Briefings in Functional Genomics, 20(4), 2021, 258–272

https://doi.org/10.1093/bfgp/elaa028
Advance Access Publication Date: 25 January 2021
Review Paper

Critical evaluation of web-based DNA N6-methyladenine site prediction tools

Md Mehedi Hasan, Watshara Shoombuatong, Hiroyuki Kurata and Balachandran Manavalan

Corresponding author: Balachandran Manavalan; E-mail: bala@ajou.ac.kr

Abstract

Methylation of DNA N6-methyladenosine (6mA) is a type of epigenetic modification that plays pivotal roles in various biological processes. The accurate genome-wide identification of 6mA is a challenging task that leads to understanding the biological functions. For the last 5 years, a number of bioinformatics approaches and tools for 6mA site prediction have been established, and some of them are easily accessible as web application. Nevertheless, the accurate genome-wide identification of 6mA is still one of the challenging works that lead to understanding the biological functions. Especially in practical applications, these tools have implemented diverse encoding schemes, machine learning algorithms and feature selection methods, whereas few systematic performance comparisons of 6mA site predictors have been reported. In this review, 11 publicly available 6mA predictors evaluated with seven different species-specific datasets (*Arabidopsis thaliana, Tolypocladium, Diospyros lotus*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Escherichia coli*). Of those, few species are close homologs, and the remaining datasets are distant sequences. Our independent, validation tests demonstrated that Meta-i6mA and MM-6mAPred models for *A. thaliana*, *Tolypocladium*, *S. cerevisiae* and *D. melanogaster* achieved excellent overall performance when compared with their counterparts. However, none of the existing methods were suitable for *E. coli*, *C. elegans* and *D. lotus*. A feasibility of the existing predictors is also discussed for the seven species. Our evaluation provides useful guidelines for the development of 6mA site predictors and helps biologists selecting suitable prediction tools.

Key words: DNA N6-methyladenine site; sequence analysis; machine learning; prediction model; web servers

Introduction

DNA methylation plays essential roles in controlling tissue-specific gene expression, gene imprinting X-chromosome inactivation, transcript synthesis and positioning and stability of nucleosome [1, 2]. Two main types of epigenetic markers in both of prokaryotes and eukaryotes are 5-methylcytosine (5mC) and N6-methyladenosine (6mA) [3]. Covalent DNA modifications on 5mC are well known to play critical epigenetic roles in regulating gene expression [4–6]. The irregular 5mC has implicated diseases including metabolic disorders, autoimmune diseases and cancer [7–9]. The 6mA is a novel DNA adenine modification, which is widespread over prokaryotes, has been recently found in the genomes of higher eukaryotes, including worms, fruit flies, mice, pigs, zebrafish, green algae and frogs [10, 11]. This modification, which is distributed in the genomes of many

Md. Mehedi Hasan received his PhD degree in bioinformatics from CAU, Beijing in 2016. He is currently as a Japan Society for the Promotion of Science (JSPS) international postdoctoral fellow in the Kyushu Institute of Technology, Japan. Before his current position, he worked as a researcher at Chinese University of Hong Kong, Hong Kong. His main research interests include protein structure prediction, machine learning, data mining, computational biology and functional genomics.

Watshara Shoombuatong is an assistant professor in the Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University. He has a very strong specialty in machine learning, data mining, bioinformatics and computational biology, and protein and peptide sequence analysis.

Hiroyuki Kurata is a professor of the Department of Bioscience and Bioinformatics in the Kyushu Institute of Technology, Japan. His research interests primarily focus on systems biology, synthetics biology, functional genomics, machine learning and their applications.

Balachandran Manavalan is an assistant professor at the Department of Physiology, Ajou University School of Medicine, Republic of Korea. He is also an associate member of Korea Institute for Advanced Study, Republic of Korea. His main research interests include prediction of protein structures, machine learning, data mining, computational biology and functional genomics.

species and amplifies genomic assortment, is responsible for regulation of the genomic imprinting, gene expression and cell developments [12–14]. The 6mAs discriminate the host DNA from foreign pathogenic DNA and defend the host genome via the several modification systems [15–18]. Investigation of the discriminating ability of host DNA from foreign DNA specifies confident modifications, such as adenine methylation, which may defend the host DNA from enzyme-mediated degradation. In common, the 6mA is roughly and evenly distributed across diverse genomes [14, 19–21]. Although much work has been done to characterize 6mAs from diverse genomes, the mechanisms by which the 6mA regulates the gene expression and cell cycle are hardly known [22, 23]. Their epigenetic modifications and functions remain largely unclear, whereas it may be associated to the modification in the size, genome intricacy and complicated epigenetic developments.

Although many tools were developed for functional representation and unbiased detection of DNA methylation sites, its abilities for solving these two tasks were not yet satisfactory. To promote the study of the 6mA, various experimental techniques have been reported, including the bisulfite sequencing of whole-genome and single-molecule real-time (SMRT) sequencing [24, 25]. Recently, with the development of deep sequencing, 6mA was found to be present in a number of plants, eukaryotes and prokaryotic, including *Chlamydomonas reinhardti* [26] fungi [27], mouse [28], zebrafish and pig [29], *Arabidopsis thaliana* [30], *Tolypocladium* [29, 31], *Diospyros lotus* [32, 33], *Saccharomyces cerevisiae* [3, 34], *Drosophila melanogaster* [25], *Caenorhabditis elegans* [35] and *Escherichia coli* [32, 36]. Therefore, the study of 6mA modification becomes pervasive and leads to an understanding of mechanisms by which 6mA regulates cellular functions in the diverse sets of genomes.

It is well recognized that high-throughput biological assays for large-scale genome sequencing is a gold-standard method in this field. However, this approach is time-consuming and expensive. Nowadays, machine learning (ML) approaches have appeared as a promising predictor that could resolve these problems and help scientists to identify 6mA sites. Therefore, it is urgently needed to develop ML-based models for rapidly and accurately predicting the potential sites of 6mA. They can be used as a complement to the experimental efforts [37–40]. To date, several computational tools have been developed for 6mA prediction [41–44] using the publicly accessible genomics database (https://www.ncbi.nlm.nih.gov/geo/). In the viewpoint of ML, the high-quality datasets could guarantee the predictive abilities of computational approaches to identify potential 6mA sites in DNA sequences as well as the prediction of 6mA modification [15]. Figure 1 shows an overview of the existing computational approaches. Although much progress has been made, there is still room for further improvement. Firstly, most of existing methods were trained by different training datasets. It is very difficult to determine the most powerful model for 6mA prediction. Secondly, their prediction performances were not validated by using independent datasets, despite the independent test being the most rigorous cross-validation (CV) method. Therefore, it could not be stated that the prediction results obtained by these methods are reliable and robust in real applications. Motivated by these considerations, an unbiased evaluation of 6mA prediction tools is carried by constructing a well-constructed independent, validation dataset.

In this work, we considered the 6mAs of the seven species of *A. thaliana*, *Tolypocladium*, *D. lotus*, *S. cerevisiae*, *D. melanogaster*, *C. elegans* and *E. coli* and discussed the specification of existing 6mA site predictors in terms of the ML algorithms, feature encoding schemes, prediction performance and webserver efficacy. In total, we examined the 11 6mA prediction tools, including Meta-i6mA [45], i6mA-Fuse [46], i6mA-stack [47], SDM6A [43], iDNA6mA-rice [48], 6mA-Finder [49], MM-6mAPred [50], i6mA-Pred [42], iDNA6mA-PseKNC [44], iDNA6mA [51] and 6mAPred-FO [52]. We constructed our validation datasets representing the overall 6mA and non-6mA patterns in the entire genome of each specific species and carrying out an unbiased assessment of these web-based 6mA prediction tools. Even though some predictors yielded outstanding performance for specific species, none of them were not able to predict the 6mA sites of *E. coli*, *D. lotus and C. elegans*. The curated analysis facilitates the improvement of the predictors for 6mA sites.

Materials and methods

Overview of computational approaches

Figure 1 shows a synopsis of the existing computational approaches for 6mA site prediction that includes four steps. In the first step, a high-quality 6mA dataset was constructed based on authenticated databases and literature searches. As the experimentally validated non-6mA sites are not available, sequence windows with 41-bp fragments having adenine at the central position are generated from the entire chromosomal DNA and discarded the fragments overlapped experimentally detected 6mA sites (i.e. positive samples). The resultant samples are considered as non-6mAs. To thwart overestimation problems, CD-HIT [53] is commonly applied to remove redundant sequences. The majority of the samples (80 or 70%) were randomly selected to train or develop the prediction model. However, the leftover sample is an independent dataset used to quantify model robustness. In the second step, feature extraction, analysis and optimization are performed. In general, a diversity of feature representation approaches are used to detect significant signals that distinguish 6mA from non-6mA samples, including composition-based features [45, 54, 55], k-mer nucleotide properties (Kmer) [56, 57], reverse complementary Kmer, enhanced nucleic acid composition (ENAC), k-space nucleotide pairs composition (KNC), pseudo-di and tri-nucleotide compositions (PseDNC and PseTNC), parallel correlation PseTNC, parallel correlation PseDNC, pseudo K-tuple nucleotide composition (PseKNC) [58, 59], series correlation PseDNC and series correlation PseTNC. The position-specific based features [50] are the mononucleotide binary encoding (MBE) [60, 61], dinucleotide binary encoding (DBE), position-specific propensity for trinucleotide [62] and accumulated nucleotide frequency (ANF). The physicochemical property-based features [49] are the ring-functions of hydrogen chemical (RFHC), electron–ion-interaction pseudopotential (EIIP), dinucleotide-based physicochemical properties (DPCP), dinucleotide-based auto-covariance, dinucleotide-based cross-correlation and physicochemical properties for trinucleotide (TPCP). The evolutionary-based features are the k-nearest neighbor (KNN)-derived features. To remove redundant features from the dataset, diverse feature optimization protocols were used, including a popular two-step feature selection [ranking followed by sequential forward search (SFS)] and recursive feature elimination (RFE) [45, 54, 55]. In the third step, based on the investigation of different methodologies and algorithms, the prediction model is developed. Precisely, from the second step, the optimum features from each encoding are inputted to several ML algorithms including extreme gradient boosting (XGBoost), support vector machine (SVM), random forest (RF)

