




i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation

Md. Mehedi Hasan^{1,2} · Balachandran Manavalan³ · Watsara Shoombuatong⁴ · Mst. Shamima Khatun¹ · Hiroyuki Kurata^{1,5} 

Received: 6 February 2020 / Accepted: 29 February 2020 / Published online: 5 March 2020
© Springer Nature B.V. 2020

Abstract

DNA N⁶-methyladenine (6 mA) is one of the most vital epigenetic modifications and involved in controlling the various gene expression levels. With the avalanche of DNA sequences generated in numerous databases, the accurate identification of 6 mA plays an essential role for understanding molecular mechanisms. Because the experimental approaches are time-consuming and costly, it is desirable to develop a computation model for rapidly and accurately identifying 6 mA. To the best of our knowledge, we first proposed a computational model named i6mA-Fuse to predict 6 mA sites from the Rosaceae genomes, especially in *Rosa chinensis* and *Fragaria vesca*. We implemented the five encoding schemes, i.e., mononucleotide binary, dinucleotide binary, k-space spectral nucleotide, k-mer, and electron–ion interaction pseudo potential compositions, to build the five, single-encoding random forest (RF) models. The i6mA-Fuse uses a linear regression model to combine the predicted probability scores of the five, single encoding-based RF models. The resultant species-specific i6mA-Fuse achieved remarkably high performances with AUCs of 0.982 and 0.978 and with MCCs of 0.869 and 0.858 on the independent datasets of *Rosa chinensis* and *Fragaria vesca*, respectively. In the *F. vesca*-specific i6mA-Fuse, the MBE and EIIP contributed to 75% and 25% of the total prediction; in the *R. chinensis*-specific i6mA-Fuse, Kmer, MBE, and EIIP contribute to 15%, 65%, and 20% of the total prediction. To assist high-throughput prediction for DNA 6 mA identification, the i6mA-Fuse is publicly accessible at <https://kurata14.bio.kyutech.ac.jp/i6mA-Fuse/>.

Key message

The existing prediction models are not suitable to identify 6mA in the Rosaceae genome because the existing algorithms are species-specific. Thus, a novel predictor is desired to be established to identify 6mA sites in the Rosaceae genome. To the best of our knowledge, we first propose a computation model named i6mA-Fuse (Identification of N6-MethylAdenine sites by Fusing multiple feature representation) to predict 6mA sites from the Rosaceae genomes, especially in *Rosa chinensis* and *Fragaria vesca*.

Keywords DNA 6 mA · Sequence analysis · Feature encoding · Machine learning

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11103-020-00988-y>) contains supplementary material, which is available to authorized users.

✉ Hiroyuki Kurata
kurata@bio.kyutech.ac.jp

¹ Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

² Japan Society for the Promotion of Science, 5-3-1 Kojimachi, Chiyoda-ku, Tokyo 102-0083, Japan

³ Department of Physiology, Ajou University School of Medicine, Suwon 443380, Korea

⁴ Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

⁵ Biomedical Informatics R&D Center, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Introduction

DNA N⁶ methyladenine (6 mA) regulates various biological functions including genomic imprinting, cell developmental, and chromosome solidity and magnifies genomic diversity in both prokaryotes and eukaryotes (Xiong et al. 2019; Zhang et al. 2018). The 6 mA plays an important role to examine the host DNA and defenses the host genome via the several modification systems (Du et al. 2019), which is evenly circulated across the genome (Liu et al. 2018b; O’Brown and Greer 2016). However, the biological functions and epigenetic modifications of 6 mA still remain unclear. Genomic 6 mA distributions are essential for revealing potential biological functions of DNA. Recently different experimental methods have been accomplished to identify 6 mA, including liquid chromatography coupled with real-time sequencing for single-molecule, and methyladenine—precise PCR (McIntyre et al. 2019; Zhang et al. 2015), but these methods were time-consuming, laborious, and expensive. The rapid development in machine learning (ML)-based algorithms have driven computational chemistry to an unprecedented revolution with the explosive growth of biological sequences in the next generation sequencing era (Chen et al. 2015; Chou 2019; Liu et al. 2016; Sun et al. 2020). Therefore, ML-based methods can be used as an alternative to experimental efforts.

Nowadays, several species-specific ML-based approaches have been established for the identification of 6 mA sites, including rice (Basith et al. 2019; Chen et al. 2019a; Huang et al. 2020; Yu and Dai 2019), and *Mus musculus* (Feng et al. 2019). Although several computational methods have been proposed for 6 mA prediction in some species (Qianfei Huang et al. 2020; Wang and Yan 2018), none of them were developed to specifically identify 6mAs in the Rosaceae genomes. Particularly, the existing prediction models are not

suitable to identify 6 mA in the Rosaceae genome because the existing algorithms are species-specific. Thus, a novel predictor is desired to be established to identify 6 mA sites in the Rosaceae genome.

To the best of our knowledge, we first propose a computation model named i6mA-Fuse (Identification of N6-MethylAdenine sites by Fusing multiple feature representation) to predict 6 mA sites from the Rosaceae genomes, especially in *Rosa chinensis* and *Fragaria vesca*. An overall framework is depicted in Fig. 1. Firstly, the five feature vectors were respectively generated by the five encoding schemes of the k-mer composition (Kmer), k-space spectral nucleotide composition (KSNC), mononucleotide binary encoding (MBE), dinucleotide binary encoding (DBE), and electron–ion interaction pseudopotentials (EIIP). Subsequently, a random forest (RF) classifier was used to build the five, single encoding-employing models. Finally, the predicted probability scores of an appropriate encoding-based models were combined through a linear regression to make a final prediction. As far as we know, the i6mA-Fuse is the first computational predictor of 6mAs within the Rosaceae genomes.

