

RESEARCH ARTICLE

Open Access



Accurate prediction of DNA N⁴-methylcytosine sites via boost-learning various types of sequence features

Zhixun Zhao¹ , Xiaocai Zhang¹, Fang Chen², Liang Fang³ and Jinyan Li^{1*}

Abstract

Background: DNA N4-methylcytosine (4mC) is a critical epigenetic modification and has various roles in the restriction-modification system. Due to the high cost of experimental laboratory detection, computational methods using sequence characteristics and machine learning algorithms have been explored to identify 4mC sites from DNA sequences. However, state-of-the-art methods have limited performance because of the lack of effective sequence features and the ad hoc choice of learning algorithms to cope with this problem. This paper is aimed to propose new sequence feature space and a machine learning algorithm with feature selection scheme to address the problem.

Results: The feature importance score distributions in datasets of six species are firstly reported and analyzed. Then the impact of the feature selection on model performance is evaluated by independent testing on benchmark datasets, where ACC and MCC measurements on the performance after feature selection increase by 2.3% to 9.7% and 0.05 to 0.19, respectively. The proposed method is compared with three state-of-the-art predictors using independent test and 10-fold cross-validations, and our method outperforms in all datasets, especially improving the ACC by 3.02% to 7.89% and MCC by 0.06 to 0.15 in the independent test. Two detailed case studies by the proposed method have confirmed the excellent overall performance and correctly identified 24 of 26 4mC sites from the *C.elegans* gene, and 126 out of 137 4mC sites from the *D.melanogaster* gene.

Conclusions: The results show that the proposed feature space and learning algorithm with feature selection can improve the performance of DNA 4mC prediction on the benchmark datasets. The two case studies prove the effectiveness of our method in practical situations.

Keywords: DNA N4-methylcytosine, Sequence feature, Feature selection, Site prediction

Background

As an essential epigenetic modification, DNA base methylation expands the DNA content and plays crucial roles in regulating various cellular processes [1–3]. According to the location where a methylated group occurs in the DNA sequence, there are many kinds of DNA base methylation. For example, 5-Methylcytosine (5mC), N6-methyladenine

(6mA) and N4-methylcytosine (4mC) are the most common types [4–6]. 5mC occurs at the C5-position of cytosine and is the dominant methylation type in eukaryotic genomes, actively involved in differentiation, gene expression, genomic imprinting, preservation of chromosome stability, aging, suppression of repetitive element, and X chromosome inactivation [7–10]. In prokaryotes, 6mA and 4mC constitute the majority of DNA base methylations [11]. 6mA occurs at the N6-position of adenine and is a marker in gene regulation, development, DNA replication, repair, and expression [12–14]. 4mC happens at the N4-amino group of cytosine and participates in the

*Correspondence: Jinyan.Li@uts.edu.au

¹Advanced Analytics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007 Sydney, Australia

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

restriction-modification system that provides a bacterial immune response against occupying DNA, DNA repair, expression, or replication [15–17]. Compared with the studies for 5mC and 6mA, biological functions of 4mC are much investigated due to the lack of sufficient detection methods.

The precise location of the DNA base methylation was a hard problem in the past for a long time. It is not affordable to locate the DNA 5mC on a large scale until the whole-genome bisulfite sequencing, and the next-generation sequencing techniques were developed [18, 19]. The detection of 6mA and 4mC in the level of whole-genome became available after the single-molecule real-time sequencing (SMRT) technology was introduced [4, 20]. Then a next-generation sequencing method called 4mC-Tet-assisted-bisulphite-sequencing and another method based on engineered transcription-activator like effectors were developed for 4mC identification [21]. However, the experimental methods were of high cost and cannot identify 4mC sites with time efficiency. Recently, the rapid development of machine learning algorithms provides a promising computational approach to address classification problems in bioinformatics, and researchers have explored using computational methods to identify 4mC sites from DNA sequences.

Collecting data from public SMRT sequencing experiments, Ye et al built the first DNA 6mA and 4mC database named MethSMRT for 156 species [22]. Chen et al [23] proposed an SVM based prediction model iDNA4mC using the nucleotide chemical property and sequential nucleotide frequency features. Recently, 4mCPred and 4mCPred-SVM were developed to improve the site prediction performance [24, 25]. In 4mCPred, He et al used two new features PSTNP and EIIP with a simple feature selection. Wei et al built 4mCPred-SVM with four kinds of sequence features and a two-step feature optimization. Recently, some other predictors have been developed to identify 4mC site in the DNA sequence for Mouse [26, 27], *Escherichia coli* [28], Rosaceae [29] and so on [30, 31].

The core idea of the previous research is to transform 4mC-contained DNA sequences into various kinds of features as the input of the machine learning algorithms. However, these features are not adequate to make the prediction methods to achieve excellent performance. Through the analysis on the sequence logos, we observe that the adjacent nucleotides' characteristics are potentially essential. Thus we extract the contiguous nucleotides sequence characteristics like k-nucleotide frequency, k-spectrum nucleotide pair frequency, and PseDNC as features to describe the sequences. Besides, two global sequence features, one-hot binary and sequential nucleotide frequency, are also merged into

our feature space. As global features have the complete information of DNA sequence and the local features can underline specific sequence patterns, the combined feature space is highly expected to improve the prediction performance.

