Meta-i6mA: An interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework

Article *in* Briefings in Bioinformatics · September 2020

DOI: 10.1093/bib/bbaa202

Briefings in Bioinformatics, 00(00), 2020, 1–16

doi: 10.1093/bib/bbaa202
Problem solving protocol

Meta-i6mA: an interspecies predictor for identifying DNA N^6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework

Md. Mehedi Hasan, Shaherin Basith, Mst. Shamima Khatun, Gwang Lee, Balachandran Manavalan and Hiroyuki Kurata

Corresponding authors: Balachandran Manavalan, Department of Physiology, Ajou University School of Medicine, Suwon 443-380, Korea. E-mail: bala@ajou.ac.kr; Hiroyuki Kurata, Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan. E-mail: kurata@bio.kyutech.ac.jp

Abstract

DNA N^6-methyladenine (6mA) represents important epigenetic modifications, which are responsible for various cellular processes. The accurate identification of 6mA sites is one of the challenging tasks in genome analysis, which leads to an understanding of their biological functions. To date, several species-specific machine learning (ML)-based models have been proposed, but majority of them did not test their model to other species. Hence, their practical application to other plant species is quite limited. In this study, we explored 10 different feature encoding schemes, with the goal of capturing key characteristics around 6mA sites. We selected five feature encoding schemes based on physicochemical and

Md. Mehedi Hasan received his PhD degree in bioinformatics from China Agricultural University, Beijing, in 2016. He is currently a Japan Society for the Promotion of Science international PD 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.

Shaherin Basith is a research assistant professor in the Department of Physiology, Ajou University School of Medicine, Republic of Korea. She obtained her PhD in Medical Sciences from Ajou University, Republic of Korea, in 2013. Before her postdoctoral research, she worked as a research trainee at Korea Institute for Advanced Study, Republic of Korea. She then joined Ewha Womans University, Republic of Korea, as a postdoctoral research fellow. Her main area of research focuses on exploring the structure–function relationships of proteins using state-of-the-art molecular modeling tools, phylogenetic analysis and biomolecular simulations. She is also actively involved in the application of machine learning tools for peptide and small molecule drug discovery.

Mst. Shamima Khatun is a PhD student in the Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Japan. She is currently working in the fields of peptide-based drug design, bioinformatics and computational biology.

Gwang Lee is a professor in the Department of Physiology, Ajou University School of Medicine, Republic of Korea. He received his PhD in 1998 from Tokyo University, Japan. He then joined National Institutes of Health/National Institute of Neurological Disorders and Stroke (MD, USA) as a postdoctoral fellow, where he focused on neurodegenerative diseases. His main research interests include the integration of triple omics (transcriptomics, metabolomics and proteomics) for nanotoxicity studies, machine learning, neurodegenerative diseases and biomolecular simulations targeting therapeutically important proteins or enzymes.

Balachandran Manavalan received his PhD in 2011 from Ajou University, Republic of Korea. After completing his PhD, he worked as a postdoctoral research fellow at the Korea Institute for Advanced Study (KIAS), Republic of Korea. Currently, he is a research professor at the Department of Physiology, Ajou University School of Medicine, Republic of Korea. He is also an associate member of KIAS, Republic of Korea. His main research interests include prediction of protein structures, machine learning, data mining, computational biology and functional genomics.

Hiroyuki Kurata is a professor of 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.

Submitted: 11 June 2020; **Received (in revised form):** 6 August 2020

position-specific information that possesses high discriminative capability. The resultant feature sets were inputted to six commonly used ML methods (random forest, support vector machine, extremely randomized tree, logistic regression, naïve Bayes and AdaBoost). The Rosaceae genome was employed to train the above classifiers, which generated 30 baseline models. To integrate their individual strength, Meta-i6mA was proposed that combined the baseline models using the meta-predictor approach. In extensive independent test, Meta-i6mA showed high Matthews correlation coefficient values of 0.918, 0.827 and 0.635 on Rosaceae, rice and *Arabidopsis thaliana,* respectively and outperformed the existing predictors. We anticipate that the Meta-i6mA can be applied across different plant species. Furthermore, we developed an online user-friendly web server, which is available at http://kurata14.bio.kyutech.ac.jp/Meta-i6mA/.

Key words: DNA N^6-methyladenine modification; prediction model; feature extraction; two-step feature optimization; meta-predictor

Introduction

DNA methylation is an important epigenetic modification present in the genomes of different species [1]. The prevalent DNA modifications include methylations on the 4th and 5th positions of the pyrimidine ring of cytosine [4-methylcytosine (4mC) and 5-methylcytosine (5mC)] and 6th position of the adenine purine ring [N^6-adenine methylation or N^6-methyladenine (6mA)] [2]. Due to the widespread distribution and multifaceted role of 5mC, this type of prevalent DNA modifications has been extensively studied. However, 6mA has been less investigated due to its sporadic distribution across genomes. The distribution and function of 6mA in eukaryotes largely remained unknown [3]. Recently, studies have shed light in multicellular eukaryotes on the contrasting regulatory functions and distribution of 6mA. 6mA DNA modifications are involved in the regulation of several biological processes, such as DNA repair and replication, gene expression and restriction-modification system [4–6]. Furthermore, recent studies have linked the abnormal status of 6mA modifications to human cancer and other diseases [7].

As the knowledge of DNA 6mA modifications and its distribution in multicellular eukaryotes remain insufficient, it is essential to develop highly sophisticated techniques for correct identification of its location in the genome. A number of experimental techniques were carried for the identification of 6mA sites, such as capillary electrophoresis with laser-induced fluorescence [8], next-generation sequencing with coupling immunoprecipitation [9], restriction enzyme-assisted sequencing with DpnI-assisted 6mA sequencing [10], DNA immunoprecipitation with 6mA antibodies [11], mass spectrometry, single-molecule real-time (SMRT) sequencing and ultrahigh-performance liquid chromatography [12]. Although these experiments are typically laborious and time consuming, they provide partial coverage of the 6mA epigenetics.

With the surging amount of novel DNA sequences and experimental intricacies, it is imperative to identify novel and efficient bioinformatics-based methods that identify methylated adenine sites in the genomes. Machine learning (ML) is such a promising approach that has been successfully implemented in addressing many biological issues [13–15]. Till date, ML-based 6mA predictors have been developed for rice, *Arabidopsis thaliana, Mus musculus* and Rosaceae including *Fragaria vesca* (FV) and *Rosa chinensis* (RC) [16–25]. Notably, a brief description of each method is provided in the method section. Although such species-specific ML-based approaches showed good performances, they did not apply their model to other species. Hence, their practical application to other species is quite limited. Therefore, they are necessitated to develop an effective predictor that can learn the knowledge buried between 6mAs and non-6mAs of a particular plant species and to successfully apply the predictor to other plant species.

In this study, we introduced Meta-i6mA, a novel and meta-approach designed to identify 6mA sites not only of Rosaceae genome but also of *A. thaliana* and rice genomes. We systematically analyzed 10 different feature encodings [Kmer frequency (Kmer), nucleic acid composition (NAC), dinucleotide composition (DNC), trinucleotide composition (TNC), *k*-space nucleotide composition (KNC), mononucleotide binary encoding (MBE), dinucleotide binary encoding (DBE), dinucleotide physicochemical properties (DPP), electron–ion interaction pseudopotentials (EIIP) of trinucleotides and nucleotide chemical properties (NCP)] using Rosaceae training dataset and identified five of them (DBE, MBE, EIIP, DPP and NCP), which indicate high discriminative capability. The optimal feature set was identified individually from five different encodings using two-step feature selection protocol and fed to six different ML classifiers [support vector machine (SVM), random forest (RF), AdaBoost (AB), logistic regression (LR), extremely randomized tree (ERT) and naïve Bayes (NB)]. We trained the 30 baseline models, whose predicted probability scores of 6mAs were further considered as a 30 dimensional (D) feature vector. Finally, the optimal feature set was identified from the 30D feature vectors for RF-based meta-predictor construction. Integration of the probability scores showed more discriminative ability than individual feature descriptors. Extensive comparative analysis demonstrated that Meta-i6mA achieved a more accurate and robust performance than the baseline models on the training dataset, and significantly outperformed the existing predictors on three independent datasets.

Materials and Methods

Overall framework of Meta-i6mA

Figure 1 shows the design and evaluation process of Meta-i6mA. It consists of the following four major steps: dataset curation, feature extraction, baseline model construction and the meta-predictor construction. At the 1st step, we collected datasets from the modification database for the rosaceae (MDR), MethSMRT and Gene Expression Omnibus (GEO) databases [26–28]. At the 2nd step, we computed the five different feature encodings that cover composition-based features, position-specific features, and physicochemical properties (PCPs) from Rosaceae training dataset and identified the most informative or optimal features using two-step feature selection procedure. At the 3rd step, these optimal feature sets were fed to six different classifiers to develop their corresponding baseline models using 5-fold cross-validation (CV). Finally, the predicted probability scores of 6mAs by the baseline models, named PSBs, were treated as an input feature vector to optimize the RF-based meta-model.

