Molecular Therapy
Nucleic Acids
Original Article

Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation

Balachandran Manavalan,[1] Shaherin Basith,[1] Tae Hwan Shin,[1,2] Leyi Wei,[3] and Gwang Lee[1,2]

[1]Department of Physiology, Ajou University School of Medicine, Suwon, Republic of Korea; [2]Institute of Molecular Science and Technology, Ajou University, Suwon, Republic of Korea; [3]School of Computer Science and Technology, Tianjin University, China

DNA N4-methylcytosine (4mC) is an important genetic modification and plays crucial roles in differentiation between self and non-self DNA and in controlling DNA replication, cell cycle, and gene-expression levels. Accurate 4mC site identification is fundamental to improve the understanding of 4mC biological functions and mechanisms. Hence, it is necessary to develop *in silico* approaches for efficient and high-throughput 4mC site identification. Although some bioinformatic tools have been developed in this regard, their prediction accuracy and generalizability require improvement to optimize their usability in practical applications. For this purpose, we here proposed Meta-4mCpred, a meta-predictor for 4mC site prediction. In Meta-4mCpred, we employed a feature representation learning scheme and generated 56 probabilistic features based on four different machine-learning algorithms and seven feature encodings covering diverse sequence information, including compositional, physicochemical, and position-specific information. Subsequently, the probabilistic features were used as an input to support vector machine and developed a final meta-predictor. To the best of our knowledge, this is the first meta-predictor for 4mC site prediction. Cross-validation results show that Meta-4mCpred achieved an overall average accuracy of 84.2% from six different species, which is ~2%–4% higher than those attainable using the state-of-the-art predictors. Furthermore, Meta-4mCpred achieved an overall average accuracy of 86% on independent datasets evaluation, which is over 4% higher than those yielded by the state-of-the-art predictors. The user-friendly webserver employed to implement the proposed Meta-4mCpred is freely accessible at http://thegleelab.org/Meta-4mCpred.

INTRODUCTION

DNA methylation is a key epigenetic mark regulating several developmental and pathological processes.[1] The most common post-replicative DNA modification is cytosine methylation, which occurs in the genomes of both prokaryotes and eukaryotes. Cytosine methylation can be mediated enzymatically by DNA methyltransferases, resulting in two epigenetic nucleobases, 5-methylcytosine (5mC) and N4-methylcytosine (4mC), or chemically by endogenous and environmental alkylation agents, resulting in 3-methylcytosine.[1,2] The most well-studied and frequently occurring cytosine methylation, 5mC plays key roles in normal development, genomic imprinting, preservation of chromosome stability, aging, suppression of repetitive element transcription and transposition, and X chromosome inactivation.[3–6] Meanwhile, the least common methylated DNA nucleobase present in bacterial DNA, namely, 4mC, is less studied and explored.[1] Like 5mC, 4mC is a part of restriction-modification systems that protects the host DNA from restriction enzyme-mediated degradation. Additionally, 4mC is involved in supplementary roles, such as correcting DNA replication errors and controlling DNA replication and the cell cycle.[7,8] However, studies on 4mC are relatively limited compared to those on 5mC; hence, its biological functions are yet to be elucidated.

For humans and other eukaryotes, there are major experimental approaches available for identifying epigenetic cytosine nucleobases in DNA. However, only a few analytical approaches are available for studies of bacterial genomes. A popular means of identifying 4mC and N6-methyladenine from unknown DNA sequences is single-molecule real-time sequencing (SMRT).[9] Due to the limited scalability and cost and time effectiveness of this approach, next-generation sequencing techniques have been used. One next-generation sequencing technique that could detect 4mC in genomic DNA is 4mC-Tet-assisted-bisulphite-sequencing.[10] Recently, another group detected 4mC selectively using engineered transcription-activator-like effectors.[1] While these experimental approaches facilitate 4mC site detection, such techniques are too laborious and expensive to be applied for large-scale genome scanning. Hence, it is necessary to develop computational methods for efficient 4mC site prediction.

Recently, computational methods, in particular machine-learning (ML) approaches have expounded efficiently for various problems,[11–14] including 4mC site prediction. Initially, Chen et al.[15]

Received 10 December 2018; accepted 22 April 2019;
https://doi.org/10.1016/j.omtn.2019.04.019.

Correspondence: Leyi Wei, School of Computer Science and Technology, Tianjin University, China
E-mail: weileyi@tju.edu.cn
Correspondence: Gwang Lee, Department of Physiology, Ajou University School of Medicine, Suwon, Republic of Korea
E-mail: glee@ajou.ac.kr

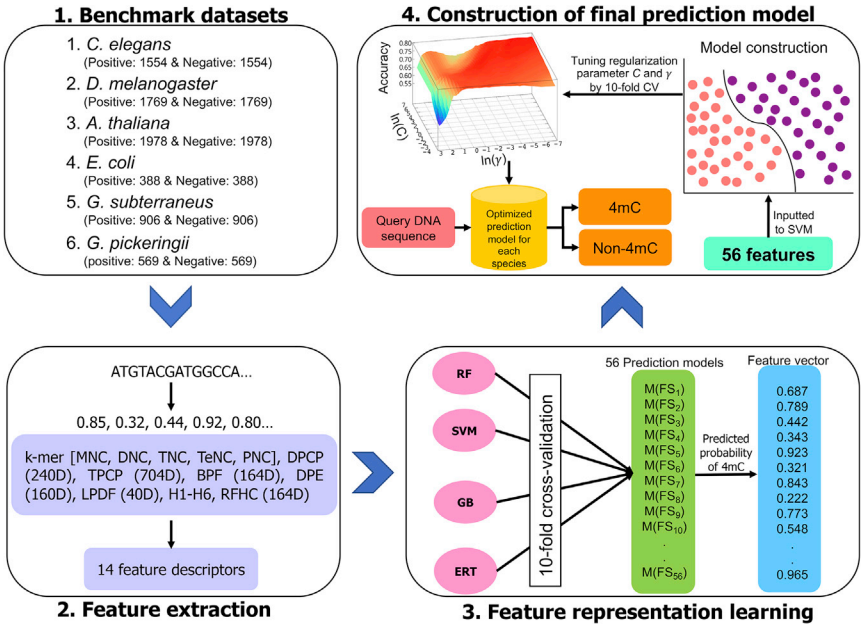

Figure 1. Overall Framework of Meta-4mCpred

Overview of the proposed methodology for predicting 4mCs in multiple species, which involves the following steps: (1) benchmark dataset construction for six different species; (2) extraction of seven feature encodings that characterize different aspects of DNA sequences and generation of 14 feature descriptors; (3) generation of a 56-dimensional feature vector using a feature representation learning scheme; and (4) construction of the final prediction model for each species that separates the input into putative 4mCs and non-4mCs.

developed a support vector machine (SVM)-based tool, iDNA4mC, where nucleotide (NT) chemical properties and frequencies were used as features to build the prediction model. The results demonstrated that the tool predicted 4mC sites from non-4mC sites effectively and showed good performance in cross-species validations. Recently, two novel predictors, 4mCPred[16] and 4mcPred-SVM,[17] were developed for 4mC site identification. In 4mCPred, the position-specific trinucleotide propensity and electron-ion interaction potential were utilized as features and predictive models were constructed using the SVM method. Meanwhile in 4mcPred-SVM, four sequence-based feature descriptors were integrated and a two-step feature optimization protocol was utilized along with an SVM classifier to construct the prediction models. Even though the above-mentioned approaches consistently perform well, they may fail in terms of generalizability, thus demanding the development of a novel predictor for effective 4mC site detection with reliable transferability.