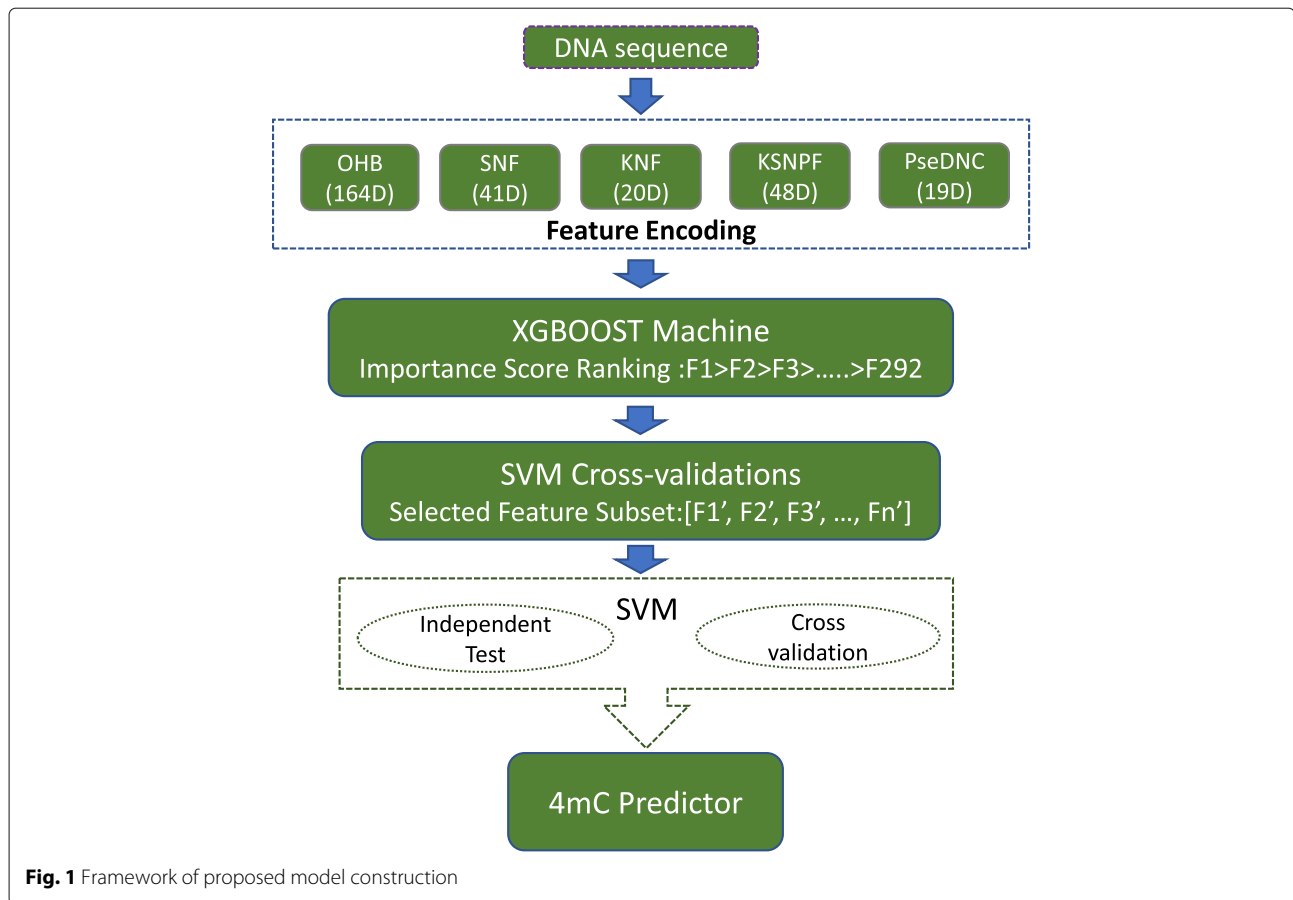
Since feature selection can reduce the feature space dimension and the modelling complexities [32], the existing 4mC prediction methods, including 4mCPred and 4mCPred_SVM, employed a feature selection scheme based on the F-score and sequential forward search (SFS) strategy [24, 25]. Although the F-score can evaluate the feature importance according to the relevance between the feature and label, the performance of the selected feature subset was still under-realized. In this paper, we propose an embedded feature selection scheme, in which features are ranked with the feature importance scores derived by the XGBoost classifier training process. Supported by information entropy theory, the feature importance here is more meaningful than F-score. Then lower-ranked features are removed one by one, each round with a cross-validation assessment on the performance of the selected feature subset.

The flowchart of our approach is shown in Fig. 1 where the new sequence feature space and feature selection scheme are depicted for DNA 4mC site prediction. First, the DNA sequence is encoded into five kinds of features, a total of 292 dimensions. Second, an XGBoost machine is trained and the feature importance scores from the training are used to rank all the features. Last, an SVM-based prediction model is built, and the parameters are optimized with 10-fold cross-validation.

In the results section, we analyze the feature importance in our feature space and show that feature selection improves the model performance significantly in the independent test. Besides, we compare the proposed method with three state-of-art methods, iDNA4mC, 4mCPred, and 4mCPred_SVM in independent test and 10-fold cross-validation on benchmark datasets, and the proposed method achieves much better performance. Two detailed case studies for 4mC site prediction on the *dlk-1* and *DSCAM* genes partly prove the effectiveness of our approach in practical situations.

Results

This section reports the feature importance scores obtained from the XGBoost machine and analyzes the influence of the feature selection on prediction performance. Then three state-of-the-art predictors are compared with the proposed method in the independent test and 10-fold cross-validation on benchmark datasets. At last, we present results from two case studies which were conducted to identify the 4mC sites in the *C.elegans* and *D.melanogaster* genes.



Feature importance analysis

As stated, five types of sequence features are created to constitute a 292-dimensional feature space. Among the 292 dimensions, OHB is from D1 to D164; SNF is from D165 to D205; KNF is from D206 to 225; KSNPF is from D226 to 273 and PseDNC is from D274 to D292. The feature importance scores are obtained from the training process of the XGBoost machine. The importance score distributions for all the datasets are illustrated in Fig. 2. Top 30 feature dimensions are reported in Table S2 of Additional File 1 and feature importance scores of all the feature dimensions are in Additional File 2.

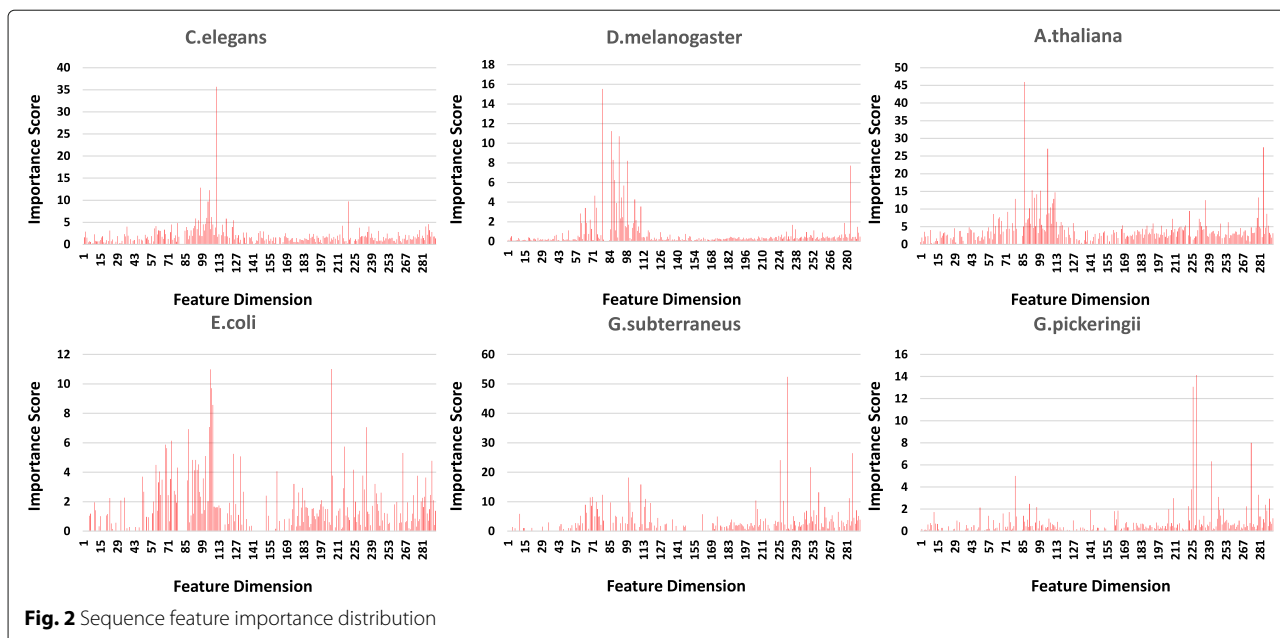
It is understood that each feature dimension has distinct importance scores in different species. OHB and PseDNC features have relatively high average scores in all species. In particular, OHB features have the highest average score in *C.elegans*, *D.melanogaster* and *A.thaliana*. KSNPF feature not only gets a high importance score in *A.thaliana*, *E.coli* and *G.subterraneus* like KNF features, but also has the highest average score in *G.pickeringii*. SNF feature just stands out in *E.coli*. The features' importance score ranges from 0 to 50 and some feature dimensions' scores are such low that they are less important in the classification and may have noise effects on model performance. Thus, the