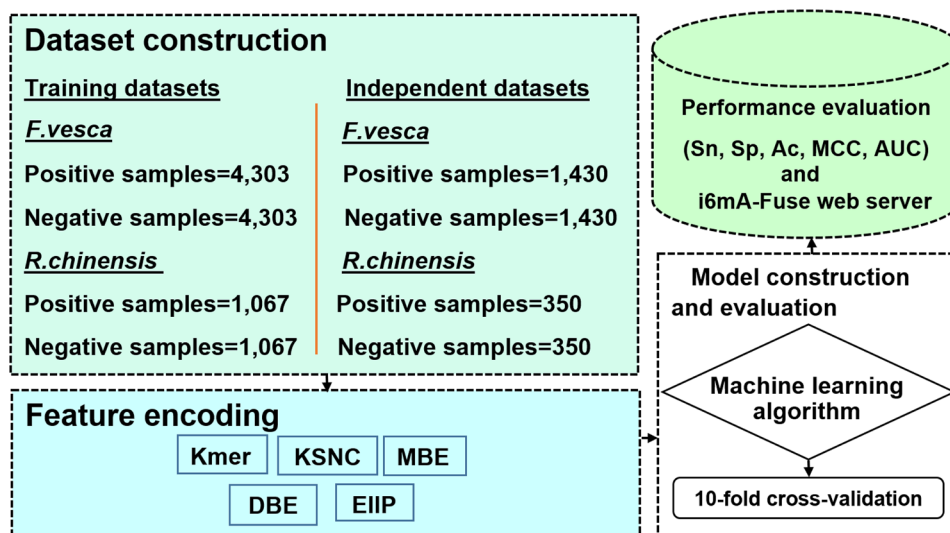
Materials and methods

An outline of i6mA-Fuse is presented in Fig. 1. Four key phases are discussed as follows: (i) datasets construction, (ii) feature extraction, (iii) probability scores calculation, and (iv) final model construction.

Datasets construction

A high quality dataset could guarantee the reliability and robustness of the proposed model (Xu et al. 2019). In this

Fig. 1 An overall framework of i6mA-Fuse. It involves three steps: (i) dataset construction based on MDR database; (ii) employing five different encoding schemes for converting nucleotides into numerical feature vectors; and (iii) model construction and evaluation using cross-validation. Subsequently, a webserver was constructed based on the proposed model (i6mA-Fuse), where it predicts putative 6 mA sites from the submitted query sequences



study, the positive samples (6mAs) were extracted from the reliable MDR database for *F. vesca* and *R. chinensis* (Liu et al. 2019). Each sample had a length of 41 base pairs with adenine nucleotide (“A”) at the center, whose modification score was ≥ 20 . To avoid any similarity bias, we utilized the CD-HIT program (Fu et al. 2012) to exclude highly similar samples by setting a cutoff threshold of 0.7. After such a screening procedure, we obtained the positive samples of 5733 and 1417 for *R. chinensis* and *F. vesca*, respectively. Meanwhile, the standard process was applied to collect negative samples as described in the previous studies (Basith et al. 2019; Lv et al. 2019a). Herein, we obtained strict and objective datasets consisting of 5733 positive and negative samples for *R. chinensis*, and 1417 positive and negative samples for *F. vesca*.

To further validate the predictive ability of the proposed model, the aforementioned datasets were randomly divided into the training and independent datasets at a 3 to 1 ratio. Finally, we obtained 4303/1067 positive and negative samples for *F. vesca/R. chinensis* as training datasets, while 1430/350 positive and negative samples for *F. vesca/R. chinensis* were treated as independent datasets. To confirm the reproducibility of models, all the curated dataset used in this study are available at <https://kurata14.bio.kyutech.ac.jp/i6mA-Fuse/help.php>.

Feature extraction

One of the vital procedures is to express DNA sequences with an operative mathematical expression that can accurately reproduce the intrinsic correlation with the anticipated objective (Yang et al. 2019a). In this study, the five encoding schemes consisting of Kmer, KSNC, MBE, DBE, and EIIP were used for constructing the i6mA-Fuse predictor. The Kmer scheme encodes the occurrence of nucleotide frequency in a DNA sequence (Liu et al. 2018a; Manavalan et al. 2018c). The DNA sequence is expressed by Kmer as: $F = f_1, f_2, f_3, \dots, f_L$, where L is the positive/negative dataset length. The F_i is a nucleotide of A, C, G, and T. Therefore, mono-, di-, tri-, and tetra-nucleotides were encoded and combined to form a 340 ($4^1 + 4^2 + 4^3 + 4^4$) dimensional feature vector. The KSNC scheme encodes the DNA nucleobase information of the curated samples using the frequency wise pair similarity search (Charoenkwan et al. 2013; Zhou et al. 2016). A space of nucleobase frequency pairs is encoded and normalized as

$$\text{Frequencypair} = \frac{N(nf_i)}{w - S - 1} \quad (1)$$

where $N(nf_i)$ is the sum of nf_i inside DNA samples with w length of sample and S is the space between two nucleotides. The KSNC generates a $4 \times 4 \times (S_{max} + 1)$ dimensional

vector for a sequence, where S_{max} was equal to 3. The MBE scheme exactly encodes the nucleotide at each position as a binary vector, where the A, G, T, and C are encoded as (1,0,0,0), (0,0,1,0), (0,1,0,0), and (0,0,0,1), respectively. The MBE encodes the sequence with length of w , a 164-dimensional vector was generated. The DPE scheme encodes 16 potential di-nucleotide as 0 or 1 as 4-dimensional vector (Manavalan et al. 2019a). For example, AA, AC, AT, and GG were coded to (0,0,0,0), (1,1,1,1), (0,0,0,1), and (0,0,1,0), respectively. Using the DPE, a DNA sample is transformed to a 160-dimension feature. The EIIP scheme expresses the electron-ion energies beside with the curated sequences, which is extensively used in bioinformatics research (Basith et al. 2019; Jia et al. 2018). The EIIP indexes of {A, C, G, T} were set to {0.1260, 0.1340, 0.0806, 0.1335} that generates a w -dimensional feature vector for a sequence.