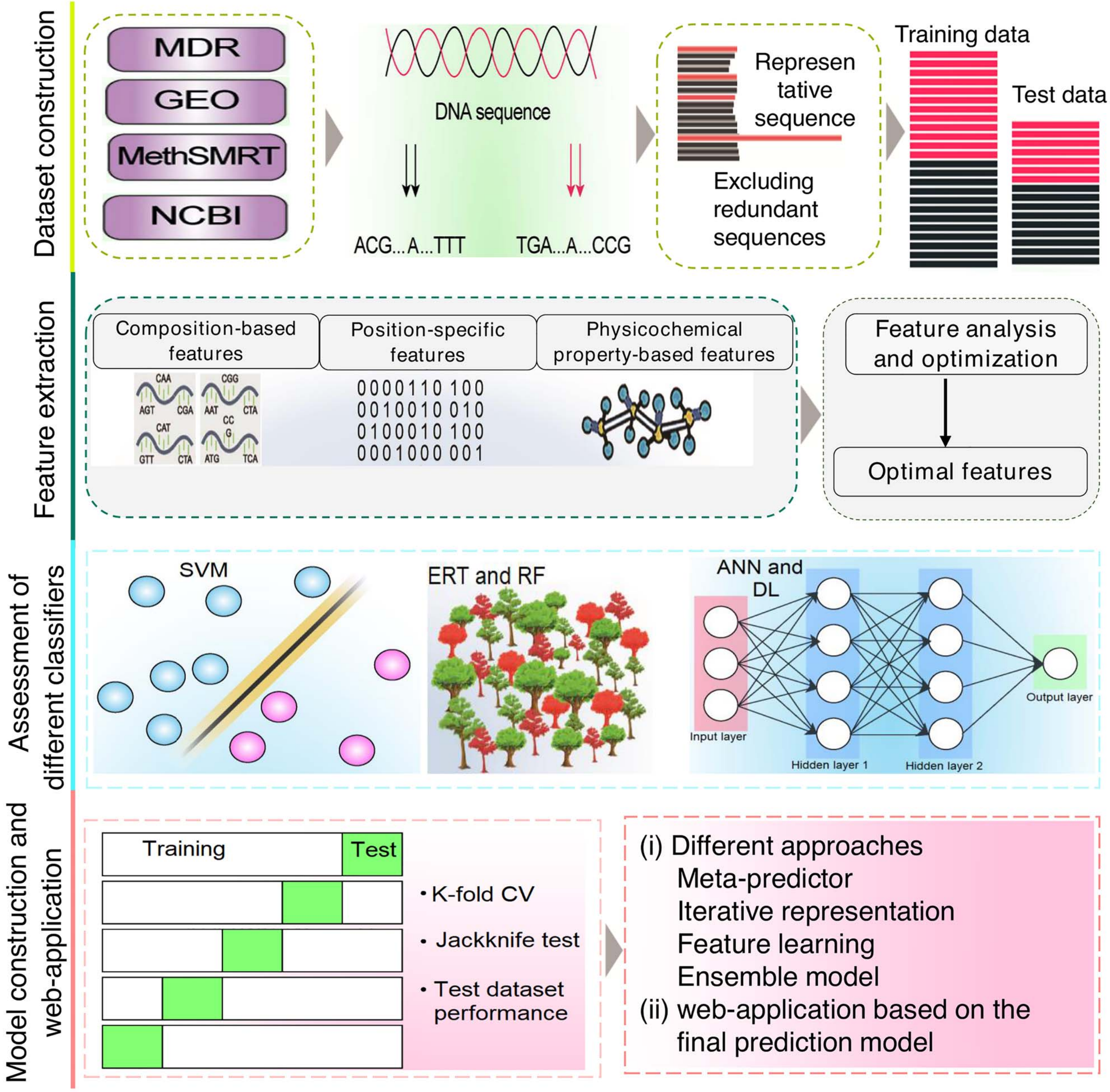

Figure 1. Synopsis of the existing computational approaches for 6mA site prediction. To develop a useful predictor, the following necessary steps are: (i) dataset construction; (ii) feature extraction, analysis and optimization; (iii) assessment of different ML models and selection of the suitable classifier and (iv) model construction based on different CV test and web application construction.

and deep learning (DL) to advance a prediction model. In the final step, the appropriate model is selected by investigating different approaches and methodologies.

Construction of validation datasets

We constructed a new validation dataset to evaluate the existing 6mA site prediction methods. Specifically, we considered seven species (*A. thaliana*, *Tolypocladium*, *D. lotus*, *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *E. coli*), whose positive samples (6mAs) were taken from the MethSMRT database [32]. We downloaded all raw data from MethSMRT, yielding the 41-bp sequence windows containing adenine (i.e. 6mA sites) at the center, with varying (modQV). Subsequently, we excluded the sequences containing no modification (ModQV) score for each species.

To construct the validation datasets of the seven species, we considered the samples with ≥20 modQV scores. Eventually, the 6mA samples indicating greater than 75% sequence identity were removed using CD-HIT to collect a high-quality species-wise dataset [53]. To build the non-6mA samples, we employed the same protocol as accessed in previous studies [45, 54, 55]. Notably, we generated a massive amount of 41 bp sequence windows containing central adenine from the entire chromosome and excluded the sequences that share greater than 75% with positive samples for each species. The validation datasets are available at http://kurata14.bio.kyutech.ac.jp/Meta-i6mA/do

Table 1. Summary of the newly constructed validation dataset

Genomes	6mAs	Non-6mAs
A. thaliana	60,700	121,400
Tolypocladium	200	1000
D. lotus	310	1550
S. cerevisiae	750	2250
C. elegans	23,100	46,200
D. melanogaster	26,700	53,400
E. coli	33,500	67,000

The first column characterizes species-wise genome names. The 2nd an 3rd columns represent the numbers of 6mA and non-6mA samples constructed in this study, respectively.

wnload_file/6mAs-BrifG.zip. Table 1 shows a statistics of the validation dataset of 6mA and non-6mA samples for each species used in this study.

Evaluation metrics

To evaluate the developed models, we employed the four commonly used statistical evaluation metrics [63–69], including sensitivity (Sn), specificity (Sp), accuracy (ACC), Matthews correlation coefficient (MCC) [60, 70–74]. The ACC, Sn, Sp and MCC are given by

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}},$$
$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$
$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},$$
$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN})}},$$

where TP, TN, FP and FN represent the number of 6mAs correctly predicted as 6mAs, the number of non-6mAs correctly predicted as non-6mAs, the number of 6mAs incorrectly predicted as non-6mAs, and the number of non-6mAs incorrectly predicted as 6mAs, respectively.

Summary of the existing 6mA prediction tools

Due to the rapid progress of high-throughput technologies, researchers studied the functional role of 6mAs extensively. With the advance of ML algorithms and the accumulation of experimental biological data, several computational predictors have been developed for identifying potential 6mA sites in genomes. The existing 6mA predictors are summarized in Table 2 that utilizes a wide range of feature encoding schemes, optimization approaches and ML classifiers. A detailed explanation of existing prediction methods is described below.