feature selection before the training is potentially useful to improve model accuracy.

Impact of feature selection on classification

We first evaluate the model performance via independent test without feature selection before model training. Then the independent test is carried out with feature selection, where the benchmark datasets divisions and SVM parameters are kept the same. Table 1 and Fig. 3 show the independent test performance before and after feature selection.

The independent test after feature selection improves the model performance in all the species. In *C.elegans*, feature selection improved Sn, Sp, ACC and MCC by 7.54%, 3.85%, 7.74% and 0.16. In *D.melanogaster*, the model performance has the most considerable improvement by 10.17%, 9.32%, 9.74% and 0.19 for Sn, Sp, ACC and MCC, respectively. For *A.thaliana*, Sp increased by 6.82% while ACC and MCC slightly increased by 2.27% and 0.05. Besides, Sp, ACC and MCC improved by 9.23%, 7.7% and 0.14 in *E.coli* dataset. In *G.subterraneus*, the metrics improvement is by 8.34% for Sn, 6.67% for Sp, 7.5% for ACC and 0.15 for MCC. As for *G.pickeringii*, the performance is improved by 5.17%, 10.73%, 7.89% and



0.15 in terms of Sn, Sp, ACC and MCC with feature selection. From Fig. 4, it's obvious that the AUCs after feature selection become better in all the species. The most massive AUC growth exists in *C.elegans* by 0.06 and the least growth is by 0.01 in *A.thaliana*. The results imply that the proposed feature selection scheme enhances the performance of the SVM model by selecting effective features from the original feature space.

Comparison with state-of-the-art predictors

Three state-of-the-art DNA 4mC prediction methods, iDNA4mC, 4mCPred, and 4mCPred_SVM are compared

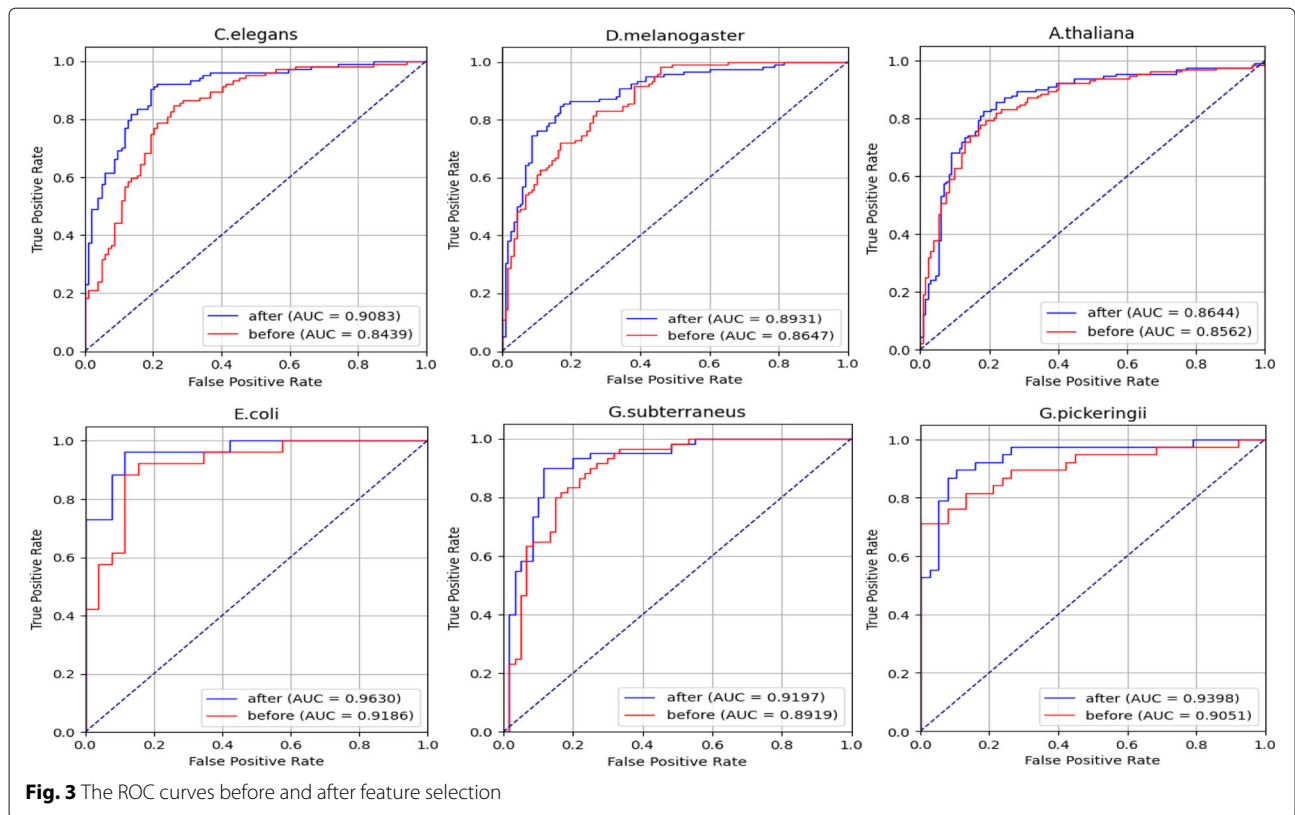
Table 1 The independent test performance before and after feature selection(Sn, Sp and ACC:%)

Datasets	Selection	Sn	Sp	ACC	MCC
<i>C.elegans</i>	before	82.69	75.00	78.85	0.58
	after	94.23	78.85	86.53	0.74
<i>D.melanogaster</i>	before	74.57	77.12	75.85	0.52
	after	84.74	86.44	85.59	0.71
<i>A.thaliana</i>	before	82.57	76.51	79.54	0.59
	after	80.30	83.33	81.81	0.64
<i>E.coli</i>	before	92.30	69.23	80.76	0.63
	after	88.46	88.46	88.46	0.77
<i>G.subterraneus</i>	before	83.33	75.00	79.17	0.59
	after	91.67	81.67	86.67	0.74
<i>G.pickeringii</i>	before	81.57	78.94	80.26	0.61
	after	86.84	89.47	88.15	0.76

with the proposed method. The comparison was conducted using the independent test and cross-validation test on the benchmark datasets.

