Epigenetics & Chromatin

METHODOLOGY

Open Access

Identification of DNA N^6-methyladenine sites by integration of sequence features

Hao-Tian Wang[1,2,3,5], Fu-Hui Xiao[1,2,3], Gong-Hua Li[1,2,3] and Qing-Peng Kong[1,2,3,4*]

Abstract

Background: An increasing number of nucleic acid modifications have been profiled with the development of sequencing technologies. DNA N^6-methyladenine (6mA), which is a prevalent epigenetic modification, plays important roles in a series of biological processes. So far, identification of DNA 6mA relies primarily on time-consuming and expensive experimental approaches. However, in silico methods can be implemented to conduct preliminary screening to save experimental resources and time, especially given the rapid accumulation of sequencing data.

Results: In this study, we constructed a 6mA predictor, p6mA, from a series of sequence-based features, including physicochemical properties, position-specific triple-nucleotide propensity (PSTNP), and electron–ion interaction pseudopotential (EIIP). We performed maximum relevance maximum distance (MRMD) analysis to select key features and used the Extreme Gradient Boosting (XGBoost) algorithm to build our predictor. Results demonstrated that p6mA outperformed other existing predictors using different datasets.

Conclusions: p6mA can predict the methylation status of DNA adenines, using only sequence files. It may be used as a tool to help the study of 6mA distribution pattern. Users can download it from https://github.com/Konglab404/p6mA.

Keywords: DNA N^6-methyladenine, Machine learning, XGBoost

Background

DNA N^6-methyladenine (6mA) is an important epigenetic modification of nucleic acid, firstly characterized in bacteria [1]. In contrast to 5mC, 6mA remains poorly studied and was previously thought to only occur in prokaryotes [2]. Accumulating evidences, however, has confirmed that it also exists in eukaryotes, including zoological and botanical species (e.g., *Arabidopsis thaliana*, *Mus musculus*, *Danio rerio*, and *Sus scrofa*). Recently, two studies found that DNA 6mA sites also exist extensively in the genomes of humans [3] and rice [4], thus deepening our understanding of this modification in high-grade organisms.

DNA 6mA plays important roles in various biological processes, such as the restriction–modification system [5, 6], DNA replication and repair [7, 8], nucleoid segregation [9, 10], and transcription [11]. To detect DNA 6mA modification, a series of experimental methods have been developed, such as methylated DNA immunoprecipitation sequencing [12], liquid chromatograph–tandem mass spectrometry [3], capillary electrophoresis and laser-induced fluorescence [13], and single-molecule real-time sequencing (SMRT-seq) [14]. However, these experimental procedures are expensive and time-consuming and thus largely limited its application in DNA 6mA study, urging the necessity for the development of bioinformatics-based approaches to predict methylated adenine sites in genomes.

Machine learning builds models by handling features to perform specific tasks and has been widely applied in biological issues, including post-transcription RNA

*Correspondence: kongqp@mail.kiz.ac.cn
[4] KIZ/CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming 650223, China
Full list of author information is available at the end of the article

identification [15–17], promoter discovery [18–20], and nucleotide modification prediction [21–23]. The occurrence of 6mA relies on the properties of its surrounding sequences, which play vital roles in methyltransferase/demethylase-dependent catalytic processes [3, 24, 25]. Recently, some machine learning-based predictors, e.g., iDNA6mA-PseKNC [26] and i6mA-Pred [27], were developed to identify 6mA sites at the genomic level. The former was trained with mouse data and achieved a high recall ratio in several datasets, whereas the latter was designed to predict 6mA sites in rice. To the best of our knowledge, however, there is no 6mA predictor trained on multi-species data.

In this study, we constructed a predictor, p6mA, to identify DNA 6mA sites by sequence-based features. The predictor was trained on dataset from four species: i.e., *Oryza sativa* (rice), *Drosophila melanogaster* (fruit fly), *Caenorhabditis elegans* (worm), and *Homo sapiens* (human). The DNA sequences were transformed into numeric vectors by extracting 172 features. We selected key features using the maximum relevance maximum distance (MRMD) method [28] and constructed the predictor using the Extreme Gradient Boosting (XGBoost) algorithm [29]. Comparison with other existing tools demonstrated that p6mA outperformed other methods in several aspects. Users can download p6mA from https://github.com/Konglab404/p6mA.