Machine learning algorithms

This study utilized an ensemble method named RF model to develop i6mA-Fuse predictor (Liaw and Wiener 2002; Schaduangrat et al. 2019; Shoombuatong et al. 2019; Su et al. 2019; Win et al. 2017). Typically, when training data of size T with Q features is given, RF builds Q subsets of the data by the bootstrap sampling, and then randomly assigns Q features to each node to optimize the trees based on the by Gini impurity. We used ‘randomForest’ implemented in R (<https://cran.r-project>) with a default cut-off tree number of 1000 to evaluate the optimum performance. This package has been successfully applied to many protein and peptide prediction problems (Hasan et al. 2019b; Lv et al. 2019b; Manavalan et al. 2018b; Zhou et al. 2016). In order to prove the effectiveness of the proposed methods, we compared our RF predictor with the well-known five ML algorithms, i.e. SVM, Adaboost (AB), Naïve Bayes (NB), artificial neural network (ANN), k-nearest neighbor (KNN). In this study, the ANN and NB models (Frank et al. 2004) were implemented from the WEKA software, while the KNN model was developed by using an in-house developed PERL language. For the SVM model, we used the SVM^{light} package with default parameters (Hasan et al. 2019c; Khatun et al. 2019b).

Fusion model

To improve the prediction performance, the RF probability scores estimated by the Kmer, KSNC, MBE, DPE, and EIIP encoding schemes were linearly combined using the following formula:

$$\begin{aligned} \text{Combined} = & w_1 * \text{Kmer} + w_2 * \text{KSNC} + w_3 * \text{MBE} \\ & + w_4 * \text{DBE} + w_5 * \text{EIIP} \end{aligned} \quad (2)$$

where w_1, w_2, w_3, w_4, w_5 are the weight values exhibiting the contribution of each encoding, where the summation of w_1, w_2, w_3, w_4, w_5 is 1. Herein, the linear fusion models of the five RF scores estimated by using the five encoding schemes are referred as Combined model. To enhance the predictive performance of the proposed model, each weight coefficient was adjusted in the range of 0–1 with an interval of 0.05 using a grid-search strategy.

Hybrid model

We investigated the effect of the hybrid feature (H) on 6 mA site prediction. The five encoding feature vectors (F) of Kmer, KSNC, MBE, DPE, and EIIP were combined as follows:

$$H = (F(\text{Kmer}), F(\text{KSNC}), F(\text{MBE}), F(\text{DPE}), F(\text{EIIP})) \quad (3)$$

where H is the sequential combination of five different feature vectors with 1406 dimensions.

Meta-predictor

We generated a meta-classifier to check its potential in 6 mA site prediction. In brief, the probability scores of 30 prediction models (6 encodings \times 5 ML classifier) were considered as a new feature vector as follows:

$$P_{met} = (P(M(1), E(1)), \dots, P(M(i), E(j)), \dots, P(M(s), E(t))) \quad (4)$$

where P_{met} is the new feature vector, $P(M(s), E(j))$ the expected probability by each ML of $M(s)$ with encoding scheme $E(j)$, i the index of the ML, j the index of the encoding scheme, s the total of ML classifiers, and t the total of encodings.

Performance assessment metrics

To assess the performances of i6mA-Fuse, we used four standard measurements consisting of accuracy (Ac), sensitivity (Sn), specificity (Sp), and the Matthews correlation coefficient (MCC) (Basith et al. 2020; Ding et al. 2016; Yang et al. 2019b):

$$Ac = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (5)$$

$$Sn = \frac{TP}{(TP + FN)} \quad (6)$$

$$Sp = \frac{TN}{(TN + FP)} \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

where TP and TN describe the number of positive samples correctly predicted and the number of negative samples correctly identified, respectively. Meanwhile, FN and FP indicate the number of positive samples falsely identified as negative ones and the number of negative samples falsely identified as positive ones, respectively. Furthermore, in order to assess the prediction performance of algorithms using threshold-independent parameters, AUC values were calculated by using the ROC curve.

Results and discussion

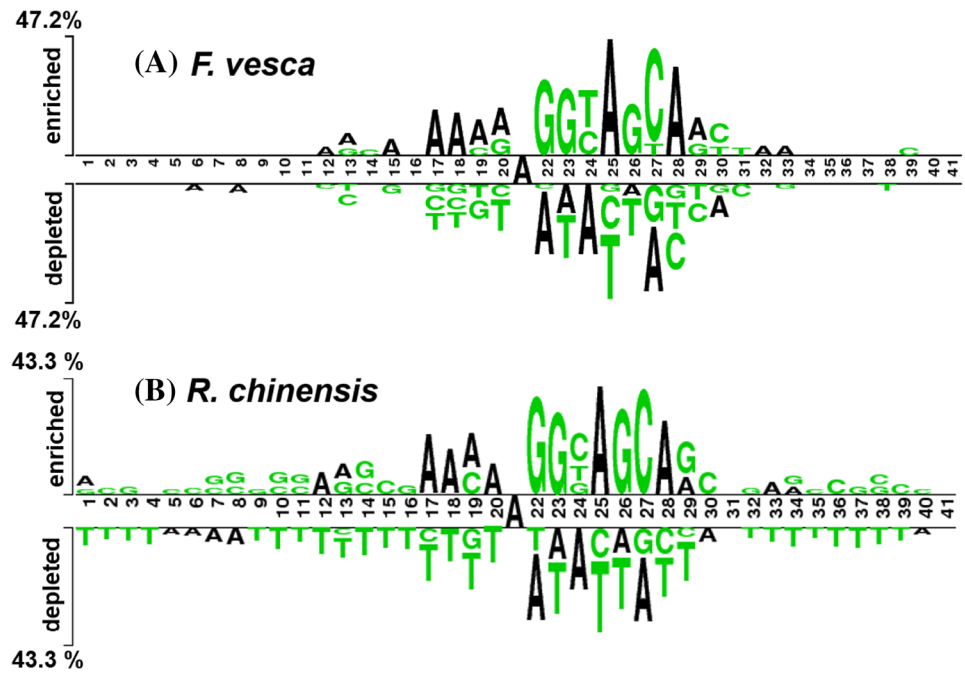
Nucleotide preference of *F. vesca* and *R. chinensis*