Construction of training and independent datasets

SMRT sequencing technique has been used to verify 4mCs and 6mAs for two species of FV [29] and RC [30], which are stored in the MDR database [26]. As our objective is to develop a 6mA site prediction method, we extracted only 6mA sites from the MDR database, where FV consists of 26 516 6mA or positive samples and RC consists of 14 666 positives samples. These two species come under category of Rosaceae family [26]. After deleting similar sequences, we combined these two genome samples and obtained 40 574 positive samples, where each sample contain 41 base pairs (bp) with adenine nucleotide ('A') at the center or 21st position. Subsequently, we excluded redundant sequences to ensure that the sequence identity of any two sequences in positive samples was less than 90%, which resulted in 36 537 positive samples. The procedure for generating the negative samples (non-6mA) was the same as widely employed in previous studies [16, 17, 23, 24]. We generated 41 bp fragment containing 'A' at 21st position from the entire chromosomes of two species and excluded the fragments that were not detected as a 6mA site by the experimental technique (SMRT), which resulted in a massive number of negative samples. To construct a balanced dataset, we randomly selected 36 537 negative samples. The window size was set to +/−20 bp after testing the performance for a series of window sizes (15, 20, 25, 30, 35 and 40). Specifically, the RF classifier was employed to develop 10 encoding-based models using a training dataset with varying window sizes. Supplementary Figure S1 shows that 41 bp achieved the best performance for all the window sizes, regardless of encodings. Therefore, 41 bp was selected as the optimal length. Our result was consistent with the previous study [31] that identified 41 bp of 4mC site the optimal length. Generally, adding more distal nucleotides improves the prediction performance, but such aspects were not observed in our evaluation. Finally, the randomly selected 80% of the samples (29 237 4mCs and 29 237 non-4mCs) were considered as a training dataset for prediction model development, whereas the remaining samples (7300 4mCs and 7300 non-4mCs) were treated as independent evaluation to check the model robustness.

In addition to the Rosaceae independent dataset, we considered two other plant species datasets (rice and *A. thaliana*) to show the prediction capability of our proposed method across plant species. In the case of rice genome, we utilized the very large dataset constructed in previous studies [20], which contains 154 000 6mA sites and 154 000 non-6mA sites. As no well-constructed *A. thaliana* dataset was available, we constructed it by extracting 31 873 6mA sites from National Center for Biotechnology Information GEO with accession number GSE81597. Positive samples were supplemented with the equal number of negative samples (non-6mAs) using the same procedure as mentioned above. To avoid the overestimation, none of samples from two plant species should be overlapped with the Rosaceae training dataset. The datasets employed in this study are summarized in Supplementary Table S1, which can be downloaded at http://kurata14.bio.kyutech.ac.jp/Meta-i6mA/help.php.

Feature extraction

Feature extraction is directly related to the performance of constructed prediction models [32, 33]. To extract meaningful information around 6mA and non-6mA sites from the training dataset, we explored 10 different feature encodings that can be categorized into three groups as follows: (i) composition-based, (ii) PCP-based and (iii) position-specific-based features.

Group 1: composition-based features

NAC: NAC computes the frequency of each bp type in a sequence. The frequency of all four bp can be computed as follows:

$$f(b) = \frac{K(b)}{L}, b \in \{A, G, C, T\} \tag{1}$$

where $K(b)$ represent the number of bp type b, and $L = 41$ bp.

DNC: DNC computes dinucleotide frequencies from a given sequence. The frequency of 16 dinucleotide pairs can be computed as follows:

$$f(m, n) = \frac{K_{m,n}}{K-1}, \quad m, n \in \{A, G, C, T\} \tag{2}$$

where $K_{m,n}$ is the dinucleotide frequencies represented by bp types m and n.

TNC: TNC gives trinucleotide frequency in a given sequence. It generates 64D feature vector that can be computed as follows:

$$f(m, n, o) = \frac{K_{m,n,o}}{K-2}, \quad m, n, o \in \{A, G, C, T\} \tag{3}$$

where $K_{m,n,o}$ is the trinucleotide frequencies represented by bp types m, n and o.

Kmer: Kmer integrated NAC, DNC, TNC, tetra- and penta-nucleotide composition and generated 1364D feature vector. Notably, tetra- and penta-nucleotide composition can be calculated in a similar manner to the Equation (3).

KNC: KNC encoding represents the DNA nucleotide information from a query sample using the frequency-wise pair similarity scores [34, 35]. A space of nucleotide frequency pair is encoded and normalized as follows:

$$\text{Frequency pair} = \frac{N\,(nf_i)}{w-d-1} \tag{4}$$

where $N(nf_i)$ is the sum of nf_i inside DNA samples, w is the length of sequence and d is the space between two nucleotides, encompassing from one to $d$max. Henceforward, the KNC generates a $4 \times 4 \times (d\text{max} + 1)$-D feature vector from a given sequence, where $d$max is set to three based on our preliminary analysis.

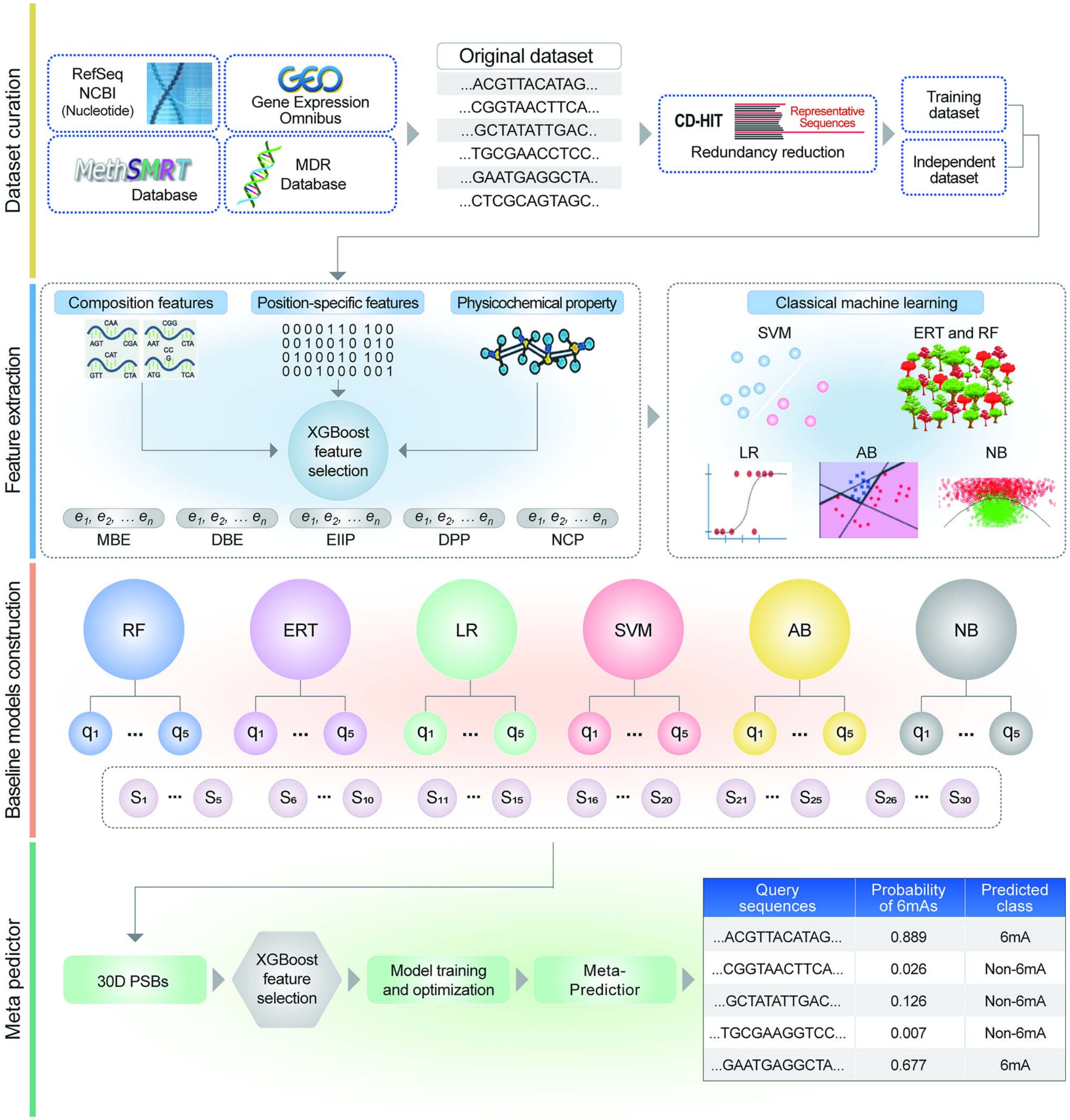
Group 2: position-specific features

MBE: MBE represents the exact nucleotide position of a given sequence. Each bp is transformed into a binary vector as follows:

$$\begin{cases} A = (1, 0, 0, 0) \\ G = (0, 1, 0, 0) \\ C = (0, 0, 1, 0) \\ T = (0, 0, 0, 1) \end{cases} \tag{5}$$

MBE generates a 164 (4×41)-D feature vector for the given DNA sequence with 41 bp.

DBE: In DBE, 16 possible dinucleotides are encoded as 0/1 (4D vector) [36]. For example, AT, AA, GG and AC are encoded



Query sequences	Probability of 6mAs	Predicted class
...ACGTTACATAG...	0.889	6mA
...CGGTAACTTCA...	0.026	Non-6mA
...GCTATATTGAC...	0.126	Non-6mA
...TGCGAAGGTCC...	0.007	Non-6mA
...GAATGAGGCTA...	0.677	6mA

Figure 1. The overall workflow for the development of Meta-i6mA. It consists of (i) data collection and curation; (ii) feature extraction; (iii) baseline model construction; and (iv) a meta-predictor construction.

as (0,0,0,1), (0,0,0,0), (1,1,1,1) and (0,0,1,0), respectively. DBE generates $(L-1\times\ 4)$-D feature vector for a given DNA sequence $(L=41)$.

Group 3: PCP-based features

EIIP: EIIP represents electron–ion energy distribution from given DNA sequences [37], whose values are 0.1260, 0.1340, 0.0806 and 0.1335 for A, C, G and T, respectively. EIIP gives 64D feature vector for a given sequence as follows:

$$D = [\mathrm{EIIP}_{\mathrm{CCC}} \times f_{\mathrm{CCC}}, \mathrm{EIIP}_{\mathrm{GGG}} \times f_{\mathrm{GGG}}, \mathrm{EIIP}_{\mathrm{AGC}} \times f_{\mathrm{AGC}}, \dots, \mathrm{EIIP}_{\mathrm{TTT}} \times f_{\mathrm{TTT}}] \tag{6}$$

Different trinucleotide combination is shown as subscripts in Equation (6); $\mathrm{EIIP}mno = \mathrm{EIIP}m + \mathrm{EIIP}n + \mathrm{EIIP}o$; where $m, n, o \in \{A, C, G, T\}$
where f and EIIPm is the value of the corresponding nucleotide and f_{mno} represents the normalized trinucleotide frequency.

