Computational and Structural Biotechnology Journal 18 (2020) 906–912

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes

Md. Mehedi Hasan [a,b], Balachandran Manavalan [c], Watshara Shoombuatong [d], Mst. Shamima Khatun [a], Hiroyuki Kurata [a,e,*]

[a] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
[b] Japan Society for the Promotion of Science, 5-3-1 Kojimachi, Chiyoda-ku, Tokyo 102-0083, Japan
[c] Department of Physiology, Ajou University School of Medicine, Suwon 443380, Republic of Korea
[d] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand
[e] Biomedical Informatics R&D Center, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

ARTICLE INFO

Article history:
Received 7 January 2020
Received in revised form 31 March 2020
Accepted 3 April 2020
Available online 8 April 2020

Keywords:
Mouse genome
Sequence analysis
Sequence encoding
Machine learning

ABSTRACT

N4-methylcytosine (4mC) is one of the most important DNA modifications and involved in regulating cell differentiations and gene expressions. The accurate identification of 4mC sites is necessary to understand various biological functions. In this work, we developed a new computational predictor called i4mC-Mouse to identify 4mC sites in the mouse genome. Herein, six encoding schemes of k-space nucleotide composition (KSNC), k-mer nucleotide composition (Kmer), mono nucleotide binary encoding (MBE), dinucleotide binary encoding, electron–ion interaction pseudo potentials (EIIP) and dinucleotide physicochemical composition were explored that cover different characteristics of DNA sequence information. Subsequently, we built six RF-based encoding models and then linearly combined their probability scores to construct the final predictor. Among the six RF-based models, the Kmer, KSNC, MBE, and EIIP encodings are sufficient, which contributed to 10%, 45%, 25%, and 20% of the prediction performance, respectively. On the independent test the i4mC-Mouse predicted the 4mC sites with accuracy and MCC of 0.816 and 0.633, respectively, which were approximately 2.5% and 5% higher than those of the existing method (4mCpred-EL). For experimental biologists, a freely available web application was implemented at http://kurata14.bio.kyutech.ac.jp/i4mC-Mouse/.

1. Introduction

In both prokaryotes and eukaryotes, N4-methylcytosine (4mC), 5-Methylcytosine (5mC), and N6-methyladenine (6 mA) alterations can regulate various functions including genomic imprinting, cell developmental, and gene expressions, and play crucial roles in the genomic diversity [1,2]. The 5mC modification is a common type of methylation alteration and well-explored that exemplifies an important role in biological developments [3,4] that are associated by the various diseases such as diabetes, neurological, and cancer [5,6]. The 4mC modification is also an effective methylation that defends the self-DNA from being degraded by restriction enzymes.

Until now, many experimental methodologies, such as mass spectrometry, methylation-precise PCR, and Single Molecule of Real-Time (SMRT) sequencing [7–10], have been efficiently used to identify the epigenetic 4mC sites. The exact dataset of modifications of 4mC sites is still limited due to the shortage of experimental identification approaches. Moreover, the aforementioned experimental approaches are labor-intensive and expensive works. Thus, computational tools are required for analysis of the accessible big data on the genome of mouse so as to allow the identification of novel 4mC sites, while shedding light on their mechanism [11,12]. Several computational approaches have been proposed by using the recently constructed database named MethSMRT [13] to predict 4mC sites from seven different species, i.e. *E. coli*, *G. subterraneus*, *A. thaliana*, *D. melanogaster*, *C. elegans*, *G. pickeringii*, and *Rosaceae genome.* [11,12,14–16]. To the best of author's knowledge, only one predictor is available for the 4mC sites in the mouse genome, named 4mCpred-EL [11]. This method implemented multiple encodings and machine learning (ML) algorithms, which was

* Corresponding author at: Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan.
E-mail address: kurata@bio.kyutech.ac.jp (H. Kurata).

https://doi.org/10.1016/j.csbj.2020.04.001
2001-0370/

applied to the dataset derived from the MethSMRT. Although the 4mCpred-EL yielded encouraging results, there is still room for further enhancement, probably because the employed feature information is not sufficient to capture the discriminative information between the two classes.