In this report, we propose a novel meta-predictor, Meta-4mCpred, for accurate 4mC site identification. The overall framework of our methodology is shown in Figure 1. First, we employed a feature representation scheme and generated 56 probabilistic features based on four ML algorithms (SVM, random forest [RF], gradient boosting [GB], and extremely randomized tree [ERT] algorithms) and seven feature encodings (*k*-mer composition, binary profile [BPF], dinucleotide binary profile encoding [DPE], local position-specific dinucleotide frequency [LPDF], ring-function-hydrogen-chemical properties [RFHC], dinucleotide physicochemical properties [DPCP], and trinucleotide physicochemical properties [TPCP]). Second, we inputted these probabilistic features into an SVM and developed a final prediction model. During cross-validation, Meta-4mCpred achieved the best average accuracy of 84.2% when compared to the state-of-the-art predictors. Furthermore, our method significantly outperformed the existing predictors on independent datasets, with an average accuracy of 86.0%. This characteristic represents the greatest advantage of our approach, highlighting the superior generalizability of our model. To the best of our knowledge, this study is the first in which a meta-based approach has been applied for 4mC site prediction. Henceforth, we believe that our approach will be useful and reliable for predicting 4mC sites and could be utilized for data from other species as well.

RESULTS AND DISCUSSION

Evaluation of Various Classifiers on Feature Learning Models

In this study, we generated 14 feature descriptors using seven different feature encodings (Table S1) that represents sequence information in different perspective. To examine each feature descriptor contribution in classifying 4mCs from non-4mCs, we conducted a 10-time randomized 10-fold cross-validation (CV) test for each feature descriptor by employing six commonly used ML algorithms or classifiers, namely, SVM, RF, ERT, GB, AdaBoost (AB), and *k*-nearest neighbor (KNN) algorithms. We obtained 84 prediction models for each species using six different ML algorithms and 14 feature descriptors. In total, 504 prediction models (84 × 6) were obtained for multiple species, whose performances are shown in Figure 2. Our results revealed that four feature sets (FSs), namely, F6 (BPF), F7 (RFHC), F8 (a combination of DPE and LPDF), and F14 (a combination of BPF and RFHC), produced significantly better performance in each species regardless of the ML algorithm, when compared to the remaining 10 features, indicating that NT profiles and ring function properties appeared to be the most powerful encodings in 4mC site prediction. However, the remaining properties also contributed to a certain extent with slightly lower accuracy (ACC), which could be still regarded as useful descriptors because they represent complementary features from a different perspective. Next, we examined the best performance of individual ML classifiers, where RF, SVM, GB, and AB algorithms achieved their highest ACC values using F6 features; however, the ERT and KNN algorithms produced their highest ACC values using F14 in multiple species. Regarding overall performance for multiple

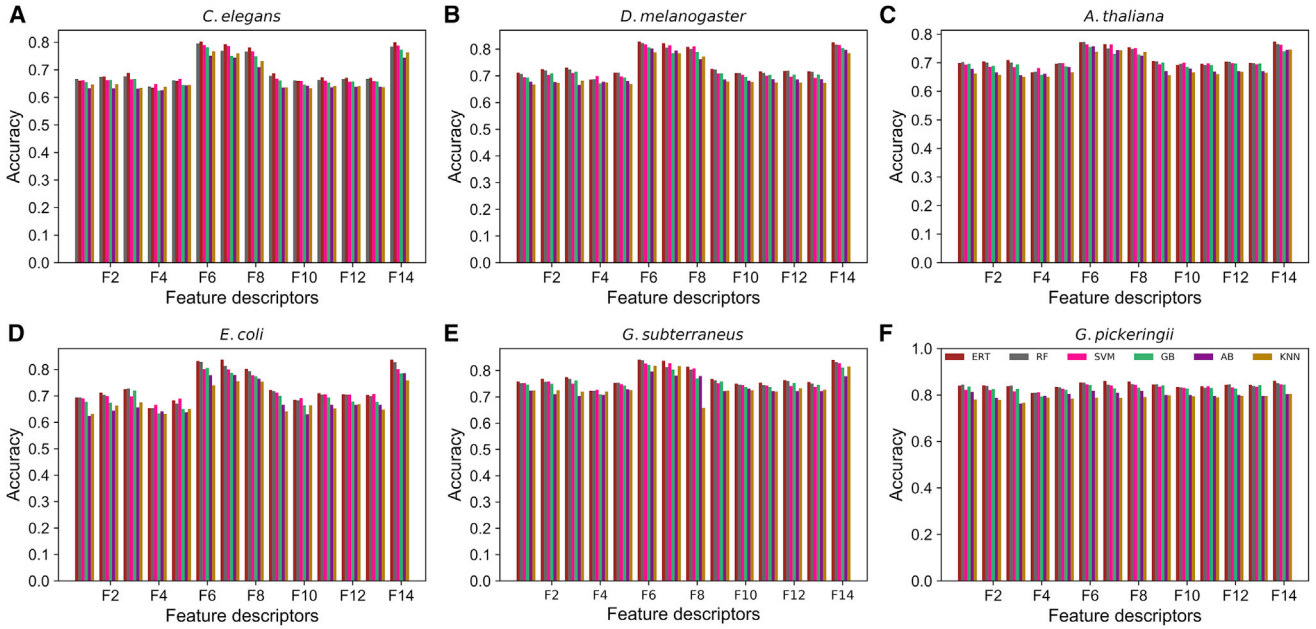

Figure 2. Accuracies of the Six Different ML Classifiers in Distinguishing between 4mCs and Non-4mCs with Respect to 14 Feature Descriptors
(A) *C. elegans*, (B) *D. melanogaster*, (C) *A. thaliana*, (D) *E. coli*, (E) *G. subterraneus*, and (F) *G. pickeringii*.

species, the ERT, RF, SVM, GB, AB, and KNN algorithms, respectively, achieved average ACC values of 82.5%, 82.0%, 81.0%, 80.2%, 78.2%, and 78.0%, indicating that the predictive model trained with the ERT classifier and F14 descriptor had more discriminative power in 4mC and non-4mC classification.

Instead of selecting the best model from Figure 2 for each species, we used all of the model outputs for meta-predictor construction and thereby considered diverse and complementary sequence information. As we employed six different ML algorithms, it was necessary to determine which algorithm-based prediction model output was better suited in developing meta-predictor. To this end, we examined the overall performance of each method. We found that the overall performance topologies of the ERT, GB, RF, and SVM algorithms were mostly similar for multiple species (Figure 2) and were better than those of the other two methods (the KNN and AB algorithms). Therefore, we considered the outputs of only four ML models (the ERT, RF, GB, and SVM models) for further analysis.

Meta-4mCpred Construction

Generally, meta-predictors take input from the outputs of different predictors under the assumption that the combined method will provide more accurate results than a single predictor.[18–21] As mentioned above, we considered only four ML-based algorithms, whose predicted 4mC site probabilities were used as inputs for meta-predictor construction. Specifically, we obtained 56 prediction models from these four methods, where each method contained exactly 14 prediction models. The predicted 4mC site probabilities acquired from these 56 models were given as inputs to the SVM algorithm, and a final model was developed for each species, whose corresponding performances are shown in Table 1. In addition to the SVM method, we explored five other ML methods (the RF, ERT, GB, AB, and KNN methods), whose performances are listed in Table S2. Unlike the baseline prediction performances, the overall performances exhibited no significant differences among the six ML algorithms; however, the SVM algorithm was slightly superior to the other methods with an overall average ACC ~1% higher than those obtained using the RF, ERT, GB, and AB algorithms and ~2% higher than that resulting from using the KNN method. Hence, we selected SVM-based model for each species and named our developed meta-predictor Meta-4mCpred.

To demonstrate the advantages of our meta-predictor, we compared its performance with that of the best model obtained from the baseline predictors. Figure 3 shows that the overall average ACC obtained using Meta-4mCpred is ~2%, 2.3%, 3.4%, 4%, 5.7%, and 6.2% higher than those resulting from using the ERT, RF, SVM, GB, AB, and KNN methods, respectively, thus highlighting the superiority of our proposed method.