DPP: we considered 15 types of PCPs [38, 39], including F-roll, F-tilt, F-twist, F-slide, F-shift, F-rise, roll, tilt, twist, slide, shift, rise, energy, enthalpy and entropy. The value for each PCP is provided in Supplementary Table S2. DPP can be computed as follows:

$$\text{DPP}(k) = \text{normalized frequency of dinucleotide}(k) \times \text{PCP}(\text{Y}_k)_l \quad (7)$$

where Y_k is the value of lth ($l = 1, 2, \ldots 15$) dinucleotide PCP and k is one of the 16 dinucleotides. Ultimately, DPP provides 240D feature vector.

NCP: four standard bp in a DNA sequence have a different property, including ring structure (RS), functional group (FG) and hydrogen bond (HB) [40]. We utilized this information to compute NCP as follows:

$$\text{RS} = \begin{cases} \text{Purine} \in \text{A, G} \\ \text{Pyrimidine} \in \text{C, T} \end{cases}, \quad (8)$$

$$\text{FG} = \begin{cases} \text{Amino} \in \text{A, C} \\ \text{Keto} \in \text{C, G} \end{cases}, \text{HB} = \begin{cases} \text{Weak} \in \text{A, T} \\ \text{Strong} \in \text{C, G} \end{cases}$$

Based on chemical properties, ‘A’ can be encoded to (1, 1, 1), ‘C’ to (0, 1, 0), ‘G’ to (1, 0, 0) and ‘T’ to (0, 0, 1), respectively. Ultimately, NCP gives 123 (3 × 41)-D feature vector for a given sequence.

Two-step feature selection approach

In general, the two-step feature selection approach is a systematic way to identify the optimal feature set from the original features [41–43]. In the two-step approach, the 1st step is to rank original feature dimension and the next step is sequential forward search. In this study, we applied the eXtreme Gradient Boosting (XGBoost) classifier [44] and computed the feature importance score that was used to rank the features. This approach is widely employed in various sequence-based function predictions [45, 46]. From the ranked features, we generated n feature subsets that contained the top ranked features ranging from top X to the top Y features with an interval of Z. The value of X, Y, Z and n depends on the original feature dimension. In the 2nd step, we sequentially inputted n feature subsets into an ML classifier and optimized the ML hyperparameters using 5-fold CV. The optimal model from different subsets were compared to select the model that achieved the highest Matthews correlation coefficient (MCC), whose corresponding feature set is considered as the optimal one.

Meta-approach

A meta-classifier was considered to develop the final prediction model. Briefly, the predicted probability scores of 6mAs by the 30 PSBs (five encodings × six ML classifiers) were used to develop a meta-predictor as follows:

$$M = P\{[L(1), E(1)], \ldots P[L(i), E(j)], \ldots, P[L(s), E(t)]\} \quad (9)$$

where M is the 30D feature vector of PSBs, $P\left[L(s), E(j)\right]$ is the PSB for a baseline model of ML $L(i)$ with encoding scheme $E(j)$, i and j are the indexes of the ML algorithm and encoding scheme and s and t are the total numbers of MLs and encodings.

Feature fusion approach

Feature fusion approach is quite popular and widely applied in computational biology [47]. Therefore, we applied this approach and checked whether or not feature fusion (F) improves the prediction performance. Briefly, 10 encoding representation vectors (E) of NAC, DNC, TNC, KNC, MBE, Kmer, DBE, DPP, EIIP and NCP were concatenated as follows:

$$F = [E(\text{NAC}), E(\text{DNC}), E(\text{TNC}), E(\text{KNC}), E(\text{MBE}), \quad (10)$$
$$E(\text{Kmer}), E(\text{DBE}), E(\text{DPP}), E(\text{EIIP}), E(\text{NCP})]$$

where F is the sequential fusion of 10 different feature vectors.

Classification algorithms and evaluation metrics

We employed six ML classifiers of RF [48], AB [49], SVM [50], ERT [51], LR and NB [52]. All these classifiers have been extensively applied in the field of computational biology and bioinformatics [42, 53, 54]. The implementation of these classifiers in the current study is the same as employed in our previous reports [24, 35, 40, 55, 56]. We optimize the hyperparameters of each ML with a search range, as shown in Supplementary Table S3. Subsequently, all possible parameter combination-based performances are evaluated using 5-fold CV. Finally, a parameter that achieves the highest MCC is considered as the optimal parameter.

Widely applied four metrics of accuracy (ACC), specificity (Sn), sensitivity (Sp) and MCC are employed to evaluate the performance of prediction models [41, 57, 58]. They are defined as follows:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}$$

where TP, FP, TN and FN denote the numbers of true positives, false positives, true negatives and false negatives, respectively.

Summary of the existing species-specific DNA 6mA site prediction methods

The number of computational methods is growing exponentially for the prediction of DNA 6mA sites from different species (Table 1). In 2018, Wang and Yan [25] developed the 1st computational method based on RF classifiers called RFAthM6A. The authors have constructed nonredundant training and independent datasets from *A. thaliana*, and considered the following encodings, namely position-specific nucleotide sequence profile, DNC, Kmer and KNC, to develop their model. It is worth to mention that nine different predictors were reported for the rice genome. Of those, Chen *et al.* developed the 1st predictor, i6mA-Pred [23], where they constructed the reduced redundancy training dataset and identified optimal features from NCP using two-step feature selection protocol and SVM. Interestingly, i6mA-Pred acted as a base for the development of the remaining eight methods, namely iDNA6mA-PseKNC [22], SDM6A [24], iDNA6mA-Rice [20], i6mA-DNCP [21], 6mA-RicePred [23], MM-6mAPred [17], SNNRice6mA [19] and 6mA-Finder

[16]. Four methods (iDNA6mA, SDM6A, i6mA-DNCP and MM-6mAPred) were applied to the same i6mA-Pred training dataset in different manners for the model development. In addition to the i6mA-Pred training dataset, three methods (iDNA6mA-Rice, SNNRice6mA and 6mA-RicePred) employed highly redundant homologous sequences for the model development with different approaches, whereas 6mA-Finder utilized a different training dataset. Briefly, iDNA6mA [59] was developed using a convolutional neural network classifier and MBE; SDM6A considered four different encodings (numerical representation of nucleotide, MBE, DBE and local position-specific dinucleotide frequency) and two different classifiers (SVM and ERT) to generate ensemble models for the final prediction; iDNA6mA-Rice considered RF and MBE [20]; SNNRice6mA employed deep learning and MBE; i6mA-DNCP [21] utilized dinucleotide-based features via classification and regression trees; 6mA-RicePred utilized four different encoding schemes (NCP, Kmer, Markov feature and MBE) and SVM algorithm; 6mA-Finder [16] employed eight different encodings (accumulated nucleotide frequency, MBE, KNC, DNC, enhanced nucleic acid composition, EIIP, NCP and pseudo DNC) and two different algorithms (RF and LR) to generate ensemble models; MM-6mAPred utilized Markov model transition probability between adjacent nucleotides.

iDNA6mA-PseKNC was proposed by Feng *et al.* [22] that identified DNA6mA sites from the mouse genome. The authors constructed the training dataset based on mouse genome and employed pseudo K-tuple nucleotide composition and SVM for the model development. Recently, our group has introduced i6mA-Fuse by fusing multiple features including Kmer, KNC, MBE, DBE and EIIP via different classifiers for identifying 6mA sites from Rosaceae subspecies [55]. We conclude that numerous efforts have been dedicated to developing computational prediction tools by exploring multiple feature encoding schemes and different classification algorithms. It remains unclear that the most prominent feature scheme with ML classifier performs well across different species. Therefore, systematic approaches with different feature encoding schemes as well as different ML classifiers are much needed. Such approaches will provide a practical guide to develop a more accurate DNA 6mA site prediction tools.

Results and Discussion

We assessed the performance of 10 different feature encoding schemes (categorized into three groups) using six different classifiers: RF, ERT, AB, SVM, LR and NB based on 5-fold CV tests. Figure 2 and Supplementary Table S4 show the prediction performances of 60 models with respect to different encodings and classifiers. We observed that two groups (position-specific- and PCPs-based encodings) achieved an excellent performance with MCC of 0.701–0.883, regardless of different classifiers. We noted the three classifiers of RF, ERT and SVM achieved a similar performance on five encodings (MBE, DBE, EIIP, DPP and NCP). Of these classifiers, RF achieved better performance in three encodings (MBE, DBE and EIIP) than the two classifiers of ERT and SVM and much better than NB, GB and LR. Surprisingly, composition-based features (NAC, DNC, TNC, Kmer and KNC) achieved an MCC value of 0.093–0.486, which was 21.5–60.8% lower than the least performing model (EIIP_LR). Overall, our comparative analysis showed that physicochemical- and position-specific encodings captured significant information around 6mA sites; hence, we considered these encodings for the further analysis. The composition-based groups were excluded from further analysis due to its difficulty in capturing meaningful information around 6mA sites.