Motivated by the aforementioned problems, in this work, we have implemented a computational tool called i4mC-Mouse for the identification of 4mCs in the genome of mouse. A workflow of the proposed i4mC-Mouse is summarized in Fig. 1. Initially, six probabilities of 4mC sites were predicted by using a random forest (RF) classifier in conjunction with the k-mer nucleotide (NT) arrangement (Kmer), k-space NT composition (KSNC), NT mono binary encoding (MBE), dinucleotide binary encoding (DBE), electron–ion pseudopotentials (EIIP), and dinucleotide physicochemical composition (DPC). Secondly, to select the successive feature vectors, the Wilcoxon rank sum test (WR) was accessed. Finally, the four (Kmer, KSNC, MBE and EIIP) models evaluated the probability scores of 4mC sites and these scores were linearly combined to develop the i4mC-Mouse. Our results on independent test showed that i4mC-Mouse outperformed the existing predictor 4mCpred-EL. Finally, for the convenience of experimental scientists, our proposed model was implemented as a web application.

2. Materials and methods

2.1. Dataset construction

To develop a sequence-based predictor of 4mCs, a reliable dataset is necessary. To make a fair comparison, we used the previous dataset [11], which was collected from MethSMRT [13]. The DNA sequence windows are set to 41 base pairs (bp) having “C” at the center. To yield a high-quality dataset, we considered the sequences with a modQV score of ≥20 and excluded the remaining sequences. It is worth mentioning that the previous study applied a CD-HIT of 80% [17] and excluded the sequences that share 80% sequence identity. To develop a more reliable model and avoid an overestimation of prediction model, we applied CD-HIT of 70% and excluded the sequences that showed greater than 70% sequence identity. After such screening procedures, we finally obtained the benchmark dataset containing 906 positive samples, which are 74 samples lower than those of the 4mCPred-EL. A subset of 906 non-4mCs were randomly extracted from the non-4mCs. After obtaining the balanced dataset consisting of 906 4mCs and 906 non-4mCs, we divided them into the training and independent sets, such as 80% samples (746 4mCs and 746 non-4mCs) and 20% samples (160 4mCs and 160 non-4mCs), respectively.

2.2. Feature encoding

The next crucial step is to represent a DNA sequence as fixed-length feature vectors [18,19]. Six encoding methods of Kmer, KSNC, MBE, DBE, EIIP and DPC were used. The potential capability of these encodings employed in many domains has already been mentioned in our previous studies [20,21].

Kmer: This encoding has been extensively used in different prediction tasks [15,22,23]. In this study, a DNA sample with L length is articulated as $D = d_1, d_2, d_3,\ldots d_L$, d_i is one of the NTs (A, C, G, T, N). Considering tri-, and tetra -nucleotides, the Kmer scheme gener-