Results

Nucleotide composition and conservation analysis

In this study, we constructed an aggregated benchmark dataset by four species' data. There are 3040 positive samples and 3040 negative samples (Table 1). All the samples are 41 nt long with an adenosine (A) in the center. We adopted Two Sample Logos [30] to visualize significantly overrepresented and underrepresented sites with a threshold of $p < 0.05$. The nucleotide enrichment status, as shown in Fig. 1a, showed that there exists nucleotide distribution bias between 6mA and non-6mA containing sequences. For example, in 6mA-containing sequences, GAGG motif was enriched in center and adenosine was enriched in the +4 nt position. The above results indicated that the surrounding nucleotide composition information can be adopted to discriminate 6mA and non-6mA sites.

Next, we investigated whether sequence bias can cause differences in conservatism. We then performed entropy analysis, aiming to determine trinucleotide-positioned conservatism differences between 6mA and non-6mA sites [31]. The entropy of trinucleotides at each position was calculated as follows:

$$\mathrm{Ent}_i = -\sum_{j-1}^{n} p(3\mathrm{mer}_j|i) \cdot \log_2 p(3\mathrm{mer}_j|i), \quad (1)$$

where n denotes the total number of trinucleotide combinations in the ith position and $p(3\mathrm{mer}_j|i)$ denotes the frequency of the jth trinucleotide at the ith position in the positive/negative samples. Two 39-dimensional numerical vectors were generated to express the entropy values at positions of positive and negative samples.

Information entropy was used to evaluate chaos in the signal processing field and help to reflect conservatism [31]. A lower entropy value, which means less chaos, indicates that the site concerned is more conserved. Figure 1b shows the comparison of trinucleotide entropy at different positions in the 6mA and non-6mA samples. Samples with 6mA sites display lower entropy values, especially at center adenine positions, than those with non-6mA sites. Our results showed that the positive samples possessed more conservatism that negative samples in specified positions, especially in regions surrounding center adenine sites.

The Two Sample Logos and entropy analysis results both supported that positioned nucleotide information is able to discriminate between 6mA and non-6mA sites, thus providing a reasonable basis for the application of positioned sequence feature extraction methods like PSTNP.

Table 1 The statistics of benchmark dataset in this study

Dataset	# Positive samples	# Negative samples	Reference genome
O. sativa	880	880	MH63
D. melanogaster	728	728	dm3
C. elegans	632	632	ce10
H. sapiens	800	800	hg38
Aggregated	3040	3040	–

Feature selection and parameter tuning

We used three methods (i.e., PSTNP, EIIP, and physicochemical properties) to extract features. Each sample was transformed into a 172-dimensional numerical vector, though the feature set also included redundant features. To reduce computational resource waste, we used MRMD score, an index positively related to feature importance, and incremental feature selection (IFS) to select optimal feature sets for each dataset. Features were ranked by MRMD score from highest to lowest. The features from the ranked list were then added one-by-one to a new set and used to construct an XGBoost-based model with default parameters. Model performance was evaluated by tenfold cross-validation and the feature set

