Empirical Comparison and Analysis of Web-Based DNA $N^4$-Methylcytosine Site Prediction Tools

Balachandran Manavalan,1,2 Md. Mehedi Hasan,2,3 Shaherin Basith,1 Vijayakumar Gosu,4 Tae-Hwan Shin,1 and Gwang Lee1,5

1Department of Physiology, Ajou University School of Medicine, Suwon 16499, Republic of Korea; 2Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan; 3Japan Society for the Promotion of Science, Chiyoda-ku, Tokyo 102-0083, Japan; 4Department of Animal Biotechnology, Jeonbuk National University, Jeonju 54896, Republic of Korea; 5Department of Molecular Science and Technology, Ajou University, Suwon 16499, Republic of Korea

DNA $N^4$-methylcytosine (4mC) is a crucial epigenetic modification involved in various biological processes. Accurate genome-wide identification of these sites is critical for improving our understanding of their biological functions and mechanisms. As experimental methods for 4mC identification are tedious, expensive, and labor-intensive, several machine learning-based approaches have been developed for genome-wide detection of such sites in multiple species. However, the predictions projected by these tools are difficult to quantify and compare. To date, no systematic performance comparison of 4mC tools has been reported. The aim of this study was to compare and critically evaluate 12 publicly available 4mC site prediction tools according to species specificity, based on a huge independent validation dataset. The tools 4mCCNN (Escherichia coli), DNA4mC-LiP (Arabidopsis thaliana), IDNA-MS (Fragaria vesca), DNA4mC-LiP and 4mCCNN (Drosophila melanogaster), and four tools for Caenorhabditis elegans achieved excellent overall performance compared with their counterparts. However, none of the existing methods was suitable for Geoalkalibacter subterraneus, Geobacter pickeringii, and Mus musculus, thereby limiting their practical applicability. Model transferability to five species and non-transferability to three species are also discussed. The presented evaluation will assist researchers in selecting appropriate prediction tools that best suit their purpose and provide useful guidelines for the development of improved 4mC predictors in the future.

INTRODUCTION

DNA methylation is one of the most common epigenetic tools used by the cell to control gene expression. It alters chromatin structure, DNA conformation, DNA-protein interactions, and DNA stability. This modification plays important roles in the regulation of several developmental and pathological processes, such as aging, carcinogenesis, genomic imprinting, repression of transposable elements, and X chromosome inactivation.2,5 Analysis of the differentiating capacity of foreign DNA and host DNA indicates that certain modifications, such as cytosine methylation, may protect the host DNA from enzyme-mediated degradation.7 Cytosine methylation is catalyzed by DNA methyl transferases and is considered a major post-replicative DNA modification in prokaryotes and eukaryotes. The most common types of enzymatic cytosine methylation are 5-methylcytosine and $N^4$-methylcytosine (4mC).8 Unlike the former, which has been extensively studied,8,9,11,12 4mC has not been thoroughly investigated. 4mC formation, predominantly found in prokaryotes, is catalyzed by $N^4$-cytosine-specific DNA methyl transferases, which methylate the amino group at the fourth position of cytosine.13 Similar to 5-methylcytosine, 4mC is an element of a restriction-modification system that protects self-DNA from enzyme-mediated degradation.14 It is also involved in the cell cycle, correction of DNA replication errors, and regulation of DNA replication.15-16

Information regarding genome-wide distribution of 4mC is crucial for deciphering the function of this modification in detail. To date, only a few experimental approaches, such as single-molecule real-time (SMRT) sequencing,17 4mC-Tet-assisted bisulfite sequencing (4mC-TAB-seq),18 and engineered transcription-activator-like effectors,3 have been developed for detecting 4mC sites in the genome. SMRT sequencing is a popular experimental approach that has been successfully implemented in the identification of 4mC modifications. Regardless of the presence or absence of an assembled genome, SMRT sequencing was designed to directly identify 4mC modification sites. However, this approach is expensive and lacks applicability to various species and large-scale genomes. Owing to this limitation, the next-generation sequencing technique 4mC-TAB-seq was designed to identify the genome-wide locations of 4mC motifs. Another group detected specific 4mC sites utilizing engineered transcription-activator-like effectors. Although these experimental methods effectively facilitate the identification of 4mC sites,
they are still too laborious and expensive for genome-wide applicability.

Computational approaches, particularly machine learning (ML) methods, have emerged as effective 4mC site prediction tools in multiple species. IDNA4mC, a pioneer ML-based method that relies on the chemical properties of nucleotides and accumulated nucleotide frequency as features to build a support vector machine (SVM)-based model, was developed in 2017. Two additional methods (4mCPred and 4mcPred-SVM) were developed in 2018; five methods (Meta-4mCpred, 4mcPred-IFL, 4mCCNN, 4mCPred-EL, and i4mCROSE) were reported in 2019. By April 2020, five more methods (iEC4mC-SVM, DNA4mC-LIP, i4mC-ROSE, i4mc-Mouse, and iDNA-MS) were reported. An overview of the existing 4mC prediction methods is provided in Figure 1. Notably, most of these methods were trained on the benchmark datasets constructed by Chen et al., and few were validated based on independent datasets constructed by Manavalan et al. Because of the recent surge in the development of 4mC prediction tools, an unbiased evaluation of these existing web-based 4mC prediction tools using our own constructed validation set, which captures the overall 4mC and non-4mC pattern in the whole-genome of each individual species. While some models achieved excellent overall performance for some species, none was suitable for Geoa. subterraneus, Geob. pickeringii, and M. musculus, limiting their practical applicability. The model transferability and non-transferability were explored, and suggestions for the design and development of new prediction tools were also presented. The presented analysis will facilitate efforts to develop improved tools for the prediction of 4mC sites.

RESULTS

In the study, an unbiased performance evaluation of existing web-based 4mC site prediction tools was performed based on validation dataset analysis. The utilized validation dataset was different from a previously reported and widely used dataset. Specifically, (1) unlike negative samples (non-4mC) analyzed before, in the new dataset, the negative samples were representative of the whole genome of each species, and (2) the new dataset was several-fold larger than the previously reported dataset. Eight species for which at least two prediction models reported are analyzed herein. Seven web servers (i.e., IDNA4mC, 4mCPred_I, 4mCPred_II, 4mcPred-IIFL, 4mcPred-SVM, Meta-4mCpred, 4mCCNN, DNA4mC-LIP, i4mC-ROSE, i4mc-Mouse, 4mCPred-EL, and iDNA-MS) were available as web servers. We carried out an unbiased evaluation of these existing web-based 4mC prediction tools using our own constructed validation set, which captures the overall 4mC and non-4mC pattern in the whole-genome of each individual species. While some models achieved excellent overall performance for some species, none was suitable for Geoa. subterraneus, Geob. pickeringii, and M. musculus, limiting their practical applicability. The model transferability and non-transferability were explored, and suggestions for the design and development of new prediction tools were also presented. The presented analysis will facilitate efforts to develop improved tools for the prediction of 4mC sites.

![Figure 1. Overview of the Current Computational Approaches for 4mC Site Prediction](image-url)
for *F. vesca*, and two web servers (4mCPred-EL and i4mC-Mouse) for *M. musculus*. All parameters were set to default values, as specified on their server. As an exception, the values reported for 4mCCNN were based on personal communication (J. Khanal, Chongbuk National University) because the web server was not active at the time of analysis. Computing the K-nearest neighbor (KNN) feature implemented in 4mcPred-IFL took several hours, leading to time-out errors during sequence submission. Therefore, 4mcPred-IFL was excluded from our evaluation.