Feature Contribution Analysis

The improved performance of Meta-4mCpred is mainly due to the features obtained through the feature learning scheme. To understand this phenomenon, we computed the t-distributed stochastic neighbor embedding (t-SNE) implemented in Scikit with the default parameters (n_components = 2, perplexity = 30, and learning rate = 1,000) for each feature encoding. Basically, we

Table 1. Performance of Meta-4mCpred on Benchmark Dataset

Species	MCC	ACC	Sn	Sp	AUC
C. elegans	0.652	0.826	0.840	0.812	0.892
D. melanogaster	0.685	0.842	0.831	0.854	0.904
A. thaliana	0.584	0.792	0.761	0.822	0.861
E. coli	0.697	0.848	0.869	0.827	0.911
G. subterruneus	0.711	0.855	0.856	0.854	0.904
G. pickeringii	0.782	0.891	0.884	0.898	0.951

MCC, Matthews correlation coefficient; ACC, accuracy; Sn, sensitivity; Sp, specificity; AUC, area under curve.

compared 56 probabilistic feature vector with the top five individual feature descriptors that exhibited consistent performance in the baseline prediction (BPF, RFHC, DPE+LPDF, DPCP, and TPCP). Figure 4 shows the distributions of the positive and negative samples in the *Geobacter pickeringii* dataset in a two-dimensional space. Figures 4A–4E depict the 4mC and non-4mC sites of five feature descriptors, where the positive and negative samples overlap in the feature space, indicating that the original feature is less capable of discriminating between the positive and negative samples. Conversely, there is a clear distinction between the positive and negative samples for the 56-dimensional vector, although a few samples overlap (Figure 4F). This result demonstrates that 4mCs and non-4mCs present in a 56-dimensional vector can be differentiated more easily than when using other feature spaces, thus enhancing the performance. Furthermore, we computed t-SNE distributions for the other five species (Figures S1–S5) and observed trends similar to those resulting from using the *G. pickeringii* dataset. Our feature learning protocol proved effective due to the easy transformation from a high-dimensional feature space into a low-dimensional one, thereby expediting the prediction process and extending its applicability to genome-wide predictions.

Comparison of Meta-4mCpred with the State-of-the-Art Predictors

We compared the performance of Meta-4mCpred with three state-of-the-art predictors, namely, iDNA4mC, 4mcPred-SVM, and 4mCPred, which were developed using the same benchmark datasets. The prediction performances reported for iDNA4mC[15] and 4mcPred-SVM[17] were utilized as such for the comparison. Meanwhile, Wei et al.[17] found that the predictions reported for 4mCPred[16] might have been over-estimates; hence, they rebuilt those models and reported the performance of 4mcPred-SVM. Therefore, we used the same values for 4mCPred as were reported for 4mcPred-SVM for the comparison.

Table S3 and Figure 5 show the performances of the various methods on the benchmark datasets, where Meta-4mCpred performed better than the existing methods both in terms of Matthews correlation coefficient (MCC) and ACC for five out of six species (*Drosophila melanogaster*, *Arabidopsis thaliana*, *Escherichia coli*, *Geoalkalibacter subterraneus*, and *G. pickeringii*). However, in the case of *Caenorhabditis elegans*, the performance of Meta-4mCpred is identical to that of 4mCPred. The most notable improvements by Meta-4mCpred are observable for four species in terms of both MCC and ACC. Our method achieved ACC and MCC values respectively 3.1% and 6.1% higher for *G. pickeringii*, 1.8% and 3.7% higher for *G. subterraneus*, 1.5% and 3.1% higher for *E. coli*, and 1.2% and 2.4% higher for *D. melanogaster* than the second-best predictor, 4mcPred-SVM. Surprisingly, all of these predictors are based on the SVM approach; however, the features used in each method are entirely different. For instance, iDNA4mC uses RFHC;[15] 4mcPred-SVM uses partial information about *k*-mer composition, BPF, DPE, and LPDF;[17] and 4mCPred uses the position-specific trinucleotide propensity.[16] Meanwhile, Meta-4mCpred uses 56 probabilistic features obtained from a feature learning scheme based on four different ML algorithms and various features, including most of the existing features (k-mer, BPF, DPE, LPDF, and RFHC) and newly explored ones (DPCP and TPCP). It is reasonable to assume that our features are more discriminative than the previously used features, enabling the key characteristics distinguishing 4mCs from non-4mCs to be captured and better prediction to be achieved.

Performance Assessment of Various Tools Based on the Independent Datasets

To check the prediction model's generalization ability or robustness, it is essential to evaluate these models on independent datasets. To make a fair comparison, we included only three methods, including Meta-4mCpred, 4mCPred, and 4mcPred-SVM, where each method has a separate prediction model for each species. The reason for excluding iDNA4mC from this evaluation is that it has only one prediction model made available in the web server.

Table 2 shows the performances of three methods on the independent datasets, where Meta-4mCpred performed better than the existing methods both in terms of MCC and ACC for four out of six species (*A. thaliana*, *D. melanogaster*, *G. subterraneus*, and *G. pickeringii*). However, in the case of *C. elegans* and *E. coli*, Meta-4mCpred and 4mCPred showed a similar performance. The most notable improvements by Meta-4mCpred are observable for three species in terms of both MCC and ACC. Our method achieved ACC and MCC values, respectively 3.9% and 7.6% higher for *G. subterraneus*, 9.2% and 18.5% higher for *G. pickeringii*, and 3.1% and 6.2% higher for *A. thaliana*, than the second-best predictor, 4mcPred-SVM. Furthermore, McNemar's chi-square test[22] was applied to find the statistical significance between Meta-4mCpred and the existing predictors. At a p value threshold of 0.05, Meta-4mCpred significantly outperformed other two methods in three species (*G. subterraneus*, *G. pickeringii*, and *A. thaliana*) and significantly outperformed only 4mCpred in the remaining two out of three species (*C. elegans* and *D. melanogaster*). In terms of overall performance, existing methods, such as 4mcPred-SVM and 4mCPred, achieved a similar performance with an average accuracy of 81.6% and 82.1%. However, the corresponding value of Meta-4mCpred is 86%, indicating significant improvement over

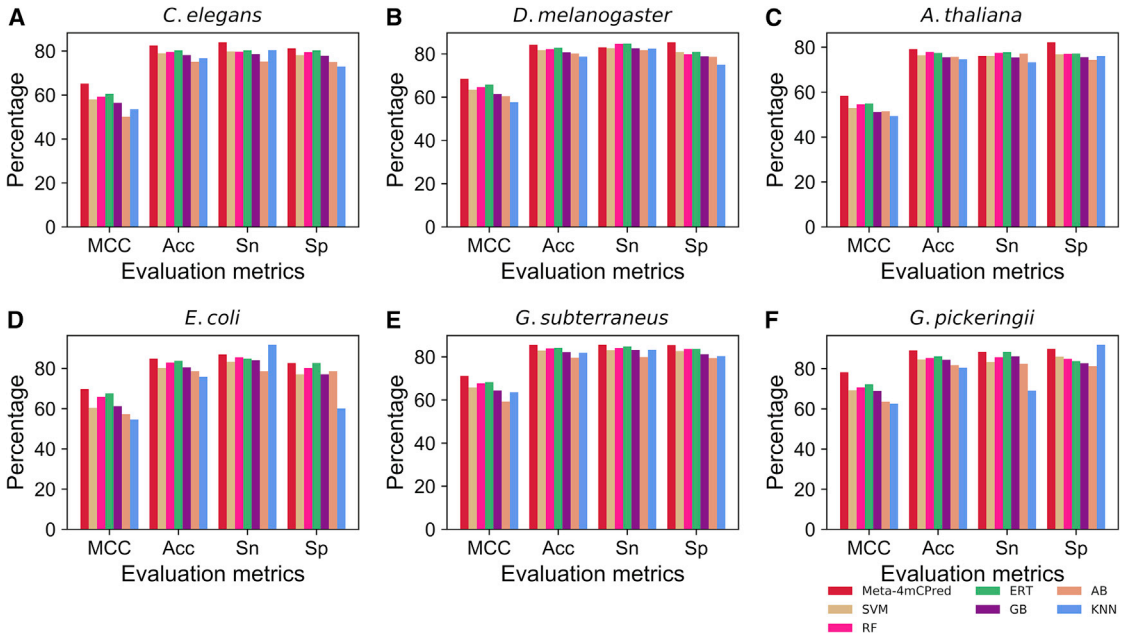

Figure 3. Performance Comparison of Meta-4mCpred and Baseline Predictors from Six Different ML Algorithms in terms of MCC, ACC, Sn, and Sp
(A) *C. elegans*, (B) *D. melanogaster*, (C) *A. thaliana*, (D) *E. coli*, (E) *G. subterraneus*, and (F) *G. pickeringii*.

the existing methods. The significant improvement of Meta-4mCpred is mainly due to the following characteristics: (1) our feature learning model integrates not only NT composition and NT position-specific information, but also physicochemical properties and ring function, which provide diverse sequence information that can be utilized to construct effective feature representation models, and (2) the final model uses 4mC site prediction probabilities from the original feature descriptors, thereby reducing the actual high-dimensional feature space into a low-dimensional feature space with more discrimination between positive and negative samples.

