu, K., Chang, H. Y., and Kool, E. T. (2015). Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification. *J. Am. Chem. Soc.* 137, 2107–2115. doi: 10.1021/ja513080v

Song, J., Tan, H., Perry, A. J., Akutsu, T., Webb, G. I., Whisstock, J. C., et al. (2012). PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS One* 7:e50300. doi: 10.1371/journal.pone.0050300

Song, J., Tan, H., Shen, H., Mahmood, K., Boyd, S. E., Webb, G. I., et al. (2010). Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* 26, 752–760. doi: 10.1093/bioinformatics/btq043

Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N. D., Webb, G. I., et al. (2018). iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinform.* doi: 10.1093/bib/bby028 [Epub ahead of print].

Su, Z.-D., Huang, Y., Zhang, Z.-Y., Zhao, Y.-W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* doi: 10.1093/bioinformatics/bty508 [Epub ahead of print].

Tang, H., Zhao, Y.-W., Zou, P., Zhang, C.-M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Wang, M., Zhao, X. M., Tan, H., Akutsu, T., Whisstock, J. C., and Song, J. (2014). Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. *Bioinformatics* 30, 71–80. doi: 10.1093/bioinformatics/btt603

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., et al. (2014). N 6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* 505, 117–120. doi: 10.1038/nature12730

Wang, X., and Yan, R. (2018). RFAthM6A: a new tool for predicting m6A sites in *Arabidopsis thaliana*. *Plant Mol. Biol.* 96, 327–337. doi: 10.1007/s11103-018-0698-9

Wei, L., Chen, H., and Su, R. (2018a). M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acids* 12, 635–644. doi: 10.1016/j.omtn.2018.07.004

Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018b). Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* 117, 212–217. doi: 10.2174/1566523218666180913110949

Wei, L., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2018c). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics*. doi: 10.1093/bioinformatics/bty824

Wei, L., Su, R., Wang, B., Li, X., and Zou, Q. (2018d). Integration of deep feature representations and handcrafted features to improve the prediction of N 6-methyladenosine sites. *Neurocomputing*. doi: 10.1016/j.neucom.2018.04.082

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018e). ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* doi: 10.1093/bioinformatics/bty451 [Epub ahead of print].

Wei, L., Tang, J., and Zou, Q. (2017). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026

Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2017a). Fast prediction of methylation sites using sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2017.2670558 [Epub ahead of print].

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z., and Zou, Q. (2017b). CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J. Proteome Res.* 16, 2044–2053. doi: 10.1021/acs.jproteome.7b00019

Xiang, S., Liu, K., Yan, Z., Zhang, Y., and Sun, Z. (2016a). RNAMethPre: a web server for the prediction and query of mRNA m6A sites. *PLoS One* 11:e0162707. doi: 10.1371/journal.pone.0162707

Xiang, S., Yan, Z., Liu, K., Zhang, Y., and Sun, Z. (2016b). AthMethPre: a web server for the prediction and query of mRNA m 6 A sites in *Arabidopsis thaliana*. *Mol. Biosyst.* 12, 3333–3337.

Yang, H., Qiu, W.-R., Liu, G., Guo, F.-B., Chen, W., Chou, K.-C., et al. (2018). iRSpot-Pse6NC: identifying recombination spots in *Saccharomyces cerevisiae*

by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci.* 14, 883–891. doi: 10.7150/ijbs.24616

Zeng, J., Li, D., Wu, Y., Zou, Q., and Liu, X. (2016). An Empirical study of features fusion techniques for protein-protein interaction prediction. *Curr. Bioinform.* 11, 4–12. doi: 10.2174/1574893611666151119221435

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017a). "Prediction and validation of disease genes using HeteSim Scores," in *Proceedings of the IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, Vol. 14, (Piscataway, NJ: IEEE), 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017b). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* 13:e1005420. doi: 10.1371/journal.pcbi.1005420

Zeng, X., Zhang, X., and Zou, Q. (2015). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinform.* 17, 193–203. doi: 10.1093/bib/bbv033

Zhang, X., Zou, Q., and Rodriguez-Paton, A. (2017). "Meta-path methods for prioritizing candidate disease miRNAs," in *Proceedings of the IEEE/ACM Transactions on Computational Biology and Bioinformatics*, (Piscataway, NJ: IEEE).

Zhou, Y., Zeng, P., Li, Y. H., Zhang, Z., and Cui, Q. (2016). SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res.* 44:e91. doi: 10.1093/nar/gkw104

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* doi: 10.1093/bib/bby090

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016a). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* 10:114. doi: 10.1186/s12918-016-0353-5

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016b). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123