Construction of baseline models based on the optimal feature set

The original feature subsets by each encoding may contain redundant and noisy information that can influence the model performance; hence, we applied feature selection protocols to extract the most relevant information from the original feature dimension [60, 61]. Here, the two-step feature selection approach (see Method section) was applied to each feature encoding to obtain their respective optimal feature set. For each encoding, XGBoost was used to rank the features, subsequently, different feature subsets were selected that contained the top ranked features ranging from the top 20 to the top 100 features with an interval of five. All these feature subsets were inputted to RF individually and their respective performances were evaluated using 5-fold CV. Finally, the feature set that achieved the highest MCC was selected as the optimal one. The two-step approach identified 100, 70, 60, 40 and 90 optimal features, respectively, for MBE, DBE, DPP, EIIP and NCP encodings, which were inputted to the remaining five classifiers to develop their corresponding models using 5-fold CV (Supplementary Table S5). To check whether the optimal feature set improved the prediction performance, we compared the performances of all classifiers in terms of MCC between the control (Supplementary Table S4) and the optimal feature set (Supplementary Table S5). In the case of the MBE encoding, MCC was similar between the control and the optimal feature sets for the three classifiers (ERT, LR and NB), whereas MCC was improved for RF, AB and SVM (Figure 3**A**). In the case of DBE, we observed an improvement for ERT and NB, whereas the remaining classifier performances were similar between the control and optimal sets (Figure 3**B**). Surprisingly, the performances of all six classifiers were similar to those of the control in the case of DPP encoding (Figure 3**C**). For the NCP encoding, a marginal improvement was observed for RF, ERT and LR, whereas the MCC value of the remaining three classifiers was similar to that of the control (Figure 3**D**). In the case of EIIP, significant improvement for LR and marginal improvement for the remaining five classifiers were observed (Figure 3**E**). Overall, the feature selection did not significantly improve the MCC, but reduced the feature dimension without affecting any control performance. Finally, the 30 optimal models, obtained from the above analysis, were defined as the baseline models.

Construction of Meta-i6mA

It is quite straightforward to create the best prediction model of 6mAs using the 30 baseline models. In this study, the probability scores predicted by each of the 30 baseline models were fused by means of meta-approach that can effectively discriminate positive and negative samples. As all the PSBs were not equally important, the two-step feature selection approach was applied again to exclude less relevant features. First, we ranked the PSBs using XGBoost (Figure 4**A**) and generated the six PSB feature subsets that contained the top ranked features ranging from top 5 to the top 30 features with an interval of five. Subsequently, all the PSB sets were inputted to just four classifiers (RF, ERT, SVM and NB) individually to develop their respective meta-predictors. The RF-based performance was consistently superior to the remaining three classifiers (SVM, NB and ERT), regardless of the PSB sets (Supplementary Figure S2); hence, we selected the RF classifier for the final model construction. Particularly, RF achieved a high performance with MCC of 0.931 using the top 15 and top 20 PSB sets (Supplementary Figure S2). However, we considered the top 15 PSB set as the optimal one to reduce the feature dimension.

Table 1. Currently available tools for DNA 6mAs prediction

Year	Predictor	ML algorithms	Features	TR/IND dataset size	Testing method	ACC TR/IND	Predictor URL	Genomes	Batch prediction
2018	RFAthM6A [25]	RF	PSNSP, DNC, Kmer, and KNC	2100 6mA; 2100 non-6mA/418 6mA; 418 non-6mA	5-fold CV and independent test	—/—	[b]https://github.com/nongdaxiaofeng/RFAthM6A	AT	NA
2019	i6mA-Pred [23]	SVM	NCP and KNC	880 6mA; 880 non-6mA/—	5-fold CV	0.831/—	[a]http://lin-group.cn/server/i6mA-Pred	Rice	Yes
	iDNA6mA [59]	CNN	MBE	880 6mA; 880 non-6mA/—	Jackknife test	0.866/—	[a]https://home.jbnu.ac.kr/NSCL/iDNA6mA.htm	Rice	NA
	SDM6A [24]	RF, ERT, GB, and SVM	NUM, MBE, DBE, and LPF	880 6mA; 880 non-6mA/221 6mA; 221 non-6mA	10-fold CV and independent test	0.882/0.880	[a]http://thegleelab.org/SDM6A/	Rice	Yes
	iDNA6mA-Rice [20]	RF	MBE, Kmer, and, NV	154000 6mA; 154000 non-6mA/880 6mA; 880 non-6mA	5-fold CV and independent test	0.917/0.940	[a]http://lin-group.cn/server/iDNA6mA-Rice	Rice	Yes
	SNNRice6mA [19]	DL	MBE	154000 6mA; 154000 non-6mA/—	5-fold CV	0.902/—	[b]https://github.com/yuht4/SNNRice6mA	Rice	NA
	i6mA-DNCP [21]	CART	DNC	880 6mA; 880 non-6mA/—	10-fold CV	0.866/—	[b]https://ww2.mathworks.cn/matlabcentral/fileexchange/72549-i6mA-dncp	Rice	NA
	iDNA6mA-PseKNC [22]	SVM, RF, PNN, and KNN	PseAAC, GDC, and KNC	1934 6mA; 1934 non-6mA /—	Jackknife test	0.967/—	[a]http://lin-group.cn/server/iDNA6mA-PseKNC	Mouse	NA
2020	6mA-RicePred [18]	SVM	NCP, Kmer, and MBE	880 6mA; 880 non-6mA/154000 6mA; 154000 non-6mA	10-fold CV Independent test	0.873/0.856	[b]https://github.com/huangqianfei0916/6ma-rice	Rice	NA
	6mA-Finder [16]	RF, SVM, KNN, and LR	ANF, MBE, KNC, DNC, ENAC, EIIP, NCP, and PseDNC	1934 6mA; 1934 non-6mA /—	10-fold CV	—/—	[a]https://bioinfo.uth.edu/6mA_Finder	Rice	Yes
	i6mA-Fuse [55]	RF	MBE, DBE, KNC, EIIP, and Kmer	FV: 4303 6mA; 4303 non-6mA/1067 6mA; 1067 non-6mA RC: 1430 6mA; 1430 non-6mA/3506mA; 350 non-6mA	10-fold CV and independent test	FV: 0.934 /0.937 RC: 0.916/929	[a]http://kurata14.bio.kyutech.ac.jp/i6mA-Fuse/	FV and RC	Yes
	MM-6mAPred [17]	Markov model	NCP	880 6mA; 880 non-6mA/—	10-fold CV	89.72/—	[a]http://www.insect-genome.com/MM-6mAPred/	Rice	Yes

Note: The 1st and 2nd columns indicate the publication year and the name of predictors. The 3rd column indicates the machine learning algorithms implemented in their method. The 4th and 5th columns represent feature encoding employed and the dataset size, respectively. Fundamentally, 5th has two types of information: the 1st portion represents the size of training dataset that is used for prediction model development and the 2nd portion represents the size of independent dataset that is used for evaluating the prediction model. The 6th column denotes types of validation evaluated in their study. The 7th, 8th, 9th and 10 represent the reported accuracy in the literature, webserver information, species specificity and batch mode response, respectively. In 8th column, [a]webservers and [b]standalone programs. If the value is not provided, it is denoted as '–'.

TR, training data; IND, independent data; PSNSP, position-specific nucleotide sequence profile; Kmer, k-mer nucleotide properties; KNC, k-space nucleotide composition; NV, natural vector; NUM, numerical representation of nucleotides; LPF, position-specific dinucleotide frequency; ENAC, enhanced nucleic acid composition; GDC, group dinucleotide properties; ANF, accumulated nucleotide frequency; PseKNC, pseudo nucleotide composition; CNN, convolution neural network; DL, deep learning; CART, classification and regression trees; PNN, probabilistic neural network; k-nearest neighbor; LR, logistic regression; SC, *Saccharomyces cerevisiae*; HS, *Homo sapiens*; AT, *Arabidopsis thaliana*.

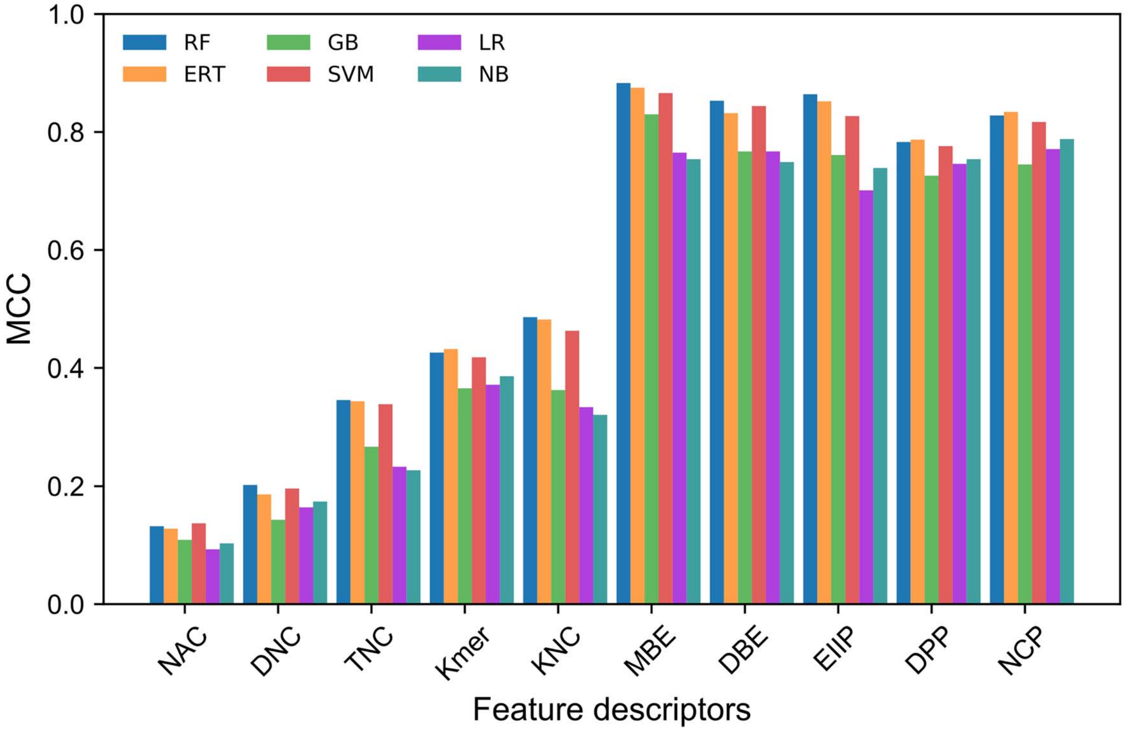


Figure 2. MCC comparison of six different classifiers (RF, ERT, GB, SVM, LR and AB) with 10 different feature encodings (NAC, DNC, TNC, Kmer, KNC, MBE, DBE, EIIP, DPP and NCP) by 5-fold CV.