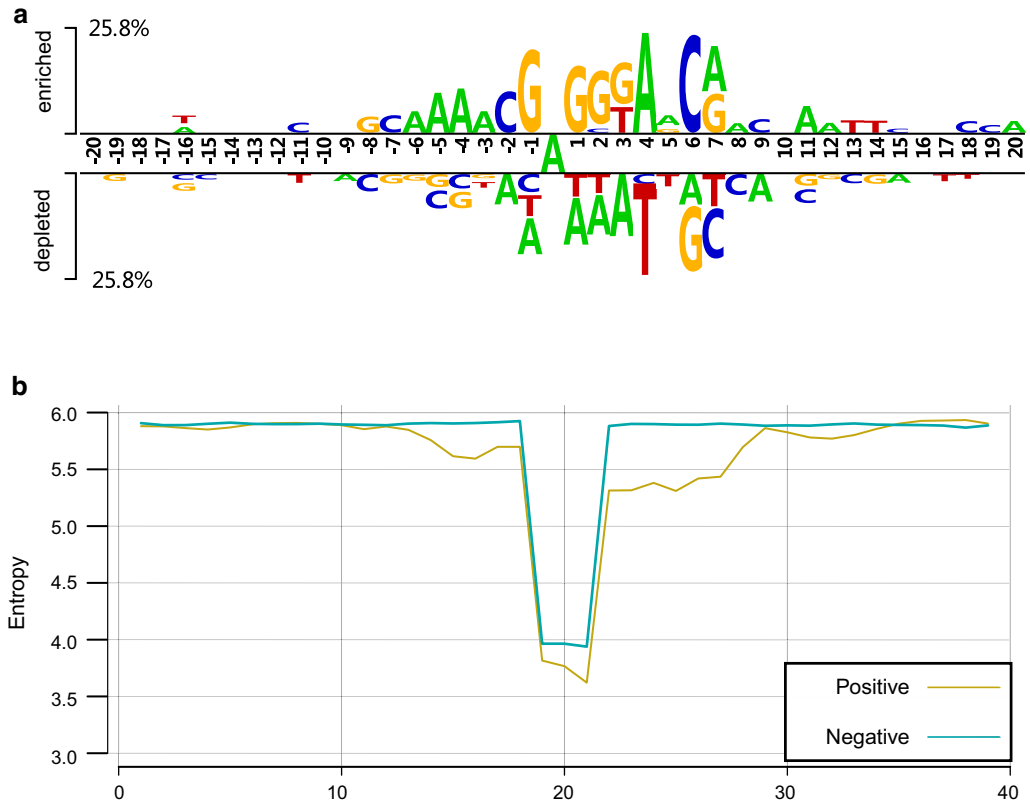

Fig. 1 Nucleotide composition of benchmark dataset. **a** Two Sample Logos result of benchmark dataset, top panel denotes the nucleotide enrichment status of 6mA-containing sequences and bottom panel is of non-6mA-containing sequences. **b** Entropy analysis of 6mA- and non-6mA-containing sequence. Red line denote 6mA-containing sequences and blue one denotes non-6mA sequences

with highest accuracy was chosen as the optimal set. As shown in Fig. 2a, the highest accuracy (82.47%) was obtained when the optimal 124 features were included. Therefore, we trained the model by its top-ranked 124 features.

We then trained models for each dataset using their optimal feature sets. To obtain better performance, a grid search strategy was used to conduct model tuning. Three parameters (i.e., gamma, eta, and max_depth) of XGBoost were optimized in the spaces [0, 0.2], [0.1, 0.5], and [2, 10] with steps of 0.1, 0.1, and 2, respectively. We trained 525 models and the parameter set with the highest tenfold cross-validation accuracy was chosen as the parameter set of the model. Accuracy scatter plots of the model based on different parameter combinations are shown in Fig. 2b. Results demonstrated that the optimal parameter set was gamma = 0.16, eta = 0.3, and max_depth = 4. Thus, we trained the model using the three optimal parameters. Then we performed jackknife test to evaluate its performance and the accuracy is 82.04%. Accordingly, a predictor named p6mA was implemented.

Comparison with existing predictors

To evaluate the prediction performance of p6mA, we compared it with three existing predictors, i.e., iDNA6mA-Pred [27], iDNA6mA-PseKNC [26] and MM-6mAPred [32]. The jackknife test result was applied to measure the predictive power of our methods.

As shown in Fig. 3a, p6mA performed better than the three predictors, it obtained the highest values among the four metrics (i.e., sensitivity, specificity, accuracy, and Matthews Correlation Coefficient). MM-6mAPred has the second highest accuracy (Acc) and second highest Matthews Correlation Coefficient (MCC), while its sensitivity (Sn) is 70.46%, which is ~10% lower than that of p6mA. iDNA6mA-PseKNC's sensitivity is 77.27%, while its specificity is only 5.95%. i6mA-Pred obtains specificity of 82.37%, while the sensitivity is 64.31%. The details of the performances can be found in Additional file 1: Table S1.

MM-6mAPred provides the prediction score for each sample, so we plotted its receiver operating characteristic (ROC) curve and compared it with ROC curve of p6mA (Fig. 3b). The area under the ROC curve (auROC) were