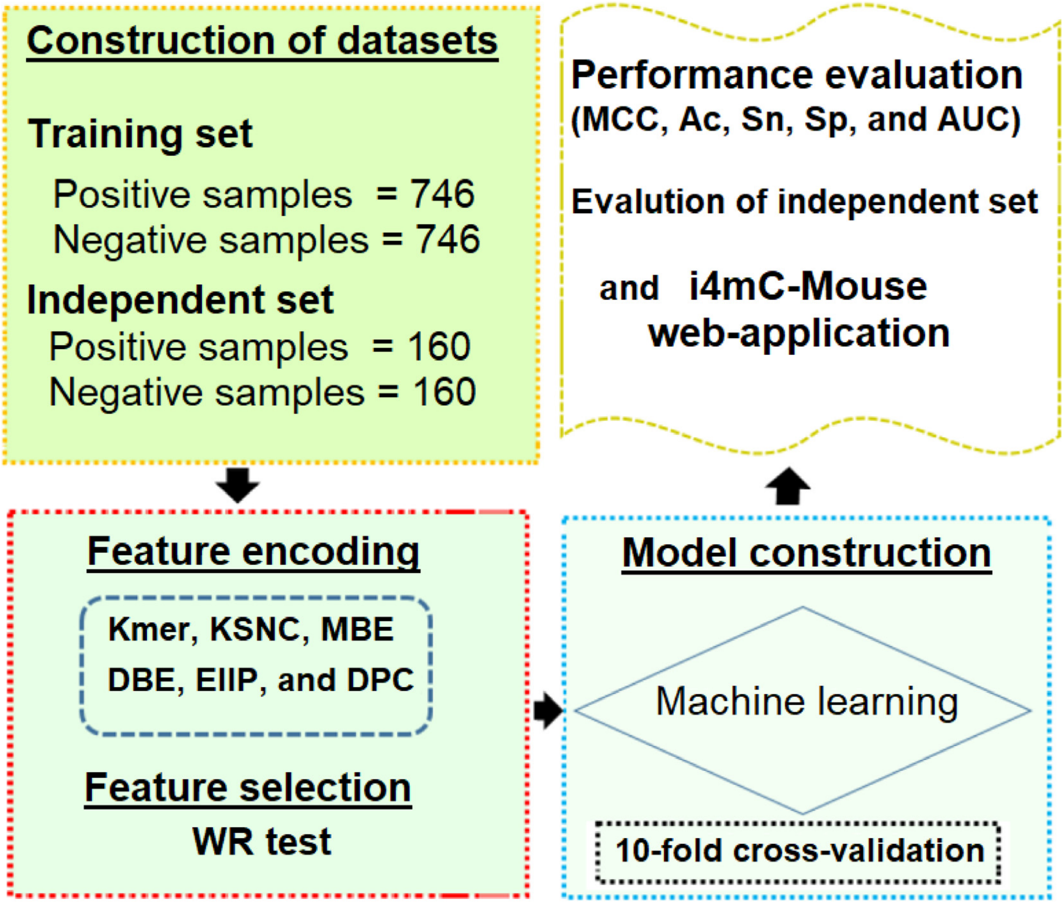

Fig. 1. A computational framework of the i4mC-Mouse. It includes three steps: (i) dataset construction; (ii) selection of six different encoding schemes that convert DNA sequences into numerical feature vectors; and (iii) model evaluation and construction using a CV test. Then, construction of a webserver for the final prediction model (i4mC-Mouse).

ated a 750 (=$5^3 + 5^4$) dimensional (D) feature vector. Here the letter 'N' signifies a non-standard nucleotide.

KSNC: This encoding signifies the frequency NTs information by using the pair-wise similarity searches [23] and widely used in bioinformatics tasks [24–26]. The NT (A, C, G, T, N) pairs ($nc_{i,}$ where i = 1, 2,. . .,25) were encoded and standardized as

$$\text{NTpair} = \frac{F(nci)}{w - d - 1} \tag{1}$$

where $F(nc_i)$ is the sum of nc_i privileged 4mC sites. The w and d are the sequence length and space length between NTs, respectively. For a range of *dmax* is 0 to 3, the KSNC signifies a 100-D feature vector.

MBE: The MBE exactly depicts the NT for the sequence of curated samples at each position, where A, T, G, C, and N are represented by (1,0,0,0,0), (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), and (0,0,0,0,1), respectively. In MBE, for a length of NT sequence, a $w \times$ 5-D vector was generated.

DBE: In the DBE scheme, the possible 16 dinucleotides are encoded as 0/1 (four-dimensional vector) [11]. For instance, AT (0,0,0,1), AA (0,0,0,0), GG (1,1,1,1), and AC (0,0,1,0) are encoded [27,28]. All N pair dinucleotides are regarded as zero. For a sequence of 4mC or non-4mC with a DBE, a 160 $\{(w - 1) \times 4\} - \text{D}$ vector was generated.

EIIP: To encode the electron–ion energies in the DNA, Nair and Sreedharan developed EIIP [29]. In this study, EIIP values were encoded as follows: A (0.1260), C (0.1340), G (0.0806), T (0.1335), and N (0.0000). The EIIP scheme transformed a sequence into a *w*-D feature vector.

DPC: Fifteen types of DPC were collected from the recent publications [20,21]. The physicochemical properties are encoded as a 375 (25 dinucleotides × 15 physicochemical properties)-D vector.

2.3. Feature selection

Inclusion of non-informative and noisy feature might cause unsatisfied prediction performances [30,31]. In fact, there are several feature selection and ranking approaches, such as Chi-square, mRMR, and WR test. In this work, the WR feature selection method was used [32].

2.4. Machine learning classifier

The computational model employed herein was constructed by using the RF algorithm [33]. The RF classifier is widely used in various biological problems [34–40]. The RF classifier is a collaborative model consisting of many regression and classification trees, and the prediction performances are enhanced by increasing the number of weak CART classifiers. In this study, the RF package 'randomForest' (https://cran.r-project.org/) was used.

It is crucial to compare the proposed RF-based models with other commonly used ML-based models, i.e. Naive Bayes (NB) [41,42], SVM [37,43], k-nearest neighbor (KNN), and AdaBoost (AB). The NB and AB classifiers were performed in R programming (https://www.r-project.org/), while the KNN classifier was implemented in our house PERL program. The SVM^{light} was used to build the SVM algorithm [38]. Notably, all these classifiers are extensively applied to various prediction problems [44–48].