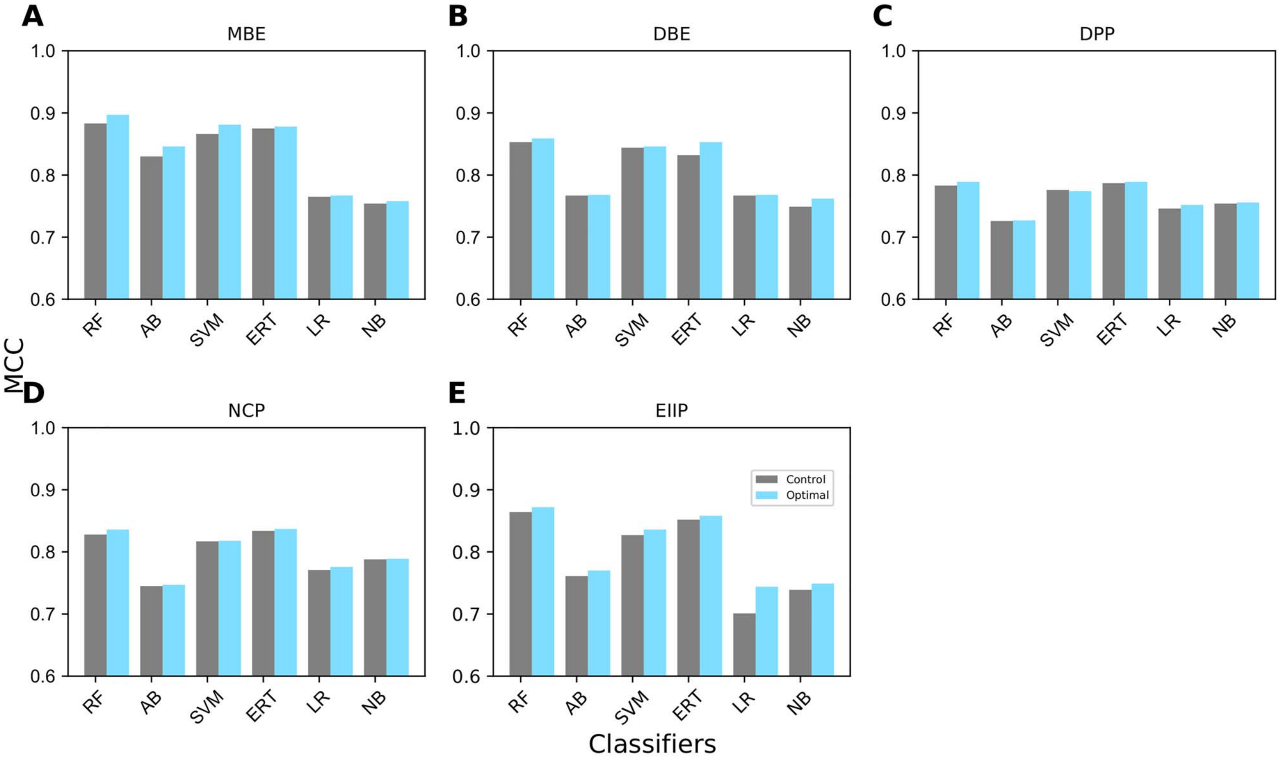


Figure 3. ACC comparison between the optimal features and original features for five different encodings (MBE, DBE, DPP, NCP and EIIP) during 5-fold CV. (**A**) MBE; (**B**) DBE; (**C**) DPP; (**D**) NCP; and (**E**) EIIP.

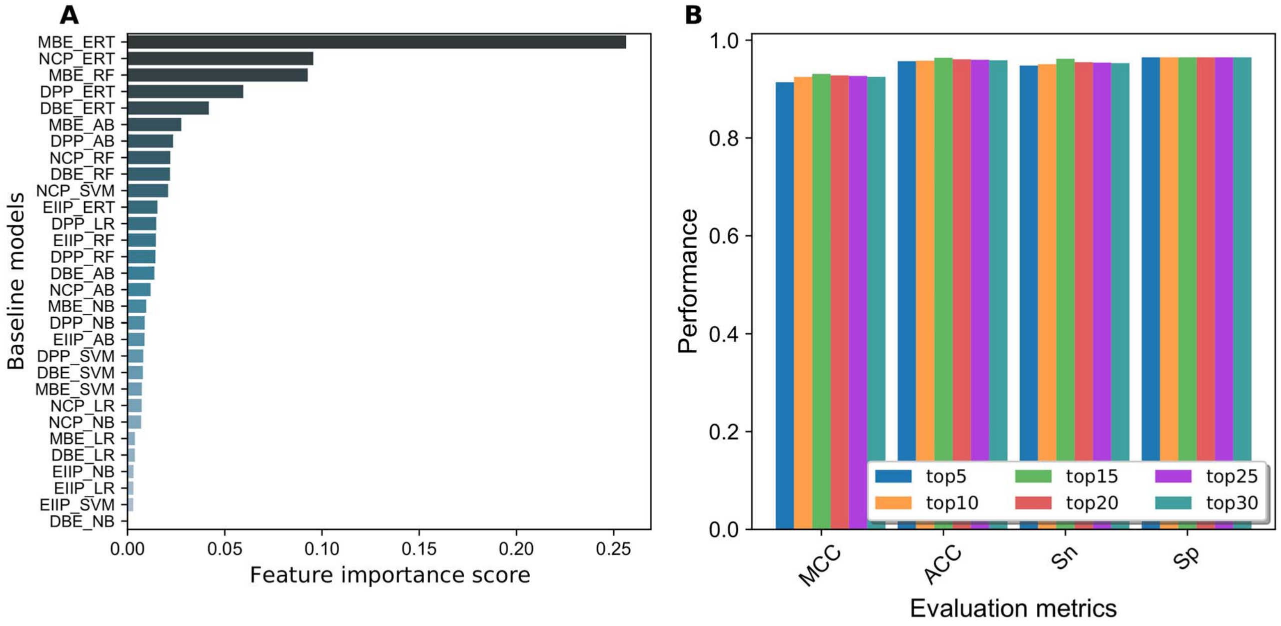


Figure 4. Performance comparison of the baseline models and meta-predictor. (**A**) Ranked baseline models by the XGBoost classifier. X and Y axes represent the feature importance score and the baseline model name, respectively. (**B**) Performance comparison of six different PSB sets in terms of MCC, ACC, Sn and Sp by 5-fold CV.

The final model was named as Meta-i6mA. The predicted probability of 6mA was optimized to define the class (6mA or non-6mA) and the following classification $\begin{cases} \text{6mA, if } P \geq 0.455 \\ \text{non} - \text{6mA, else} \end{cases}$ was obtained. The Meta-i6mA achieved MCC, ACC, Sn and Sp of 0.931, 0.964, 0.962 and 0.965, respectively (Figure 4**B**).

To show the advantage of our approach, we compared Meta-i6mA with the 60 baseline models that included all feature encoding scheme-based models shown in Figure 2. Figure 5**A** shows that the MCC of Meta-i6mA is 5.0–83.8% higher than that of the baseline models, indicating that our approach significantly improves the performance. Further comparison was made for independent evaluation (Figure 5**B**–**D**), which will be discussed later.

Comparison of Meta-i6mA with the feature fusion approach

Recently, several studies have demonstrated some advantages of feature fusion approaches in various prediction problems [53, 62–64]. Therefore, we employed the feature fusion approach to check whether it can improve the prediction performance compared to the meta-approach. All successive optimal feature vectors of each encoding were concatenated in a row to generate a 474D (including NAC:4D, DNC:16D, TNC:25D, Kmer:40D, KNC:35D, MBE:90D, DBE:70D, DPP:50D, EIIP:64D and NCP:80D) feature vector. By applying XGBoost, we ranked the features to generate different feature sets that contained the top ranked features ranking from 25 to 450 features with an interval of 25. Each feature set was inputted to RF to evaluate its performance and to select the optimal feature set that show higher performances than their counterparts. As 200D feature vector achieved superior performance, we inputted this feature vector into six different classifiers (RF, AB, SVM, ERT, LR and NB) to develop their corresponding prediction models using 5-fold CV. The RF-based feature fusion model achieved the best performance with MCC, ACC, Sn, Sp and AUC of 0.891, 0.936, 0.921, 0.950 and 0.958, respectively (Supplementary Table S6). Interestingly, the RF-based feature fusion model significantly outperformed the AB-, LR- and NB-based feature fusion models with a *P*-value threshold of 0.01, and the RF-based feature fusion model were better than SVM- and ERT-based models. Furthermore, when we compared the RF-based feature fusion model with the Meta-i6mA, we observed significant improvement of Meta-i6mA, with 2.7% increased MCC and 2.1% increased ACC. It indicates that only the meta-approach was responsible for the improved performance.