The independent test results by iDNA4mC and 4mCPred were reported in [24], and we cannot find the independent test results of 4mCPred_SVM method. Since 4mCPred_SVM only provides the final prediction model, it's not available to rebuild the independent test. Thus, here we compare our method with iDNA4mC and 4mCPred in independent test under the same division of training and testing data. The results of independent test are presented in Table 2. Our method outperforms the other methods in all species. Generally, the proposed method improves ACC from 3.02% to 7.89% and increases MCC from 0.06 to 0.15. Especially, a significant improvement of our approach can be observed in *G.pickeringii* (improving Sn by 5.26%, Sp by 10.52%, ACC by 7.89%, and MCC by 0.15).

We performed a 10-fold cross-validation with the same process as the existing methods. The cross-validation results of the three state-of-the-art predictors were reported in the publication of 4mCPred_SVM [25], where the reported performance of 4mCPred has been modified by solving the over-estimated problem. The summary of cross-validations are illustrated in Table 3. Except for the four evaluation metrics, we also list the sample count of TP (True Positive), FN (False Negative), FP (False Positive) and TN (True Negative). As shown in the table, in *D.melanogaster*, *A.thaliana* and *Gpickeringii*, our method has the most TP and TN counts, increasing ACC by 0.7% to 1.7% and MCC by 0.015 to 0.033. In *G.subterraneus*, our method has the highest TN, improving more ACC



and MCC by 1% and 0.02% than 4mC_SVM which has the second-best performance. Additionally, the TP and TN of our method are not the highest in C.elegans and E.coli, but our method slightly improve the ACC and MCC by 1% and 0.02 in E.coli and has a comparative performance with 4mCPred, better than other two methods in C.elegans.

It's clear that our method achieves better overall performance than the existing predictors in independent and cross-validation tests. The improvement of ACC indicates

that our method accurately identifies more 4mC sites and the increase of MCC means that our method has more balanced performance for classifying positive and negative samples. Therefore, our method is more effective to identify DNA 4mC sites than the existing predictors.

Case studies

To confirm the effectiveness of our method to solve practical problems, two detailed case studies are conducted.

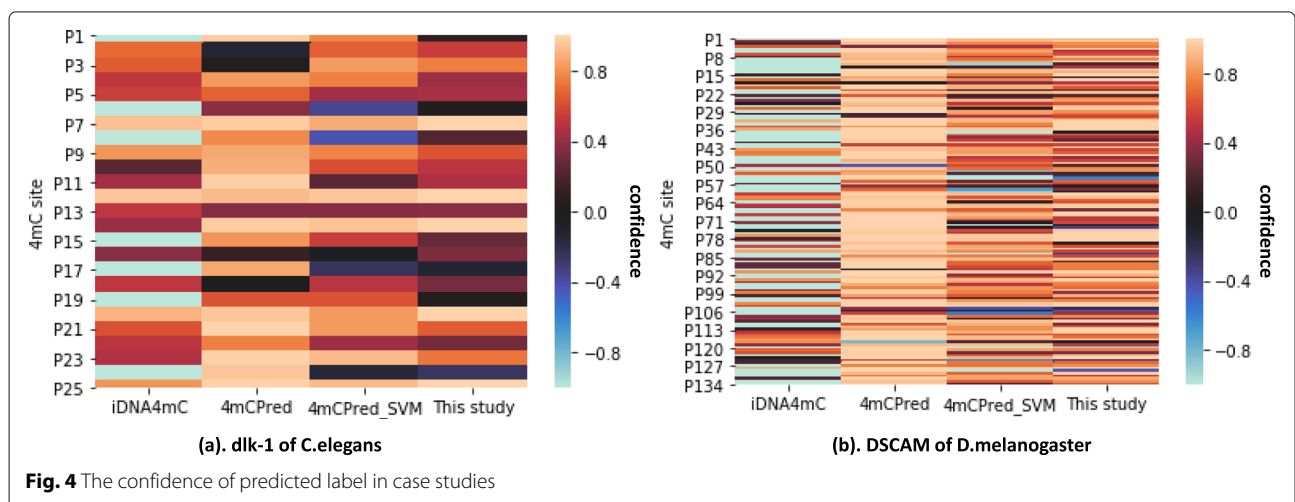


Table 2 Independent Test Results on Benchmark Datasets (Sn, Sp and ACC:%)

Methods	Datasets	Sn	Sp	ACC	MCC
iDNA4mC	C.elegans	80.77	73.08	76.92	0.54
	D.melanogaster	74.58	77.97	76.27	0.53
	A.thaliana	80.3	77.27	78.79	0.58
	E.coli	96.15	69.23	82.69	0.68
	G.subterraneus	85.00	76.67	80.83	0.62
	G.pickeringii	81.58	78.95	80.26	0.61
4mCPred	C.elegans	85.58	78.85	82.21	0.65
	D.melanogaster	83.90	81.36	82.63	0.65
	A.thaliana	76.52	76.52	76.52	0.53
	E.coli	84.62	80.77	82.69	0.65
	G.subterraneus	91.67	75.00	83.33	0.68
	G.pickeringii	86.84	68.42	77.63	0.56
this study	C.elegans	94.23	78.85	86.53	0.74
	D.melanogaster	84.74	86.44	85.59	0.71
	A.thaliana	80.30	83.33	81.81	0.64
	E.coli	88.46	88.46	88.46	0.77
	G.subterraneus	91.67	81.67	86.67	0.74
	G.pickeringii	86.84	89.47	88.15	0.76

C.elegans and *D.melanogaster* are model organisms widely applied in human disease-related research works, like Parkinson and human aging research investigations [33–36]. As 4mC plays critical roles in DNA expression and replication in these models, we describe how our method can help identify 4mC sites more accurately in the related genes. We focus on the *dlk-1* gene which can promote mRNA stability and local translation in *C.elegans* [37], and on *DSCAM* gene which can contribute to the specificity of neuronal connectivity in *D.melanogaster* [38].

The 26 and 137 validated 4mC sites in *dlk-1* and *DSCAM* gene are collected from the MethSMRT database. The collected 4mC-contained DNA sequences are all 41-bit, that can be directly submitted into the web tools of three state-of-the-art methods. The prediction result are depicted in Fig. 4 and Table 4. Figure 4 shows the label confidence predicted by these four predictors, where the positive confidence refers that the corresponding site is predicted to be 4mC site and the negative confidence means the site is predicted to be a non-4mC site. As shown in the figure, iDNA4mC achieves the worst performance in both two cases, and half of the predictions are incorrect in *DSCAM* gene. 4mCPred, 4mCPred_SVM and the proposed method have similar performance in the *DSCAM* gene case, while the results made by 4mCPred and our proposed method on the *dlk-1* gene are better than 4mCPred_SVM.