The position-specific preferences of nucleotide compositions were analyzed by the two-sample logo software (Vacic et al. 2006) as depicted in Fig. 2. Figure 2a, b display the nucleotide preferences on the DNA sequences having a length of 41 base pairs and the A base at the center for *F. vesca* and *R. chinensis*, respectively. We examined the DNA preferences of the sequences adjacent the A bases of *F. vesca* and *R. chinensis*, while the enriched nucleotides indicate a statistical significance at a level of $p < 0.05$ (two-sample t -test). In case of *F. vesca*, Figure 2a shows that the A base is enriched at positions 12, 13, 15, 17–20, 25, 28, 29, 32, and 33, while the G base is more enriched at positions 13, 20, 22, 23, 26, and 29 than other nucleotides. The T base was significantly depleted at position 13, 17–20, 23, 25, 26, 28, 29, and 38. In case of *R. chinensis*, the G base was enriched at positions 1, 3, 7–11, 13, 14, 16, 22–24, 26, 29, 32, 34, 37, and 38, while the C base was more enriched at positions 2, 5–8, 10, 11, 14, 15, 19, 24, 27, 30, 35, 36, and 38–40 than other nucleotides. As seen in Fig. 2b, the T base was significantly depleted at position 1–4, 9–20, 22, 23, 25, 26, 28, 29, and 32–39. Enrichment and depletion of nucleotides at a specific position might be significant information for discriminating positives from and negatives on both the *F. vesca* and *R. chinensis* samples. Thus, in the present study, the mentioned significant position-specific preferences of nucleotide compositions are used as input features to develop the i6mA-Fuse.

Performance comparison by cross-validation test

We carried out a series of comparative simulations using RF models with the five feature vectors of Kmer, KSNC, MBE, DBE, and EIIP encodings, and evaluated their

Fig. 2 Nucleotide preferences of the surrounding positive samples compared to negative samples. **a** *F. vesca*. **b** *R. chinensis*. The level of Y-axis is dissimilar due to the different datasets. Only nucleotides that are significantly enriched or depleted (*t*-test, $P < 0.05$) nearby the centered positive and negative samples are shown



performances on the training dataset by tenfold cross-validation. The cross-validation results are listed in Table 1 and Fig. 3. The previous studies suggested several ways to integrate multiple prediction models for improving the performances, including meta-predictor (Boopathi et al. 2019; Manavalan et al. 2019b, c, d), ensemble approach, and linear regression (Hasan et al. 2019a, d; Khatun et al. 2019a). Herein, the i6mA-Fuse linearly combined the five probability scores evaluated by the five, single encoding-employing RF models.

For *F. vesca*, MBE encoding achieved the highest prediction results with Ac = 0.925, Sn = 0.891, Sp = 0.958, MCC = 0.971, and AUC = 0.858. Meanwhile, the second highest prediction result was obtained by EIIP encoding, which gave Ac = 0.915, Sn = 0.879, Sp = 0.950, MCC = 0.839, and AUC = 0.963. These two encoding schemes yielded Ac values of 0.915–0.925, which were 11.6–24.6% higher than Ac values of the other three encodings. To construct the i6mA-Fuse, the optimal weight coefficients of Kmer, KSNC, MBE, DBE, and EIIP were 0.00, 0.00, 0.75, 0.00, and 0.25. While the three encodings of

Table 1 Cross-validation results of the proposed predictors and other five encodings

Genome	Method ^a	MCC	Ac	Sn	Sp	AUC	^b P-value
<i>F. vesca</i>	Kmer	0.529	0.703	0.542	0.864	0.846	<0.001
	KSNC	0.398	0.674	0.492	0.855	0.804	<0.001
	MBE	0.858	0.925	0.891	0.958	0.971	0.293
	DBE	0.648	0.804	0.704	0.904	0.926	<0.001
	EIIP	0.839	0.915	0.879	0.950	0.963	0.068
	i6mA-Fuse	0.873	0.934	0.908	0.957	0.981	–
<i>R. chinensis</i>	Kmer	0.559	0.734	0.548	0.921	0.877	<0.001
	KSNC	0.477	0.721	0.528	0.912	0.844	<0.001
	MBE	0.839	0.912	0.872	0.952	0.956	0.244
	DBE	0.608	0.793	0.684	0.902	0.902	<0.001
	EIIP	0.816	0.900	0.866	0.933	0.945	0.028
	^a i6mA-Fuse	0.851	0.916	0.881	0.950	0.968	–

^aThe i6mA-Fuse is constructed by the linear combination of the probability scores derived from RF models in conjunction with Kmer, KSNC, MBE, DBE, and EIIP encodings and their weight coefficients (*F. vesca* and *R. chinensis*) are (0.00 and 0.15), (0.00, 0.00), (0.75, 0.65), (0.00, 0.00) and (0.25, 0.20), respectively