Web Server Implementation

Generally, user-friendly web servers have been helpful for experimentalists, where they can do the prediction without going through mathematical equations, and also it represents the future direction for developing novel and more useful predictors.[23] Indeed, it has been demonstrated by a series of publications.[24–27] Therefore, we established a user-friendly webserver, Meta-4mCpred, for use by a wider research community. This web server is freely accessible at http://thegleelab.org/Meta-4mCpred. Below, we provide step-by-step guidelines on how to use our web server to obtain the predicted outcomes. First, the user chooses the desired species. Second, the user enters the query sequences into the input box. Note that the input sequences should be in FASTA format. Examples of FASTA-formatted sequences can be seen by clicking on the FASTA format button located above the input box. Finally, clicking on the "submit" button provides the predicted results as output.

Conclusions

In this study, we developed a novel meta-predictor for 4mC site prediction called Meta-4mCpred. To build an efficient predictive model, we applied a feature representation learning scheme and generated 56 probabilistic features based on four different ML algorithms and seven feature encodings covering diverse sequence information, including compositional, physicochemical, and NT position-specific information. Subsequently, these features were used as SVM input and a final meta-predictor was developed. Indeed, this is the first meta-predictor for 4mC site prediction. Furthermore, the 56 features obtained from the feature learning scheme are more capable of discriminating between 4mC and non-4mC in the feature space, thus providing significant improvement compared to several currently available feature descriptors.

We further compared the performance of the proposed predictor with those of three state-of-the art predictors (iDNA4mC, 4mcPred-SVM, and 4mCPred) both on a benchmark and independent datasets. The results show that the overall performance of Meta-4mCpred was better than those of the other methods on the benchmark datasets and significantly better in independent evaluation, indicating that the proposed method is more effective and promising for 4mC site identification. As an application of this work, we made our web server publicly available for the wider community to use. We expect that Meta-4mCpred will be a useful and reliable computational tool for predicting 4mC sites and facilitating DNA methylation analysis. The scheme employed in our current method is a general one that can be employed to address various sequence-based prediction problems, including enhancer

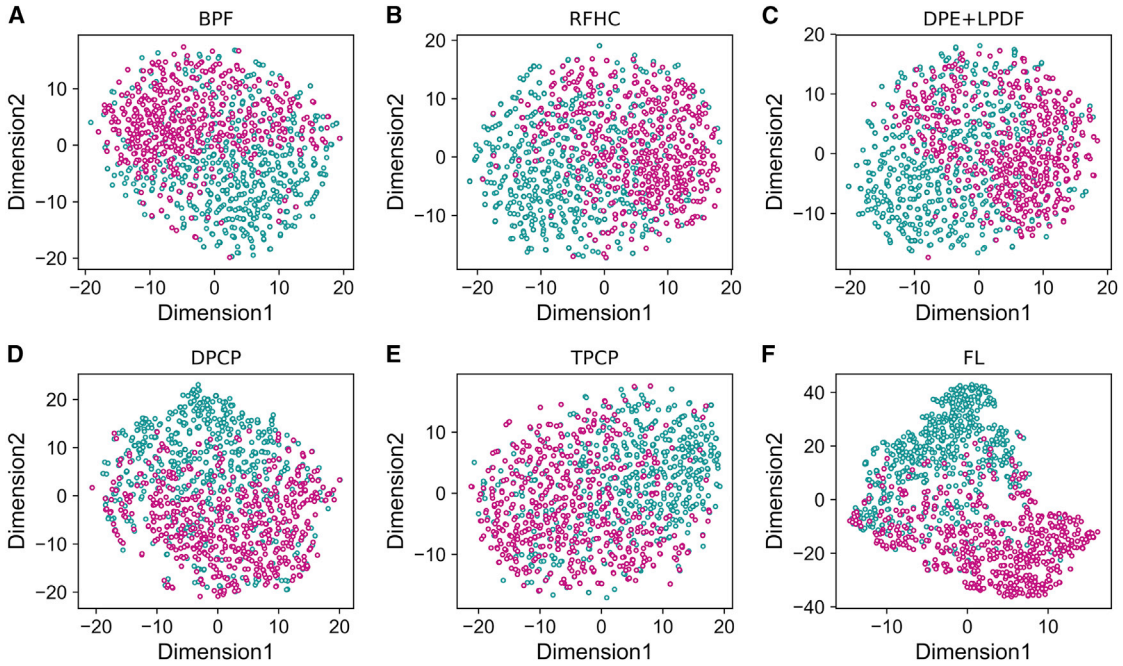

Figure 4. t-SNE Visualization of the *G. pickeringii* Dataset in a Two-Dimensional Feature Space

The orange circles and sky-blue diamonds represent 4mCs and non-4mCs, respectively. (A) BPF, (B) RFHC, (C) DPE+LPDF, (D) DPCP, (E) TPCP, and (F) the 56-dimensional feature obtained by feature learning (FL)

prediction,[28] recombination hotspot prediction,[29] transcriptional terminator prediction,[30] and protein function prediction.[31,32] Furthermore, our method could be integrated with genomic features extracted from RNA-sequencing (RNA-seq)[33] and chromatin immunoprecipitation (ChIP)-seq,[34] and exploring other powerful ML algorithms[35] will greatly improve the 4mC predictions.

MATERIALS AND METHODS

A flowchart of the Meta-4mCpred methodology is shown in Figure 1 and consists of four major steps: (1) benchmark dataset construction; (2) extraction of features that represent the different aspects of the sequence information; (3) feature representation learning; and (4) construction of the meta-predictor for each species. These major steps are described individually in the following sections.

Dataset Construction

We utilized the datasets constructed by Chen et al.,[15] which were specifically used to classify 4mCs and non-4mCs. The reasons for considering these datasets are as follows: (1) the authors constructed reliable datasets based on the MethSMRT database;[36] (2) the datasets are nonredundant and none of the sequences share more than 80% of their pairwise sequence identities with other sequences, thereby avoiding overestimation in the computational model; and (3) these datasets enabled fair comparison between the proposed method and the existing method, which was developed using the same datasets. These datasets contain 14,328 sequences derived from six different species. Of those, *C. elegans*, *D. melanogaster*, *A. thaliana*, *E. coli*, *G. subterraneus*, and *G. pickeringii* contain equal numbers of positive (4mC 1554, 1769, 1978, 388, 906, and 569, respectively) and negative (non-4mC) samples. All of the positive and negative samples are 41 bp long with cytosine located at the central position. It should be noted that we excluded one positive sample from *G. subterraneus* because it had a non-standard bp and considered the remaining 14,327 sequences.