Predictors proposed in 2019

Eight ML-based predictors were reported in 2019. Of these, seven predictors were developed simultaneously with diverse methodologies using the rice genome dataset. The remaining predictor was constructed using the mouse genome dataset. Notably, both the rice and mouse 6mA sites were taken from the MethSMRT database [32].

iDNA6mA-PseKNC

Feng *et al.* [44] developed the first bioinformatics tool iDNA6mA-PseKNC to predict 6mAs based on the mouse dataset. They used the PseKNC descriptor and SVM for model development. During the Jackknife test, iDNA6mA-PseKNC achieved an Sn, Sp, ACC and MCC values of 93.28%, 100%, 96.73% and 0.930, respectively. Instead of evaluating their prediction model with the same species, they evaluated with eight other species (*C. elegans*, *A. thaliana*, *E. coli*, *Acidobacteria bacterium*, *Alteromonadaceae bacterium*, *Polycyclovorans algicola*, *Ruminococcus flavefaciens* and *Sphingomonas melonis*) and showed the success rate with a range of 82.2–99.44%. This method has been established at http://lin-group.cn/server/iDNA6mA-PseKNC.

i6mA-Pred

Chen *et al.* [42] developed i6mA-Pred based on the rice genome, where they utilized the RFHC descriptor and SVM for the model development. The authors have applied a two-step feature selection technique to identify the optimal feature set during model development, where the features were ranked with the maximum relevance maximum distance (MRMD) method followed by SFS using an SVM classifier. i6mA-Pred achieved an MCC, ACC, Sp and Sn of 0.66, 83.1%, 83.3% and 83.0%, respectively, during the jackknife test using a nonredundant (nr) training dataset containing 880 6mAs and 880-non-6mA. However, this model did not show model robustness on independent datasets. Indeed, this first method acted as a base for the development of later methods. Notably, the training dataset utilized in i6mA-Pred hereafter is mentioned as the i6mA-Pred dataset. The program of i6mA-Pred is accomplished at http://lin-group.cn/server/i6mA-Pred.

iDNA6mA

Tahir *et al.* [51] developed iDNA6mA based on the DL approach and utilized i6mA-Pred dataset. They employed MBE encoding to convert DNA sequence into a 164-dimensional feature vector and created the final model. iDNA6mA achieved an MCC, ACC, Sp and Sn of 0.730, 86.6%, 86.6% and 86.7%, respectively, and showed it significantly outperformed the existing predictor i6mA-Pred. Similar to i6mA-Pred, the authors did not evaluate their model robustness using an independent dataset. The iDNA6mA program is available at https://home.jbnu.ac.kr/NSCL/iDNA6mA.htm.

iDNA6mA-Rice

Lv *et al.* [48] constructed an nr training dataset containing 154,000 6mAs and 154,000 non-6mAs and explored three different encodings (MBE, KNC and natural vector) and their combinations using the RF classifier. They demonstrated that the MBE-based RF model achieved an MCC, ACC, Sp and Sn of 0.835, 91.7%, 93.0% and 90.5%, respectively, and named this model as iDNA6mA-Rice. Unlike the existing methods, they evaluated their model using an independent dataset and showed the MCC, ACC, Sn and Sp of 0.891, 94.6%, 95.8% and 93.3%, respectively. Notably, the training dataset utilized in iDNA6mA-Rice hereafter is referred as the iDNA6mA-Rice dataset. iDNA6mA-Rice is available at http://lin-group.cn/server/iDNA6mA-Rice.

SDM6A

Basith *et al.* [43] utilized the i6mA-Pred dataset and developed an integrated approach called SDM6A. They employed five different encodings, namely RFHC, numerical information of nucleotides, MBE, DBE and KNN-derived features, and generated their respective optimal model using SVM and extremely randomized tree (ERT). They employed a two-step feature selection approach, where the features were ranked by the F-score, followed by the

Table 2. A statistics of present available tools for DNA 6mAs prediction

Year	Predictor name	Feature encoding	Algorithm	TR/IND dataset size	Web-server URL	Testing method	ACC (%) TR/IND	Active web server	Option of batch prediction	Genomes
2019	iDNA6mA-PseKNC	PseKNC	SVM	1934 6mA; 1934 non-6mA/—	http://lin-group.cn/server/iDNA6mA-PseKNC	Jackknife test	96.70/—	Yes	NA	Mouse
	SD6MA	NUM, MBE, DBE, LPF	RF, ERT, GB, SVM	880 6mA; 880 non-6mA/221 6mA; 221 non-6mA	http://thegleelab.org/SD6MA/	10-fold CV and IND test	88.22/88.01	Yes	Yes	Rice
	iDNA6mA	MBE	CNN	880 6mA; 880 non-6mA/—	https://home.jbnu.ac.kr/NSCL/iDNA6mA.htm	Jackknife test	86.60/—	Yes	Yes	Rice
	iDNA6mA-Rice	MBE	RF	154 000 6mA; 154 000 non-6mA/880 6mA; 880 non-6mA	http://lin-group.cn/server/iDNA6mA-Rice	5-fold CV and IND test	91.71/94.0	Yes	Yes	Rice
	SNNRice6mA	MBE	DL	154 000 6mA; 154 000 non-6mA/880 6mA; 880 non-6mA	https://github.com/yuht4/SNNRice6mA	5-fold CV	90.2/—	No	NA	Rice
	MM-6mAPred	NCP	Markov model	880 6mA; 880 non-6mA/—	http://www.insect-genome.com/MM-6mAPred/	10-fold CV	89.72/—	Yes	Yes	Rice
	i6mA-Pred	NCP KNC	SVM	880 6mA; 880 non-6mA/—	http://lin-group.cn/server/i6mA-Pred/	Jackknife test	83.10/—	Yes	Yes	Rice
	i6mA-DNCP	DNC	CART	880 6mA; 880 non-6mA/221 6mA; 221 non-6mA	https://ww2.mathworks.cn/matlabcentral/fileexchange/72549-i6mA-dncp	10-fold CV	86.6/—	No	NA	Rice
2020	p6mA	EIIP, PseTNP, PP, MRMD	GB	3040 6mA; 3040 non-6mA/	https://github.com/Konglab404/p6mA	10-fold CV	82.01/76.81	No	NA	Fruit fly, worn and human
	6mA-RicePred	NCP, Kmer, MBE	SVM	880 6mA; 880 non-6mA/154000 6mA; 154 000 non-6mA	https://github.com/huangqianfei0916/6mA-rice	10-fold CV IND test	87.3/85.6	No	NA	Rice

Continued

Table 2. Continued

Year	Predictor name	Feature encoding	Algorithm	TR/IND dataset size	Web-server URL	Testing method	ACC (%) TR/IND	Active web server	Option of batch prediction	Genomes
	6mA-Finder	ANF, MBE, KNC, DNC, ENAC, EIIP, NCP, PseDNC	RF, SVM, KNN, LR	1934 6mA; 1934 non-6mA/—	https://bioinfo.uth.edu/6mA_Finder	10-fold CV	—/—	Yes	Yes	Rice and Mouse
	6mAPred-FO	NPS and PseDNC	SVM	880 6mA; 880 non-6mA/—	http://server.malab.cn/6mAPred-FO/	10-fold CV/—	87.44/—	Yes	Yes	Rice
	i6mA-Fuse	MBE, DBE, KNC, EIIP and Kmer	RF	FV: 4303 6mA; 4303 non-6mA/1067 6mA; 1067 non-6mA RC: 1430 6mA; 1430 non-6mA/3506mA; 350 non-6mA	http://kurata14.bio.kyutech.ac.jp/i6mA-Fuse/	10-fold CV and IND test	FV: 93.40 93.70 RC: 91.61 92.91	Yes	Yes	FV and RC
	Meta-i6mA	MBE, KNC, DNC, ENAC, EIIP, NCP, extreme GB	RF, ERT, NB, LR, SVM	RG: 29 237 6mA; 29 237 non-6mA/7300 6mAs/7300 non-6mAs Rice: 154000 6mA; 154 000 non-6mA sites AT: 31873 6mA; 31 873 non-6mA sites	http://kurata14.bio.kyutech.ac.jp/Meta-i6mA/	10-fold CV and IND test	94.40/89.31	Yes	Yes	RG
	i6mA-stack	ONF, BE, RFHC, EIIP, DPCP, TPCP	SVM, RF, LR, GNB	FV: 1966 6mA; 1966 non-6mA/347 6mA; 347 non-6mA RC: 813 6mA; 813 non-6mA/143 6mA; 143 non-6mA	http://nsclbio.jbnu.ac.kr/tools/i6mA-stack/	5-fold CV and IND test	FV: 93.81 91.50 RC: 90.80/91.1	Yes	Yes	FV and RC

The 1st and 2nd columns indicate the publication year and existing predictors. The 3rd column signifies the feature encoding schemes. The 4th column indicates the ML classifiers employed. The 5th column indicates training and independent dataset information. Fundamentally, it has two types of information: the first portion represents the size of training dataset that is used to develop the prediction model and the second portion represents the independent dataset size that is used for evaluating the prediction model. A study without information is represented as '—.' The sixth column denotes the web link information. The 7th column represents cross-validation and independent validation tests. The 8th represents the reported accuracy on the training and independent dataset. The 9th and 10th columns signify web server activity and batch prediction information, respectively. The last column represents the genome information.

LPF, position-specific dinucleotide frequency; CNN, convolution neural network; GNB, Gaussian Naive Bayes; TR, training; IND, independent; Ac, accuracy; HS, *Homo sapiens*; AT, *Arabidopsis thaliana*; FV, *Fragaria vesca*; RC, *Rosa chinensis*.

sequential SFS. Finally, all these models were assigned with different weights to make the final prediction. During cross-validation, SDM6A achieved Sn, Sp, ACC and MCC values of 87.5%, 88.3%, 87.9% and 0.758, respectively. They have also constructed an independent dataset and showed their model transferability with ACC and MCC of 88.2% and 0.765. The program for SDM6A is publicly available at http://thegleelab.org/SDM6A.

MM-6mAPred

Pian *et al.* [50] proposed MM-6mAPred by using the i6mA-Pred dataset and Markov model with the nucleotide chemical properties (NCP) encoding [48]. The performances on the training dataset presented Sn, Sp, ACC and MCC values of 89.3%, 90.1%, 89.72% and 0.789, but MM-6mAPred did not consider any independent dataset. The proposed program for MM-6mAPred is freely accessible at http://www.insect-genome.com/MM-6mAPred/.