**Evaluation of Species-Specific Performance of the Existing 4mC Prediction Tools**

*Geob. pickeringii*

The constructed validation set was 56-fold larger than the training set used to evaluate seven different methods (see Materials and Methods for details of the datasets and methods used). The performances of different models are shown in Figure 2A and Table S1. iDNA4mC achieved the best performance for four out of five metrics tested, i.e., Matthews correlation coefficient (MCC), balanced accuracy (BACC), specificity (Sp), and area under the curve (AUC) (0.118, 0.610, 0.687, and 0.653, respectively). These values were higher than those of the second-best method (4mCPred_I) by 0.046, 3.48%, 16.4%, and 5.5%, respectively. 4mCPred_I performance was similar to that of 4mCPred_I. The remaining four predictors (4mCCNN, 4mcPred-SVM, Meta-4mCpred, and DNA4mC-LIP) had much lower BACC values (below 50.0%), and even the random classifier performed better than these four predictors. To obtain statistically meaningful differences between the top model and other methods, AUC values for any two methods were compared, and the p value was computed based on the two-tailed t-test. At p < 0.01, iDNA4mC significantly outperformed the other methods (Table S1).

Most of the existing method web servers implemented a species-specific prediction model. However, the iDNA4mC web server implemented a single model based on a combined training dataset from six different species. Therefore, the improved performance of iDNA4mC may be associated with the larger training dataset. Although iDNA4mC showed superior performance, it was still far from satisfactory because, with a BACC of 61.0%, it was only slightly better than the random classifier. Recently, Tang et al. evaluated seven methods using a smaller validation dataset (200 4mCs and 200 non-4mCs) and demonstrated a reasonable performance for most tested methods (accuracy [ACC] range, 75.0%–85.0%). However, in our study, such performance was not replicated when using a larger validation set.

**E. coli**

The validation dataset used was 153-fold larger than the training dataset. As shown in Figure 2B and Table S1, 4mCCNN achieved the best performance for four out of five metrics, i.e., MCC, BACC, Sp, and AUC (0.110, 0.793, 0.805, and 0.873, respectively). The values were slightly higher than those of the second to fourth best predictors
(DNA4mC-LIP, 4mCPred_II, and 4mcPred-SVM, accordingly), by 0.013–0.005, 0.5%–1.1%, 0.6%–8.3%, and 0.7%–1.5%, respectively. Of the remaining three methods, Meta-4mCpred and 4mCPred_I showed reasonable performance, with a BACC value of 77.0%, and iDNA4mC showed the worst performance. At p < 0.01, the performance of the best method (4mCCNN) was similar to that of the other methods (DNA4mC-LIP, 4mCPred_I, 4mCPred_II, and DNA4mC-LIP). This method significantly outperformed iDNA4mC. Although all these methods were developed using a smaller training dataset than that used in iDNA4mC (388 4mCs and 388 non-4mCs) and with different approaches, most of them performed reasonably well with respect to the genome-wide detection of 4mC sites.

**C. elegans**

The validation set used was 53-fold larger than the training dataset. As shown in Figure 2C and Table S1, the methods were classified based on performance into two groups, i.e., (1) four methods (4mCCNN, 4mCPred_I, 4mCPred_II, and DNA4mC-LIP), which showed similar performance in terms of MCC, BACC, and AUC values (0.453–0.486, 81.9%–82.5%, and 89.2%–90.4%, respectively), with no single best-performing method; and (2) the remaining three methods (Meta-4mCpred, 4mcPred-SVM, and IDNA4mC), which also achieved reasonable performance, with MCC and BACC values in the ranges of 0.376–0.405 and 76.3%–78.1%, respectively. However, the performance of group 2 methods was lower than that of group 1 methods. A reasonably sized training dataset was used in the model development for all methods (1,554 4mCs and 1,554 non-4mCs), and all methods were relatively robust for genome-wide analyses.

**A. thaliana**

The validation set used was 64-fold larger than the training dataset. As shown in Figure 2D and Table S1, DNA4mC-LIP showed the best performance for all five metrics tested, i.e., MCC, BACC, sensitivity (Sn), Sp, and AUC (0.561, 0.796, 0.778, 0.813, and 0.863, respectively). These values were higher than those of the second-best method (4mCPred_I) by 0.33, 1.6%, 1.01%, 1.5%, and 1.6%, respectively. The other four methods performed similarly well, with MCC, BACC, and AUC values in the ranges of 0.509–0.520, 77.0%–77.6%, and 84.3%–84.9%, respectively. iDNA4mC ranked at the bottom, with a BACC of 0.609, which was slightly better than that of the random predictor. At p < 0.01, the best method (DNA4mC-LIP) significantly outperformed the other six methods tested. Notably, DNA4mC-LIP considered the output of six different methods (evaluated in the current study) for the final prediction. Interestingly, this was the first time in the evaluation that a combined approach significantly outperformed the individual predictor.

**D. melanogaster**

The validation set was 72-fold larger than the training dataset. As shown in Figure 2E and Table S1, DNA4mC-LIP and 4mCCNN had similar performances, with MCC, BACC, Sn, Sp, and AUC values in the ranges of 0.498–0.515, 81.7%–82.2%, 88.5%–91.3%, 72.1–75.8, and 89.5%–89.6%, respectively. Accordingly, it was difficult to select the best method. Four methods (4mCPred_I, Meta-4mCpred, 4mcPred-SVM, and 4mCPred_II), ranked from third to sixth (accordingly) showed similar performance with MCC, BACC, and AUC values in the ranges of 0.449–0.490, 79.0%–80.9%, and 86.2%–88.5%, respectively. iDNA4mC ranked last, with a BACC value of 64.3%, was only slightly better than that of the random predictor. All methods (except for iDNA4mC) used a reasonably sized training dataset (1,769 4mCs and 1,769 non-4mCs), and these methods performed exceptionally well in the evaluation, especially when a larger validation set was used.

**Geoa. subterraneus**

The validation set used was 16-fold larger than the training dataset. The performances of the different models are shown in Figure 2F and Table S1. iDNA4mC showed the best performance for four out of five metrics, i.e., MCC, BACC, Sp, and AUC (0.150, 0.575, 0.701, and 0.611, respectively). These values were higher than those of the second-best method (4mCPred_I) by 0.053, 2.6%, 21.9%, and 3.7%, respectively. Notably, the performance of 4mCCNN was similar to that of the second-best predictor, 4mCPred_II. The remaining four predictors (4mCPred_I, 4mcPred-SVM, Meta-4mCpred, and DNA4mC-LIP) showed similar performances, with MCC and BACC values in the ranges of 0.043–0.076 and 52.2%–53.9%, respectively. At p < 0.01, iDNA4mC significantly outperformed the other six methods. As mentioned above, the improved iDNA4mC performance may be related to the larger training dataset size.

Although iDNA4mC was superior to other methods, it is not an adequate method because the BACC value (57.5%) was only slightly higher than that of the random classifier. Recently, Tang et al.27 evaluated seven different methods (excluding 4mCCNN) using a smaller validation dataset (350 4mCs and 350 non-4mCs); most of the methods showed reasonable performance (ACC range, 80.0%–88.0%). However, such performance was not replicated when evaluating larger validation sets in the current study.

**M. musculus**

The constructed validation set was 155-fold larger than the training dataset, and it was used to evaluate 4mCPred-EL and i4mC-Mouse. As shown in Figure 3A and Table S1, both methods showed a similar performance, with MCC, BACC, and AUC values in the ranges of 0.018–0.020, 57.1%–57.8%, and 0.612–0.633, respectively. Furthermore, based on the Sn metric, both methods performed exceptionally well, with values exceeding 77.0%. However, the Sp value was quite low, indicating a potential drawback of the training dataset. Hence, the practical applicability of the two methods was limited because of higher false positives.

**F. vesca**

The constructed validation set was 32-fold larger than the training dataset, and it was used to evaluate i4mC-Rose and IDNA-MS. Detailed performance information is shown in Figure 3B and Table S1. Notably, IDNA-MS showed the best performance for four out of...
five metrics, i.e., MCC, BACC, Sn, Sp, and AUC (0.417, 0.847, 0.876, and 0.930, respectively). These values were higher than those of i4mC-ROSE by 0.031, 4.5%, 11.8%, and 4.0%, respectively. At p < 0.01, iDNA-MS significantly outperformed i4mC-ROSE. Interestingly, iDNA-MS used a slightly smaller training dataset and a simpler approach (single-feature model) than i4mC-Rose, yet it showed a relatively better generalization capability.