To evaluate our prediction models along with the existing methods, we constructed the independent datasets for six different species using the same protocol as mentioned in previous study.[15] The positive samples for six species obtained from MethSMRT, where each positive sample containing modification QV score greater than 30, indicating a position as modified. Finally, we obtained 750, 1,000, 1,250, 134, 350, and 200 4mCs, respectively, from *C. elegans*, *D. melanogaster*, *A. thaliana*, *E. coli*, *G. subterraneus*, and *G. pickeringii* genomes. Furthermore, the positive samples were supplemented with equal numbers of negative samples for each species using the same procedure as mentioned in a previous study.[15] Notably, none of these positive and negative samples from each species share a sequence identity of greater than 70% within each species of independent dataset and also benchmark dataset.

DNA Feature Representation

An NT sequence is represented as

$$D = b_1, b_2, b_3, \ldots b_L, \qquad \text{(Equation 1)}$$

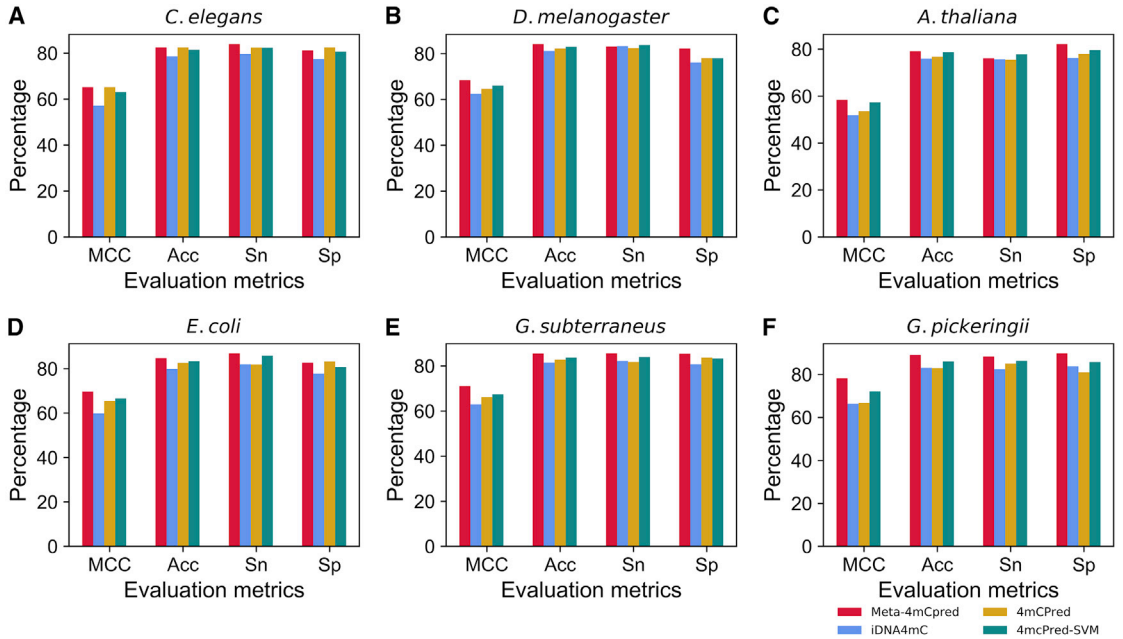

Figure 5. Performance Comparison of Meta-4mCpred and Three State-of-the-Art Predictors on Six Benchmark Datasets from Multiple Species
(A) *C. elegans*, (B) *D. melanogaster*, (C) *A. thaliana*, (D) *E. coli*, (E) *G. subterraneus*, and (F) *G. pickeringii*.

where b_1, b_2, and b_3, respectively, denote the first, second, and third base pairs in the DNA sequence, and so forth, and L denotes the NT sequence length. Note that base pair b_i is an element of the standard NTs (adenine [A], thymine [T], guanine [G], and cytosine [C]). In this study, we explored various features, including k-mer composition, BPF, DPE, LPDF, RFHC, DPCP, and TPCP, which cover various aspects of the sequence information and can be described as follows.

k-mer NT Composition

Generally, the frequency of a k-tuple of NTs is one way of representing DNA sequences that has been widely used as an input feature in various prediction problems.[37–39] In this study, we considered mono- (MNC), di- (DNC), tri- (TNC), tetra- (TeNC), and penta-nucleotide compositions (PNC), respectively encoded as vectors containing 4, 16, 64, 256, and 1,024 elements.

BPF

As mentioned above, there are four different NTs in the standard DNA alphabet. Each NT type is encoded with a feature vector (FV) composed of 0 and 1. Specifically, A is encoded as P(A) = (1, 0, 0, 0), T is encoded as P(T) = (0, 1, 0, 0), G is encoded as P(G) = (0, 0, 1, 0), and C is encoded as P(C) = (0, 0, 0, 1). Subsequently, for a given DNA sequence D with a length of k (k = 41),[17,40] the base pairs can be encoded using the following FV:

$$BPF(k) = [P(b_1), P(b_2), P(b_3), \ldots P(b_L)]. \quad \text{(Equation 2)}$$

Thus, the dimension of $BFP(k)$ is 4 × 41 = 164 features.

DPE

In DPE,[17,40] each dinucleotide type is encoded as a four-dimensional vector containing 0 and 1. For instance, AA is encoded as (0, 0, 0, 0), AC is encoded as (0, 0, 1, 0), AT is encoded as (0, 0, 0, 1), and so on. Therefore, the dimension of DPE for a given DNA sequence is a 160 (4 × 40)-dimensional vector.

LPDF

The LPDF can be calculated as follows:

$$f = \frac{1}{|N_i|} C(X_{i-1}X_i), 2 \le i \le L, \quad \text{(Equation 3)}$$

where $|N_i|$ is the length of the i^{th} prefix string $\{X_1X_2X_3 \ldots X_i\}$ in the given sequence and $C(X_{i-1}X_i)$ is the occurrence number of dinucleotide $X_{i-1}X_i$ in position i of the i^{th} prefix string. The LPDF is encoded as 40-dimensional vector for a given DNA sequence.[17,40]

RFHC

DNA consists of four NTs (A, T, G, and C) that have different chemical properties based on their rings, functional groups, and hydrogen bonds.[15,21,41–43] In terms of ring structure, the purines (A and G) and pyrimidines (C and G), respectively, contain two rings and one ring. In terms of secondary structures, A and T form weak hydrogen bonds and are allotted to one group, whereas C and G form strong hydrogen bonds and are allotted to another group. Regarding chemical functionality, A and C can be assigned to the amino group, while G and T can be assigned to the keto group. To convert these properties into FVs, three coordinates (x, y, z) were used to represent the

Table 2. Performances of the Proposed Meta-4mCpred and Two State-of-Art Predictors, 4mCPred and 4mcPred-SVM, on Six Independent Datasets from Different Species

Species	Predictors	MCC	ACC	Sn	Sp	TP	FN	FP	TN	p Value
C. elegans	4mCPred	0.731	0.865	0.883	0.849	666	84	118	632	0.670
	4mcPred-SVM	0.684	0.842	0.828	0.856	621	129	108	642	0.001*
	Meta-4mCpred	0.741	0.870	0.843	0.897	632	118	77	673	–
D. melanogaster	4mCPred	0.803	0.900	0.933	0.868	933	67	132	868	0.465
	4mcPred-SVM	0.771	0.886	0.886	0.885	886	114	115	885	0.030*
	Meta-4mCpred	0.812	0.906	0.913	0.899	913	87	101	899	–
A. thaliana	4mCPred	0.632	0.816	0.842	0.789	1,053	197	264	986	<0.00001*
	4mcPred-SVM	0.649	0.824	0.842	0.806	1,053	197	242	1,008	<0.00001*
	Meta-4mCpred	0.711	0.855	0.876	0.834	1,095	155	207	1,043	–
E. coli	4mCPred	0.634	0.817	0.851	0.784	114	20	29	105	0.887
	4mcPred-SVM	0.569	0.784	0.746	0.821	100	34	24	110	0.132
	Meta-4mCpred	0.650	0.825	0.806	0.843	108	26	21	113	–
G. subterruneus	4mCPred	0.578	0.789	0.757	0.820	265	85	63	287	<0.00001*
	4mcPred-SVM	0.624	0.811	0.783	0.840	274	76	56	294	<0.00001*
	Meta-4mCpred	0.701	0.850	0.817	0.883	286	64	41	309	–
G. pickeringii	4mCPred	0.503	0.742	0.610	0.875	122	78	25	175	<0.00001*
	4mcPred-SVM	0.515	0.758	0.750	0.765	150	50	47	153	<0.00001*
	Meta-4mCpred	0.700	0.850	0.835	0.865	167	33	27	173	–