SNNRice6mA

Yu and Dai [41] utilized the iDNA6mA-Rice dataset and developed SNNRice6mA, where they employed the MBE encoidng and DL architecture. This method achieved Sn, Sp, ACC and MCC values of 94.3%, 89.7%, 92.0% and 0.84, respectively, on the training data via 10-fold CV test. They did not consider any independent samples. The standalone program package of SNNRice6mA is available at https://github.com/yuht4/SNNRice6mA.

i6mA-DNCP

Kong and Zhang [75] utilized the i6mA-Pred dataset and developed i6mA-DNCP, where they considered the dinucleotide composition properties (DNCP) with the following properties, including F-twist, slide, energy and enthalpy. The authors have explored different ML algorithms and identified that the classification and regression trees (CART) was suitable for the prediction. i6mA-DNCP achieved the Sn, Sp, ACC and MCC values of 84.09%, 88.07%, 86.08% and 0.722, respectively. It considred the validation datasets of *A. thaliana*, *Fragaria vesca* (FV) and *Rosa chinensis* (RC), but i6mA-DNCP did not consider any independent data of the rice genome.

Predictors proposed in 2020

Seven predictors were reported for different species in 2020. Details in each predictor are described as below.

6mA-RicePred

Hung *et al.* [13] employed i6mA-Pred dataset and proposed a fusion-based prediction model 6mA-RicePred. They employed four encoding schemes (MBE, NCP, Kmer and Markov model) and SVM classifier. 6mA-RicePred achieved Sn, Sp, ACC and MCC values of 84.89%, 89.66%, 84.77% and 0.695, respectively. Unlike the existing methods, they evaluated their model using a large independent dataset (the iDNA6mA-Rice dataset) and achieved Sn, Sp, ACC and MCC values of 95.97%, 75.33%, 85.65% and 0.73, respectively, and has been shown to perform slightly better than any individual predictors on the independent test. 6mA-RicePred is accessible at https://github.com/huangqianfei0916/6mA-rice.

p6mA

Wang *et al.* [76] proposed p6mA by using sequence-based features. The p6mA predictor was trained on the combined dataset of the four species [*Oryza sativa* (rice), *C. elegans* (worm), *D. melanogaster* (fruit fly) and *Homo sapiens* (human)]. After deleting similar sequences, p6mA considered 3040 6mA and 3040 non-6mA samples. Three types of feature encoding approaches of EIIP, position-specific triple-nucleotide propensity and PseKNC were used. The MRMD method was applied to find the optimal feature set. Finally, the XGBoost-based model achieved Sn, Sp, ACC and MCC of 84.3%, 80.6%, 76.8% and 0.538, respectively. The proposed p6mA program is publicly available at https://github.com/Konglab404/p6mA.

6mA-Finder

Xu *et al.* [49] developed 6mA-Finder for rice genome. They employed the seven encoding schemes of ENAC, ANF, KNC, dinucleotide composition (DNC), EIIP, NCP and PseDNC with the three ML algorithms of RF, SVM and KNN. 6mA-Finder used 1934 6mA and 1934 non-6mA samples as a training dataset, but did not consider any independent data. 6mA-Finder outperformed the previous models in terms of the AUC value during 10-fold CV test. The proposed model is available at https://bioinfo.uth.edu/6mA_Finder.

i6mA-Fuse

Hasan *et al.* [46] developed the first predictor i6mA-Fuse based the Rosaceae (RG) genome datasets of two species (RC and FV). The datasets of i6mA-Fuse were constructed from the MDR database [77]. The sequences with 65% sequence identity with other samples were removed by CD-HIT. Six different feature encodings (Kmer, DPCP, EIIP, MBE, KNC and TPCP) and a RF classifier were considered and developed their respective model. Subsequently, the predicted probabilities of 6mAs were combined using a linear regression approach. The i6mA-Fuse (FV) achieved Sn, Sp, ACC and MCC of 90.8%, 95.7%, 0.93.4% and 0.873 for *F. vesca*, respectively. The corresponding metrics for i6mA-Fuse (RV) were 88.1%, 95.0%, 91.6% and 0.851. The web-application of i6mA-Fuse is publicly available at http://kurata14.bio.kyutech.ac.jp/i6mA-Fuse/.

6mAPred-FO

Cai *et al.* [52] developed 6mAPred-FO by using the feature fusion and optimization protocols. They considered the nucleotide positional specificity (NPS) and PseDNC encodings. Afterward, the features were enhanced by a filter method of analysis of variance to obtain the best feature. To train the predictive model, the resulting feature vectors are fed into the SVM classifier. 6mAPred-FO achieved Sn, Sp, ACC and MCC of 84.4%, 85.4%, 84.9% and 0.70, respectively, but this predictor did not consider any independent evaluation datasets. The web-application of 6mAPred-FO is publicly available at http://server.malab.cn/6mAPred-FO.

Meta-i6mA

Hasan *et al.* [45] developed a predictor for plant genome, termed Meta-i6mA, by exploiting informative features in an integrative machine-learning framework. They considered 10 types of encoding schemes of NAC, KNC, TNC, DNC, Kmer, MBE, DBE, EIIP, dinucleotide physicochemical properties and NCP. Subsequently, six commonly ML methods were used such as RF, SVM, ERT, Naïve Bayse (NB) and AdaBoost. To train the above classifiers, the RG genome datasets including 29,237 6mA and 29,237 non-6mA samples were employed. The final prediction model combined the 30 optimal baseline models. *A. thaliana* and rice datasets were employed for validation. The Meta-i6mA achieved Sn, Sp, ACC and MCC of 96.2%, 96.5%, 96.4% and 0.931 on the training data, respectively, and the resulting performances indicated 96.0%, 95.7%, 95.8% and 0.918 on the validation data, respectively. The web-application of Meta-i6mA is publicly accessible at http://kurata14.bio.kyutech.ac.jp/Meta-i6mA/.

i6mA-stack

Khanal *el al.* [47] proposed a predictor i6mA-stack by an ensemble-based approach. They considered five types of encoding schemes of BE, RFHC, EIIP, DPCP and TPCP. Subsequently, the four commonly ML methods of SVM, RF, logistic regression (LR) and gradient boosting (GB) were used. To train the classifier, *F. vesca* (1966 6mA and 1966 non-6mA training data; 347 6mA and 347 non-6mA independent data) and *R. chinensis* (813 6mAs and 813 non-6mA training data; 143 6mA and 143 non-6mA independent data) were used. The two-step feature selection approaches were applied via the RFE algorithm. i6mA-stack (FV) achieved ACC values of 93.8% and 91.5% on the training and independent datasets, respectively. The corresponding metrics for i6mA-stack (RC) were 90.8% and 91.1%. The proposed i6mA-stack is publicly available at http://nsclbio.jbnu.ac.kr/tools/i6mA-stack/.

In summary, 15 types of ML methods have been developed as active 6mA site prediction tools in multiple species. iDNA6mA-PseKNC [44] was developed in 2019, which is a pioneer ML-based method that relies on the PseKNC as the feature to build an SVM-based model. Three additional methods (i6mA-Pred [42], 6mA-RicePred [13] and 6mAPred-FO [52]) were based on SVM classifiers with multiple features encoding approaches; two methods (iDNA6mA [51] and SNNRice6mA [41]) were based on the DL framework; three methods (iDNA6mA-Rice, i6mA-DNCP [75] and i6mA-Fuse [46]) were based on the tree-based classifiers; MM-6mApred [50] was based on the Markov model; p6mA [76] was developed using GB classifier; and four methods (SD6MA [43], 6mA-Finder [49], Meta-i6mA [45] and i6mA-stack [47]) were based on the ensemble of ML-classifiers. A summary of the existing 6mA prediction tools is provided in Table 2. Particularly, most of the methods were trained on the different training datasets and few were validated based on independent datasets. Due to the recent surge in the advance of 6mA prediction tools, an unbiased assessment of these methods using a well-constructed validation dataset is essential.

Results and discussion

Publicly available 6mA predictors

Our objective was to conduct an unbiased performance evaluation of the existing tools based on our newly constructed validation datasets. Our validation datasets consisted of the three plant (*A. thaliana*, *Tolypocladium* and *D. lotus*), three eukaryotic (*S. cerevisiae, C. elegans* and *D. melanogaster*) and one prokaryotic (*E. coli*) species. Notably, positive and negative samples were different from the previously reported datasets (Table 1). Subsequently, these datasets were submitted to the publicly available 11 servers, namely Meta-i6mA, i6mA-Fuse, i6mA-stack, SDM6A, iDNA6mA-rice, 6mA-Finder, MM-6mAPred, i6mA-Pred, iDNA6mA-PseKNC, iDNA6mA and 6mAPred-FO, with the default parameters mentioned in a particular server. Of those, two servers i6mA-Fuse and i6mA-stack containing two prediction models and included both models for this evaluation. In total, 13 prediction models are evaluated in this study. Notably, six methods (iDNA6mA, SNNRice6mA, i6mA-DNCP, iDNA6mA-PseKNC and 6mA-RicePred and p6mA) were excluded from the current evaluation because their methods or servers are publicly inaccessible during our assessment.