**Rationale for Model Transferability**

To understand the generalization of the above methods, two-sample logos were used, the statistically significant position-specific composition of 4mCs and non-4mCs was determined, and comparative analysis between training and validation sets was performed. Interestingly, few nucleotide stretches or single nucleotides at specific positions shared the same location in the training and validation datasets for *E. coli*, *C. elegans*, *D. melanogaster*, and *F. vesca*. For *E. coli*, nucleotides at positions 15–18 and 22–24 in positive samples, and those at positions 15, 16, and 22–24 in negative samples, were in the same position in the two sets (Figures 4A and 4B). In *C. elegans*, nucleotides at positions 15–18 and 22–33 in positive samples, and those at positions 17, 19, 20, and 22–32 in negative samples, were in the same position in the two sets (Figures 4C and 4D). In *A. thaliana*, nucleotides at positions 15–17, 22–25, and 27–31 in positive samples, and those at positions 17–20, 22–31, and 36–41 in negative samples, were in the same position in the two sets (Figures 4E and 4F). For *D. melanogaster*, nucleotides at positions 9–12, 19, 20, and 22–25 in positive samples, and those at positions 16, 17, and 22–25 in negative samples, were in the same position in the two sets (Figures 4G and 4H). For *F. vesca*, nucleotides at positions 22–30 and 33–41 in positive samples, and those at positions 22–29 in negative samples, were in the same position in the two sets (Figures 4I and 4J). The analysis revealed that the training dataset utilized for five species (*E. coli*, *C. elegans*, *A. thaliana*, *D. melanogaster*, and *F. vesca*) covered position-specific informative sequence patterns (4mCs and non-4mCs) from their respective genomes or representative samples from the entire genome. Hence, prediction models developed using these smaller training datasets showed robustness or transferability during the evaluation. It is therefore evident that the variations in method ranking depend on the features used, the choice of classifier, and the specific approach used.

**Comparison of 4mC Site Prediction Web Servers**

Because the user experience of web servers is important for experimentalists, user-friendliness of the web servers was then evaluated. Several limitations were noted. First, most of the existing web servers (except for iDNA4mC and 4mCpred-EL) could only handle 41-bp sequences with a cytosine in a central location; this limited the application, particularly for genome-wide analyses. Second, the number of FASTA sequences handled during a single request varied. In particular, Meta-4mCpred, 4mCpred-EL, iDNA-MS, i4mC-Mouse, i4mC-ROSE, and iDNA4mC handle up to 10,000 sequences; 4mCpred_I and 4mCpred_HII handle up to 25,000 sequences; DNA4mC-LIP handles up to 400 sequences; and 4mCpred-SVM handles up to 5,000 sequences. Third, for batch processing, half of the servers did not support the upload of the FASTA sequence files. Only Meta-4mCpred, 4mCpred-SVM, DNA4mC-LIP, i4mC-Mouse, iDNA-MS, and i4mC-Rose offered the option of uploading FASTA sequence files. Fourth, the run times varied, ranging from 3 to 20 min. 4mCpred_I and 4mCpred_HII were the only two servers that could handle large numbers of sequences in a single run and return the prediction results quickly (within 3 min). Finally, a user lacking programming knowledge would not be able to use most of...
the existing methods because the whole genome has to be processed into 41-bp fragments prior to analysis.

**DISCUSSION**

In the current study, eight species-specific computational tools for predicting 4mC sites were surveyed and assessed herein. Some methods for *E. coli*, *A. thaliana*, *F. vesca*, *D. melanogaster*, and *C. elegans* showed an excellent overall performance. However, none of the existing methods was suitable for *Geo. subterraneus*, *Geo. pickeringii*, and *M. musculus*, limiting their applicability. Model robustness for five species and non-transferability to three species could be explained by the position-specific compositional information between training and validation sets. During our evaluation, we observed that current methods have limitations. Seven predictors used the same training dataset to develop six species-specific prediction models, and 4mCpred-EL and i4mC-Mouse used a similar training dataset. However, the training dataset between iDNA-MS and i4mC-Rose was entirely different. Generally, construction of a high-quality dataset is the first and most important step of ML-based prediction model development.31 Surprisingly, none of the ensuing methods tested the training dataset quality of the preceding methods, and there was no attempt to enhance the training dataset quality. Based on the evaluation, the proposed training dataset for *Geo. subterraneus*, *Geo. pickeringii*, and *M. musculus* did not cover the informative sequence patterns around 4mC and non-4mC sites in the whole genomes. Hence, no existing methods for these three species replicated the training performance during our evaluation, indicating that a high-quality dataset construction is important for the development of future predictors. Interestingly, some methods for the remaining five species replicated the training results during evaluation. Although the training dataset covered the informative sequence patterns around 4mC and non-4mC sites for the entire genome, it was quite small compared with the validation dataset generated in the current study. Therefore, apart from dataset quality, dataset size is another important factor that should be considered when developing future predictors.

Based on single-feature encoding and SVM, Chen et al.19 reported the optimal fragment length to 41 bp. Subsequent studies used this length regardless of the species analyzed. In the future, different feature-encoding schemes and classifiers should be explored to identify the optimal fragment length. Based on the non-transferability performance of species-specific models tested with other species, Chen et al.19 concluded the essentials of species-specific models for 4mC site prediction. Subsequent studies followed species-specific model concepts rather than attempting to develop a common 4mC site

---

**Figure 4. Statistically Significant Position-Specific Composition of 4mC and Non-4mC Sites**

The compositional preferences of 4mCs and non-4mCs are denoted above and below the axis, respectively. Training and validation datasets for each species are compared. (A and B) *E. coli*; (C and D) *C. elegans*; (E and F) *A. thaliana*; (G and H) *D. melanogaster*; (I and J) *F. vesca*; (K and L) *Geo. subterraneus*; (M and N) *Geo. pickeringii*; (O and P) *M. musculus*. (A, C, E, G, I, K, M, and O) Training datasets. (B, D, F, H, J, L, N, and P) Validation datasets.
predictor using a multiple-species training dataset. Future studies in this direction are needed to understand the necessity of species-specific models rather than a common predictor. Furthermore, in terms of web server utility, most of the existing methods are not user-friendly because they can handle only a small number of sequences in a single run, require a long processing run time, lack the options for result file download and genome-wide 4mC site screening, and sometimes show time-out errors.

To improve the prediction performance, several points should be considered. First, highly homologous sequences should be excluded to reduce bias in the training dataset. Second, feature-encoding algorithms focusing on position-specific information may be useful for differentiating 4mCs from non-4mCs. Third, several feature-representation schemes have been recently proposed, such as iterative feature representation,22 adaptive feature learning,32 effective feature representation,3 and fused multi-view information.31 Application of two or more schemes to the same dataset and selection of the most appropriate one may improve model robustness. Fourth, the reliable and robust performance of deep-learning algorithm is often dependent on a huge training dataset.31-37 Considering the large size of the dataset constructed in the current study, it would be interesting to apply the deep-learning algorithm to 4mC site prediction by exploring different feature representation schemes, and to ultimately compare the performance of deep-learning and conventional ML-based models. Overall, this study will assist scientists with an interest in the field of developing improved tools for the prediction of 4mC sites.