2.5. Combined model

To increase the prediction performance of the proposed model, we linearly combined the probability scores of the six, single encoding-based models, as given by:

$$\text{Combined }(s) = \sum_{i=1}^{6} w_i x_i(s), \ \sum_{i=1}^{6} w_i = 1 \tag{2}$$

where Combined (s) specifies the combination of the 6 scores evaluated by the single encoding scheme-employing MLs, w_i characterizes the weight of the i-th encoding model and $xi(s)$ specifies the ML scores of sample s based on the i-th encoding model. These weight values were adjusted based on the AUC values via 10-fold cross-validation (CV) tests.

2.6. Evaluation metrics

Four statistical metrics: Matthews correlation coefficient (MCC), accuracy (Ac), sensitivity (Sn), and specificity (Sp) were used to evaluate the performance of the predictors as follows [39,49–52]:

$$\text{MCC} = \frac{\text{n(TP)} \times \text{n(TN)} - \text{n(FP)} \times \text{n(FN)}}{\sqrt{[\text{n(TN)} + \text{n(FN)}] \times [\text{n(TP)} + \text{n(FP)}] \times [\text{n(TN)} + \text{n(FP)}] \times [\text{n(TP)} + \text{n(FN)}]}} \tag{3}$$

$$\text{Ac} = \frac{\text{n(TP)} + \text{n(TN)}}{\text{n(TP)} + \text{(FN)} + \text{n(FP)} + \text{n(TN)}} \tag{4}$$

$$\text{Sn} = \frac{\text{n(TP)}}{\text{n(TP)} + \text{n(FN)}} \tag{5}$$

$$\text{Sp} = \frac{\text{n(TN)}}{\text{n(TN)} + \text{n(FP)}} \tag{6}$$

where n(TP) and n(TN) specify the numbers correctly predicted samples of 4mCs and non-4mCs, respectively. n(FP) and n(FN) specify the numbers incorrectly predicted samples of 4mCs and non-4mCs, respectively.

3. Results and discussion

3.1. Nucleotide preference analysis

We aim to develop a computational model for discriminating 4mC samples from non-4mC ones. Therefore, we sought to determine the composition of sequence preferences between the 4mC and non-4mC samples by using the pLogo software [53]. The pLogo examines the statistically significant differences in position-specific NTs (p < 0.05). As seen in Fig. 2, the C base was overrepresented compared to the other bases in the 4mC samples and the A base was under-represented compared to the other bases, while the G and T bases were observed at both the over- and underrepresented positions. In summary, the over- and under-represented A and C bases were considerably varied between the 4mC and non-4mC samples, suggesting the importance of position-specific preferences of nucleotide base pairs, which is consistent with the previous study [11].

3.2. Performance evaluation of i4mC-Mouse

First, the training dataset was converted into feature vectors by using six schemes (Kmer, KSNC, MBE, DBE, EIIP, and DPC) and individually inputted to a RF classifier. Second, we evaluated the successive feature vectors for the six, single encoding models by 10-fold CV tests. To reduce the feature dimension and improve the prediction performance, we carried out the WR test approach to select an optimal feature set on each encoding and compared its performance with the control. As shown in Table S1, the feature selection improved the performance on the three encodings (Kmer (160D), KSNC (80D) and DPC (110D)), while the remaining three

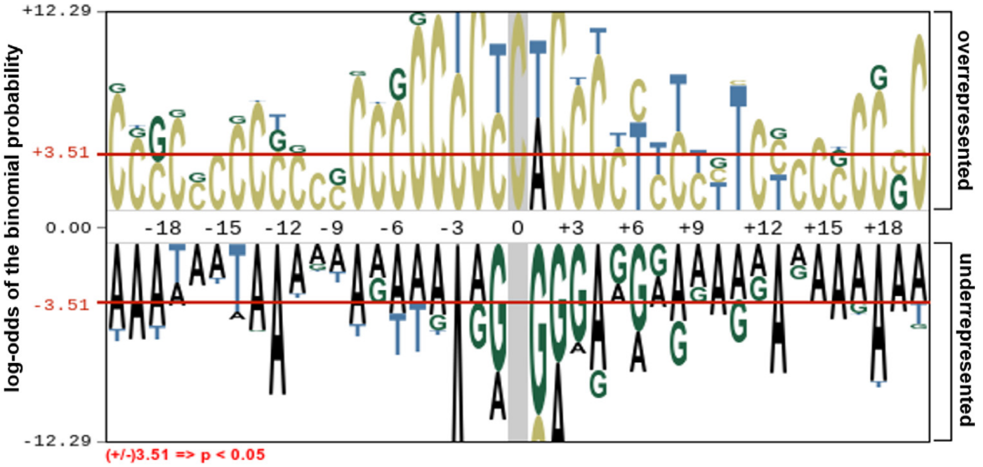

Fig. 2. Sequence logo representation of 4mC samples. The 20 upstream and 20 downstream DNA residues surrounding the mouse 4mC site were analyzed.