^bP-value was calculated with the AUC values and two-sample *t*-test

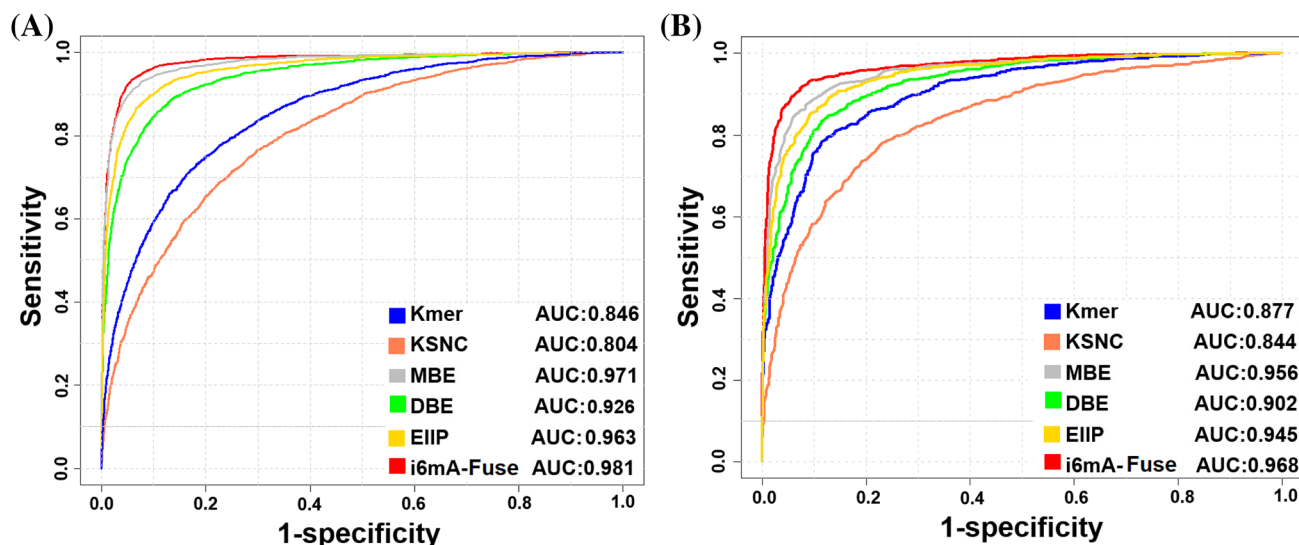


Fig. 3 ROC curves of i6mA-Fuse and the single encoding-employing models as evaluated by 10-fold cross-validation. **a** *F. vesca*. **b** *R. chinensis*

Kmer, KSNC, and DBE did not contribute to any prediction performance, the MBE and EIIP contributed to 75% and 25% of the total prediction. The ROC curves of the i6mA-Fuse and the single encoding-employing models, evaluated by 10-fold cross-validation, are presented in Fig. 3a. The i6mA-Fuse yielded AUC=0.981, MCC=0.873, Ac=0.934, Sn=0.908, and Sp=0.957, surpassing all the single-employing models (Table 1). According to a *P*-value of 0.05, the i6mA-Fuse significantly outperformed the three models employing the single encoding of Kmer, KSNC, and DBE.

For *R. chinensis*, MBE achieved the best performance with an AUC value of 0.956 and Ac of 0.912 for all the single encoding-employing models. Meanwhile, the second highest prediction with an AUC of 0.945 and Ac of 0.900 was obtained by using EIIP encoding. To construct the i6mA-Fuse, we combined the RF scores with optimal weight coefficients of Kmer, KSNC, MBE, DBE, and EIIP of 0.15, 0.00, 0.65, 0.00, and 0.20, respectively, indicating that Kmer, MBE, and EIIP contribute to 15%, 65%, and 20% of the total prediction, respectively. The i6mA-Fuse yielded a peak AUC value of 0.968, while indicating MCC=0.851, Ac=0.916, Sn=0.881, and Sp=0.950 (Table 1). The i6mA-Fuse significantly outperformed the four models employing the single encoding of Kmer, KSNC, DBE, and EIIP (*P*-value of 0.05 by two-sample *t*-test).

Performance comparison among different ML algorithms by cross-validation test

To validate the effectiveness of the RF classifier in the i6mA-Fuse, we compared its performance with the five ML classifiers of SVM, AB, NB, ANN, and KNN on the training dataset. To make a fair comparison, we implemented the

five ML classifiers in the same manners as the RF classifier. Figure S1 shows that the RF model provided better results than the other ML classifiers, while the prediction results of the SVM model were comparable to RF model in both two genomes. In *F. vesca*, the i6mA-Fuse achieved ~2–6% higher AUCs than any other ML-based combined models (Figure S1A). Meanwhile, the AUC values of the i6mA-Fuse were ~3–6% higher than those of the other combined models in *R. chinensis* (Figure S1B), thus demonstrating the superiority of RF.

Performance comparison of i6mA-Fuse with hybrid model and meta-predictor

We compared the linear regression model, employed by the i6mA-Fuse, with the two different models namely a hybrid model and meta-predictor. First, we concatenated the five feature encoding vectors of Kmer, KSNC, MBE, DPE, and EIIP and obtained 1406-dimensional feature vector. We generated the hybrid model for both species (*F. vesca* and *R. chinensis*), inputted these features to six different classifiers (RF, SVM, AB, NB, ANN, and KNN), and evaluated their performances by tenfold CV on the training datasets. In case of *F. vesca*, the hybrid model of RF, SVM, AB, NB, ANN, and KNN achieved AUC values of 0.978, 0.963, 0.957, 0.944, 0.919 and 0.933, respectively (Figure S2). Similarly, for *R. chinensis*, the hybrid model of RF, SVM, AB, NB, ANN, and KNN algorithms achieved AUCs of 0.958, 0.962, 0.927, 0.942, 0.918 and 0.922, respectively (Figure S2). Furthermore, we constructed the meta-predictor as described elsewhere (Manavalan et al. 2018a, 2019d; Wei et al. 2019). In *F. vesca* the i6mA-Fuse achieved higher AUCs than the hybrid model and meta-predictor (Figures S2 & S3). In *R.*

chinensis, the i6mA-Fuse showed ~1.0–8.0% higher AUCs than the hybrid model and meta-predictor (Figures S2 & 3).

Performance of i6mA-Fuse by independent test

The i6mA-Fuse was evaluated by independent tests for two genomes. We compared the prediction performance of the i6mA-Fuse with that of the five, single encoding-employing models (Kmer, KSNC, MBE, DBE, and EIIP) by using an independent dataset of *F. vesca*. MBE encoding gave higher performance than any other single-encoding employing models for two genomes, as shown in Fig. 4a, b. Moreover, the ROC curves displayed that the i6mA-Fuse achieved better than all the single encoding-employing methods. The i6mA-Fuse achieved outstanding performances (MCC=0.858, Ac=0.929, Sn=0.915, Sp=0.943, and AUC=0.978) and (MCC=0.869, Ac=0.937, Sn=0.928, Sp=0.948, and AUC=0.982) for *R. chinensis* and *F. vesca*, respectively, on the independent sets (Table 2).