MATERIALS AND METHODS

A General Framework of the Existing Computational Approaches for 4mC Site Prediction

The general framework for the existing computational approaches is shown in Figure 1 and comprises four steps. As the first step, a high-quality 4mC dataset based on validated databases38,39 and literature search is constructed. Because not many experimentally characterized non-4mC sites are available, 41-bp fragments are generated from chromosomal DNA, containing cytosine at the central position. These fragments should not overlap with 4mC sites detected experimentally and are considered negative (non-4mC) samples. To avoid overestimation of the predictions, CD-HIT is generally applied to discard redundant sequences.45 From the final dataset, 80% or 75% of samples are randomly selected and treated as the training dataset for prediction model development, and the remaining samples are treated as an independent dataset to check model robustness. The second step involves feature representation and optimization. For the former, a variety of feature descriptors are generally used to capture meaningful information that distinguishes between positive and negative samples, including composition-based features31-44 (Kmer nucleotide frequency [Kmer], reverse complementary Kmer, enhanced nucleic acid composition, composition of k-spaced nucleic acid pairs, pseudo-dinucleotide composition [PseDNC], pseudo-trinucleotide composition (PseTNC), pseudo-k-tuple composition, parallel correlation PseDNC, parallel correlation PseTNC, series correlation PseDNC, and series correlation PseTNC), position-specific-based features45 (mononucleotide binary encoding [MBE], dinucleotide binary encoding [DBE], numerical representation of nucleotides, position-specific trinucleotide propensity, electron-ion interaction pseudopotential [EIIP]), physicochemical property-based features46 (ring-function hydrogen chemical properties and accumulated nucleotide frequency [RFHCP], dinucleotide physicochemical properties [DPCP], trinucleotide physicochemical properties [TPCP], dinucleotide-based autocovariance [DAC], and dinucleotide-based cross-correlation [DCC]), and evolutionary-based features (KNN features). To discard irrelevant and redundant features from the original feature descriptor, feature optimization using sequential forward search (SFS) or other approaches is generally performed.47-49 In the third step, the prediction model is constructed based on the exploration of different classifiers and different approaches. Specifically, the optimal features from each descriptor (from the second step) are input to several ML classifiers (SVM, extreme gradient boosting, deep learning, and random forest) to develop a prediction model. In the fourth step, while developing the prediction model, different types of approaches are explored, including meta-predictor,50 feature representation learning,51,52 iterative feature representation,22,23 single descriptor-based model,54,55 ensemble model,56-58 and stacking approach.59,60 Finally, the web server is constructed based on the final prediction model, to predict whether a specific sequence represents a 4mC or non-4mC sample.

Validation Dataset Construction

To evaluate the performance of the existing 4mC prediction tools and to enable a fair comparison between them, positive and negative samples were constructed in the current study. First, positive samples were extracted for seven species (C. elegans, D. melanogaster, A. thaliana, Geoa. subterraneus, Geob. pickeringii, E. coli, and M. musculus) from the MethSMRT database,38 which includes data for 156 species (149 prokaryotes and seven eukaryotes). Epigenetic modification (4mC or N6-methyladenine [6mA]) data denoted in MethSMRT were compiled either from the Sequence Read Archive or NCBI Gene Expression Omnibus. Second, the downloaded raw data from MethSMRT were processed. Sequences lacking relevant information (i.e., with 6mA and 4mC sites lacking any modification [modQV] score) were excluded, yielding 41-bp sequences containing central cytosine (i.e., 4mC sites) with varying modQV scores. For F. vesca, sorted data (41-bp fragments with different modQV scores) were downloaded from the MDR database.29 Selecting the optimal modQV threshold was challenging, as thresholds of 2031 and 3052 corresponding to p values of 0.01 and 0.001, respectively, are recommended in different studies. The existing methods utilized positive samples with a modQV greater than 30. However, average cutoffs greater than or equal to 25 were used for six species (C. elegans, D. melanogaster, A. thaliana, Geoa. subterraneus, Geob. pickeringii, and F. vesca), and thresholds greater than or equal to 20 were used for the remaining two species (E. coli and M. musculus). As the analysis of the latter two species at the original cutoff (≥25) resulted in a very small positive sample size, a slightly smaller modQV cutoff value was used. Finally, positive samples for each species that shared greater
The first column represents species name. The second and third columns respectively represent positive and negative samples constructed in this study. *modQV cutoff slightly reduced for the positive samples.

A statistical summary of positive and negative samples for each species is shown in Table 1. On average, the numbers of positive and negative samples generated in the current study were 14-fold and 137-fold higher than the numbers of respective samples used in the training dataset used by the existing methods, the above generated positive samples, and positive samples with low modQV values were excluded. The validation dataset can be downloaded at http://theoglelab.org/Meta-4mCpred/EvaluationData.html.

To construct the negative samples, the same protocol as that used in previous studies was followed. Briefly, the entire single circular chromosome was considered for three species (E. coli, Geoa. subterraneus, and Geob. pickeringii), and 41-bp fragments with a central cytosine were generated. For species with multiple chromosomes, similar fragments were generated and 50% of sequences from each chromosome were randomly selected. In this manner, all chromosomes for the remaining species (C. elegans, D. melanogaster, A. thaliana, M. musculus, and F. vesca) were covered. For species with multiple chromosomes, similar fragments were generated and 50% of sequences from each chromosome were randomly selected. In this manner, all chromosomes for the remaining species (C. elegans, D. melanogaster, A. thaliana, M. musculus, and F. vesca) were covered. Subsequently, sequences that shared more than 75% sequence identity with the training dataset used by the existing methods, the above generated positive samples, and positive samples with low modQV values were excluded. The validation dataset can be downloaded at http://theoglelab.org/Meta-4mCpred/EvaluationData.html.

Table 1. Summary of the Newly Constructed Validation Dataset

| Species               | Positive | Negative |
|-----------------------|----------|----------|
| E. coli*              | 670      | 118,266  |
| C. elegans            | 19,289   | 144,553  |
| D. melanogaster       | 45,809   | 209,690  |
| A. thaliana           | 73,785   | 178,711  |
| Geoa. subterraneus    | 17,573   | 11,404   |
| Geob. pickeringii     | 4,445    | 60,005   |
| M. musculus*          | 942      | 247,823  |
| F. vesca              | 21,350   | 290,857  |

The first column represents species name. The second and third columns respectively represent positive and negative samples constructed in this study. *modQV cutoff slightly reduced for the positive samples.

A statistical summary of positive and negative samples for each species is shown in Table 1. On average, the numbers of positive and negative samples generated in the current study were 14-fold and 137-fold higher than the numbers of respective samples used in the training dataset.

**Existing 4mC Prediction Methods**

ML algorithms have been widely applied in various fields, such as post-replication DNA modification site predictions, particularly 4mC, 6mA, and 5-hydroxymethylcytosine sites. Table 2 summarizes the existing 4mC prediction methods that utilize a wide range of ML algorithms, feature encodings, and different approaches. The methods are described in detail below, according to the year of publication.

**4mC Prediction Methods Developed in 2017**

**iDNA4mC.** Chen et al. proposed iDNA4mC, the first method for 4mC site prediction. First, the authors constructed a nonredundant training dataset for six species, where each species was represented by an equal number of positive (4mCs) and negative (non-4mCs) samples. The positive samples were obtained from the MethSMRT database; the fragment length was 41 bp, with a cytosine base at position 21 and a modQV score greater than 30. Negative samples were constructed from the respective genomes; the fragment length and cytosine positioning were the same as in the positive samples. Most importantly, the fragments were not detected by the SMRT sequencing technology. To reduce the bias and over-fitting, the authors applied CD-HIT and excluded sequences that shared more than 80% sequence identity with sequences in each species. They then experimented with the different sequence lengths and evaluated the model performance. Eventually, they identified 41 bp as the optimal length for consistently obtaining the best performance, regardless of species. Surprisingly, the same optimal sequence length has been used in later studies of other species. Ultimately, a set of 3,108, 3,538, 3,956, 776, 1,812, and 1,138 samples for C. elegans, D. melanogaster, A. thaliana, E. coli, Geoa. subterraneus, and Geob. pickeringii species, respectively, was generated. Herein, this dataset is referred to as the "Chen dataset."

**4mC Prediction Methods Reported in 2018**

Two methods were proposed in 2018, relying on different approaches to predict 4mC sites in different species. Both methods utilize the Chen dataset and SVM for model construction.