encodings (MBE, DBE and EIIP) did not outperform their controls. Therefore, we used three optimal feature set-based models for the subsequent analysis. Fig. 3 and Table 1 show the prediction performances of the six, single encoding-based models and the combined model (i4mC-Mouse). The six, single encoding-based models of Kmer, KSNC, MBE, DBE, EIIP and DPC provided AUCs of 0.869, 0.882, 0.851, 0.814, 0.840 and 0.822, respectively. In terms of Ac and MCC, the KSNC encoding outperformed the other encodings, where the AUC of the KSNC was approximately ~1–7% higher than the AUCs of the other encodings.

In the combined model, a linear regression model was used to integrate the six RF probability scores, as mentioned in the method section, where the weight coefficients of the Kmer, KSNC, MBE, DBE, EIIP, and DPC schemes are 0.10, 0.45, 0.25, 0.00, 0.20 and 0.00, respectively. Notably, our approach excluded the two models (DBE and DPC) by assigning weight 0.00 and considered the remaining four models. The contribution of Kmer, KSNC, MBE and EIIP are 10%, 45%, 25%, and 20%, respectively, in the final prediction. As noticed in Table 1, at a Sp control of 90.42%, the i4mC-Mouse yielded MCC, Ac, Sn, and Sp of 0.651, 79.30% 68.31%, and 90.42% respectively. To show the advantage of our approach, we computed the statistically significant differences between the i4mC-Mouse and each single encoding-based model using two-tailed *t*-test [54]. The i4mC-Mouse outperformed the five models at a *p*-value of <0.05, except the KSNC model at a *p*-value of 0.063.

Table 1
Prediction performances of the i4mC-Mouse model and the single encoding-based RF models.

Methods	MCC	Ac (%)	Sn (%)	Sp (%)	AUC	P-value
Kmer	0.566	74.81	59.53	90.10	0.869	0.011
KSNC	0.602	76.90	63.42	90.30	0.882	0.063
MBE	0.486	71.20	53.81	88.61	0.851	0.006
DBE	0.432	69.13	48.11	90.10	0.814	0.001
EIIP	0.473	70.80	52.31	89.21	0.840	0.001
DPC	0.428	69.21	49.91	88.52	0.822	0.001
i4mC-Mouse	0.651	79.30	68.31	90.20	0.904	–

* i4mC-Mouse specifies the linear arrangement of the RF scores for Kmer, KSNC, MBE, DBE, EIIP, and DPC encodings and their weight values are 0.10, 0.45, 0.25, 0.00, 0.20, and 0.00, respectively.

3.3. Effect of ML algorithms on prediction performances of the combined model

We applied the above procedure (the construction of six encoding-based models and combined models) to other commonly used four classifiers (NB, SVM, AB and KNN) and compared their