Performance evaluation of existing species-specific 6mA prediction tools

A. thaliana

We used a validation set containing 60,700 6mAs and 121,400 non-6mAs with the ratio of 1:2 samples. Table 3 and Figure 2A shows that Meta-i6mA achieved the best performance with the MCC and ACC of 0.603 and 82.86%, respectively. Notably, the Meta-i6mA performance is ∼4% (in terms of MCC) higher than the performance of the second-best method MM-6mAPred. We observed six methods (i6mA-stack(FV), MM-6mAPred, SDM6A, i6mA-stack (RC), i6mA-Fuse (FV) and iDNA6mA) ranked from 2 to 7, achieved a similar performance with MCC values of 0.541–0.563. iDNA6mA-PseKNC achieved the worst performance for all the compared methods. Interestingly, none of the existing methods were trained with *A. thaliana* dataset. Still, most of the methods achieved a good performance, indicating that the 6mA pattern is probably similar between three species (RC, FV and rice) and *A. thaliana*. Prediction performance in terms of ranking depends on the author's computational approaches, employed feature encodings and the classifier choice.

Tolypocladium

The validation dataset containing 200 6mAs and 1000 non-6mAs with a ratio of 1:5 was used to evaluate the publicly available methods. Table 3 and Figure 2B show that Meta-i6mA achieved the best performance with MCC and ACC of 0.522 and 87.17%, respectively. Specifically, the corresponding MCC and ACC values are 3.5–42.9% and 0.12–26.42% higher than those of the other methods (excluding iDNA6mA-PseKNC), respectively. The five methods ranked as 2-6 (i6mA-stack (FV), i6mA-Fuse (FV), i6mA-stack (RC), i6mA-Fuse (RC) and iDNA6mA-rice) achieved similar performances with MCC of 0.487 to 0.417. Simultaneously, the remaining seven methods (SDM6A, i6mA-Pred, MM-6mAPred, 6mAPred-FO, iDNA6mA-PseKNC, iDNA6mA and 6mA-Finder) performances were significantly lower than the performances of the top six methods. Interestingly, the top five methods were trained with either RC or FV or a combination of RC and FV. They performed reasonably well with the *Tolypocladium* dataset, indicating that the RG genome may share a similar 6mA pattern with the *Tolypocladium* genome. However, most of the models trained with rice genome could not capture 6mA sites from the *Tolypocladium* genome.

D. lotus

Our constructed *D. lotus* validation set contained 310 6mAs and 1550 non-6mAs (a 1:5 ratio of positive to negative samples). Table 3 and Figure 2C show the performances of different predictors. i6mA-Pred achieved the best performance with MCC and ACC of 0.174 and 75.18%. Unfortunately, these metrics are far from satisfactory. If we apply the balanced accuracy metrics [(Sn + SP)/2)] all the predictor performance is closer to the random prediction performance. The existing methods' failure may

Table 3. Performance of existing predictors on validation set

Species	Methods	Sp (%)	Sn (%)	ACC (%)	MCC
A. thaliana	Meta-i6mA	91.20	66.22	82.86	0.603
	i6mA-stack (FV)	93.71	57.81	81.73	0.563
	MM-6mAPred	77.62	81.21	78.82	0.560
	SD6MA	81.91	75.61	78.37	0.552
	iDNA6mA	89.95	53.80	77.90	0.546
	i6mA-stack (RC)	91.21	57.30	79.91	0.542
	i6mA-Fuse (FV)	88.21	56.21	77.53	0.541
	i6mA-Fuse (RC)	87.41	55.31	76.71	0.508
	i6mA-Pred	78.22	71.50	75.98	0.488
	iDNA6mA-rice	77.12	68.20	74.15	0.463
	6mA-Finder	74.45	69.95	71.44	0.421
	6mAPred-FO	87.14	40.92	71.73	0.366
	iDNA6mA-PseKNC	3.50	77.01	40.25	-0.253
Tolypocladium	Meta-i6mA	93.30	56.51	87.17	0.522
	i6mA-stack (FV)	94.62	45.63	86.45	0.487
	i6mA-Fuse (FV)	87.11	48.41	80.65	0.462
	i6mA-stack (RC)	85.22	57.11	80.53	0.444
	i6mA-Fuse (RC)	82.72	54.11	77.95	0.432
	iDNA6mA-rice	81.80	67.50	76.41	0.417
	iDNA6mA	92.32	27.5	81.51	0.236
	SD6MA	73.52	53.5	70.18	0.225
	i6mA-Pred	76.80	46.50	71.75	0.196
	MM-6mAPred	61.6	56.50	60.75	0.138
	6mAPred-FO	85.20	25.00	75.16	0.103
	6mA-Finder	63.20	54.20	61.74	0.093
	iDNA6mA-PseKNC	3.05	69.5	14.02	-0.369
D. lotus	i6mA-Pred	84.51	28.54	75.18	0.174
	MM-6mAPred	77.64	32.56	70.12	0.152
	SD6MA	76.64	32.56	69.12	0.111
	6mA-Finder	80.36	28.77	71.76	0.107
	Meta-i6mA	90.87	10.18	77.53	0.034
	iDNA6mA-PseKNC	0.033	96.60	0.188	0.00
	i6mA-Fuse (FV)	90.03	6.14	76.04	-0.011
	i6mA-stack (FV)	91.33	6.84	77.21	-0.020
	iDNA6mA	90.04	6.88	75.04	-0.020
	i6mA-Fuse (RC)	87.60	5.58	73.96	-0.021
	i6mA-stack (RC)	81.60	6.55	69.09	-0.029
	iDNA6mA-rice	71.11	21.91	62.91	-0.056
	6mAPred-FO	81.05	5.47	68.45	-0.108
S. cerevisiae	Meta-i6mA	94.44	39.33	80.67	0.459
	i6mA-stack (FV)	93.66	35.50	79.12	0.446
	i6mA-Fuse (FV)	88.65	31.53	74.37	0.402
	i6mA-stack (RC)	89.02	36.20	75.82	0.392
	i6mA-Fuse (RC)	83.65	34.53	71.37	0.371
	SD6MA	76.42	57.66	72.23	0.329
	MM-6mAPred	76.09	54.40	71.16	0.302
	i6mA-Pred	79.11	50.66	72.00	0.287
	6mA-Finder (rice)	73.28	57.86	69.44	0.284
	iDNA6mA-rice	82.76	44.67	73.23	0.278
	iDNA6mA	91.12	20.5	73.46	0.259
	6mAPred-FO	72.44	29.46	61.70	0.019
	iDNA6mA-PseKNC	3.12	70.8	20.00	-0.381
C. elegans	Meta-i6mA	94.21	25.88	71.43	0.266
	SD6MA	75.53	46.01	65.69	0.242
	i6mA-Fuse (FV)	90.60	22.11	67.77	0.218
	i6mA-stack (FV)	93.68	17.44	68.27	0.202
	iDNA6mA-rice	85.33	32.18	67.61	0.200
	i6mA-stack (RC)	90.82	18.01	66.55	0.181
	i6mA-Fuse (RC)	89.82	17.80	65.81	0.177
	i6mA-Pred	79.17	33.51	63.95	0.174

Continued

Table 3. Continued

Species	Methods	Sp (%)	Sn (%)	ACC (%)	MCC
	MM-6mAPred	65.62	48.22	59.82	0.151
	iDNA6mA	90.13	13.31	64.52	0.142
	6mA-Finder	65.42	43.44	58.09	0.103
	6mAPred-FO	71.83	25.51	56.39	-0.012
	iDNA6mA-PseKNC	3.36	76.01	47.75	-0.296
D. melanogaster	Meta-i6mA	86.77	69.23	80.92	0.561
	i6mA-stack (FV)	86.60	59.13	77.44	0.499
	SD6MA	77.63	76.11	77.13	0.488
	MM-6mAPred	82.73	65.26	76.90	0.476
	i6mA-Pred	76.44	75.01	75.96	0.463
	6mA-Finder	77.30	70.41	75.00	0.452
	iDNA6mA-rice	83.13	58.09	74.77	0.431
	i6mA-Fuse (FV)	84.01	55.11	74.37	0.426
	iDNA6mA	90.0 3	39.15	73.03	0.417
	i6mA-stack (RC)	83.70	57.01	74.80	0.401
	i6mA-Fuse (RC)	82.81	53.11	72.91	0.398
	6mAPred-FO	74.60	18.71	55.97	0.071
	iDNA6mA-PseKNC	1.55	67.70	22.97	-0.339
E. coli	Meta-i6mA	90.40	16.29	65.69	0.071
	iDNA6mA	91.26	8.03	63.52	0.061
	i6mA-Fuse (FV)	88.30	14.91	63.84	0.052
	MM-6mAPred	77.16	22.14	58.83	0.047
	SD6MA	81.25	20.83	58.39	0.041
	i6mA-stack (FV)	87.30	15.10	63.57	0.031
	i6mA-Fuse (RC)	85.81	14.01	61.87	0.011
	i6mA-stack (RC)	81.30	13.91	58.83	0.010
	i6mA-Pred	82.31	17.66	60.76	0.001
	6mA-Finder	80.76	18.23	59.91	0.001
	iDNA6mA-rice	72.70	21.81	55.74	-0.020
	6mAPred-FO	71.07	6.55	49.56	-0.214
	iDNA6mA-PseKNC	3.12	72.8	26.34	-0.366