MCC, Matthews correlation coefficient; ACC, accuracy; Sn, sensitivity; Sp, specificity; TP, true positive; FN, false negative; FP, false positive; TN, true negative. The last column represents McNemar's Chi-squared test, which was used to evaluate the performance between Meta-4mCpred and other methods. *A p value < 0.05 was considered to indicate a statistically significant difference between Meta-4mCpred and the selected method.

chemical properties of the four NTs and values of 0 and 1 were assigned to the coordinates. The three coordinates respectively describe the ring structure, hydrogen bond, and chemical functionality, where each NT can be encoded as follows:

$$x_i = \begin{cases} 1, & if\ S_i \in \{A, G\} \\ 0, & if\ S_i \in \{T, C\} \end{cases}, y_i = \begin{cases} 1, & if\ S_i \in \{A, T\} \\ 0, & if\ S_i \in \{C, G\} \end{cases}, \quad z_i = \begin{cases} 1, & if\ S_i \in \{A, C\} \\ 0, & if\ S_i \in \{T, G\} \end{cases}, . \qquad \text{(Equation 4)}$$

Therefore, A, C, G, and T can be represented by the coordinates (1, 1, 1), (0, 0, 1), (1, 0, 0), and (0, 1, 0), respectively.

To include the NT compositions surrounding 4mC or non-4mC sites, the density method was employed to measure the importance between frequency and position, using the following definition:

$$d_i = \frac{1}{|N_i|} \sum_{j=1}^{L} f(n_j), f(n_j) = \begin{cases} 1, & if\ n_j = q \\ 0, & otherwise \end{cases}, \qquad \text{(Equation 5)}$$

where d_i is the density of NT i, $|N_i|$ is the length from the current NT position to the first NT, and q is any one of the four standard NTs. By integrating the NT chemical properties and NT composition (combining Equations 4 and 5), a 41-NT sequence will be encoded as a 164 (4 × 41)-dimensional vector.

DPCP

In this study, we used 15 physicochemical properties: PC1, F-roll; PC2, F-tilt; PC3, F-twist; PC4, F-slide; PC5, F-shift; PC6, F-rise; PC7, roll; PC8, tilt; PC9, twist; PC10, slide; PC11, shift; PC12, rise; PC13, energy; PC14, enthalpy; and PC15, entropy. Table S4 summarizes the values of these 15 physicochemical properties for each dinucleotide, which were normalized to the range of [0, 1] according to the formula described in Manavalan et al. [44] prior to the following calculation. The DPCP can be formulated as follows:

$$DPCP(i) = normalized\ frequency\ of\ dinucleotide(i) \times PC(X_i), \qquad \text{(Equation 6)}$$

where X is one of the 15 physicochemical properties, and i is one of the 16 dinucleotides. The DPCP are encoded as a 240 (16 × 15)-dimensional vector.

TPCP

We used the following 11 physicochemical properties: PC1, bendability (DNase); PC2, bendability (consensus); PC3, trinucleotide GC content; PC4, nucleosome positioning; PC5, consensus (roll); PC6, consensus (rigid); PC7, DNase I (rigid); PC8, molecular weight (daltons); PC9, nucleosome (rigid); PC10, nucleosome; and PC11, DNase I. Table S5 shows the values of these 11 physicochemical properties for each trinucleotide, which were normalized as described

above prior to the following calculation. The TPCP can be formulated as follows:

$$TPCP(i) = normalized\ frequency\ of\ trinucleotide(i) \times PC(X_i), \quad \text{(Equation 7)}$$

where X is one of 11 physicochemical properties, and i is one of the trinucleotides. The TPCP are encoded as a 704 (64 × 11)-dimensional vector.

ML Algorithms Implemented in Meta-4mCpred

Meta-4mCpred utilizes four different ML algorithms, namely, the SVM, RF, ERT, and GB algorithms, which were implemented using the Scikit-Learn package (v0.18).[45] Brief descriptions of these methods and how they were used in this study are provided in the following sections.

SVM

The SVM algorithm is one of the most widely used ML algorithms in computational biology.[20,39,42,43,46–51] It finds the optimal hyperplane with the largest margin that minimizes the misclassification rate.[52] Basically, the given input features are mapped into a high-dimensional space using kernel functions, and a hyperplane is found that maximizes the distance between the hyperplane and two classes. We experimented with different kernel functions, including linear functions, polynomial functions, and Gaussian radial basis functions (RBFs) and found that the RBF kernel was appropriate for this problem. Two critical parameters, C (controls the trade-off between the training error and margin) and γ (controls how peaked Gaussians are centered on the support vectors), require optimization in the RBF-SVM algorithm. Therefore, we optimized these parameters using the following ranges:

$$\begin{cases} 2^{-5} \le C \le 2^{15}\ with\ step\ \Delta C = 2 \\ 2^{-15} \le \gamma \le 2^{-5}\ with\ step\ \Delta\gamma = 2^{-1} \end{cases} \quad \text{(Equation 8)}$$

RF

The RF algorithm[53] is one of the most popular ML algorithms and has been widely applied in computational biology and bioinformatics.[44,49,54–57] It utilizes an ensemble of decision trees to perform both classification and regression. In the RF algorithm, three key parameters are the number of trees (*ntree*), the number of randomly selected features (*mtry*), and the minimum number of samples required to split an internal node (*nsplit*). A grid search was employed to fine-tune these parameters with the following search space:

$$\begin{cases} 50 \le ntree \le 2000\ with\ step\ \Delta ntree = 25 \\ 1 \le mtry \le 15\ with\ step\ \Delta mtry = 1 \\ 1 \le nsplit \le 12\ with\ step\ \Delta nsplit = 1 \end{cases}. \quad \text{(Equation 9)}$$

ERT

The ERT algorithm is a commonly used ML algorithm and utilizes an ensemble of decision trees to solve classification and regression problems.[58] It has been applied to solve numerous biological problems.[49,55,59,60] The objective of the ERT algorithm is to decrease the prediction model variance further by considering randomization techniques. Although the working principle of the ERT algorithm is similar to that of the RF algorithm, it has the following differences: (1) the ERT algorithm utilizes all of the input data to construct a tree instead of the bagging procedure applied in the RF algorithm and (2) unlike in the RF algorithm, the node selection for splitting is fully random in the ERT algorithm. Grid searches were performed by evaluating various combinations of three regularization parameters, namely, *ntree*, *mtry*, and *nsplit*, using the benchmark dataset and 10-fold CV. The search space for *ntree*, *mtry*, and *nsplit* is as follows:

$$\begin{cases} 40 \le ntree \le 1000\ with\ step\ \Delta ntree = 20 \\ 1 \le mtry \le 15\ with\ step\ \Delta mtry = 1 \\ 1 \le nsplit \le 10\ with\ step\ \Delta nsplit = 1 \end{cases}. \quad \text{(Equation 10)}$$

GB

GB[61] is a forward learning ensemble approach, which is suitable for both classification and regression problems. The final strong prediction models given by GB based on ensembles of weak models (decision trees) have been widely used in bioinformatics.[55,62] GB consecutively fits new models to provide more accurate response variable estimates than other ensemble methods, such as the RF and ERT algorithms. In GB, the three most influential parameters are *ntree*, *mtry*, and *nsplit*, which were optimized using the following search space:

$$\begin{cases} 50 \le ntree \le 1000\ with\ step\ \Delta ntree = 25 \\ 1 \le mtry \le 10\ with\ step\ \Delta mtry = 1 \\ 1 \le nsplit \le 6\ with\ step\ \Delta nsplit = 1 \end{cases}. \quad \text{(Equation 11)}$$

CV

In general, three CV methods are often used to evaluate the anticipated success rate of a predictor: independent dataset, sub-sampling (or *k*-fold CV), and jackknife tests. Among these, the jackknife test is recognized as the least arbitrary and most objective one, as demonstrated by Equations 28–32 in Chou[63], and hence has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors.[15,46,64–74] In the jackknife test, each sequence in the training dataset is singled out as an independent test sample in turn and all of the rule parameters are calculated, excluding the one being identified. To reduce the computational time, we adopted 10-fold CV, as employed in previous studies.[17,55,75,76] In 10-fold CV, a dataset is first randomly partitioned into 10 subsets of equal size. Of these, nine subsets are chosen as training data to train a predictive model, while the remaining subset is retained as validation data to test the model. This process is repeated 10 times, with each of the 10 subsets used exactly once as the validation data. Finally, the 10 results are averaged to obtain a final prediction.