Evaluation of Meta-i6mA and the existing predictors using independent datasets

During training, we like to exclude a possibility of over-fitting problems to achieve the highest ACC. Therefore, it is highly recommended to evaluate the developed model on the independent dataset to check whether the model is applicable to unseen data [41]. Notably, the independent dataset never contains the data used during the model development (i.e. training). The following results are organized according to the species:

(i) Rosaceae

The prediction performance of Meta-i6mA was compared with that of the baseline models, followed by the existing predictors listed in Table 1. We excluded higher eukaryotic species model (Mouse) and considered the remaining 11 models, which were developed using training datasets of plant species such as *A. thaliana*, Rosaceae and rice. It should be noted that the iDNA6mA and iDNA6mA-PseKNC servers are no longer functional. The four models (RFAthM6A, SNNRice6mA, i6mA-DNCP and 6mA-RicePred) did not provide any web implementation for testing their algorithms. These six applications were therefore not characterized. Of those existing six models (i6mA-Fuse, iDNA6mA-Rice, MM-6mAPred, 6mA-Finder, i6mA-Pred and SDM6A), the i6mA-Fuse contains two variants of i6mA-Fuse (RC) and i6mA-Fuse (FV) that are developed using RC and FV training

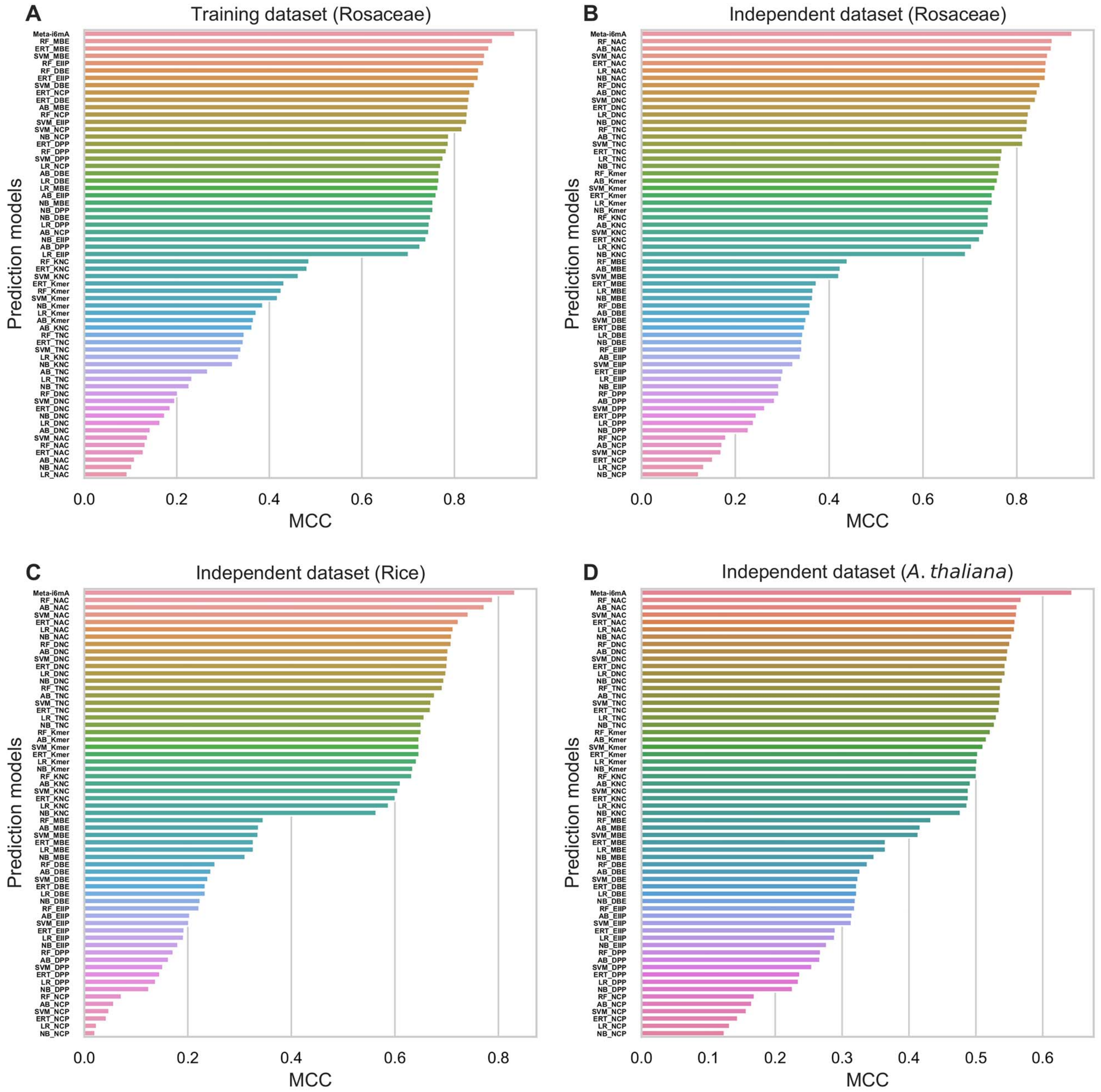


Figure 5. MCC comparison of Meta-i6mA and 60 baseline models. (**A**) The MCC of each predictor on the Rosaceae training dataset. Performances on the three independent datasets of (**B**) Rosaceae, (**C**) rice and (**D**) *A. thaliana* genomes.

datasets, respectively. Table 2 shows that Meta-i6mA achieves MCC, ACC, Sn and Sp of 0.918, 0.958, 0.960 and 0.957, respectively. Specifically, MCC of the proposed method was 4.20–79.6% higher than the baseline models (Figure 5**B**), indicating that meta-approach-based integration of multiple PSBs significantly improved the performance not only on training dataset but also on independent test.

Table 2 shows the performance comparison between Meta-i6mA and the existing methods. Among the existing methods, the two variants of i6mA-Fuse used a part of Meta-i6mA training dataset. Meta-i6mA indicates 10.2–25.2% higher MCC than the existing methods. Using a *P*-value threshold of 0.01, Meta-i6mA significantly outperformed the existing predictors. The remaining methods used rice datasets for the model development; hence, it is a good choice to validate the rice model as other plant species.

(ii) Rice

In the rice genome dataset, Meta-i6mA achieved MCC, ACC, Sn and Sp of 0.827, 0.899, 0.892 and 0.905, respectively (Table 3). Specifically, the MCC of Meta-i6mA was 4.3–81.1% higher than that of the 60 baseline models (Figure 5**C**). Furthermore, we observed that MCC of Meta-i6mA was 8.4–33.6% higher than that of the existing predictors. Interestingly, Meta-i6mA

Table 2. Performance comparison for 6mA site predictors on Rosaceae independent dataset

Method	MCC	ACC	Sn	Sp
Meta-i6mA	0.918	0.958	0.960	0.957
i6mA-Fuse (FV)	0.816	0.897	0.852	0.942
i6mA-Fuse (RC)	0.770	0.874	0.847	0.901
SD6MA	0.758	0.867	0.911	0.823
iDNA6mA-Rice	0.737	0.860	0.916	0.803
6mA-Finder	0.699	0.846	0.927	0.764
MM-6mAPred	0.686	0.839	0.940	0.738
i6mA-Pred	0.666	0.832	0.884	0.779

Table 3. Performance comparison for 6mA site predictors on rice dataset

Method	MCC	ACC	Sn	Sp
Meta-i6mA	0.827	0.899	0.892	0.905
i6mA-Fuse (FV)	0.743	0.869	0.828	0.910
MM-6mAPred	0.701	0.840	0.914	0.766
SDM6A	0.697	0.838	0.917	0.759
iDNA6mA-Rice	0.639	0.802	0.908	0.696
6mA-Finder	0.635	0.808	0.928	0.687
i6mA-Fuse (RC)	0.564	0.783	0.693	0.872
i6mA-Pred	0.491	0.744	0.815	0.672

significantly outperformed the existing predictors ($P < 0.01$). Of the eight methods, Meta-i6mA and two variants of i6mA-Fuse were developed using Rosaceae genome or a part of Rosaceae genome dataset; the remaining five models were developed using the rice dataset. The five methods trained by the rice dataset showed very high ACC, but such a high prediction ACC was not reproduced on our independent evaluation. Surprisingly, Meta-i6mA performance was sufficiently high when applying to the rice dataset, indicating that large training dataset and the meta-approach are the major reasons for such improved performance, especially on other plant species.

(iii) *Arabidopsis thaliana*

In the *A. thaliana* dataset, Meta-i6mA achieved MCC, ACC, Sn and Sp of 0.635, 0.809, 0.708 and 0.910, respectively (Table 4). Specifically, the MCC of Meta-i6mA was 7.6–38.9% higher than that of the 60 baseline models (Figure 5**D**). Table 4 compared the performance of Meta-i6mA with the existing predictors. Although the existing methods achieved a reasonable ACC of 71.8–76.0%, the improvement of Meta-i6mA was significant. Specifically, the MCC and ACC of Meta-i6mA were 11.5–21.4% and 4.9–9.8% higher than those of the existing predictors. Interestingly, Meta-i6mA significantly outperformed the existing predictors ($P < 0.01$). Overall, our analysis showed that Meta-i6mA trained with Rosaceae genome predicted 6mA sites of the other two plant species with higher ACC than their counter parts, indicating Meta-i6mA has a great potential of predicting 6mAs across plant species genomes.

Feature contribution analysis

Meta-i6mA has significantly improved the performance when compared to the baseline models not only on the training dataset but also on independent datasets of different plant species. To understand the reason, we computed t-distributed stochastic neighbor embedding implemented in scikit-learn v0.22.1 with the following parameters (n_component = 2, perplexity = 50 and learning rate = 300) on five optimal feature sets (MBE, DBE, EIIP, DPP and NCP) and the optimal PSBs. As shown in Figure 6**A**–**E**, the positive and negative samples for the five feature encodings were mixed up together, indicating that they have limited capability to distinguish true 6mAs from non-6mAs in the feature space. Unfortunately, the actual feature may not be sufficient enough to discriminate positive and negative samples in the feature space. Conversely, we observed two distinct clusters for positive and negative samples in the feature space of the PSBs, with some overlaps (Figure 6**F**), indicating that PSBs uncovers the underlying patterns between positive and negative samples. Overall, the above analysis can explain why our optimal PSB representations can provide a discriminative power for the prediction of 6mAs.