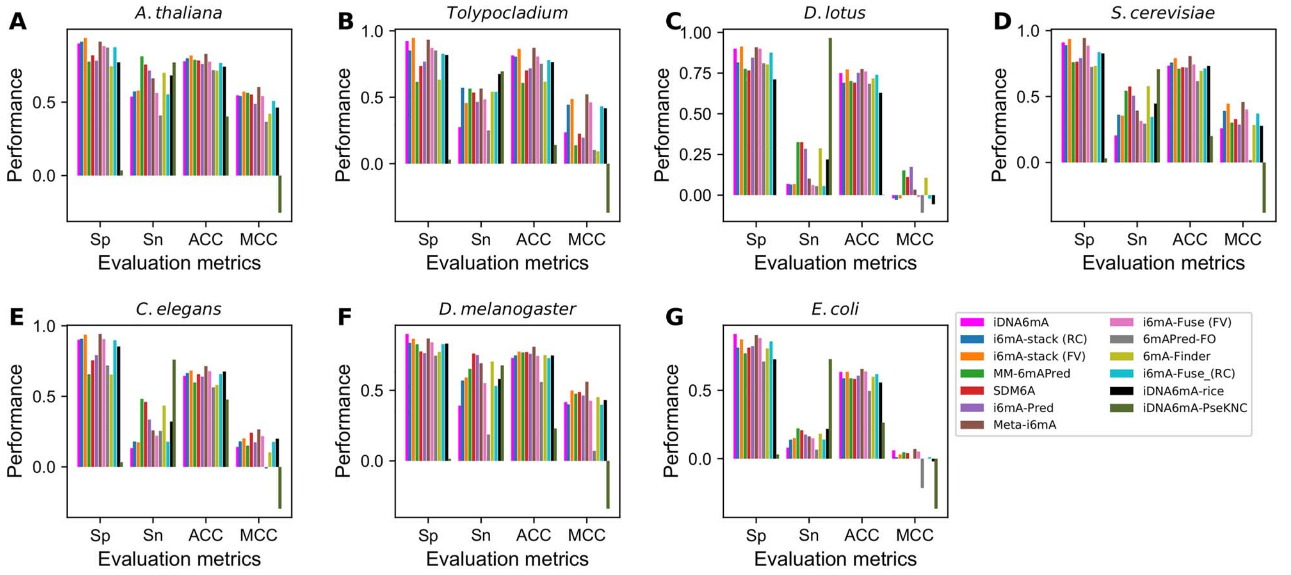

Figure 2. Comparison of the prediction performance of 13 models on seven species-specific validation datasets. (**A**) *A. thaliana*, (**B**) *Tolypocladium*, (**C**) *D. lotus*, (**D**) *S. cerevisiae*, (**E**) *C. elegans*; (**F**) *D. melanogaster* and (**G**) *E. coli*.

be that 6mA site patterns are entirely different between *D. loctus* and the other species. Hence, the practical applicability of the current methods for predicting 6mA sites in *D. loctus* is limited. Therefore, it is essential to develop a species-wise prediction model for *D. lotus*.

S. cerevisiae

S. cerevisiae's validation dataset consisted of 750 6mAs and 2250 non-6mAs (a 1:3 ratio) (Table 1). Table 3 and Figure 2D show comprehensive performance information. Meta-i6mA and i6mA-stack (FV) are the top two methods that achieved a similar performance with MCC and ACC in the ranges of 0.446-0.459 and 79.12-80.67%, respectively. Specifically, the MCC improvement of Meta-i6mA is enormous, which is 1.3-44.1% higher than the MCC of the other methods. Similar to *Tolypocladium* observation, the prediction model developed with the RG training dataset achieved a good performance. Interestingly, the top five methods performed exceptionally well in identifying non-6mAs, which resulted in a high Sp. However, they could not replicate similar performance, while identifying 6mAs, which resulted in a low Sn. A probable reason is that the upstream and downstream sequence around 6mA sites is different among *S. cerevisiae* and other species. Hence these methods predicted most of 6mAs as non-6mAs (high false-negative rate).

C. elegans

We compared the 13 prediction models on *C. elegans* validation dataset that contain a 1: 2 ratio of positive to negative (23,100 6mA to 46,200 non-6mAs) samples. Table 3 and Figure 2E show that Meta-i6mA and SDM6A achieved a similar performance in MCC in the range of 0.262–0.242. The remaining methods (except 6mAPred-FO and iDNA6mA-PseKNC) achieved similar performances and significantly lower than the top two methods. The existing methods performed excellently the prediction of non-6mAs, which resulted in high Sp in the range of 65.42–94.21%. However, there is a problem in accurately identifying 6mAs, whose Sp ranges from 17.44 to 48.22%, indicating nucleotides around 6mA sites may be entirely different among three species (rice, RC and FV) and *C. elegans*. The above observation shows that none of the existing methods is suitable for predicting 6mA sites in *C. elegans* species.

D. melanogaster

A validation dataset containing 26,700 6mA and 53,400 non-6mAs was considered to evaluate 13 models. As shown in Table 3 and Figure 2F, Meta-i6mA archived best performances in MCC and ACC values of 0.561 and 80.92%, respectively. Specifically, the corresponding metrics of Meta-i6mA is 3.48–24.95 and 6.2–49.0% higher than those of the other methods. Except 6mAPred-FO and iDNA6mA-PseKNC, the remaining methods performed reasonably well with MCC and ACC ranges of 0.398–0.499 and 72.91–77.44%, respectively. Our evaluation shows that a model developed using a specific training dataset still can predict accurately the 6mAs in other species. The method ranking is primarily dependent on the training dataset's size, computational approaches, and integration of various feature encoding schemes.

E. coli

A validation dataset containing 33,500 6mAs and 67,000 non-6mAs (see Table 1) was used to evaluate the existing predictors. As revealed in Figure 2G and Table 3, all of the predictors achieved a lower performance for the four metrics: Sp, Sn, ACC and MCC. Table 3 shows that all MCC values are closer to zero, indicating none of the methods is suitable for genome-wide prediction of 6mA sites from *E. coli*. Understandably, genome distribution, including epigenetic modification sites, is entirely different between plant genome and prokaryotes. As a result, none of the methods developed with plant species genome indicated high performance when applying to the *E. coli*. Therefore, it is urgent to create a species-wise prediction model for *E. coli*.

Comparison of 6mA site prediction web servers

We next evaluated whether the servers are user friendly. Web application servers are quite important for experimental scientists. We noted that there were several limitations to the existing web servers, as follows. First, the existing web servers handle only the sequences with a length of 41 bp with adenine at the center; hence, it may limit practical application to genome-wide investigation. Second, the number of FASTA sequences that could be handled in a single request varies among the prediction models. In particular, SDM6A, DNA-Finder and 6mAPred-FO handled up to 10,000 sequences; i6mA-Fuse and Meta-6mA handled up to 30,000 sequences; MM-6mAPred, iDNA-6mA-Rice and i6mA-Pred handled up to 5000 samples. Third, for a batch processing, a half of the servers did not support any FASTA sequence files (Table 2). SDM6A, 6mA-Finder, i6mA-Fuse and Meta-i6mA have the option for users to upload their FASTA sequence files. Fourth, the different models showed varying run times from 3 to 20 min. Meta-i6mA could handle large numbers of FASTA sequences in a single run and returned the prediction results quickly (within 3 min). Generally, the predicted probability score of a given sample is important for experimentalists to make a decision. In this regard, the three methods (iDNA6mA-Rice, i6mA-stack and i6mAPred) did not provide such information while returning their prediction. Finally, researchers without programming knowledge could not use most of the existing predictors because the whole genome had to be handled into fragments (with a length of 41 bp) before submission.

We may conclude from the above discussion; numerous efforts have been dedicated to the computational prediction of 6mAs by exploiting various feature encoding algorithms, classifiers and different approaches. Nonetheless, it remains unclear which features and ML algorithms are the most instructive for different species. Thus, systematic analysis of features contribution and different classifiers' analytical ability upon distinct feature(s) are much desirable. Toward a more accurate prediction of DNA 6mAs, such a study will provide a practical guide.

Limitations of current methods and future improvements

Recently, the six prediction tools used the same rice genome training dataset, including SDM6A, i6mA-DNCP, i6mA-Pred, iDNA6mA, MM-6mAPred and 6mAPred-FO. Among these methods, only two predictors of SDM6A and i6mA-DNCP used independent evaluation datasets. The remaining four methods (p6mA, i6mA-stuck, i6mA-Fuse and Meta-i6mA) used entirely different training datasets for their model development. Usually, to develop an ML-based prediction model, the construction of a high-quality dataset is the first essential step. Unexpectedly, none of the successive approaches checked the quality of earlier method training datasets, and there was no effort to increase the training dataset quality. Based on our evaluation, we noted that the proposed validation datasets for the three species (*E. coli*, *C. elegans* and *D. lotus*), whose 6mA site patterns seem to be entirely different compared to rice genome or RG. As a result, the existing methods did not work well in predicting the majority of 6mAs as non-6mAs. In this regard, either a species-specific model or developing a single model using a combined multiple-species dataset is highly recommended. Importantly, our validation dataset can be integrated into the existing dataset for future prediction model development. To improve the representation of the true 'negative', to improve the model robustness and to avoid overfitting are still crucial challenges. Generally, adding more distal nucleotides improves the prediction performance. However, such improvement was not observed in 6mA evaluation of existing models. Recently, Meta-i6mA has investigated different window sizes (31, 41, 51, 61, 71 and 81) and found that 41 bp optimized the performance. In the future study, researchers may investigate more distal nucleotides to improve the prediction performances of 6mAs.