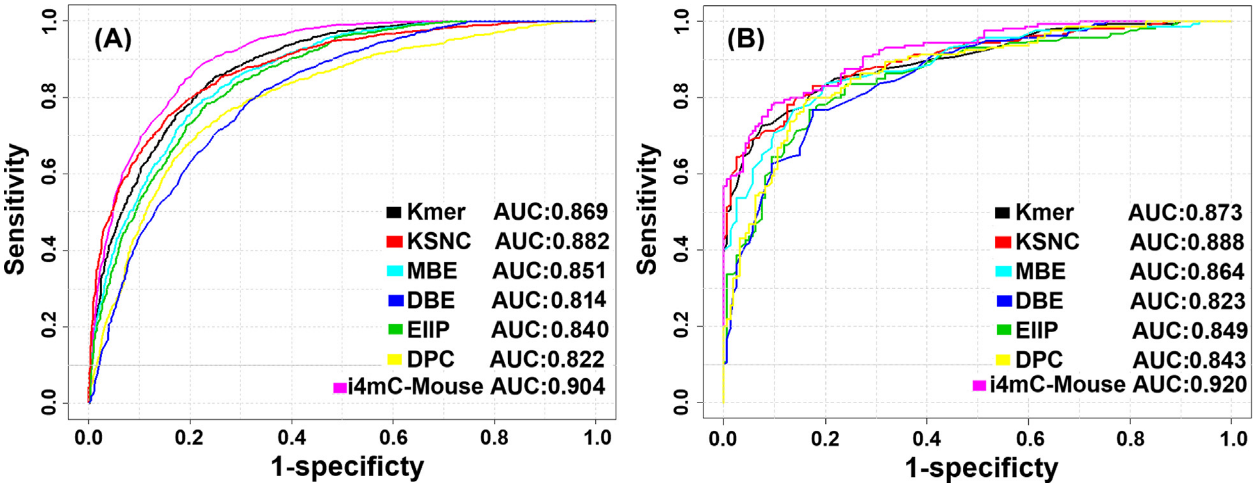

Fig. 3. Performance comparisons of single encoding-based models and i4mC-Mouse. The ROC curves were evaluated on the training dataset by a 10-fold CV test (A) and independent dataset (B).

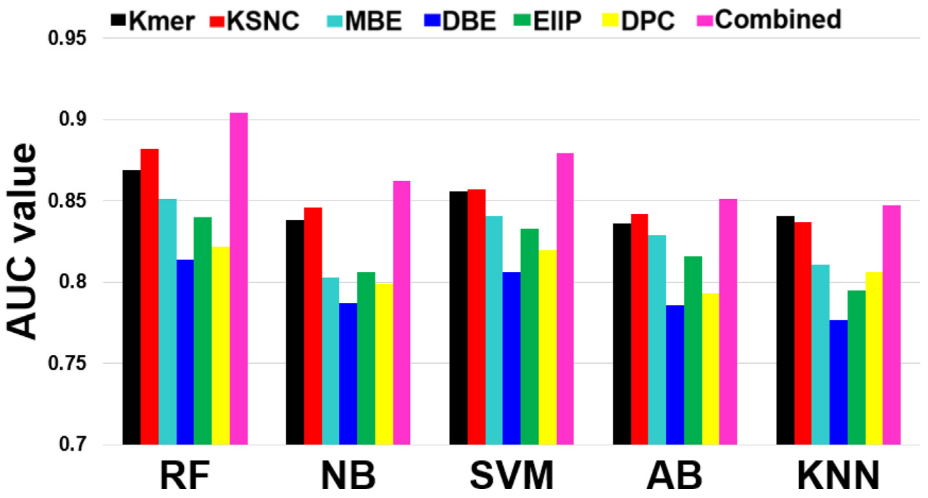

Fig. 4. Effect of different ML algorithms on the AUC values of the six single encoding-based models and i4mC-Mouse. The performances were evaluated on the training datasets by a 10-fold CV test.

performances with the RF-based models. Instead of selecting default ML parameters, 10-fold CV was employed to optimize their respective ML parameters on each encoding-based classifier. Finally, an optimal model was obtained for each classifier, whose performances are shown in Fig. 4. We noted that the combined model for each classifier performed better than the individual encoding-based model, indicating the integration of multiple information is effective in achieving the best performance. Furthermore, comparison among the combined models with five different classifiers showed that the RF achieved the best performance, while the SVM was comparable to the RF model. Specifically, AUCs of the RF (i.e. i4mC-Mouse) were ~1–5% higher than those of any other combined models, demonstrating that the RF model is the most suitable for the i4mC prediction.

3.4. Comparison of i4mC-Mouse with 4mCpred-EL on the independent dataset

We compared the proposed i4mC-Mouse with the existing method (4mCpred-EL) on the same independent dataset consisting of 160 4mCs and 160 non-4mCs, as shown in Table 2. We directly submitted to the independent dataset to the 4mCpred-EL web server. The 4mCpred-EL yielded 79.10% Ac, 75.72% Sn, 82.51% Sp, 0.584 MCC, and 0.881 AUC, while the i4mC-Mouse provided 81.61% Ac, 80.71% Sn, 82.52% Sp, 0.633 MCC, and 0.920 AUC. The i4mC-Mouse outperformed the 4mCpred-EL with increased ratios of >3%, >5% and >5% on Ac, Sn and MCC, respectively. The better performance of the i4mC-Mouse would be due to the followings: selection of an appropriate classifier, a linear combination of single encoding-based models, and reduction of dataset redundancy.

3.5. i4mC-Mouse web server

A user-friendly and freely accessible web application was established for the prediction of mouse genome at http://kurata14.bio.kyutech.ac.jp/i4mC-Mouse/. The manuals are as follows: (i) select the exact 41 bp DNA 4mC genome (ii) browse or enter the query sequences from users' own file (FASTA format) to the input page, where a sample is shown our server page, (iii) push the 'Submit' button. The server completes the query tasks with the probability scores within one min.