Meta-i6mA generalizability to other plant species

Generally, a model developed based on species-specific dataset performs well on the particular species, and it may perform well on closest homologs. To understand the result that Meta-i6mA is applicable to other plant species, we computed the position-specific nucleic acid preference of three datasets (Rosaceae, rice and *A. thaliana*) using the two-sample logo program [65]. Figure 7**A** and **B** shows that downstream of 6mA sites from position 1 to 11 (G/C, G/C, G/T, A, G, C, A, A/G, C, A, A) and downstream of non-6mA sites from position 1 to 9 (A, T/A, A, T/C, T, A/G, C/T, C/T, A) are highly conserved between RF and rice datasets. Similarly, Figure 7**A** and **C** shows that upstream of 6mA sites from position −4 to −1 (A, A, C/A, G/A) and downstream of non-6mA sites from position 1 to 9 (A/T, A/T, A, T/C, T, G/A, C/T, C/T, A) are highly conserved between *A. thaliana* and Rosaceae. Notably, the previous predictors that were developed using either rice dataset or Rosaceae family dataset were not able to capture the consensus information well when tested on other plant species. However, the Meta-i6mA successfully used the learnt information to identify significant patterns around 6mA site in other two plant species. This

Table 4. Performance comparison for 6mA site predictors on Arabidopsis dataset

Method	MCC	ACC	Sn	Sp
Meta-i6mA	0.635	0.809	0.708	0.910
MM-6mAPred	0.520	0.760	0.778	0.742
SD6MA	0.505	0.752	0.711	0.792
i6mA-Fuse (FV)	0.466	0.723	0.603	0.843
iDNA6mA-Rice	0.461	0.733	0.813	0.652
i6mA-Fuse (RC)	0.451	0.718	0.616	0.820
i6mA-Pred	0.447	0.722	0.773	0.671
6mA-Finder	0.421	0.711	0.715	0.706

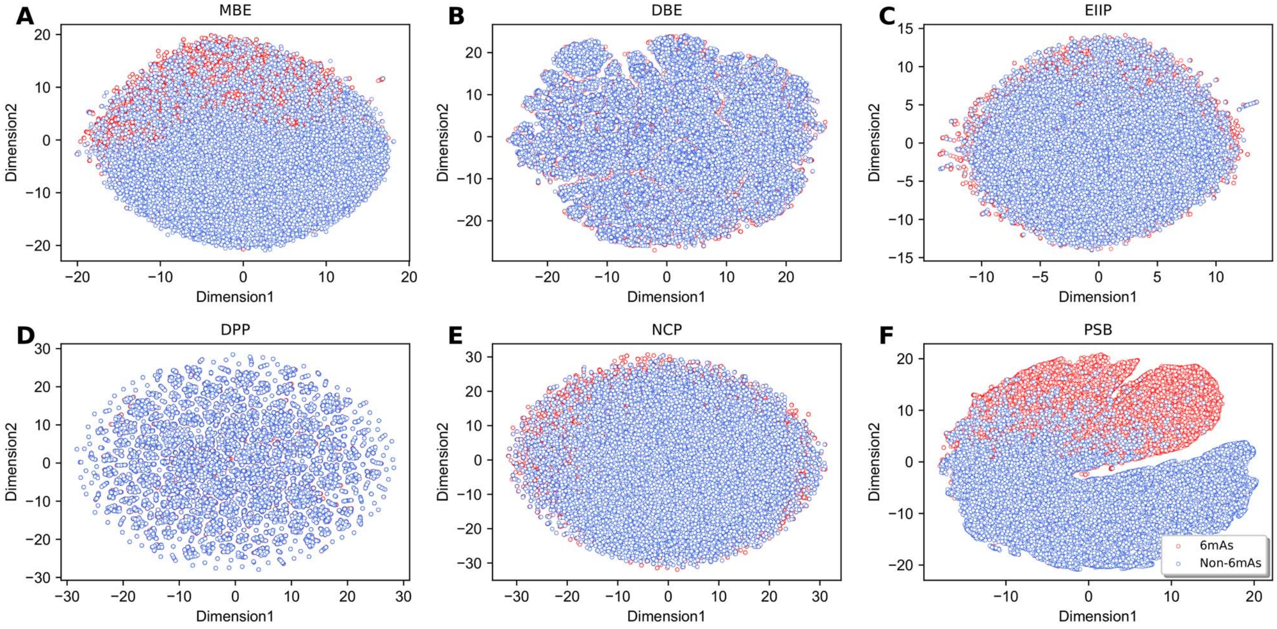


Figure 6. t-distributed stochastic neighbor embedding (t-SNE) distribution of the five selected different encodings and PSBs. (**A**–**E**) Distribution of the five individual optimal feature descriptors: MBE, DBE, EIIP, DPP and NCP. (**F**) Distribution of the optimal PSBs.

led to improved performance of Meta-i6mA over the existing methods.

Conclusions

A novel predictor, Meta-i6mA, has been proposed that can accurately predict 6mA sites of not only Rosaceae species but also *A. thaliana* and rice species. Meta-i6mA integrated PSBs generated by five different feature encoding schemes and six different popular classifiers. Subsequently, the optimal feature set was identified from the 30D feature vectors and inputted to RF to develop the robust meta-predictor. Our empirical studies, based on independent evaluation using three plant species datasets, demonstrated the superiority of Meta-i6mA, which consistently outperformed the existing predictors. The excellent performance of Meta-i6mA was mainly due to the following reasons: (i) a large high-quality dataset is employed, (ii) XGBoost-based feature ranking and sequential forward search can effectively identify the optimal feature set on both the baseline models and meta-predictor, (iii) the knowledge buried between positive and negative samples can be effectively learnt, and (iv) the learnt knowledge from Rosaceae genomes can capture common patterns to other plant species.

Overall, we demonstrated that the prediction model built based on Rosaceae species can be applied to the rice and Arabidopsis species. We expect that Meta-i6mA can be used to other plant species for genome-wide detection of 6mA sites. A user-friendly web server for Meta-i6mA is freely accessible at http://kurata14.bio.kyutech.ac.jp/Meta-i6mA. Due to the increasing demand of epigenetic modification site predictions, the applicability of the proposed predictors of other DNA sites, such as DNA N^4-methylcytosine [40, 66], 5mC and 5-hydroxymethylcytosine [61], will be investigated in the future studies. In general, the performance of ML-based models is directly proportional to the effective feature information amount learnt from the training data. Thus, it is a challenging task to propose novel effective learning approaches. Recently, several novel computational frameworks have been proposed for sequence-based function prediction problems [15, 42, 67–69]. It would be interesting to apply such approaches to 6mA site prediction problems and investigate whether these approaches improve the prediction performances.

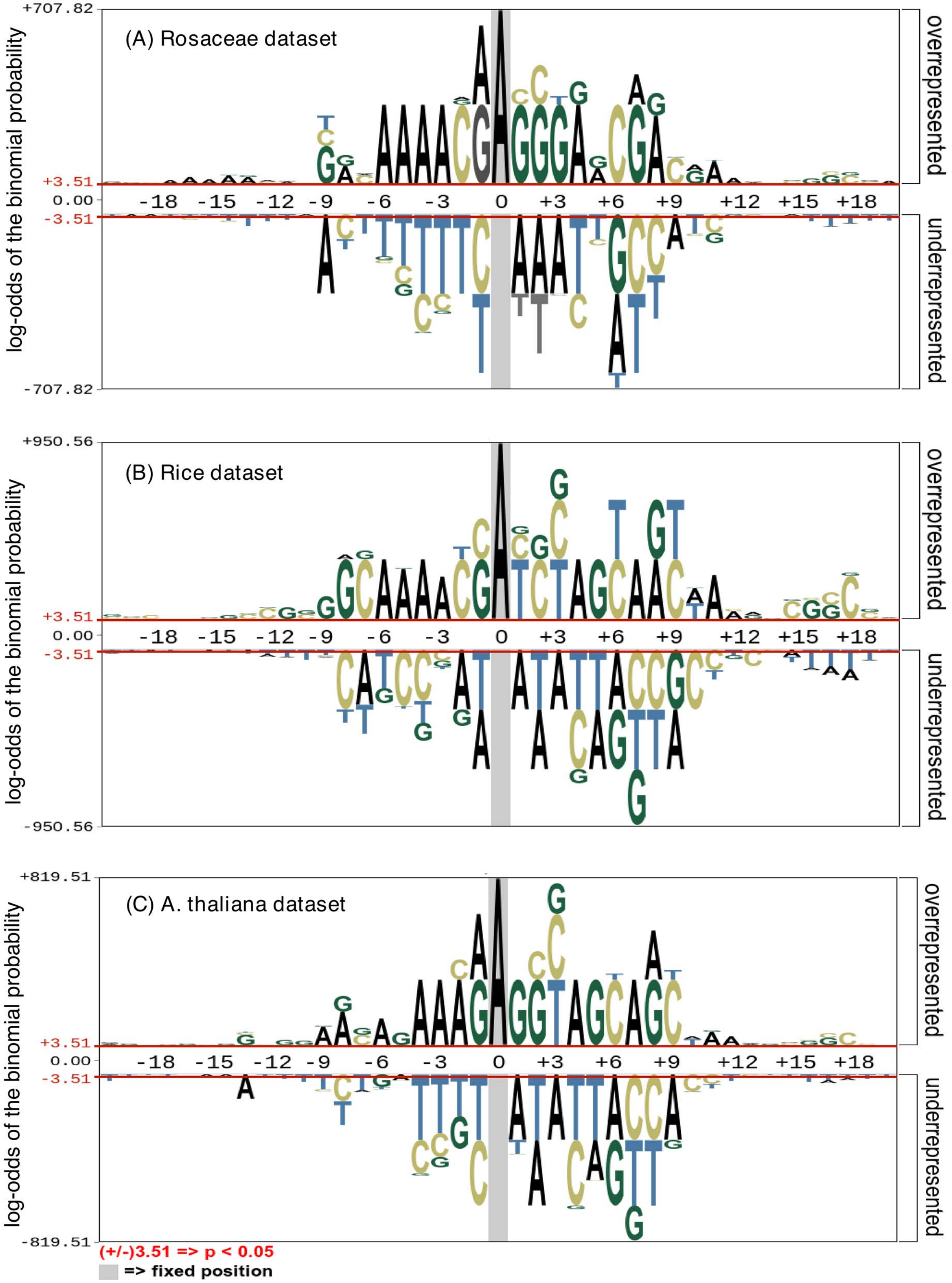


Figure 7. Nucleotide preference around 6mA and non-6mA sites: (**A**) Rosaceae dataset, (**B**) rice dataset and (**C**) *A. thaliana* dataset. In each figure, the top panel shows the overrepresented nucleotides for 6mA site-containing sequences; the bottom panel indicates the underrepresented nucleotides for non-6mA site-containing sequences. Notably, the nucleotides were significantly overrepresented or underrepresented (*t*-test, $P < 0.05$) nearby the centered 6mA and in non-6mA samples.