To further improve the prediction performance, we have the following suggestions. First, decreasing bias in the training dataset, excluding highly homologous sequences, and representing 'true negative' are needed. Such a dataset will be helpful for developing more reliably trained models. Secondly, exploring multiple ML classifiers and different feature encoding algorithms is highly recommended. Thirdly, systematic evaluation of different classical computational approaches (adaptive feature learning, iterative representation feature, meta-classifier representation, stacking-framework and fusing with multi-view evidence) on the same training dataset is recommended more than a single feature encoding-based prediction model. Integration of multiple feature encodings and ML algorithms evolves model accuracy and robustness [55, 78–83]. Finally, bioinformaticians should develop a webserver while considering the difficulties that experimentalists face when using computational methods. Considering a large size of benchmark datasets, researchers may apply the DL frameworks with different feature representation schemes to further improve the prediction performances and to ultimately compare the performances of the DL and conventional ML-based models.

Conclusion

Accurate genome-wide identification of 6mA sites is essential due to the critical roles of 6mA in many biological developments for essentially revealing its regulatory mechanism and providing important clues for drug development [84–86]. The reliable and effective computational approaches can help biologists making an experimental plan. In this work, to assess currently available 6mA site prediction algorithms, we used seven species-specific datasets (*A. thaliana*, *Tolypocladium*, *D. lotus*, *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *E. coli*) from the recent large-scale genome sequencings. We systematically compared the prediction performances by using our validation datasets. To help users choosing best tools, the advantages and disadvantages of the existing webservers and stand alone software were discussed from different aspects.

The major remark from our investigation is that no universal best web tools are available for predicting 6mA sites for all the seven genomes. In particular, none of the existing predictors was suitable for predicting 6mA on *D. lotus, C. elegans* and *E. coli* datasets. Even though the performance of existing predictors on *A. thaliana, Tolypocladium, D. melanogaster* and *S. cerevisiae* was satisfactory, there is further room to advance the prediction performance. Overall, we hope that this study supports researchers with interest in this field to develop new prediction tools for the 6mA sites.

Key Points

- We conducted a comprehensive review and assessed 11 publicly available 6mA site prediction tools using a newly constructed validation dataset that captured the overall pattern of 6mAs and non-6mAs from the entire genome for each species.
- Our results demonstrated that Meta-i6mA achieved the best performance for four species (*Arabidopsis thaliana, Tolypocladium*, *S. cerevisiae* and *D. melanogaster*) when compared with their counterparts. However, none of the existing methods was suitable for *E. coli, C. elegans* and *D. lotus*, limiting these methods' practical applicability.
- Our analysis could be helpful to wet-lab researchers to select the appropriate tools for identifying putative 6mAs. It also gave directions to the computational biologists for the development of next-generation species-specific 6mA site prediction tools.

Funding

This work was supported by the Grant-in-Aid for Scientific Research (B) [19H04208] and partially supported by the Grant-in-Aid for JSPS Research Fellow [19F19377] from JSPS and the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science and ICT [2018R1D1A1B07049572].

Conflict of Interest

The authors have declared no competing interests.

References

1. Molinie B, Giallourakis CC. Genome-wide location analyses of N6-Methyladenosine modifications (m(6)A-Seq). *Methods Mol Biol* 2017;**1562**:45–53.
2. Nye TM, van Gijtenbeek LA, Stevens AG, *et al.* Methyltransferase DnmA is responsible for genome-wide N6-methyladenosine modifications at non-palindromic recognition sites in Bacillus subtilis. *Nucleic Acids Res* 2020;**48**: 5332–5348.
3. Lv H, Dao FY, Zhang D, *et al.* iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* 2020;**23**:100991.
4. Barros-Silva D, Lobo J, Guimaraes-Teixeira C, *et al.* VIRMA-dependent N6-Methyladenosine modifications regulate the expression of Long non-coding RNAs CCAT1 and CCAT2 in prostate cancer. *Cancers (Basel)* 2020;**12**(4):771. doi: 10.3390/cancers12040771.
5. Chen J, Fang X, Zhong P, *et al.* N6-methyladenosine modifications: interactions with novel RNA-binding proteins and roles in signal transduction. *RNA Biol* 2019;**16**:991–1000.
6. Matsuzawa S, Wakata Y, Ebi F, *et al.* Development and validation of monoclonal antibodies against N6-methyladenosine for the detection of RNA modifications. *PLoS One* 2019;**14**:e0223197.
7. Vaidyanathan PP, AlSadhan I, Merriman DK, *et al.* Pseudouridine and N(6)-methyladenosine modifications weaken PUF protein/RNA interactions. *RNA* 2017;**23**:611–618.
8. Wang X, Xie H, Ying Y, *et al.* Roles of N(6) -methyladenosine (m(6) A) RNA modifications in urological cancers. *J Cell Mol Med* 2020;**24**(18):10302–10310. doi: 10.1111/jcmm.15750.
9. Wang Y, Li Y, Yue M, *et al.* Publisher correction: N(6)-methyladenosine RNA modification regulates embryonic

neural stem cell self-renewal through histone modifications. *Nat Neurosci* 2018;**21**:1139.

10. Du K, Zhang S, Chen W, *et al*. Epigenetic DNA modification N(6)-methyladenine inhibits DNA replication by Sulfolobus solfataricus Y-family DNA polymerase Dpo4. *Arch Biochem Biophys* 2019;**675**:108120.
11. Ratel D, Ravanat JL, Berger F, *et al*. N6-methyladenine: the other methylated base of DNA. *Bioessays* 2006;**28**:309–315.
12. Xiong J, Ye TT, Ma CJ, *et al*. N 6-Hydroxymethyladenine: a hydroxylation derivative of N6-methyladenine in genomic DNA of mammals. *Nucleic Acids Res* 2019;**47**:1268–1277.
13. Luo GZ, He C. DNA N(6)-methyladenine in metazoans: functional epigenetic mark or bystander? *Nat Struct Mol Biol* 2017;**24**:503–506.
14. Hong T, Yuan Y, Wang T, *et al*. Selective detection of N6-methyladenine in DNA via metal ion-mediated replication and rolling circle amplification. *Chem Sci* 2017;**8**: 200–205.
15. Kweon SM, Chen Y, Moon E, *et al*. An adversarial DNA N(6)-Methyladenine-sensor network preserves polycomb silencing. *Mol Cell* 2019;**74**:1138–1147 e1136.
16. Du K, Zhang X, Zou Z, *et al*. Epigenetically modified N(6)-methyladenine inhibits DNA replication by human DNA polymerase eta. *DNA Repair (Amst)* 2019;**78**:81–90.
17. Zhang Q, Liang Z, Cui X, *et al*. N(6)-Methyladenine DNA methylation in japonica and indica rice genomes and its association with gene expression, plant development, and stress responses. *Mol Plant* 2018;**11**:1492–1508.
18. Xie Q, Wu TP, Gimple RC, *et al*. N(6)-methyladenine DNA modification in Glioblastoma. *Cell* 2018;**175**:1228–1243 e1220.
19. Liu X, Lai W, Zhang N, *et al*. Predominance of N(6)-Methyladenine-specific DNA fragments enriched by multiple Immunoprecipitation. *Anal Chem* 2018;**90**:5546–5551.
20. O'Brown ZK, Greer EL. N6-Methyladenine: a conserved and dynamic DNA mark. *Adv Exp Med Biol* 2016;**945**:213–246.
21. Singer B, Antoccia A, Basu AK, *et al*. Both purified human 1,N6-ethenoadenine-binding protein and purified human 3-methyladenine-DNA glycosylase act on 1,N6-ethenoadenine and 3-methyladenine. *Proc Natl Acad Sci U S A* 1992;**89**:9386–9390.
22. Wang Y, Li Y, Yue M, *et al*. N(6)-methyladenosine RNA modification regulates embryonic neural stem cell self-renewal through histone modifications. *Nat Neurosci* 2018; **21**:195–206.
23. Yao M, Dong Y, Wang Y, *et al*. N(6)-methyladenosine modifications enhance enterovirus 71 ORF translation through METTL3 cytoplasmic distribution. *Biochem Biophys Res Commun* 2020;**527**:297–304.
24. McIntyre ABR, Alexander N, Grigorev K, *et al*. Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nat Commun* 2019;**10**:579.
25. Zhang G, Huang H, Liu D, *et al*. N6-methyladenine DNA modification in drosophila. *Cell* 2015;**161**:893–906.
26. Fu Y, Luo GZ, Chen K, *et al*. N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* 2015;**161**:879–892.
27. Mondo SJ, Dannebaum RO, Kuo RC, *et al*. Widespread adenine N6-methylation of active genes in fungi. *Nat Genet* 2017;**49**:964–968.
28. Wu TP, Wang T, Seetin MG, *et al*. DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* 2016;**532**:329–333.
29. Liu J, Zhu Y, Luo GZ, *et al*. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nat Commun* 2016;**7**:13052.
30. Liang Z, Shen L, Cui X, *et al*. DNA N(6)-adenine methylation in *Arabidopsis thaliana*. *Dev Cell* 2018;**45**:406–416 e403.
31. Luo GZ, Blanco MA, Greer EL, *et al*. DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? *Nat Rev Mol Cell Biol* 2015;**16**:705–710.
32. Ye P, Luan Y, Chen K, *et al*. MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res* 2017;**45**:D85–D89.
33. Zhu X, He J, Zhao S, *et al*. A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of *Saccharomyces cerevisiae*. *Brief Funct Genomics* 2019;**18**:367–376.
34. Clancy MJ, Shambaugh ME, Timpte CS, *et al*. Induction of sporulation in *Saccharomyces cerevisiae* leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic Acids Res* 2002;**30**:4509–4518.
35. Greer EL, Blanco MA, Gu L, *et al*. DNA methylation on N6-adenine in C. elegans. *Cell* 2015;**161**:868–878.
36. O'Brown ZK, Boulias K, Wang J, *et al*. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics* 2019;**20**:445.
37. Chou KC. Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr Med Chem* 2019;**26**:4918. https://doi.org/10.2174/0929867326666190507082559.
38. Liu B, Fang L, Long R, *et al*. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* 2016;**32**: 362–369.
39. Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol Biosyst* 2015;**11**:2620–2634.
40. Sun S, Wang C, Ding H, *et al*. Machine learning and its applications in plant molecular studies. *Brief Funct Genomics* 2020;**19**:40–48.
41. Yu H, Dai Z. SNNRice6mA: a deep learning method for predicting DNA N6-Methyladenine sites in Rice genome. *Front Genet* 2019;**10**:1071.
42. Chen W, Lv H, Nie F, *et al*. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 2019;**35**:2796–2800.
43. Basith S, Manavalan B, Shin TH, *et al*. SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the Rice genome. *Mol Ther Nucleic Acids* 2019;**18**: 131–141.
44. Feng P, Yang H, Ding H, *et al*. iDNA6mA-PseKNC: identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* 2019;**111**:96–102.
45. Hasan MM, Basith S, Khatun MS, *et al*. Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform* 2020, bbaa202. doi: 10.1093/bib/bbaa202.
46. Hasan MM, Manavalan B, Shoombuatong W, *et al*. i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. *Plant Mol Biol* 2020;**103**:225–234.