**4mCPred.** He et al. proposed 4mCPred that consists of two prediction models (4mCPred_I and 4mCPred_II) built using different input features. 4mCPred_I was developed using single-feature descriptors, namely, position-specific trinucleotide propensity (PSTNP) features that achieved accuracies of 87.0%, 86.94%, 82.25%, 94.46%, 89.95%, and 90.69% for C. elegans, D. melanogaster, A. thaliana, E. coli, Geoa. subterraneus, and Geob. pickeringii, respectively. 4mCPred_II was developed using hybrid features (a combination of PSTNP and EIIP), where the optimal feature set was identified by a two-step feature selection protocol, with the features ranked based on F-scores, followed by SFS using SVM. 4mCPred_I achieved accuracies of 87.71%, 87.79%, 83.37%, 94.97%, 91.04%, and 90.89% for C. elegans, D. melanogaster, A. thaliana, E. coli, Geoa. subterraneus, and Geob. pickeringii, respectively. The performance of both...
was overestimated because of over-fitting. Hence, they rebuilt the 4mCpred model. The reported metrics for the six species, i.e., the average ACC, MCC, Sn, and Sp, were 0.637, 0.817, 0.815, and 0.818, respectively. Overall, the performance of 4mCpred-SVM was superior to that of the above two methods. 4mCpred-SVM is publicly accessible at http://server.malab.cn/4mCpred-SVM.

### 4mC Prediction Methods Reported in 2019

Five methods were proposed in 2019. Of these, three methods were proposed simultaneously using different approaches that predict 4mC sites in six different species. Notably, all of these methods essentially relied on the same Chen dataset for prediction model development.

#### Meta-4mCpred

Manavalan et al.\(^8\) reported the first meta-predictor, called Meta-4mCpred. First, 14 feature descriptors were generated by the listed tool addresses are as follows: iDNA4mC, http://lin-group.cn/server/iDNA4mC/; 4mCPred, http://server.malab.cn/4mCPred/; 4mCpred-SVM, http://server.malab.cn/4mCpred-SVM; Meta-4mCpred, http://thegleelab.org/Meta-4mCpred/; 4mcPred-IFL, http://server.malab.cn/4mCpred-IFL/; 4mcCCNN, https://home.jbnu.ac.kr/NSCL/4mcCCNN. Tools contain six species-specific prediction models, namely A. thaliana, C. elegans, D. melanogaster, E. coli, Geoa. subterraneus, and Geob. pickeringii. Web server is not functional. Tool contains one prediction model to compute 4mC site from specific species. Tool contains two prediction models for F. vesca and Rosa chinensis. Tool contains four different species-specific models, namely F. vesca, Casuaria equisetifolia, Saccharomyces cerevisiae, and Ts. SUP5-1.

**Table 2. List of Currently Available Tools for 4mC Sites Prediction Assessed in This Study**

| Year | Tool\(^a\) | Classifier | Training/Independent Dataset Size | Features | Web Server | Evaluation Strategy | File Upload |
|------|------------|------------|-----------------------------------|----------|------------|--------------------|------------|
| 2017 | iDNA4mC\(^b\) | SVM        | Chen dataset/–                     | RFHCP    | yes        | LOOCV              | no         |
| 2018 | 4mCPred\(^c\) | SVM        | Chen dataset/–                     | PSTNP, EIIP | yes        | LOOCV              | no         |
|      | 4mCPred-SVM\(^d\) | SVM       | Chen dataset/–                     | Kmer, MBE, DBE, LPDF | yes        | 10-fold CV        | yes        |
| 2019 | Meta-4mCpred\(^e\) | RF, ERT, GB, SVM | Chen dataset/Manavalan dataset | Kmer+MBE, DBE+LPDF, PCPs, PseDNC, KNN, EIIP, MMI, RFHCP | yes        | 10-fold CV        | yes        |
| 2019 | 4mcPred-IFL\(^d\) | SVM        | Chen dataset/–                     | Kmer, MBE, DBE+LPDF, PCPs, PseDNC, KNN, EIIP, MMI, RFHCP | yes        | 10-fold CV        | yes        |
| 2019 | 4mcCCNN\(^d\) | CNN        | Chen dataset/–                     | MBE      | yes\(^f\) | 10-fold CV        | –          |
| 2019 | 4mCpred-EL\(^d\) | RF, GB, ERT, SVM | (800 4mCs and 800 non-4mCs)/ (180 4mCs and 180 non-4mCs) | Kmer, DPE+LPDF, RFHC, EIIP, MBE, DPCP, TPCC | yes        | 10-fold CV        | yes        |
| 2019 | i4mC-ROSE\(^d\) | RF         | (4854 4mCs and 4854 non-4mCs)/ (1617 4mCs and 1617 non-4mCs) | KSNC, MBE, EIIP | yes        | 10-fold CV        | yes        |
| 2019 | iEC4mC-SVM\(^d\) | SVM        | (388 4mCs and 388 non-4mCs)/ (134 4mCs and 134 non-4mCs) | MBE, RFHC, DAE, X-k-YCF, Kmer | yes        | 10-fold CV        | –          |
| 2019 | DNA4mC-LIP\(^d\) | –          | –/Manavalan dataset | integration of six existing predictors | yes        | independent evaluation | yes        |
| 2020 | 4mCDeep-CBId | CNN, BLSTM | (1,173 4mCs and 6,635 non-4mCs)/ (180 4mCs and 180 non-4mCs) | Kmer, MBE, DBE, LPDF, RFHCP, X-k-YCF | yes        | 10-fold CV        | yes        |
| 2020 | iDNA-MS\(^d\) | RF         | 7,899 samples/7,898 samples | Kmer, RFHCP, MBE | yes        | 5-fold CV         | yes        |
| 2020 | i4mC-Mouse\(^d\) | RF         | (746 4mCs and 746 non-4mCs)/ (160 4mCs and 160 non-4mCs) | Kmer, KSNC, MBE, EIIP | yes        | 10-fold CV       | yes        |

Chen dataset contains C. elegans (4mCs, 1,554; non-4mCs, 1,554), D. melanogaster (4mCs, 1,769; non-4mCs, 1,769), A. thaliana (4mCs, 1,978; non-4mCs, 1,978), E. coli (4mCs, 388, non-4mCs, 388), Geoa. subterraneus (4mCs, 906; non-4mCs, 906), and Geob. pickeringii (4mCs, 569; non-4mCs, 569). Manavalan dataset contains C. elegans (4mCs, 750; non-4mCs, 750), D. melanogaster (4mCs, 1,000; non-4mCs, 1,000), A. thaliana (4mCs, 1,250; non-4mCs, 1,250), E. coli (4mCs, 134; non-4mCs, 134), Geoa. subterraneus (4mCs, 350; non-4mCs, 350), and Geob. pickeringii (4mCs, 200; non-4mCs, 200). SVM, support vector machine; RF, random forest; GB, gradient boosting; CNN, convolutional neural network; BLSTM, bidirectional long short-term memory network; ERT, extremely randomized tree; RFHCP, ring-function-hydrogen-chemical properties, PSTNP, position-specific trinucleotide pseudo-penalty; EIIP, electron-ion interaction pseudopotential; Kmer, Kmer nucleotide frequency; MBE, mononucleotide binary encoding, DBE, dinucleotide binary encoding, LPDF, local position-specific dinucleotide frequency; DPCP, dinucleotide physicochemical properties; TPCP, trinucleotide physicochemical properties; PseDNC, pseudo-dinucleotide composition; KNN, k-nearest neighbor; KSNC, k-space nucleotide composition; DAC, dinucleotide physicochemical properties autocorrelation; X-k-YCF, Xmer-kGap-Ymer composition frequency; ANF, accumulated nucleotide frequency; LOOCV, leave-one-out cross-validation; CV, cross-validation.

\(^a\)The listed tool addresses are as follows: iDNA4mC, http://lin-group.cn/server/iDNA4mC/; 4mCPred, http://server.malab.cn/4mCPred/; 4mcPred-SVM, http://server.malab.cn/4mCPred-SVM/; Meta-4mCpred, http://thegleelab.org/Meta-4mCpred/; 4mcPred-IFL, http://server.malab.cn/4mCPred-IFL/; 4mcCCNN, https://home.jbnu.ac.kr/NSCL/4mcCCNN.html; 4mCpred-EL, http://theagleelab.org/4mCpred-EL/; i4mC-ROSE, http://kurata14.bio.kyutech.ac.jp/i4mC-ROSE/; DNA4mC-LIP, http://iueest.edu.cn/DNA4mC-LIP/; iDNA-MS, http://lin-group.cn/server/iDNA-MS/; i4mC-Mouse, http://kurata14.bio.kyutech.ac.jp/i4mC-Mouse/.