More details of the prediction are presented in Table 4. Since the testing data in the case study only contains positive samples, there are only TP and FN counts in the results. For the *dlk-1* case, 4mCPred has only one wrong prediction and the proposed method has made two false predictions out of 26 samples, while iDNA4mC and 4mCPred_SVM have 7 and 6 incorrect predictions respectively. For the *DSCAM* case, there are 137 4mC sites tested, and our proposed method has made 126 correct predictions (i.e., only 11 incorrect predictions). 4mCPred and 4mCPred_SVM have 16 and 15 false predictions, while iDNA4mC has made 67 false predictions. More detailed results can be found at the supplementary Additional file 3.

Discussion

To improve the performance, we have focused on choosing more efficient features for 4mC site prediction, including extracting better sequence feature and feature selection before model learning. However, there are also some limitations in the study: first, the feature are mostly from the content of sequence, not the biological characters; second, the size of training data is limited.

In the future, we will continue to optimize our feature space with novel sequence features of important biological characteristics. Furthermore, we will expand the size of the benchmark datasets to enhance the model's accuracy and generalization ability. Also, since the number of 4mC is much smaller than non-4mC sites in practical situations, the data imbalance will be considered in the next research. At last, we will apply our method to solve other sequence site prediction problems.

Conclusions

The 4mC site prediction is a typical sequence site classification problem. The state-of-the-art research work have made some explorations, but their performance still needs improvement. For this purpose, we propose to construct a more effective feature space, integrating five types of sequence features, and suggest to use a novel learning algorithm with XGBoost based feature selection scheme. The results show that the feature selection improves the performance, and the prediction model outperforms the other three existing predictors in the independent tests and the cross-validations.

Methods

Based on the benchmark datasets, this paper proposed a new sequence feature space and a machine learning algorithm with feature selection scheme. In the sequence encoding, five types of sequence features are integrated to form a 292-dimension feature space, representing both global and local sequence characteristics. Then a feature selection scheme is applied, by which feature importance

Table 3 Cross Validation Result on Benchmark Datasets (Sn, Sp and ACC.%; TP: true positive, FN: false negative, FP: false positive, TN: true negative)

Datasets	Methods	Sn	Sp	ACC	MCC	TP	FN	FP	TN
C.elegans	iDNA4mC	79.7	77.5	78.6	0.572	1328	316	349	1205
	4mCPred	82.5	82.6	82.6	0.652	1282	272	270	1284
	4mCPred_SVM	82.4	80.7	81.5	0.631	1280	274	300	1254
	this study	84.9	80.4	82.6	0.653	1319	235	305	1249
D.melanogaster	iDNA4mC	83.3	79.1	81.2	0.625	1474	295	369	1400
	4mCPred	82.4	82.1	82.2	0.646	1458	311	317	1452
	4mCPred_SVM	83.8	82.2	83.0	0.661	1483	286	314	1455
	this study	85.4	83.2	84.3	0.686	1510	259	297	1472
A.thaliana	iDNA4mC	75.7	76.2	76.0	0.519	1498	480	471	1507
	4mCPred	75.5	78.0	76.8	0.536	1494	484	435	1543
	4mCPred_SVM	77.8	79.6	78.7	0.573	1538	440	404	1574
	this study	78.3	80.5	79.4	0.589	1549	429	385	1593
E.coli	iDNA4mC	82.0	77.8	79.9	0.598	318	70	86	302
	4mCPred	81.9	83.2	82.6	0.655	318	70	65	302
	4mCPred_SVM	85.8	80.7	83.3	0.666	333	51	67	321
	this study	86.1	82.5	84.3	0.686	334	54	68	320
G.subterraneus	iDNA4mC	82.2	80.8	81.5	0.630	745	161	174	732
	4mCPred	81.8	83.7	82.8	0.662	742	164	148	758
	4mCPred_SVM	84.0	83.4	83.7	0.674	760	145	150	755
	this study	83.6	85.7	84.7	0.694	757	148	129	776
G.pickeringii	iDNA4mC	82.4	83.8	83.1	0.663	469	100	92	477
	4mCPred	85.0	81.0	83.0	0.668	484	85	108	461
	4mCPred_SVM	86.3	85.8	86.0	0.721	491	78	81	488
	this study	86.3	89.1	87.7	0.754	491	78	62	507

score derived from the training process of the XGBoost machine is taken as the criterion of feature selection. Then an SVM prediction model is trained under the selected features and optimized by 10-fold cross-validation.

Benchmark datasets

From the DNA 4mC database MethSMRT, Chen et al constructed the benchmark databases containing *Caenorhabditis elegans* (*C.elegans*), *Drosophila melanogaster* (*D.melanogaster*), *Arabidopsis thaliana* (*A.thaliana*), *Escherichia coli* (*E.coli*), *Geobacter subterraneus* (*G.subterraneus*) and *Geobacter pickeringii* (*G.pickeringii*) [23]. The benchmark datasets are obtained from Chen’s work. According to the reference, the 41-bit 4mC-centred DNA sequences were obtained from MethSMRT with a Modification QV threshold of 30. The CD-HIT software was used to remove the redundant positive samples. The same number of negative samples were selected randomly to construct a balanced dataset. The negative samples were also 41-bit cytosine-centered DNA sequences and were not detected by SMRT. To compare

with the existing predictors, we use the same division of the datasets for independent tests. The summary of benchmark datasets is listed in Table 5.

Feature space construction

To visualize the difference between the positive and negative sequences, the sequence logos of all the six

Table 4 4mC site identification in case studies (TP: True Positive, FN: False Negative)

Case	Methods	Total	TP	FN
dlk-1	iDNA4mC	26	19	7
	4mCPred	26	25	1
	4mCPred_SVM	26	20	6
	This study	26	24	2
DSCAM	iDNA4mC	137	70	67
	4mCPred	137	121	16
	4mCPred_SVM	137	122	15
	This study	137	126	11

Table 5 Summary of six benchmark datasets

Species	Positive Sample	Negative Sample	Total
<i>C.elegans</i>	1554	1554	3108
<i>D.melanogaster</i>	1769	1769	3538
<i>A.thaliana</i>	1978	1978	3956
<i>E.coli</i>	388	388	776
<i>G.subterraneus</i>	906	906	1812
<i>G.pickeringii</i>	569	569	1138

species are plotted using the web tool ‘two sample logos’ [39]. See Fig. 5. It is clear that the sequence characteristics are distinct among the six species; especially positions near the 4mC sites exhibit different patterns in positive and negative samples. In addition, the adjacent nucleotide and spectrum nucleotide across the entire sequence have specific patterns in different label groups. Thus an expanded feature space combining global and local patterns is good to construct accurate models for all the species. Among the existing methods, iDNA4mC only use nucleotide chemical property and frequency feature, which cannot extract the local adjacent nucleotide patterns; in 4mCPred and 4mCPred_SVM, the features mainly focus on the trinucleotide or dinucleotide nucleotide patterns, ignoring the spectrum nucleotide patterns. In this study, the feature space covers five types of features, one-hot 4-bit binary feature (OHB), sequential nucleotide frequency (SNF), k-nucleotide frequency

(KNF), k-spectrum nucleotide pair frequency (KSNPF) and PseDNC. The OHB and SNF feature possess the information of the whole sequence and represent the global sequential properties, while KNF, KSNPF, and PseDNC features capture the local sequence patterns.