Key Points

- Meta-i6mA is a powerful tool to predict DNA N^6-methyladenine sites of plant genomes.
- We explored 10 different feature encoding schemes, with the goal of capturing key characteristics around 6mA sites.
- Meta-i6mA integrated the PSB sets generated by five different feature encoding schemes and six different popular classifiers (30 baseline models).
- We demonstrated that the model built based on Rosaceae species can be applied to the rice and Arabidopsis species.
- Meta-i6mA provides a convenient web server for researchers to thoroughly explore 6mA sites of plant genomes.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work was supported by the Grant-in-Aid for Scientific Research (B) [19H04208] and partially supported by the Grant-in-Aid for Japan Society for the Promotion of Science (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 (MSIT) [2018R1D1A1B07049572, 2019R1I1A1A01062260, 2020R1A4A4079722 and 2020M3E5D9080661].

Conflict of Interest

The authors have declared no competing interests.

References

1. Shi DQ, Ali I, Tang J, *et al.* New insights into 5hmC DNA modification: generation, distribution and function. *Front Genet* 2017;**8**:100.
2. Liang Z, Shen L, Cui X, *et al.* DNA N(6)-adenine methylation in *Arabidopsis thaliana*. *Dev Cell* 2018;**45**:406–16.e3.
3. Feng S, Cokus SJ, Zhang X, *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* 2010;**107**:8689–94.
4. Au KG, Welsh K, Modrich P. Initiation of methyl-directed mismatch repair. *J Biol Chem* 1992;**267**:12142–8.
5. Campbell JL, Kleckner N. *E. coli* oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. *Cell* 1990;**62**:967–79.
6. Robbins-Manke JL, Zdraveski ZZ, Marinus M, *et al.* Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase- and mismatch repair-deficient *Escherichia coli*. *J Bacteriol* 2005;**187**:7027–37.
7. Tang W, Wan S, Yang Z, *et al.* Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 2018;**34**:398–406.
8. Krais AM, Cornelius MG, Schmeiser HH. Genomic N(6)-methyladenine determination by MEKC with LIF. *Electrophoresis* 2010;**31**:3548–51.
9. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet* 2013;**14**:204–20.
10. Luo GZ, Wang F, Weng X, *et al.* Characterization of eukaryotic DNA N(6)-methyladenine by a highly sensitive restriction enzyme-assisted sequencing. *Nat Commun* 2016;**7**:11301.
11. Zhang G, Huang H, Liu D, *et al.* N6-methyladenine DNA modification in Drosophila. *Cell* 2015;**161**:893–906.
12. Fang G, Munera D, Friedman DI, *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat Biotechnol* 2012;**30**:1232–9.
13. Li J, Wei L, Guo F, *et al.* EP3: an ensemble predictor that accurately identifies type III secreted effectors. *Brief Bioinform* 2020;bbaa008. https://doi.org/10.1093/bib/bbaa008.
14. 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**:3350–6.
15. Rao B, Zhou C, Zhang G, *et al.* ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. *Brief Bioinform* 2019;bbz088. https://doi.org/10.1093/bib/bbz088.
16. 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**:3257–9.
17. Pian C, Zhang G, Li F, *et al.* MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics* 2020;**36**:388–92.
18. 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.
19. Yu H, Dai Z. SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Front Genet* 2019;**10**:1071.
20. 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.
21. 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**:828. doi: 10.3390/genes10100828.
22. 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.
23. Chen W, Lv H, Nie F, *et al.* i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 2019;**35**:2796–800.
24. 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**: 131–41.
25. Wang X, Yan R. RFAthM6A: a new tool for predicting m(6)a sites in *Arabidopsis thaliana*. *Plant Mol Biol* 2018;**96**:327–37.
26. 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.
27. 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**:D85–9.

28. Clough E, Barrett T. The Gene Expression Omnibus database. *Methods Mol Biol* 2016;**1418**:93–110.
29. Edger PP, VanBuren R, Colle M, *et al.* Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. *Gigascience* 2018;**7**:1–7.
30. Raymond O, Gouzy J, Just J, *et al.* The Rosa genome provides new insights into the domestication of modern roses. *Nat Genet* 2018;**50**:772–7.
31. Chen W, Yang H, Feng P, *et al.* iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 2017;**33**:3518–23.
32. Xu ZC, Feng PM, Yang H, *et al.* iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics* 2019;**35**:4922–9.
33. Yang H, Lv H, Ding H, *et al.* iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens. *J Comput Biol* 2018;**25**:1266–77.
34. 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**: 906–12.
35. 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**:752–58.
36. Manavalan B, Basith S, Shin TH, *et al.* 4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse genome. *Cell* 2019;**8**:1332. doi: 10.3390/cells8111332.
37. Nair AS, Sreenadhan SP. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation* 2006;**1**:197–202.
38. 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.
39. Liu K, Chen W. iMRM:a platform for simultaneously identifying multiple kinds of RNA modifications. *Bioinformatics* 2020;**36**:3336–42.
40. Manavalan B, Basith S, Shin TH, *et al.* Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol Ther Nucleic Acids* 2019;**16**:733–44.
41. 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. doi: 10.1002/med.21658.
42. Chen Z, Liu X, Li F, *et al.* Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief Bioinform* 2019;**20**:2267–90.
43. Manavalan B, Basith S, Shin TH, *et al.* AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput Struct Biotechnol J* 2019;**17**:972–81.
44. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: Cornell University, 2016, p. 785–94.
45. Jia C, Bi Y, Chen J, *et al.* PASSION: an ensemble neural network approach for identifying the binding sites of RBPs on circRNAs. *Bioinformatics* 2020;btaa522. https://doi.org/10.1093/bioinformatics/btaa522.
46. Yu J, Shi S, Zhang F, *et al.* PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization. *Bioinformatics* 2019;**35**:2749–56.
47. Niu M, Zhang J, Li Y, *et al.* CirRNAPL: a web server for the identification of circRNA based on extreme learning machine. *Comput Struct Biotechnol J* 2020;**18**:834–42.
48. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
49. Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. In: *European Conference on Computational Learning Theory*. Berlin, Heidelberg: Springer, 1995, 23–37. https://doi.org/10.1007/3-540-59119-2_166.
50. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.
51. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn* 2006;**63**:3–42.
52. Hand DJ, Yu K. Idiot's Bayes—not so stupid after all? *Int Stat Rev* 2001;**69**:385–98.
53. Li F, Chen J, Ge Z, *et al.* Computational prediction and interpretation of both general and specific types of promoters in *Escherichia coli* by exploiting a stacked ensemble-learning framework. *Brief Bioinform* 2020;bbaa049. doi: 10.1093/bib/bbaa049.
54. Zhu Y-H, Hu J, Ge F, *et al.* Accurate multistage prediction of protein crystallization propensity using deep-cascade forest with sequence-based features. *Brief Bioinform* 2020;bbaa076. doi: 10.1093/bib/bbaa076.
55. 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**:225–34. doi: 10.1007/s11103-020-00988-y.
56. Manavalan B, Basith S, Shin TH, *et al.* mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 2019;**35**:2757–65.
57. Khatun MS, Hasan MM, Kurata H. PreAIP: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. *Front Genet* 2019;**10**:129.
58. Ding H, Yang W, Tang H, *et al.* PHYPred: a tool for identifying bacteriophage enzymes and hydrolases. *Virol Sin* 2016;**31**:350–2.
59. Tahir M, Tayara H, Chong KT. 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.
60. 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**:412–20.
61. 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.
62. Elbasir A, Mall R, Kunji K, *et al.* BCrystal: an interpretable sequence-based protein crystallization predictor. *Bioinformatics* 2020;**36**:1429–38.
63. Yu B, Qiu W, Chen C, *et al.* SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics* 2020;**36**:1074–81.
64. Zhang Y, Xie R, Wang J, *et al.* Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform* 2019;**20**:2185–99.

65. O'Shea JP, Chou MF, Quader SA, *et al*. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods* 2013;**10**:1211–2.
66. Wei L, Su R, Luan S, *et al*. Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 2019;**35**:4930–7.
67. Chen Z, Zhao P, Li F, *et al*. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. *Brief Bioinform* 2019;bbz112. doi: 10.1093/bib/bbz112.
68. Li F, Chen J, Leier A, *et al*. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics* 2019;**36**:1057–65. doi: 10.1093/bioinformatics/btz721.
69. Qiang X, Zhou C, Ye X, *et al*. CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief Bioinform* 2018;**21**:11–23. https://doi.org/10.1093/bib/bby091.