Validation of i6mA-Fuse with other species datasets

To further examine the generalization of i6mA-Fuse, the proposed i6mA-Fuse was applied to identify 6 mA sites in other species, i.e. rice and mouse genomes. We collected the rice genome dataset from SDM6A (Basith et al. 2019), which contains 221 positive and 221 negative samples, and the mouse genome dataset from iDNA6mA-PseKNC (Feng et al. 2019), where we randomly selected 200 positive and 200 negative samples. The prediction performances of AUC,

MCC, Sp, Sn, and Ac were shown in Table S1. The *R. chinensis*- and *F. vesca*-specific i6mA-Fuse yielded AUC values of 0.870 and 0.928 for the rice genome, respectively. For the mouse genome, they provided AUC values of 0.748 and 0.769, respectively. The i6mA-Fuse can be applicable to the rice genome, especially the *F. vesca*-specific i6mA-Fuse (Table S1). It suggests that the sequences surrounding 6 mA sites of Rosaceae genomes have common characteristic patterns to rice genomes.

Conclusions

The accurate prediction of 6 mA sites is one of the challenging tasks in bioinformatics. Because the experimental approaches are time-consuming and costly, it is desirable to develop a computational model for rapidly and accurately identifying 6 mA sites. Although several computational methods have been proposed in some species (Basith et al. 2019; Chen et al. 2019a; Feng et al. 2019; Lv et al. 2019a; Yu and Dai 2019), none of them were developed to specifically identify 6 mA sites in the Rosaceae genomes. In this study, we developed the first species-specific predictor named i6mA-Fuse for identifying 6 mA sites of the Rosaceae genomes, especially in *R. chinensis* and *F. vesca*. We constructed the random forest (RF) models with the five encoding schemes of Kmer, KSNC, MBE, DPE, and EIIP, and then combined the predicted probability scores of the five models through a linear regression. The resultant species-specific i6mA-Fuse achieved remarkably high

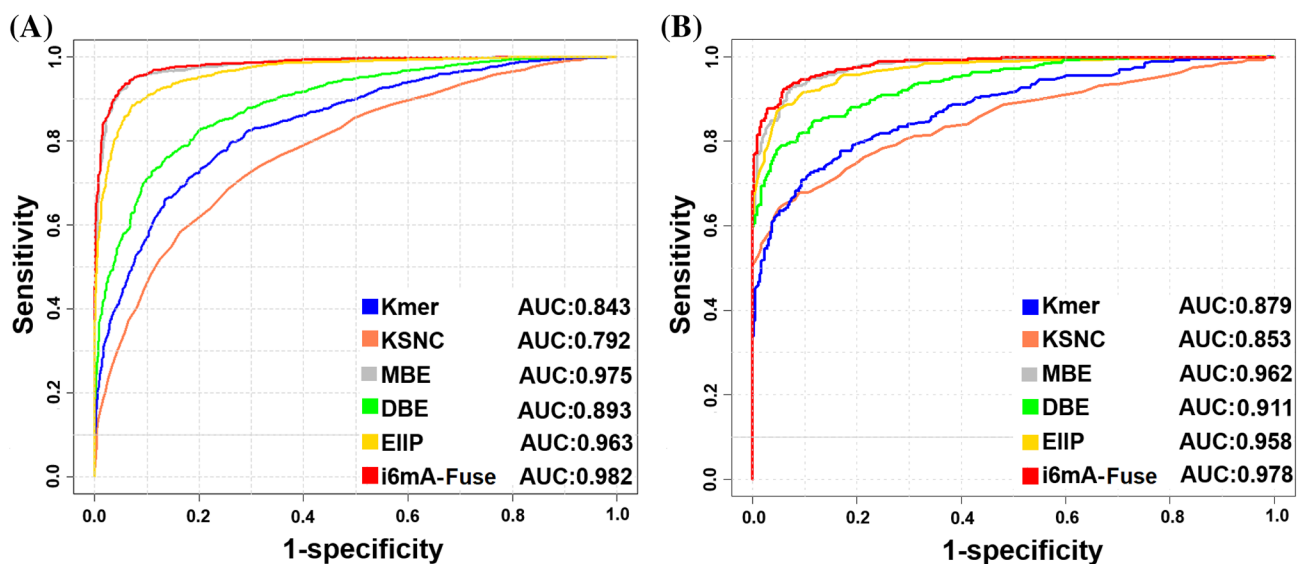


Fig. 4 ROC curves of i6mA-Fuse and the single encoding-employing models as evaluated by independent test. **a** *F. vesca*. **b** *R. chinensis*

Table 2 Independent test of the proposed predictors and other five encodings

Genome	Method	MCC	Ac	Sn	Sp	AUC
<i>F. vesca</i>	Kmer	0.532	0.761	0.655	0.857	0.843
	KSNC	0.434	0.708	0.582	0.836	0.792
	MBE	0.862	0.932	0.917	0.947	0.975
	DBE	0.643	0.810	0.704	0.893	0.893
	EIIP	0.842	0.916	0.878	0.948	0.963
	i6mA-Fuse	0.869	0.937	0.928	0.948	0.982
<i>R. chinensis</i>	Kmer	0.613	0.737	0.732	0.874	0.879
	KSNC	0.565	0.782	0.687	0.877	0.853
	MBE	0.835	0.918	0.901	0.935	0.962
	DBE	0.644	0.803	0.713	0.893	0.911
	EIIP	0.819	0.909	0.884	0.934	0.958
	^a i6mA-Fuse	0.858	0.929	0.915	0.943	0.978

^aThe i6mA-Fuse is constructed by the linear combination of the probability scores derived from RF models in conjunction with Kmer, KSNC, MBE, DBE, and EIIP encodings and their weight coefficients (*F. vesca* and *R. chinensis*) are (0.00 and 0.15), (0.00, 0.00), (0.75, 0.65), (0.00, 0.00) and (0.25, 0.20), respectively

performances with AUCs of 0.982 and 0.978 and with MCCs of 0.869 and 0.858 on the independent datasets of *Rosa chinensis* and *Fragaria vesca*, respectively. In the *F. vesca*-specific i6mA-Fuse, the MBE and EIIP contributed to 75% and 25% of the total prediction; in the *R. chinensis*-specific i6mA-Fuse, Kmer, MBE, and EIIP contribute to 15%, 65%, and 20% of the total prediction. Interestingly, the i6mA-Fuse can be applicable to rice genome. To show the superiority of the linear regression, we compared it with the two combination methods of the hybrid model and meta-classifier. To further improve the prediction performance, we may use recently proposed integration methods (Chen et al. 2019b; Li et al. 2019; Zhang et al. 2019) and various modes of Chou's pseudo information (Chen et al. 2016; Chou 2011; Liu et al. 2015). To assist high-throughput identification for DNA 6 mA sites, the i6mA-Fuse is publicly accessible at <https://kurata14.bio.kyutech.ac.jp/i6mA-Fuse/>.