One-hot binary feature

The one-hot binary feature is the most widely used sequence representation feature. It converts each of the nucleotides in the DNA sequence into a 4-bit vector, which contains only one ‘1’. The length of the OHB feature is related to the number of nucleotide types and length of the sequence. Since the DNA sample sequence here is 41-bit and has four types of nucleotide, the one-hot binary feature is 164 bits. The encoding rules in this study are as follows: ‘A’-(1,0,0,0), ‘G’-(0,1,0,0), ‘T’-(0,0,1,0), ‘C’-(0,0,0,1). From the rule, it is obvious that the OHB feature is sparser than 2-bit or 3-bit binary features. But, the one-hot binary feature makes it more reasonable to calculate the importance score for each dimension in feature space and to discover local motifs.

Sequential nucleotide frequency

The sequential nucleotide frequency, also known as nucleotide density, is the frequency that the corresponding nucleotide occurs before the current position. SNF is commonly used together with the binary encoding feature as a global density feature. For an n-bit long sequence, SNF calculates n values for each position in the sequence and

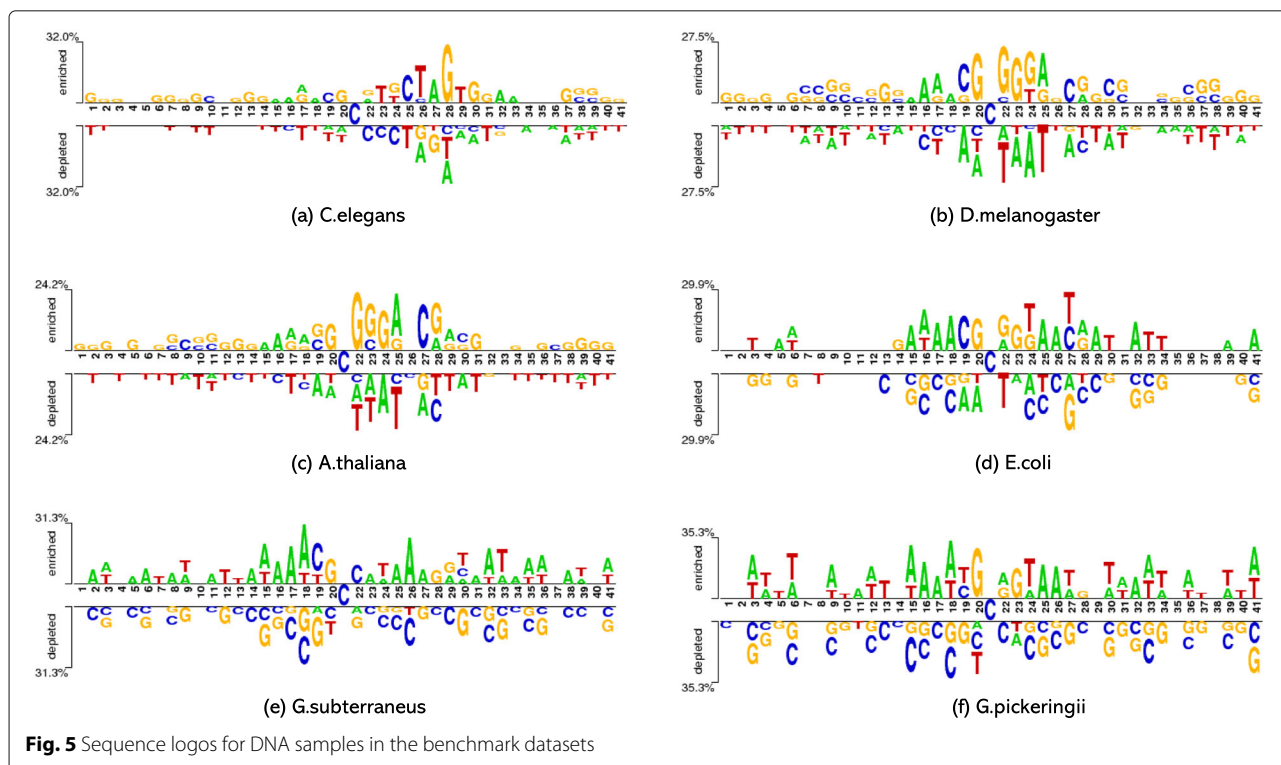


Fig. 5 Sequence logos for DNA samples in the benchmark datasets

produces an n -dimensional feature that starts with '1'. The SNF feature d_i is defined as:

$$d_i = \frac{1}{|S_i|} \sum_{j=1}^i f(s_j) f(s_j) = \begin{cases} 1 & s_j = s_i \\ 0 & s_j \neq s_i \end{cases} \quad (1)$$

where S_i denotes the length of sequence before the current position i and s_i is the nucleotide at position i . For example, a sequence like 'AACGTACT' can be converted into the SNF feature vector (1, 0.5, 0.33, 0.25, 0.2, 0.5, 0.28, 0.25).

k-nucleotide frequency

The k-nucleotide (k-mer) frequency is a classic concept in DNA sequence encoding. KNF feature is the frequency that adjacent k nucleotides occur in the whole sequence. The length of the KNF feature vector is 4^k , determined by the parameter k . The calculation of KNF is as below:

$$F(n_1 n_2 \dots n_k) = \frac{C(n_1 n_2 \dots n_k)}{S - k + 1} \quad (2)$$

where $n_1 n_2 \dots n_k$ donates the adjacent k nucleotides and $n_i \in (A, C, G, T)$. F and C is the feature value and total count of the adjacent nucleotides, while S is the length of sequence. When $k = 1$, the KNF is a vector like (F_A, F_C, F_G, F_T) ; when $k=2$, the KNF of a sequence is like $(F_{AA}, F_{AC}, F_{AG}, F_{AT}, F_{CA}, F_{CC}, F_{CG}, F_{CT}, F_{GA}, F_{GC}, F_{GG}, F_{GT}, F_{TA}, F_{TC}, F_{TG}, F_{TT})$ with a dimension of $4^2 = 16$.

k-spectrum nucleotide pair frequency

The KSNPF feature depicts the sequence context by calculating the frequency of k-spaced nucleotide pairs (e.g., AXXT is a two-spaced nucleotide pair, and CXXXG is a three-spaced nucleotide pair). Like the adjacent nucleotides pair above, the feature dimension of the KSNPF is 16 for each k . The calculation of this feature is as follows:

$$F(n_1 X \dots X n_2) = \frac{C(n_1 X \dots X n_2)}{S - k - 1} \quad (3)$$

where $n_1 X \dots X n_2$ donates the k-spaced nucleotides pair and $n_i \in (A, C, G, T)$.