\(^b\)Tools contain six species-specific prediction models, namely A. thaliana, C. elegans, D. melanogaster, E. coli, Geoa. subterraneus, and Geob. pickeringii.

\(^c\)Web server is not functional.

\(^d\)Tool contains one prediction model to compute 4mC site from specific species.

\(^e\)Tool contains two prediction models for F. vesca and Rosa chinensis.

\(^f\)Tool contains four different species-specific models, namely F. vesca, Casuaria equisetifolia, Saccharomyces cerevisiae, and Ts. SUP5-1.
| Species          | Methods     | Training          | Independent test          |
|------------------|-------------|-------------------|----------------------------|
|                  |             | MCC   | ACC   | Sn    | Sp    | MCC   | ACC   | Sn    | Sp    |
|                  |             |        |        |       |       |        |        |       |       |
| **C. elegans**   | iDNA4mC     | 0.572 | 0.786 | 0.797 | 0.775 | –      | –      | –      | –      |
|                  | 4mCPred_I   | 0.740 | 0.870 | 0.871 | 0.869 | –      | –      | –      | –      |
|                  | 4mCPred_II  | 0.750 | 0.877 | 0.875 | 0.879 | –      | –      | –      | –      |
|                  | 4mcPred-SVM | 0.631 | 0.815 | 0.824 | 0.807 | –      | –      | –      | –      |
|                  | Meta-4mCpred| 0.652 | 0.826 | 0.840 | 0.812 | 0.741  | 0.870 | 0.843  | 0.897  |
|                  | 4mcPred-IFL | 0.761 | 0.880 | 0.890 | 0.871 | –      | –      | –      | –      |
|                  | 4mCCNN      | 0.694 | 0.842 | 0.895 | 0.825 | –      | –      | –      | –      |
|                  | DNA4mC-LIP  | –     | –     | –     | –     | 0.786  | 0.893 | 0.885  | 0.901  |
|                  | 4mcDeep-CBI | 0.850 | 0.929 | 0.949 | 0.894 | –      | –      | –      | –      |
| **D. melanogaster** | iDNA4mC     | 0.625 | 0.812 | 0.833 | 0.791 | –      | –      | –      | –      |
|                  | 4mCPred_I   | 0.740 | 0.869 | 0.869 | 0.869 | –      | –      | –      | –      |
|                  | 4mCPred_II  | 0.760 | 0.878 | 0.876 | 0.880 | –      | –      | –      | –      |
|                  | 4mcPred-SVM | 0.661 | 0.830 | 0.838 | 0.822 | –      | –      | –      | –      |
|                  | Meta-4mCpred| 0.685 | 0.842 | 0.831 | 0.854 | 0.812  | 0.906 | 0.913  | 0.899  |
|                  | 4mcPred-IFL | 0.745 | 0.873 | 0.865 | 0.88  | –      | –      | –      | –      |
|                  | 4mCCNN      | 0.687 | 0.854 | 0.864 | 0.854 | –      | –      | –      | –      |
|                  | DNA4mC-LIP  | –     | –     | –     | –     | 0.849  | 0.924 | 0.943  | 0.905  |
| **A. thaliana**  | iDNA4mC     | 0.519 | 0.760 | 0.757 | 0.762 | –      | –      | –      | –      |
|                  | 4mCPred_I   | 0.650 | 0.823 | 0.813 | 0.832 | –      | –      | –      | –      |
|                  | 4mCPred_II  | 0.670 | 0.834 | 0.830 | 0.838 | –      | –      | –      | –      |
|                  | 4mcPred-SVM | 0.573 | 0.787 | 0.778 | 0.796 | –      | –      | –      | –      |
|                  | Meta-4mCpred| 0.584 | 0.792 | 0.761 | 0.822 | 0.711  | 0.855 | 0.876  | 0.834  |
|                  | 4mcPred-IFL | 0.644 | 0.822 | 0.803 | 0.840 | –      | –      | –      | –      |
|                  | 4mCCNN      | 0.622 | 0.797 | 0.804 | 0.792 | –      | –      | –      | –      |
|                  | DNA4mC-LIP  | –     | –     | –     | –     | 0.720  | 0.859 | 0.883  | 0.836  |
| **E. coli**      | iDNA4mC     | 0.598 | 0.799 | 0.820 | 0.778 | –      | –      | –      | –      |
|                  | 4mCPred_I   | 0.890 | 0.945 | 0.956 | 0.933 | –      | –      | –      | –      |
|                  | 4mCPred_II  | 0.900 | 0.950 | 0.951 | 0.949 | –      | –      | –      | –      |
|                  | 4mcPred-SVM | 0.666 | 0.833 | 0.858 | 0.807 | –      | –      | –      | –      |
|                  | Meta-4mCpred| 0.697 | 0.848 | 0.869 | 0.827 | 0.650  | 0.825 | 0.806  | 0.843  |
|                  | 4mcPred-IFL | 0.789 | 0.894 | 0.907 | 0.881 | –      | –      | –      | –      |
|                  | 4mCCNN      | 0.688 | 0.859 | 0.881 | 0.789 | –      | –      | –      | –      |
|                  | DNA4mC-LIP  | –     | –     | –     | –     | 0.676  | 0.837 | 0.803  | 0.871  |
|                  | iEC4mC-SVM  | 0.711 | 0.854 | 0.820 | 0.889 | 0.665  | 0.832 | 0.851  | 0.813  |
| **Geoa. subterraneus** | iDNA4mC    | 0.630 | 0.815 | 0.822 | 0.808 | –      | –      | –      | –      |
|                  | 4mCPred_I   | 0.800 | 0.900 | 0.899 | 0.900 | –      | –      | –      | –      |
|                  | 4mCPred_II  | 0.820 | 0.910 | 0.912 | 0.909 | –      | –      | –      | –      |
|                  | 4mcPred-SVM | 0.674 | 0.837 | 0.840 | 0.834 | –      | –      | –      | –      |
|                  | Meta-4mCpred| 0.711 | 0.855 | 0.856 | 0.854 | 0.701  | 0.850 | 0.817  | 0.883  |
|                  | 4mcPred-IFL | 0.773 | 0.887 | 0.886 | 0.887 | –      | –      | –      | –      |
|                  | 4mCCNN      | 0.704 | 0.860 | 0.852 | 0.843 | –      | –      | –      | –      |
|                  | DNA4mC-LIP  | –     | –     | –     | –     | 0.674  | 0.837 | 0.814  | 0.860  |

(Continued on next page)
for which a corresponding optimal feature set-based SVM model was developed; (2) concatenation of the predicted probabilities of 4mCs from eight models (step 1), which was considered a new feature vector; and (3) training of the 8D vector (step ) using SVM, which generated a new model for which the predicted probabilities of 4mCs were combined with the original 8D to establish a new 9D vector. This process was repeated until the performance reached convergence. For the six species, 4mCpred-IFL achieved the average MCC, ACC, Sn, and Sp values of 0.754, 0.877, 0.876, and 0.878, respectively. The performance of 4mCpred-IFL was better than that of iDNA4mC, 4mCpred, and 4mcPred-SVM. This indicated that the features learned through the iterative process had a reasonable capacity to discriminate between positive and negative samples. 4mcPred-IFL is publicly accessible at http://server.malab.cn/4mcPred-IFL.

4mCCNN. Khanal et al.23 proposed 4mCCNN, the first deep-learning-based method, based on MBE encoding and a convolutional neural network (CNN) classifier. The authors applied 10-fold cross-validation and optimized CNN regularization parameters. For the six species, 4mCCNN achieved the average MCC, ACC, Sn, and Sp values of 0.754, 0.877, 0.876, and 0.878, respectively. The performance of 4mCCNN was superior to that of iDNA4mC, 4mCpred, and 4mcPred-SVM. The authors also constructed a web server, accessible at https://home.jbnu.ac.kr/NSCL/4mCCNN.htm. However, the 4mCCNN server was out of service during the preparation of the current manuscript.