Acknowledgements This work is supported by the Grant-in-Aid for JSPS Research Fellow (19F19377) from Japan Society for the Promotion of Science (JSPS), partially supported from Japan Society for the Promotion of Science by Grant-in-Aid for Scientific Research (B) (19H04208) and by the developing key technologies for discovering and manufacturing pharmaceuticals used for next-generation treatments and diagnoses both from the Ministry of Economy, Trade and Industry, Japan (METI) and from Japan Agency for Medical Research and Development (AMED).

Author contributions MMH and HK conceived the project. MMH and KMS collected and analyzed the datasets. MMH drafted the manuscript. HK, MMH, MB, SW and KMS thoroughly revised the manuscript. All authors approved and read the final manuscript.

References

- Basith S, Manavalan B, Shin TH, Lee G (2019) SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol Ther Nucleic Acids* 18:131–141. <https://doi.org/10.1016/j.omtn.2019.08.011>
- Basith S, Manavalan B, Shin TH, Lee G (2020) Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med Res Rev*. <https://doi.org/10.1002/med.21658>
- Boopathi V, Subramaniam S, Malik A, Lee G, Manavalan B, Yang DC (2019) mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int J Mol Sci*. <https://doi.org/10.3390/ijms20081964>
- Charoenkwan P, Shoombuatong W, Lee HC, Chaijaruanich J, Huang HL, Ho SY (2013) SCMCRYST: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. *PLoS ONE* 8:e72368. <https://doi.org/10.1371/journal.pone.0072368>
- Chen W, Lin H, Chou KC (2015) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol BioSyst* 11:2620–2634. <https://doi.org/10.1039/c5mb00155b>
- Chen W, Lv H, Nie F, Lin H (2019a) i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 35:2796–2800. <https://doi.org/10.1093/bioinformatics/btz015>
- Chen W, Tang H, Ye J, Lin H, Chou KC (2016) iRNA-PseU: identifying RNA pseudouridine sites. *Mol Ther Nucleic Acids* 5:e332. <https://doi.org/10.1038/mtna.2016.37>
- Chen Z et al (2019b) Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. *Brief Bioinform*. <https://doi.org/10.1093/bib/bbz112>
- Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. *J Theor Biol* 273:236–247. <https://doi.org/10.1016/j.jtbi.2010.12.024>
- Chou KC (2019) Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr Med Chem*. <https://doi.org/10.2174/092986732666190507082559>
- Ding H, Yang W, Tang H, Feng PM, Huang J, Chen W, Lin H (2016) PHYPred: a tool for identifying bacteriophage enzymes and