PseDNC

As an essential sequence feature, PseDNC combines global and local structural properties and has been widely used in sequence site prediction problems [40]. For a DNA sequence, the PseDNC feature is a vector:

$$F_{PseDNC} = [d_1, d_1, \dots, d_{16} d_{16} \dots d_{16+\lambda}]^T \quad (4)$$

where,

$$d_k = \begin{cases} \frac{f_k}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (1 \leq k \leq 16) \\ \frac{w \theta_{\mu-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (16 < k \leq 16 + \lambda) \end{cases} \quad (5)$$

where f_k denotes the normalized frequency of two adjacent nucleotide pairs; w is the weight factor, and θ is the correlation factor of j-tier, representing the correlation of all j-tier from the sequence. The definition of θ is:

$$\theta_j = \frac{1}{L - j - 1} \sum_{i=1}^{L-j-1} \Theta_{i,i+j} (j = 1, 2, \dots, \lambda; \lambda < L) \quad (6)$$

where Θ is the correlation function and given by:

$$\Theta_{i,i+j} = \frac{1}{\mu} \sum_{u=1}^{\mu} [P_u(R_i R_{i+1}) - P_u(R_j R_{j+1})]^2 \quad (7)$$

where μ is the length of sequence; $P_u(R_i R_{i+1})$ is the numerical value of the u-th DNA local property for the adjacent nucleotide pair $R_i R_{i+1}$ at position i . In this study, PseDNC feature is computed by a python package 'repDNA' [41] and the λ value is default to 3. The names of 38 DNA local properties utilized in the definitions here are detailed in the supplementary Table S1 of Additional File 1.

Feature selection scheme

Feature selection can reduce the dimension of feature space and speed up the model training. A lot of feature selection strategies have been employed in machine learning [42]. In particular, a filter feature selection scheme has been used to improve the prediction performance. The filter feature selection scheme has two steps: first, F-score is calculated for each dimension in feature space according to the relevance between feature and label; second, a selection strategy called SFS is adopted to ascertain the feature subset. In this study, we proposed an embedded feature selection method also with two steps. However, we rank features with importance scores produced from the XGBoost training process [43] and select the top features with cross-validations.

In our method, XGBoost is the predefined classifier to analyze the feature importance. XGBoost has been proved to be an efficient tool in data science. In the training process, the XGBoost classifier calculates the feature importance score for each dimension based on the dimension location and the split efficiency in the boosting tree. In this study, XGBoost is implemented with a python package 'xgboost' of version 0.90. The feature importance scores are obtained through the function 'get_score'. According to the calculation method, the feature importance score has 5 types: 'weight', 'gain', 'cover', 'total_weight', 'total_gain' and here we use the default 'weight' importance score.

With the importance scores derived by the XGBoost classifier, feature dimensions are ranked from the highest to the lowest. Then the lower-ranked features are removed from the feature space one by one, and the feature subset performance is evaluated by 10-fold cross-validation with a support vector machine. The feature subset with the best

performance is taken as the final feature space for 4mC prediction.

Support vector machine

Support vector machine (SVM) is a popular machine learning classifier and has been proved to be more efficient than the other algorithms for DNA 4mC prediction in the state-of-the-art researches [25]. In this study, SVM is implemented by using the python package 'scikit - learn(vision0.22)' [44]. The kernel function of the SVM prediction model is set as a radial basis kernel function (RBF). The hyperparameter C and γ are optimized by a grid search with cross-validations and the search ranges are listed below:

$$\begin{cases} 2^{-5} \leq C \leq 2^{10} & \text{step} = 2 \\ 2^{-15} \leq \gamma \leq 2^2 & \text{step} = 2^{-1} \end{cases} \quad (8)$$

With the output of the probability scores, the ROC curve can be plotted. The threshold of probability score is set as 0.5 to obtain the predicted label. Here, we compare SVM with other three traditional machine learning methods, such as Random Forest, Naive Bayes and Neural Network, and the results are reported in Table S3 of Additional File 1.

Performance evaluation metrics

To compare with the existing predictors, the evaluation metrics in this study are consistent with the state-of-the-art methods, including Sensitivity(Sn), Specificity(Sp), Accuracy(ACC) and Matthews correlation coefficient(MCC). The definitions of these four metrics are as follows:

$$\begin{cases} Sn = \frac{TP}{TP+FN} \times 100\% \\ Sp = \frac{TN}{TN+FP} \times 100 \\ Acc = \frac{TP+TN}{TP+FN+TN+FP} \times 100 \\ Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \end{cases} \quad (9)$$

Sn shows the model capability of identifying positive samples, while Sp tells the capacity of classifying negative samples; ACC is the prediction accuracy of all samples; MCC evaluates the overall performance of a predictor. In this study, the receiver operating characteristic(ROC) curve is also used to analyze model performance. The ROC curve is plotted in a coordinate graph where the x-axis is the false positive rate(1-Sp) and the y-axis is the true positive rate(Sn). The area under the curve(AUC) evaluates the classification performance, and larger AUC means better performance.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-020-07033-8>.

Additional file 1: Supplementary tables. Table S1: The 38 physico-chemical property in PseDNC feature calculation of repDNA package. Table S2: The top 30 feature dimensions of feature importance score ranking in six species.

Additional file 2: Feature importance scores. In Additional File 2, the importance scores for each feature dimension are reported. The feature scores are listed by the position number in all six species.

Additional file 3: Case study results. In Additional File 3, the 4mC-contained DNA sequence and the prediction results of four predictors are reported.

Abbreviations

4mC: DNA N4-methylcytosine(4mC); OHB: one hot binary; SNF: sequential nucleotide frequency; KNF: k-nucleotide frequency; KSNPF: k-spectrum nucleotide pair frequency; ROC: the receiver operating characteristic curve; MCC: Matthews correlation coefficient; ACC: Accuracy

Acknowledgements

Not Applicable.

Authors' contributions

ZZ: Collected data, proposed predictor model, designed experiment and wrote the paper. XZ: Reviewed and edited the paper. LF: reviewed and edited the paper. FC: reviewed and edited the paper. JL: Designed experiment, reviewed and edited paper. All authors have read and approved the manuscript.

Funding

Publication of this supplement was funded by Faculty of Engineering and Information Technology, University of Technology Sydney. This research work was partially supported by the China Scholarship Council (CSC).