Table 2
Comparison between the i4mC-Mouse and 4mCpred-EL.

Method	MCC	Ac (%)	Sn (%)	Sp (%)	AUC
4mCpred-EL	0.584	79.10	75.72	82.51	0.881
i4mC-Mouse	0.633	81.61	80.71	82.52	0.920

The performances were evaluated on the independent dataset.

4. Conclusions

4mC plays an important role in the DNA modifications and is involved in regulating cell differentiations and gene expression levels. Therefore, accurate identification of 4mC sites is an essential step to understand the exact biological functions. To date, several computational prediction tools have been developed to identify 4mC sites from different species [11,12,14–16,20,55,56], but only one method is available for mouse species. In this study, we have developed a new computational model, called i4mC-Mouse, for improving the prediction of 4mCs in the mouse genome. We employed six encoding schemes of Kmer, KSNC, MBE, DBE, EIIP and DPC to cover various aspects of DNA sequences and optimized the successive features via the WR feature selection method. The final constructed i4mC-Mouse was a linear combination of the predicted probabilities by four, single encoding-based RF-models, where the Kmer, KSNC, MBE and EIIP encodings contributed to 10%, 45%, 25%, and 20%, respectively. On the independent test the i4mC-Mouse outperformed the existing method (4mCpred-EL). The i4mC-Mouse is demonstrated to be the most accurate predictor. Finally, a freely available web application was implemented.

5. Author statement

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

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.04.001.

References

[1] Rathi P, Maurer S, Summerer D. Selective recognition of N4-methylcytosine in DNA by engineered transcription-activator-like effectors. Philos Trans R Soc London Ser B, Biol Sci 2018;373(1748).

[2] Jeltsch A, Jurkowska RZ. New concepts in DNA methylation. Trends Biochem Sci 2014;39(7):310–8.

[3] Akalin A, Garrett-Bakelman FE, Kormaksson M, Busuttil J, Zhang L, Khrebtukova I, et al. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. PLoS Genet 2012;8(6):e1002781.

[4] Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 2008;9(6):465–76.

[5] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012;13(7):484–92.

[6] Ling C, Groop L. Epigenetics: a molecular link between environmental factors and type 2 diabetes. Diabetes 2009;58(12):2718–25.

[7] Doherty R, Couldrey C. Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. Front Genet 2014;5:126.

[8] Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 2010;7(6):461–5.

[9] Boch J, Bonas U. Xanthomonas AvrBs3 family-type III effectors: discovery and function. Annu Rev Phytopathol 2010;48:419–36.

[10] Buryanov YI, Shevchuk TV. DNA methyltransferases and structural-functional specificity of eukaryotic DNA modification. Biochem Biokhimiia 2005;70 (7):730–42.

[11] Manavalan B, Basith S, Shin TH, Lee DY, Wei L, Lee G. 4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse genome. Cells 2019;8(11).

[12] He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics 2019;35(4):593–601.

[13] Ye P, Luan Y, Chen K, Liu Y, Xiao C, Xie Z. MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Res 2017;45(D1):D85–9.

[14] Wei L, Luan S, Nagai LAE, Su R, Zou Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. Bioinformatics 2019;35(8):1326–33.

[15] Hasan MM, Manavalan B, Khatun MS, Kurata H. i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. Int J Biol Macromol 2019.

[16] Chen W, Yang H, Feng P, Ding H, Lin H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 2017;33(22):3518–23.

[17] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28(23):3150–2.

[18] Xu ZC, Feng PM, Yang H, Qiu WR, Chen W, Lin H. iRNAD: a computational tool for identifying D modification sites in RNA sequence. Bioinformatics 2019.

[19] Yang H, Lv H, Ding H, Chen W, Lin H. iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens. J Comput Biol 2018;25 (11):1266–77.

[20] Manavalan B, Basith S, Shin TH, Wei L, Lee G. Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. Mol Ther Nucleic Acids 2019;16:733–44.

[21] Manavalan B, Shin TH, Lee G. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. Oncotarget 2018;9(2):1944.

[22] Yang H, Yang W, Dao FY, Lv H, Ding H, Chen W, et al. A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. Brief Bioinf 2019.

[23] Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res 2016;44(10):e91.

[24] Hasan MM, Guo D, Kurata H. Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. Mol BioSyst 2017;13(12):2545–50.

[25] Hasan MM, Yang S, Zhou Y, Mollah MN. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. Mol BioSyst 2016;12(3):786–95.

[26] Khatun MS, Hasan MM, Kurara H. PreAIP: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. Front Genet 2019.