- hydrolases. *Virol Sin* 31:350–352. <https://doi.org/10.1007/s12250-016-3740-6>
- Du K et al (2019) Epigenetically modified N(6)-methyladenine inhibits DNA replication by human DNA polymerase ϵ . *DNA Repair* 78:81–90. <https://doi.org/10.1016/j.dnarep.2019.03.015>
- Feng P, Yang H, Ding H, Lin H, Chen W, Chou KC (2019) iDNA6mA-PseKNC: identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* 111:96–102. <https://doi.org/10.1016/j.ygeno.2018.01.005>
- Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. *Bioinformatics* 20:2479–2481. <https://doi.org/10.1093/bioinformatics/bth261>
- Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Hasan MM, Khatun MS, Kurata H (2019a) Large-scale assessment of bioinformatics tools for lysine succinylation sites. *Cells*. <https://doi.org/10.3390/cells8020095>
- Hasan MM, Manavalan B, Khatun MS, Kurata H (2019b) i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. *Int J Biol Macromol*. <https://doi.org/10.1016/j.ijbiomac.2019.12.009>
- Hasan MM, Manavalan B, Khatun MS, Kurata H (2019c) Prediction of S-nitrosylation sites by integrating support vector machines and random forest. *Mol Omics* 15:451–458. <https://doi.org/10.1039/c9mo00098d>
- Hasan MM, Rashid MM, Khatun MS, Kurata H (2019d) Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. *Sci Rep* 9:8258. <https://doi.org/10.1038/s41598-019-44548-x>
- Huang Q, Zhang J, Wei L, Guo F, Zou Q (2020) 6mA-RicePred: a method for identifying DNA N(6)-methyladenine sites in the rice genome based on feature fusion. *Front Plant Sci* 11:4. <https://doi.org/10.3389/fpls.2020.00004>
- Jia C, Yang Q, Zou Q (2018) NucPosPred: predicting species-specific genomic nucleosome positioning via four different modes of general PseKNC. *J Theor Biol* 450:15–21. <https://doi.org/10.1016/j.jtbi.2018.04.025>
- Khatun MS, Hasan MM, Kurata H (2019a) PreAIP: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. *Front Genet* 10:129. <https://doi.org/10.3389/fgene.2019.00129>
- Khatun S, Hasan M, Kurata H (2019b) Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. *FEBS Lett* 593:3029–3039. <https://doi.org/10.1002/1873-3468.13536>
- Li F et al (2019) DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btz721>
- Liaw A, Wiener M (2002) Classification and regression by random forest. *R News* 2:18–22
- Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA RNA, and protein sequences. *Nucleic Acids Res* 43:W65–71. <https://doi.org/10.1093/nar/gkv458>
- Liu B, Fang L, Long R, Lan X, Chou KC (2016) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* 32:362–369. <https://doi.org/10.1093/bioinformatics/btv604>
- Liu B, Li K, Huang DS, Chou KC (2018a) iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* 34:3835–3842. <https://doi.org/10.1093/bioinformatics/bty458>
- Liu X, Lai W, Zhang N, Wang H (2018b) Predominance of N(6)-methyladenine-specific DNA fragments enriched by multiple immunoprecipitation. *Anal Chem* 90:5546–5551. <https://doi.org/10.1021/acs.analchem.8b01087>
- Liu ZY et al (2019) MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. *Hortic Res* 6:78. <https://doi.org/10.1038/s41438-019-0160-4>
- Lv H et al (2019a) iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Front Genet* 10:793. <https://doi.org/10.3389/fgene.2019.00793>
- Lv Z, Jin S, Ding H, Zou Q (2019b) A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features. *Front Bioeng Biotechnol* 7:215. <https://doi.org/10.3389/fbioe.2019.00215>
- Manavalan B, Shin TH, Lee G (2018) DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 9:1944–1956. <https://doi.org/10.18632/oncotarget.23099>
- Manavalan B, Shin TH, Kim MO, Lee G (2018) AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. *Front Pharmacol* 9:276. <https://doi.org/10.3389/fphar.2018.00276>
- Manavalan B, Basith S, Shin TH, Wei L, Lee G (2018a) mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/bty1047>
- Manavalan B, Basith S, Shin TH, Lee DY, Wei L, Lee G (2019a) 4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse. *Genome Cells*. <https://doi.org/10.3390/cells8111332>
- Manavalan B, Basith S, Shin TH, Wei L, Lee G (2019b) AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. *Comput Struct Biotechnol J* 17:972–981. <https://doi.org/10.1016/j.csbj.2019.06.024>
- Manavalan B, Basith S, Shin TH, Wei L, Lee G (2019c) mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35:2757–2765. <https://doi.org/10.1093/bioinformatics/bty1047>
- Manavalan B, Basith S, Shin TH, Wei L, Lee G (2019d) Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol Ther Nucleic Acids* 16:733–744. <https://doi.org/10.1016/j.omtn.2019.04.019>
- McIntyre ABR, Alexander N, Grigorev K, Bezdán D, Sichtig H, Chiu CY, Mason CE (2019) Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nat Commun* 10:579. <https://doi.org/10.1038/s41467-019-08289-9>
- O’Brown ZK, Greer EL (2016) N6-methyladenine: a conserved and dynamic DNA mark. *Adv Exp Med Biol* 945:213–246. https://doi.org/10.1007/978-3-319-43624-1_10
- Qianfei Huang F, Zhang Z, Wei L, Guo F, Zou Q (2020) 6mA-RicePred: a method for identifying DNA N6-methyladenine sites in the rice genome based on feature fusion. *Front Plant Sci*. <https://doi.org/10.3389/fpls.2020.00004>
- Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombua-tong W (2019) Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int J Mol Sci*. <https://doi.org/10.3390/ijms20225743>
- Shoombua-tong W, Schaduangrat N, Pratiwi R, Nantasenamat C (2019) THPeP: a machine learning-based approach for predicting tumor homing peptides. *Comput Biol Chem* 80:441–451. <https://doi.org/10.1016/j.compbiolchem.2019.05.008>
- Su R, Hu J, Zou Q, Manavalan B, Wei L (2019) Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief Bioinform*. <https://doi.org/10.1093/bib/bby124>

- Sun S, Wang C, Ding H, Zou Q (2020) Machine learning and its applications in plant molecular studies. *Brief Funct Genom* 19:40–48. <https://doi.org/10.1093/bfpg/elz036>
- Vacic V, Iakoucheva LM, Radivojac P (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 22:1536–1537. <https://doi.org/10.1093/bioinformatics/btl151>
- Wang X, Yan R (2018) RFathM6A: a new tool for predicting m(6)A sites in *Arabidopsis thaliana*. *Plant Mol Biol* 96:327–337. <https://doi.org/10.1007/s11103-018-0698-9>
- Wei L, Su R, Luan S, Liao Z, Manavalan B, Zou Q, Shi X (2019) Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btz408>
- Win TS, Malik AA, Prachayasittikul V, Wikberg SJE, Nantasenamat C, Shoombuatong W (2017) HemoPred: a web server for predicting the hemolytic activity of peptides. *Future Med Chem* 9:275–291. <https://doi.org/10.4155/fmc-2016-0188>
- Xiong J, Ye TT, Ma CJ, Cheng QY, Yuan BF, Feng YQ (2019) N6-Hydroxymethyladenine: a hydroxylation derivative of N6-methyladenine in genomic DNA of mammals. *Nucleic Acids Res* 47:1268–1277. <https://doi.org/10.1093/nar/gky1218>
- Xu ZC, Feng PM, Yang H, Qiu WR, Chen W, Lin H (2019) iRNAD: a computational tool for identifying D modification sites in RNA sequence. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btz358>
- Yang W, Zhu XJ, Huang J, Ding H, Lin H (2019) A brief survey of machine learning methods in protein sub-Golgi localization. *Curr Bioinform* 14:234–240
- Yang H, Yang W, Dao FY, Lv H, Ding H, Chen W, Lin H (2019) A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*. *Brief Bioinform*. <https://doi.org/10.1093/bib/bbz123>
- Yu H, Dai Z (2019) SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Front Genet* 10:1071. <https://doi.org/10.3389/fgene.2019.01071>
- Zhang G et al (2015) N6-methyladenine DNA modification in *Drosophila*. *Cell* 161:893–906. <https://doi.org/10.1016/j.cell.2015.04.018>
- Zhang Q et al (2018) N(6)-methyladenine DNA methylation in Japonica and Indica rice genomes and its association with gene expression, plant development, and stress responses. *Mol Plant* 11:1492–1508. <https://doi.org/10.1016/j.molp.2018.11.005>
- Zhang Y et al (2019) PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btz629>
- Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q (2016) SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. *Nucleic Acids Res* 44:e91. <https://doi.org/10.1093/nar/gkw104>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.