Availability of data and materials

The benchmark datasets analyzed during the current study are available in 4mCPred website (<http://server.malab.cn/4mCPred/data.jsp>). The data used in case study are collected from MethSMRT database (<http://sysbio.gzcc.com/methsmrt/index.html>). The three state-of-art predictors are available: iDNA4mC at (<http://lin-group.cn/server/iDNA4mC.php>), 4mCPred at (<http://server.malab.cn/4mCPred/index.jsp>) and 4mCPred_SVM at (<http://server.malab.cn/4mCPred-SVM/>). The source code and datasets in this study are freely available at <https://github.com/Zhixun-Zhao/4mCPrediction>.

Ethics approval and consent to participate

Not applicable. All utilized data sets are publicly and freely available which do not require any ethics approval and consent to participate.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Advanced Analytics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007 Sydney, Australia. ²Data Science Institute, University of Technology Sydney, PO Box 123, Broadway, NSW 2007 Sydney, Australia. ³School of Computer, National University of Defense Technology, 410073 Changsha, China.

Received: 6 May 2020 Accepted: 27 August 2020

Published online: 11 September 2020

References

- Rathi P, Maurer S, Summerer D. Selective recognition of N 4-methylcytosine in DNA by engineered transcription-activator-like effectors. *Philos Trans R Soc B Biol Sci.* 2018;373(1748):20170078.

2. Stoiber MH, Quick J, Egan R, Lee JE, Celniker SE, Neely R, Loman N, Pennacchio L, Brown JB. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *BioRxiv*. 2016094672.
3. Chen K, Zhao BS, He C. Nucleic acid modifications in regulation of gene expression. *Cell Chem Biol*. 2016;23(1):74–85.
4. Davis BM, Chao MC, Waldor MK. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr Opin Microbiol*. 2013;16(2):192–8.
5. Korlach J, Turner SW. Going beyond five bases in DNA sequencing. *Curr Opin Struct Biol*. 2012;22(3):251–61.
6. Roberts RJ, Vincze T, Posfai J, Macelisz D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res*. 2015;43(D1):298–9.
7. Robertson KD. DNA methylation and human disease. *Nat Rev Genet*. 2005;6(8):597–610.
8. Jin B, Li Y, Robertson KD. DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes & Cancer*. 2011;2(6):607–17.
9. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 2012;13(7):484–92.
10. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*. 2009;324(5929):930–5.
11. Heyn H, Esteller M. An adenine code for DNA: a second life for N6-methyladenine. *Cell*. 2015;161(4):710–3.
12. Fu Y, Luo G-Z, Chen K, Deng X, Yu M, Han D, Hao Z, Liu J, Lu X, Doré LC, Weng X. N6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. *Cell*. 2015;161(4):879–92.
13. Greer EL, Blanco MA, Gu L, Sendinc E, Liu J, Aristizábal-Corales D, Hsu C-H, Aravind L, He C, Shi Y. DNA methylation on N6-adenine in *C. elegans*. *Cell*. 2015;161(4):868–78.
14. Zhang G, Huang H, Liu D, Cheng Y, Liu X, Zhang W, Yin R, Zhang D, Zhang P, Liu J, Li C. N6-methyladenine DNA modification in *Drosophila*. *Cell*. 2015;161(4):893–906.
15. Cheng X. DNA modification by methyltransferases. *Curr Opin Struct Biol*. 1995;5(1):4–10.
16. Modrich P. Mechanisms and biological effects of mismatch repair. *Annu Rev Genet*. 1991;25(1):229–53.
17. Messer W, Noyer-Weidner M. Timing and targeting: the biological functions of Dam methylation in *E. coli*. *Cell*. 1988;54(6):735–7.
18. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. Shotgun bisulfite sequencing of the arabidopsis genome reveals DNA methylation patterning. *Nature*. 2008;452(7184):215–9.
19. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*. 2008;133(3):523–36.
20. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*. 2010;7(6):461.
21. Yu M, Ji L, Neumann DA, Chung D, Groom J, Westpheling J, He C, Schmitz RJ. Base-resolution detection of N⁴-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite-sequencing. *Nucleic Acids Res*. 2015;43(21):148.
22. 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*. 2016;950.
23. 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):3518–33.
24. He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics*. 2019;35(4):593–601.
25. 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.
26. Manavalan B, Basith S, Shin TH, Lee DY, Wei L, Lee G, et al. 4mCPred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome. *Cells*. 2019;8(11):1332.
27. Hasan MM, Manavalan B, Shoombuatong W, Khatun MS, Kurata H. i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. *Comput Struct Biotechnol J*. 2020;18:906–912.
28. Lv Z, Wang D, Ding H, Zhong B, Xu L. Escherichia coli DNA N4-methylcytosine site prediction accuracy improved by light gradient boosting machine feature selection technology. *IEEE Access*. 2020;8:14851–9.
29. 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*. 2020;157:752–758.
30. 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.
31. Wei L, Su R, Luan S, Liao Z, Manavalan B, Zou Q, Shi X. Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*. 2019;35(23):4930–7.
32. Wei H-L, Billings SA. Feature subset selection and ranking for data dimensionality reduction. *IEEE Trans Pattern Anal Mach Intell*. 2006;29(1):162–6.
33. Feany MB, Bender WW. A drosophila model of parkinson's disease. *Nature*. 2000;404(6776):394–8.
34. Auluck PK, Chan HE, Trojanowski JQ, Lee VM-Y, Bonini NM. Chaperone suppression of α -synuclein toxicity in a *Drosophila* model for parkinson's disease. *Science*. 2002;295(5556):865–8.
35. Van Ham TJ, Thijssen KL, Breiting R, Hofstra RM, Plasterk RH, Nollen EA. *C. elegans* model identifies genetic modifiers of α -synuclein inclusion formation during aging. *PLoS Genet*. 2008;4(3):1000027.
36. Feng Z, Li W, Ward A, Piggott BJ, Larkspur ER, Sternberg PW, Xu XS. A *C. elegans* model of nicotine-dependent behavior: regulation by TRP-family channels. *Cell*. 2006;127(3):621–33.
37. Yan D, Wu Z, Chisholm AD, Jin Y. The DLK-1 kinase promotes mRNA stability and local translation in *C. elegans* synapses and axon regeneration. *Cell*. 2009;138(5):1005–18.
38. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*. 2000;101(6):671–84.
39. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188–90.
40. Chen W, Feng P-M, Lin H, Chou K-C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res*. 2013;41(6):68.
41. Liu B, Liu F, Fang L, Wang X, Chou K-C. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*. 2015;31(8):1307–9.
42. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: A data perspective. *ACM Comput Surv*. 2017;50(6):1–45.
43. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*; 2016. p. 785–94.
44. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. *Journal of machine learning research*. 2011;12:2825–30.

Publisher's Note

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