47. Khanal J, Lim DY, Tayara H, *et al.* i6mA-stack: a stacking ensemble-based computational prediction of DNA N6-methyladenine (6mA) sites in the Rosaceae genome. *Genomics* 2020, S0888–7543(20)31362–8. doi: 10.1016/j.ygeno.2020.09.054.
48. Lv H, Dao FY, Guan ZX, *et al.* iDNA6mA-Rice: a computational tool for detecting N6-Methyladenine sites in Rice. *Front Genet* 2019;**10**:793.
49. Xu H, Hu R, Jia P, *et al.* 6mA-Finder: a novel online tool for predicting DNA N6-methyladenine sites in genomes. *Bioinformatics* 2020;**36**:3257–3259.
50. Pian C, Zhang G, Li F, *et al.* MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics* 2020;**36**:388–392.
51. Tahir M, Tayara H, Chong K. iDNA6mA (5-step rule): identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. *Chemom Intel Lab Syst* 2019;**189**:96–101.
52. Cai J, Wang D, Chen R, *et al.* A bioinformatics tool for the prediction of DNA N6-Methyladenine modifications based on feature fusion and optimization protocol. *Front Bioeng Biotechnol* 2020;**8**:502.
53. Fu L, Niu B, Zhu Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–3152.
54. Huang Q, Zhang J, Wei L, *et al.* 6mA-RicePred: a method for identifying DNA N (6)-Methyladenine sites in the Rice genome based on feature fusion. *Front Plant Sci* 2020;**11**:4.
55. Manavalan B, Hasan MM, Basith S, *et al.* Empirical comparison and analysis of web-based DNA N4-methylcytosine site prediction tools. *Molecular Therapy-Nucleic Acids* 2020;**22**:406–420. doi: 10.1016/j.omtn.2020.09.010
56. Zhang ZY, Yang YH, Ding H, *et al.* Design powerful predictor for mRNA subcellular location prediction in *Homo sapiens*. *Brief Bioinform* 2020, bbz177. doi: 10.1093/bib/bbz177.
57. Yang H, Yang W, Dao FY, *et al.* A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*. *Brief Bioinform* 2019;**21**(5):1568–1580. doi: 10.1093/bib/bbz123.
58. Feng CQ, Zhang ZY, Zhu XJ, *et al.* iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 2019;**35**:1469–1477.
59. Dao FY, Lv H, Wang F, *et al.* Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics* 2019;**35**:2075–2083.
60. Lv H, Dao F-Y, Guan Z-X, *et al.* Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. *Brief Bioinform* 2020, bbaa255. doi: 10.1093/bib/bbaa255.
61. Dao FY, Lv H, Zulfiqar H, *et al.* A computational platform to identify origins of replication sites in eukaryotes. *Brief Bioinform* 2020, bbaa017. doi: 10.1093/bib/bbaa017.
62. Dao FY, Lv H, Yang YH, *et al.* Computational identification of N6-methyladenosine sites in multiple tissues of mammals. *Comput Struct Biotechnol J* 2020;**18**:1084–1091.
63. Khatun MS, Hasan MM, Kurata H. PreAIP: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. *Front Genet* 2019;**10**: 129.
64. Khatun MS, Hasan MM, Shoombuatong W, *et al.* ProIn-Fuse: improved and robust prediction of proinflammatory peptides by fusing of multiple feature representations. *J Comput Aided Mol Des* 2020;**34**(12):1229–1236. doi: 10.1007/s10822-020-00343-9.
65. Manavalan B, Shin TH, Lee G. PVP-SVM: sequence-based prediction of phage Virion proteins using a support vector machine. *Front Microbiol* 2018;**9**:476.
66. Manavalan B, Shin TH, Lee G. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 2018;**9**:1944–1956.
67. Manavalan B, Subramaniyam S, Shin TH, *et al.* Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J Proteome Res* 2018;**17**:2715–2726.
68. Shoombuatong W, Schaduangrat N, Nantasenamat C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. *EXCLI J* 2018;**17**:734–752.
69. Shoombuatong W, Schaduangrat N, Pratiwi R, *et al.* THPep: a machine learning-based approach for predicting tumor homing peptides. *Comput Biol Chem* 2019;**80**:441–451.
70. Basith S, Manavalan B, Hwan Shin T, *et al.* Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med Res Rev* 2020;**40**:1276–1314.
71. Basith S, Manavalan B, Shin TH, *et al.* iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput Struct Biotechnol J* 2018;**16**:412–420.
72. Charoenkwan P, Kanthawong S, Nantasenamat C, *et al.* iDPPIV-SCM: a sequence-based predictor for identifying and analyzing dipeptidyl peptidase IV (DPP-IV) inhibitory peptides using a scoring card method. *J Proteome Res* 2020;**19**:4125–4136.
73. Charoenkwan P, Yana J, Nantasenamat C, *et al.* iUmami-SCM: a novel sequence-based predictor for prediction and analysis of umami peptides using a scoring card method with propensity scores of dipeptides. *J Chem Inf Model* 2020. doi: 10.1021/acs.jcim.0c00707.
74. Zhang D, Xu ZC, Su W, *et al.* iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. *Bioinformatics* 2020, btaa702. doi: 10.1093/bioinformatics/btaa702.
75. Kong L, Zhang L. i6mA-DNCP: computational identification of DNA N(6)-Methyladenine sites in the Rice genome using optimized dinucleotide-based features. *Genes (Basel)* 2019;**10**(10):828. doi: 10.3390/genes10100828.
76. Wang HT, Xiao FH, Li GH, *et al.* Identification of DNA N(6)-methyladenine sites by integration of sequence features. *Epigenetics Chromatin* 2020;**13**:8.
77. Liu ZY, Xing JF, Chen W, *et al.* MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. *Hortic Res* 2019;**6**:78.
78. Hasan MM, Manavalan B, Khatun MS, *et al.* i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. *Int J Biol Macromol* 2020;**157**:752–758.
79. Hasan MM, Manavalan B, Shoombuatong W, *et al.* i4mC-mouse: improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. *Comput Struct Biotechnol J* 2020;**18**:906–912.
80. Hasan MM, Schaduangrat N, Basith S, *et al.* HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* 2020;**36**:3350–3356.
81. Hasan MM, Khatun MS, Kurata H. iLBE for computational identification of linear B-cell epitopes by integrating sequence and evolutionary features. *Genomics Proteomics*

Bioinformatics 2020, S1672–0229(18)30274–2. doi: 10.1016/j.gpb.2019.04.004.

82. Khatun S, Hasan M, Kurata H. Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. *FEBS Lett* 2019;**593**: 3029–3039.
83. Hasan MM, Guo D, Kurata H. Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. *Mol Biosyst* 2017;**13**: 2545–2550.
84. Abakir A, Giles TC, Cristini A, *et al.* N(6)-methyladenosine regulates the stability of RNA:DNA hybrids in human cells. *Nat Genet* 2020;**52**:48–55.
85. Liang Z, Kidwell RL, Deng H, *et al*. Epigenetic N6-methyladenosine modification of RNA and DNA regulates cancer. *Cancer Biol Med* 2020;**17**:9–19.
86. Luan MW, Chen W, Xing JF, *et al*. DNA N6-Methyladenosine modification role in transmitted variations from genomic DNA to RNA in Herrania umbratica. *BMC Genomics* 2019; **20**:508.