4mCpred-EL. Manavalan et al.22 proposed 4mCpred-EL, the first 4mC site prediction method for M. musculus. First, the training and independent datasets were constructed, comprising 1,600 and 320 samples, respectively; each dataset contained an equal number of positive and negative samples. Notably, none of the samples shared greater than 80% sequence identity with other samples in the dataset.
Subsequently, seven different feature descriptors (MBE, DPCP, EIIP, Kmer, a combination of DPE and LPDF, RFHCP, and TPCP) and four different ML classifiers (RF, GB, ERT, and SVM) were used to develop a prediction model. The model was developed with the following steps: (1) development of 28 prediction models based on seven descriptors and four classifiers, combining the predicted probabilities of 4mCs, and considering them a new feature vector; and (ii) inputting the 28D vector into four classifiers again, with the majority voting from the ensemble classifiers considered as the final output. 4mCpred-EL yielded MCC, ACC, Sn, and Sp values of 0.591, 0.795, 0.804, and 0.787, respectively, during cross-validation. The corresponding metrics for independent evaluation were 0.596, 0.798, 0.804, and 0.792, respectively. This method is publicly accessible at http://thegleelab.org/Meta-4mCpred/.

i4mC-ROSE. Hasan et al.25 proposed i4mC-ROSE, the first 4mC site prediction method for F. vesca and Rosa chinensis. Because the current study focuses on F. vesca, only information for this species model is provided. The authors constructed nonredundant training and independent datasets comprising 9,708 and 3,234 samples, respectively; positive samples were obtained from the MDR database,39 and negative samples were derived as described for iDNA4mC.15 Notably, none of the samples shared greater than 65% sequence identity with other samples. Subsequently, six different feature descriptors (K-space nucleotide composition [KSNC], MBE, EIIP, Kmer, DPCP, and TPCP) and RF classifier were used for model construction, which was achieved as follows: (1) development of RF-based six descriptor models; (2) integration of the predicted probabilities of 4mCs using a linear regression approach; and (3) choosing three models, KSNC, MBE, and EIIP, which contributed 25%, 45%, and 30%, respectively, for the final prediction. i4mC-ROSE achieved MCC, ACC, Sn, and Sp values of 0.545, 0.767, 0.635, and 0.899, respectively, during cross-validation. The corresponding metrics for independent evaluations were 0.601, 0.797, 0.721, and 0.873, respectively (Table 3). i4mC-ROSE is freely accessible at http://kurata14.bio.kyutech.ac.jp/i4mC-ROSE/.

4mC Prediction Methods Developed in 2020
As of April 2020, five additional methods were published for different species and approaches. A brief description of each method is given below.

iEC4mC-SVM. Lv et al.26 developed a predictor for E. coli, termed iEC4mC-SVM. The authors used the E. coli dataset of Chen et al.19 for model building and that of Manavalan et al.8 for checking model transferability. Using the training dataset, the authors considered six different descriptors, namely, MBE, DBE, DAC, DCC, RFHCP, and Xmer-kGap-Ymer composition frequency. All of these descriptors were concatenated to generate a 10060D feature vector. Subsequently, a light gradient-boosting machine was applied to rank the features; the top 250 features were selected from the sorted features, ranked from highest to lowest. SFS was applied to these 250D features. An optimal feature set (187D) was thus obtained, with the MCC, ACC, Sn, and Sp values of 0.711, 0.854, 0.820, and 0.889, respectively, during cross-validation. The corresponding metrics for independent evaluations were 0.665, 0.832, 0.851, and 0.813, respectively. The performance of iEC4mC-SVM was similar to that of Meta-4mCpred, but the metrics were significantly worse than those of the other three methods (4mCpred, 4mcPred-IFL, and 4mcPred-SVM) during cross-validation. However, iEC4mC-SVM was slightly better than 4mCpred and Meta-4mCpred, and significantly better than 4mcPred-IFL and 4mcPred-SVM during independent assessment. Unfortunately, iEC4mC-SVM is not publicly available.

DNA4mC-LIP. Tang et al.27 proposed a novel meta-predictor, called DNA4mC-LIP. Instead of developing a prediction model based on a training dataset, the authors utilized an independent dataset (for six species) of Manavalan et al.8 and generated predicted probabilities of 4mC scores from the six existing models, namely, iDNA4mC, 4mCpred_1, 4mCpred_2, 4mCpred-IFL, and Meta-4mCpred. Subsequently, the predicted probabilities of 4mCs were integrated using an optimal weight to make a final prediction. For each species, the number of existing predictor contributions varied significantly. In DNA4mC-LIP, the authors did not use any ML classifiers. Instead, they used existing ML-based prediction methods; hence, the DNA4mC-LIP method is classified as an ML-based method in the current study. For the six species, DNA4mC-LIP achieved the average MCC, ACC, Sn, and Sp values of 0.721, 0.860 0.850, and 0.870, respectively. The predictor performs slightly better than the individual predictors. DNA4mC-LIP is accessible at http://i.uestc.edu.cn/DNA4mC-LIP/.

4mcDeep-CBI. Zeng et al.1 developed 4mcDeep-CBI, a predictor for E. coli. The authors constructed a new imbalanced dataset containing 11,173 4mCs and 6635 non-4mCs. Subsequently, the authors used eight feature descriptors (as for 4mCpred-IFL) input into CNN and a bidirectional long short-term memory network for the development of the final prediction model. 4mcDeep-CBI achieved MCC, ACC, Sn, and Sp values of 0.850, 0.929, 0.949, and 0.894 during cross-validation, with a significantly better performance than 4mCpred-IFL. Although the authors provided a standalone version of 4mcDeep-CBI (https://github.com/mat310/4mcDeep), they did not provide detailed information for executing this program in the README.md file. Therefore, this method was excluded from the current evaluation.

iDNA-MS. Lv et al.2 proposed iDNA-MS, a method that predicts not only 4mC sites but also 5-hydroxymethylcytosine and 6mA sites. In total, the authors reported 17 prediction models for different species. Considering the F. vesca-based model, the authors generated 7,899 and 7,898 samples, and used them for training and independent evaluations, respectively. Positive samples were extracted from MDR database, and negative samples were generated as in the iDNA4mC study.19 None of the sequences in the dataset shared greater than 80% sequence identity with other sequences. The RF-based MBE descriptor achieved MCC, ACC, Sn, and Sp values of 0.868, 0.934, 0.943, and 0.925, respectively. The corresponding metrics for...
The comparison of ML algorithms and their performances with those of the existing methods (Tables 2 and 3) revealed the following similarities and dissimilarities. (1) Five methods (iDNA4mC, 4mCPred, 4mcPred-SVM, 4mCPred-IFL, and iEC4mC-SVM) used the SVM classifier; three algorithms (i4mC-ROSE, iDNA-MS, and i4mC-Mouse) used RF; two methods (4mCCNN and 4mcDeep-CBI) applied deep learning; and the remaining methods (Meta-4mCPred and 4mCPred-EL) utilized multiple ML classifiers. (2) Two methods (iDNA4mC and 4mCCNN) used single-feature encoding, whereas the remaining methods (except for DNA4mC-LIP) explored multiple-feature encodings. While developing a prediction model, most methods used feature-selection techniques and identified the optimal feature set during model development. (3) Three existing models (Meta-4mCPred, 4mCPred-EL, i4mC-Mouse, iDNA-MS, and i4mC-Rose) evaluated model transferability using an independent dataset, whereas the remaining methods focused on improving the training ACC. (4) Based on the training ACC reported in the literature, 4mcDeep-CBI achieved the best performance for C. elegans; 4mCPred_II had the best performance for five different species (D. melanogaster, A. thaliana, E. coli, Geoa. subterraneus, and Geob. pickeringii); 4mCPred-EL and i4mC-Mouse achieved similar performance for M. musculus; and iDNA-MS achieved the best performance for F. vesca.