Feature Representation Learning Scheme

Feature learning scheme has been successfully implemented in various sequence-based prediction problems, including anticancer peptide,[20] cell-penetrating peptide,[19] quorum-sensing peptide,[77] and antihypertensive peptide[18] predictions. The same protocol was employed in this study, representing its first application to DNA sequences, as described in the following sections.

Initial Feature Pool Generation

As mentioned above, we extracted seven feature encoding schemes based on the composition, physicochemical properties, and profiles, including *k*-mer composition, BPF, DPE, LPDF, RFHC, DPCP, and TPCP. For *k*-mer composition, there were five different FSs; MNC, DNC, TNC, TeNC, and PNC). Most of these features were used as such, and a set of hybrid features was generated based on different combination of the above feature encodings. Finally, we generated 14 FSs, which are listed in Table S1. For clarity, the j^{th} FS is represented as FS*j* (j = 1, 2, 3, …, 14).

Feature Learning Models

For each FS*j* (j = 1, 2, 3, …, 14), the following four ERT-, RF-, SVM-, and GB-based prediction models were developed, represented as ML(FS*j*), using the benchmark dataset and 10-fold CV. Generally, one application of 10-fold CV could produce biased ML parameters. Therefore, we applied 10-fold CV three more times by random partitioning and considered the median values as the optimal ML parameters. Finally, we obtained 56 prediction models (14 × 4 ML algorithms) and considered them as the baseline models.

Learning a New FV for Meta-Predictor Construction

For a given DNA sequence *D*, we used each baseline model ML(FS_j) to predict the probability of 4mCs, whose value was between 0 and 1. The probability predicted using each model was subsequently employed as a feature. In our experiment, predicted probabilities ≥ 0.5 were designated as 4mCs, and the others were non-4mCs. Finally, *D* was encoded with a new FV by concatenating all of t*he* features generated by the 56 models, which can be represented as

$$FV(D) = Y(P, ML(FS_1)), Y(P, ML(FS_2)), \ldots Y(P, ML(FS_{S6})) \ . \quad \text{(Equation 12)}$$

Here, $FV(D)$ is the FV for a given *D*, and $Y(P, ML(FS_j))$ is the prediction probability of each model for *D*. Finally, FV contains 56 probabilistic features, which was subsequently used as input to the SVM and developed the final meta predictor separately for each species.

Performance Evaluation

We used four different measures that are commonly used in binary classification tasks to evaluate the performances of the models:[46,65,78–80] sensitivity, *Sn*; specificity, *Sp*; accuracy, *ACC*; and the Matthews correlation coefficient, *MCC*. These measures can be calculated as follows:

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{ACC} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} \\ \mathrm{MCC} = \dfrac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \end{cases}, \quad \text{(Equation 13)}$$

where *TP* is the number of true positives, i.e., 4mCs classified correctly as 4mCs; *TN* is the number of true negatives, i.e., non-4mCs classified correctly as non-4mCs; *FP* is the number of false positives, i.e., 4mCs classified incorrectly as non-4mCs; and *FN* is the number of false negatives, i.e., non-4mCs classified incorrectly as 4mCs.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtn.2019.04.019.

AUTHOR CONTRIBUTIONS

B.M., L.W., and G.L. conceived the project and designed the experiments. B.M., S.B., and T.S. performed the experiments and analyzed the data. B.M., S.B., L.W., and G.L. wrote the manuscript. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education, Science, and Technology (2018R1D1A1B07049572 and 2018R1D1A1B07049494); the Ministry of Information and Communication Technology and Future Planning (2016M3C7A1904392); a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI); the Ministry of Health & Welfare, Republic of Korea (HI16C0992); the National Natural Science Foundation of China (61701340); and the Natural Science Foundation of Tianjin City (18JCQNJC00500).

REFERENCES

1. Rathi, P., Maurer, S., and Summerer, D. (2018). Selective recognition of *N*4-methylcytosine in DNA by engineered transcription-activator-like effectors. Philos. Trans. R. Soc. Lond. B Biol. Sci. *373*, 20170078.
2. Pataillot-Meakin, T., Pillay, N., and Beck, S. (2016). 3-methylcytosine in cancer: an underappreciated methyl lesion? Epigenomics *8*, 451–454.
3. Robertson, K.D. (2005). DNA methylation and human disease. Nat. Rev. Genet. *6*, 597–610.
4. Casadesús, J., and Low, D. (2006). Epigenetic gene regulation in the bacterial world. Microbiol. Mol. Biol. Rev. *70*, 830–856.

5. Jin, B., Li, Y., and Robertson, K.D. (2011). DNA methylation: superior or subordinate in the epigenetic hierarchy? Genes Cancer *2*, 607–617.

6. Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

7. Modrich, P. (1991). Mechanisms and biological effects of mismatch repair. Annu. Rev. Genet. *25*, 229–253.

8. Cheng, X. (1995). DNA modification by methyltransferases. Curr. Opin. Struct. Biol. *5*, 4–10.

9. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., and Turner, S.W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods *7*, 461–465.

10. Yu, M., Ji, L., Neumann, D.A., Chung, D.H., Groom, J., Westpheling, J., He, C., and Schmitz, R.J. (2015). Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite- sequencing. Nucleic Acids Res. *43*, e148.

11. Zou, Q., Chen, L., Huang, T., Zhang, Z., and Xu, Y. (2017). Machine learning and graph analytics in computational biomedicine. Artif. Intell. Med. *83*, 1.

12. Xu, Y., Wang, Y., Luo, J., Zhao, W., and Zhou, X. (2017). Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. Nucleic Acids Res. *45*, 12100–12112.

13. Wei, L., Wan, S., Guo, J., and Wong, K.K. (2017). A novel hierarchical selective ensemble classifier with bioinformatics application. Artif. Intell. Med. *83*, 82–90.

14. Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif. Intell. Med. *83*, 67–74.

15. Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics *33*, 3518–3523.

16. He, W., Jia, C., and Zou, Q. (2019). 4mCPred: Machine Learning Methods for DNA N4-methylcytosine sites prediction. Bioinformatics *35*, 593–601.

17. Wei, L., Luan, S., Nagai, L.A.E., Su, R., and Zou, Q. (2018). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics. Published online September 19, 2018. https://doi.org/10.1093/bioinformatics/bty824.

18. Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2018). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics. Published online December 24, 2018. https://doi.org/10.1093/bioinformatics/bty1047.

19. Qiang, X., Zhou, C., Ye, X., Du, P.F., Su, R., and Wei, L. (2018). CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. Brief. Bioinform. Published online September 17, 2018. https://doi.org/10.1093/bib/bby091.

20. 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.

21. Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics. Published online January 8, 2019. https://doi.org/10.1093/bioinformatics/btz015.

22. McNEMAR, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika *12*, 153–157.

23. Chou, K.-C., and Shen, H.-B. (2009). Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. *1*, 63–92.

24. Dao, F.Y., Lv, H., Wang, F., Feng, C.Q., Ding, H., Chen, W., and Lin, H. (2018). Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics. Published online November 14, 2018. https://doi.org/10.1093/bioinformatics/bty943.