[27] Hasan MM, Manavalan B, Shoombuatong W, Khatun MS, Kurata H. i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. Plant Mol Biol 2020.

[28] Lv H, Zhang ZM, Li SH, Tan JX, Chen W, Lin H. Evaluation of different computational methods on 5-methylcytosine sites identification. Brief Bioinf 2019.

[29] Nair AS, Sreenadhan SP. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation 2006;1(6):197–202.

[30] Feng CQ, Zhang ZY, Zhu XJ, Lin Y, Chen W, Tang H, et al. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics 2019;35(9):1469–77.

[31] Dao FY, Lv H, Wang F, Feng CQ, Ding H, Chen W, et al. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics 2019;35(12):2075–83.

[32] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38.

[33] Liaw A. Wiener: classification and regression by random forest. R news 2002;2:18–22.

[34] Boopathi V, Subramaniyam S, Malik A, Lee G, Manavalan B, Yang DC. mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. Int J Mol Sci 2019;20(8).

[35] Manavalan B, Shin TH, Kim MO, Lee G. PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions. Front Immunol 2018;9:1783.

[36] Manavalan B, Basith S, Shin TH, Wei L, Lee G. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 2018.

[37] Hasan MM, Zhou Y, Lu X, Li J, Song J, Zhang Z. Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. PLoS ONE 2015;10(6):e0129635.

[38] Khatun S, Hasan M, Kurata H. Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. FEBS Lett 2019;593(21):3029–39.

[39] Hasan MM, Khatun MS, Kurata H. Large-scale assessment of bioinformatics tools for lysine succinylation sites. Cells 2019;8(2).

[40] Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. ACPred: a computational tool for the prediction and analysis of anticancer peptides. Molecules 2019;24(10).

[41] Hasan MM, Rashid MM, Khatun MS, Kurata H. Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. Sci Rep 2019;9(1):8258.

[42] Feng PM, Ding H, Chen W, Lin H. Naive Bayes classifier with feature selection to identify phage virion proteins. Comput Math Methods Med 2013;2013:530696.

[43] Lai HY, Zhang ZY, Su ZD, Su W, Ding H, Chen W, et al. iProEP: a computational predictor for predicting promoter. Mol Ther Nucleic acids 2019;17:337–46.

[44] Manavalan B, Basith S, Shin TH, Wei L, Lee G. AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees. Comput Struct Biotechnol J 2019;17:972–81.

[45] Hasan MM, Manavalan B, Khatun MS, Kurata H. Prediction of S-nitrosylation sites by integrating support vector machines and random forest. Mol Omics 2019;15(6):451–8.

[46] Charoenkwan P, Yana J, Schaduangrat N, Nantasenamat C, Hasan MM, Shoombuatong W. iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. Genomics 2020. https://doi.org/10.1016/j.ygeno.2020.03.019.

[47] Shoombuatong W, Schaduangrat N, Pratiwi R, Nantasenamat C. THPep: a machine learning-based approach for predicting tumor homing peptides. Comput Biol Chem 2019;80:441–51.

[48] Win TS, Schaduangrat N, Prachayasittikul V, Nantasenamat C, Shoombuatong W. PAAP: a web server for predicting antihypertensive activity of peptides. Future Med Chem 2018;10(15):1749–67.

[49] Yang W, Zhu XJ, Huang J, Ding H, Lin H. A brief survey of machine learning methods in protein sub-Golgi localization. Curr Bioinform 2019;14:234–40.

[50] Ding H, Yang W, Tang H, Feng PM, Huang J, Chen W, et al. PHYPred: a tool for identifying bacteriophage enzymes and hydrolases. Virologica Sinica 2016;31 (4):350–2.

[51] Basith S, Manavalan B, Hwan Shin T, Lee G. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. Med Res Rev 2020.

[52] Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics 2020.

[53] O'Shea JP, Chou MF, Quader SA, Ryan JK, Church GM, Schwartz D. pLogo: a probabilistic approach to visualizing sequence motifs. Nat Methods 2013;10(12):1211–2.

[54] Raju TN, Gosset WS, Silverman WA. Two "students" of science. Pediatrics 2005;116(3):732–5.

[55] Zeng F, Fang G, Yao L. A deep neural network for identifying DNA N4-methylcytosine sites. Front Genet 2020;11:209.

[56] Lv H, Dao FY, Zhang D, Guan ZX, Yang H, Su W, et al. iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. ISCIENCE 2020. https://doi.org/10.1016/j.isci.2020.100991.