**Evaluation Metrics**

To quantify and evaluate the performance of the developed models, five commonly used evaluation metrics were used:\(^1\)\(^-\)\(^5\)\(^,\)\(^7\)\(^-\)\(^11\)\( Sn, Sp, ACC, BAAC, and MCC. Each metric is defined as follows:

\[
\begin{align*}
ACC &= \frac{TP + TN}{TP + TN + FP + FN} \\
Sn &= \frac{TP}{TP + FN} \\
Sp &= \frac{TN}{TN + FP} \\
BAAC &= \frac{Sn + Sp}{2} \\
MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}
\end{align*}
\]

where TP is the number of 4mCs correctly predicted as 4mCs, TN is the number of non-4mCs correctly predicted as non-4mCs, FP is the number of 4mCs incorrectly predicted as non-4mCs, and FN is the number of non-4mCs incorrectly predicted as 4mCs.
cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. J. Biol. Chem. 287, 34801–34808.

7. Bart, A., van Passel, M.W., van Amsterdam, K., and van der Ende, A. (2005). Direct detection of methylation in genomic DNA. Nucleic Acids Res. 33, e124.

8. Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2019). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. Mol. Ther. Nucleic Acids 16, 733–744.

9. Rahi, P., Maurer, S., and Summerer, D. (2018). Selective recognition of N4-methylcytosine in DNA by engineered transcription-activator-like effectors. Philos. Trans. R. Soc. Lond. B Biol. Sci. 373, 20170078.

10. Patail-lat-Meakin, T., Pillay, N., and Beck, S. (2016). 3-Methylcytosine in cancer: an underappreciated methyl lesion? Epigenomics 8, 451–454.

11. Robertson, K.D. (2005). DNA methylation and human disease. Nat. Rev. Genet. 6, 597–610.

12. Casadesús, J., and Low, D. (2006). Epigenetic gene regulation in the bacterial world. Microbiol. Mol. Biol. Rev. 70, 830–856.

13. Timmuskas, A., Butkus, V., and Janulaitis, A. (1995). Sequence motifs characteristic for DNA [cytosine-N4] and DNA [adenine-N6] methyltransferases. Classification of all DNA methyltransferases. Gene 157, 3–11.

14. Schweizer, H. (2008). Bacterial genetics: past achievements, present state of the field, and future challenges. Biotechniques 44, 633–634, 636–641.

15. Iyer, L.M., Abhiman, S., and Aravind, L. (2011). Natural history of eukaryotic DNA methylation systems. Prog. Mol. Biol. Transl. Sci. 101, 25–104.

16. Modrich, P. (1991). Mechanisms and biological effects of mismatch repair. Annu. Rev. Biochem. 60, 441–465.

17. Ye, P., Luan, Y., Chen, K., Liu, Y., Xiao, C., and Xie, Z. (2017). MethSMART: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Res. 45 (D1), D85–D89.

18. Liu, Z.Y., Xing, J.F., Chen, W., Luan, M.W., Xie, R., Huang, J., Xie, S.Q., and Xiao, C.L. (2019). MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. Hortic. Res. 6, 78.

19. Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

20. Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.

21. Sessions, V., and Valtorta, M. (2006). The effects of data quality on machine learning algorithms. In Proceedings of the 11th International Conference on Information Quality, pp. 485–498.

22. Cao, R., Bhattacharya, D., Hou, J., and Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics 17, 495.

23. Cao, R., Adhikari, B., Bhattacharya, D., Sun, M., Hou, J., and Cheng, J. (2017). QAcom: single model quality assessment using protein structural and contact information with machine learning techniques. Bioinformatics 33, 586–588.

24. Conover, M., Staples, M., St, D., Sun, M., and Cao, R. (2019). AngularQA: protein model quality assessment with LSTM networks. Comput. Math. Biophys. 7, 1–9.

25. Smith, J., Conover, M., Stephenson, N., Eickholt, J., St, D., Sun, M., and Cao, R. (2020). TopQA: a topological representation for single-model protein quality assessment with machine learning. Int. J. Comput. Biol. Drug Des. 13, 144–153.
50. Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2019). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 35, 2757–2765.

51. Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2020). Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Brief. Bioinformatics 21, 106–119.

52. Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics 34, 4007–4016.

53. Xu, Y., Zhao, X., Liu, S., Liu, S., Niu, Y., Zhang, W., and Wei, L. (2019). LncPred-IEL: a long non-coding RNA prediction method using iterative ensemble learning. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 555–562.

54. Manavalan, B., Subramanyam, S., Shin, T.H., Kim, M.O., and Lee, G. (2018). Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. J. Proteome Res. 17, 2715–2726.

55. Tan, J.X., Li, S.H., Zhang, Z.M., Chen, C.X., Chen, W., Tang, H., and Lin, H. (2019). Identification of hormone binding proteins based on machine learning methods. Math. Biosci. Eng. 16, 2466–2480.

56. Basith, S., Manavalan, B., Shin, T.H., and 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.

57. Wang, J., Li, J., Yang, B., Leier, A., Hayashida, M., Akutsu, T., Zhang, Y., Chou, K.C., et al. (2019). Bastion3: a two-layer ensemble predictor of type III secreted effectors. Bioinformatics 35, 2017–2028.

58. Wang, J., Yang, B., Leier, A., Marquez-Lago, T.T., Hayashida, M., Rocker, A., Zhang, Y., Akutsu, T., Chou, K.C., Strugnell, R.A., et al. (2018). Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. Bioinformatics 34, 2546–2555.

59. Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. Bioinformatics 36, 3028–3034.

60. Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D.Q. (2018). PredT4SE-Stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. Front. Microbiol. 9, 2571.

61. Kang, X., Hu, L., Shen, P., Li, R., and Liu, D. (2017). SMRT sequencing revealed mitogenome characteristics and mitogenome-wide DNA modification pattern in Ophiocordyceps sinensis. Front. Microbiol. 8, 1422.

62. Liu, G., Jiang, Y.M., Liu, Y.C., Han, L.L., and Feng, H. (2020). A novel DNA methylaton motif identified in Bacillus pumilus BA06 and possible roles in the regulation of gene expression. Appl. Microbiol. Biotechnol. 104, 3445–3457.

63. Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., and Chen, Z. (2017). ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. Molecules 22, 1732.

64. Si, D., Moritz, S.A., Pfläb, J., Hou, J., Cao, R., Wang, L., Wu, T., and Cheng, J. (2020). Deep learning to predict protein backbone structure from high-resolution cryo-EM density maps. Sci. Rep. 10, 4282.

65. Chen, W., Feng, P., Liu, T., and Jin, D. (2019). Recent advances in machine learning methods for predicting heat shock proteins. Curr. Drug Metab. 20, 224–228.

66. Chen, W., Feng, P., Song, X., Lv, H., and Lin, H. (2019). miRNA-m7G: identifying N7-methylguanosine sites by fusing multiple features. Mol. Ther. Nucleic Acids 18, 269–274.

67. Liu, K., and Chen, W. (2020). i6mA-Fuse: improved and robust prediction of DNA 6mA sites in the Rosaceae genome by fusing multiple feature representation. Plant Mol. Biol. 103, 225–234.

68. Liu, Q., Chen, J., Wang, Y., Li, S., Jia, C., Song, J., and Li, F. (2020). DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Brief. Bioinform. Published online July 1, 2020. https://doi.org/10.1093/bib/bbaa124.

69. Xu, H., Jia, P., and Zhao, Z. (2020). Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. Brief. Bioinform. Published online June 24, 2020. https://doi.org/10.1093/bib/bbaa099.

70. Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020). Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. Med. Res. Rev. 40, 1276–1314.

71. Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2020). Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Brief. Bioinform. 21, 408–420.

72. Zhu, X.J., Feng, C.Q., Lai, H.Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowl. Based Syst. 163, 787–793.

73. Liu, M.L., Su, W., Guan, Z.X., Zhang, D., Chen, W., Liu, L., and Ding, H. (2020). An overview on predicting protein subchloroplast localization by using machine learning methods. Curr. Protein Pept. Sci. Published online January 17, 2020. https://doi.org/10.2174/1389203721666200117153412.