25. Liu, B., Weng, F., Huang, D.S., and Chou, K.C. (2018). iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC. Bioinformatics *34*, 3086–3093.

26. Bhattacharya, D., Nowotny, J., Cao, R., and Cheng, J. (2016). 3Drefine: an interactive web server for efficient protein structure refinement. Nucleic Acids Res. *44* (W1), W406-9.

27. Cao, R., and Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. Sci. Rep. *6*, 23990.

28. Liu, B., Fang, L., Long, R., Lan, X., and Chou, K.-C. (2016). iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics *32*, 362–369.

29. Liu, B., Liu, Y., Jin, X., Wang, X., and Liu, B. (2016). iRSpot-DACC: a computational predictor for recombination hot/cold spots identification based on dinucleotide-based auto-cross covariance. Sci. Rep. *6*, 33483.

30. Feng, C.Q., Zhang, Z.Y., Zhu, X.J., Lin, Y., Chen, W., Tang, H., and Lin, H. (2019). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics *35*, 1469–1477.

31. Basith, S., Manavalan, B., Shin, T.H., and Lee, G. (2018). iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. Comput. Struct. Biotechnol. J. *16*, 412–420.

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

33. Ma, Q., Liu, B., Zhou, C., Yin, Y., Li, G., and Xu, Y. (2013). An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. Bioinformatics *29*, 2261–2268.

34. Liu, B., Yang, J., Li, Y., McDermaid, A., and Ma, Q. (2018). An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data. Brief. Bioinform. *19*, 1069–1081.

35. Zou, Q., Mrozek, D., Ma, Q., and Xu, Y. (2017). Scalable Data Mining Algorithms in Computational Biology and Biomedicine. BioMed Res. Int. *2017*, 5652041.

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

37. Lee, D., Karchin, R., and Beer, M.A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. *21*, 2167–2180.

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

39. Manavalan, B., Shin, T.H., and Lee, G. (2017). DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. Oncotarget *9*, 1944–1956.

40. Qiang, X., Chen, H., Ye, X., Su, R., and Wei, L. (2018). M6AMRFS: Robust Prediction of N6-Methyladenosine Sites With Sequence-Based Features in Multiple Species. Front. Genet. *9*, 495.

41. Bari, A.T.M.G., Reaz, M.R., Choi, H.J., and Jeong, B.S. (2013). DNA encoding for splice site prediction in large DNA sequence. In Database Systems for Advanced Applications. DASFAA 2013. Lecture Notes in Computer Science, B. Hong, X. Meng, L. Chen, W. Winiwarter, and W. Song, eds. (Springer), pp. 46–58.

42. Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.C. (2019). iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics *111*, 96–102.

43. Wei, L., Chen, H., and Su, R. (2018). M6APred-EL: A Sequence-Based Predictor for Identifying N6-methyladenosine Sites Using Ensemble Learning. Mol. Ther. Nucleic Acids *12*, 635–644.

44. Manavalan, B., Lee, J., and Lee, J. (2014). Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. PLoS ONE *9*, e106542.

45. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. *8*, 14.

46. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.C. (2018). iRNA-3typeA: Identifying Three Types of Modification at RNA's Adenosine Sites. Mol. Ther. Nucleic Acids *11*, 468–474.

47. Cao, R., Wang, Z., and Cheng, J. (2014). Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment. BMC Struct. Biol. *14*, 13.

48. Cao, R., Wang, Z., Wang, Y., and Cheng, J. (2014). SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC Bioinformatics *15*, 120.

49. Manavalan, B., Shin, T.H., Kim, M.O., and Lee, G. (2018). PIP-EL: A New Ensemble Learning Method for Improved Proinflammatory Peptide Predictions. Front. Immunol. *9*, 1783.

50. Chen, W., Feng, P., Liu, T., and Jin, D. (2018). Recent advances in machine learning methods for predicting heat shock proteins. Curr. Drug Metab. Published online October 30, 1018. https://doi.org/10.2174/1389200219666181031105916.

51. Usmani, S.S., Bhalla, S., and Raghava, G.P.S. (2018). Prediction of Antitubercular Peptides From Sequence Information Using Ensemble Classifier and Hybrid Features. Front. Pharmacol. *9*, 954.

52. Noble, W.S. (2006). What is a support vector machine? Nat. Biotechnol. *24*, 1565–1567.

53. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

54. Wei, L., Xing, P., Su, R., Shi, G., Ma, Z.S., and Zou, Q. (2017). CPPred-RF: A Sequence-based Predictor for Identifying Cell-Penetrating Peptides and Their Uptake Efficiency. J. Proteome Res. *16*, 2044–2053.

55. Manavalan, B., Govindaraj, R.G., Shin, T.H., Kim, M.O., and Lee, G. (2018). iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction. Front. Immunol. *9*, 1695.

56. Khatun, M.S., Hasan, M.M., and Kurata, H. (2019). PreAIP: Computational Prediction of Anti-inflammatory Peptides by Integrating Multiple Complementary Features. Front. Genet. *10*, 129.

57. Hasan, M.M., and Kurata, H. (2018). GPSuc: Global Prediction of Generic and Species-specific Succinylation Sites by aggregating multiple sequence features. PLoS ONE *13*, e0200283.

58. Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Mach. Learn. *63*, 3–42.

59. Manavalan, B., Subramaniyam, 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.

60. Šícho, M., de Bruyn Kops, C., Stork, C., Svozil, D., and Kirchmair, J. (2017). FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity. J. Chem. Inf. Model. *57*, 1832–1846.

61. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Ann. Stat. *29*, 1189–1232.

62. Rawi, R., Mall, R., Kunji, K., Shen, C.H., Kwong, P.D., and Chuang, G.Y. (2018). PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. Bioinformatics *34*, 1092–1098.

63. Chou, K.-C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. *273*, 236–247.

64. Chen, W., Feng, P.M., Lin, H., and Chou, K.C. (2013). iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. *41*, e68.

65. Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K.C. (2016). iRNA-PseU: Identifying RNA pseudouridine sites. Mol. Ther. Nucleic Acids *5*, e332.

66. Feng, P.M., Chen, W., Lin, H., and Chou, K.C. (2013). iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal. Biochem. *442*, 118–125.

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

68. Lin, H., Ding, C., Song, Q., Yang, P., Ding, H., Deng, K.J., and Chen, W. (2012). The prediction of protein structural class using averaged chemical shifts. J. Biomol. Struct. Dyn. *29*, 643–649.

69. Lin, H., Liang, Z.Y., Tang, H., and Chen, W. (2017). Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM Trans. Comput. Biol. Bioinform. Published online February 8, 2017. https://doi.org/10.11/09/TCBB.2017.2666141.

70. Yang, H., Tang, H., Chen, X.X., Zhang, C.J., Zhu, P.P., Ding, H., Chen, W., and Lin, H. (2016). Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition. BioMed Res. Int. *2016*, 5413903.

71. Zhao, Y.W., Su, Z.D., Yang, W., Lin, H., Chen, W., and Tang, H. (2017). IonchanPred 2.0: A Tool to Predict Ion Channels and Their Types. Int. J. Mol. Sci. *18*, E1838.

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

73. 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.

74. 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*, E1732.

75. Manavalan, B., Basith, S., Shin, T.H., Choi, S., Kim, M.O., and Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget *8*, 77121–77136.

76. Manavalan, B., and Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. Bioinformatics *33*, 2496–2503.

77. Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2018). Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Brief. Bioinform. Published online October 31, 2018. https://doi.org/10.1093/bib/bby107.

78. Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., and Chou, K.C. (2017). iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC. Mol. Ther. Nucleic Acids *7*, 155–163.

79. Liu, B., Yang, F., and Chou, K.C. (2017). 2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function. Mol. Ther. Nucleic Acids *7*, 267–277.

80. Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2019). Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Brief. Bioinform. Published online January 10, 2019. https://doi.org/10.1093/bib/bby124.