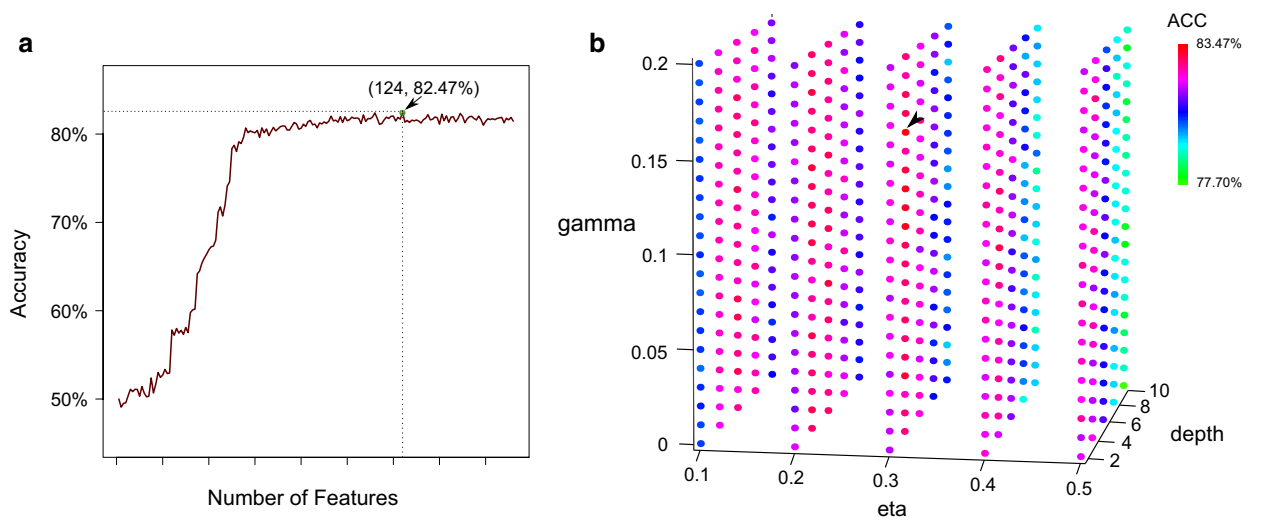

Fig. 2 Feature selection and parameters tuning. **a** IFS curve of feature selection. **b** Grid-search results of parameter tuning

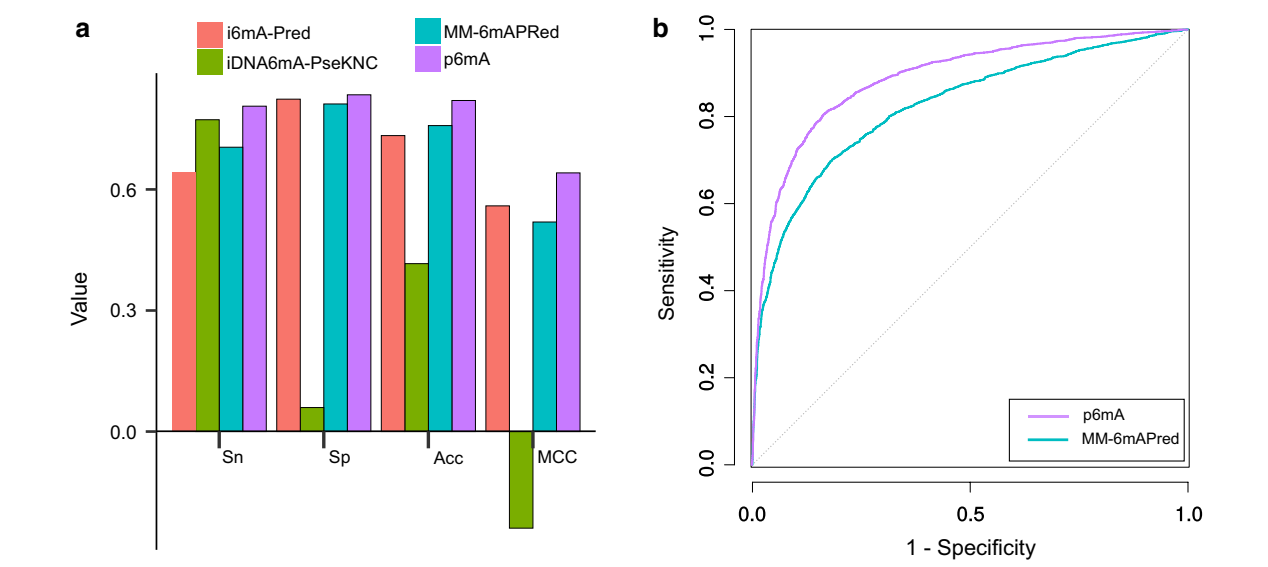

Fig. 3 Comparison between p6mA and other existing predictors by benchmark dataset. **a** The performance of the 4 predictors. **b** ROC curves of p6mA and MM-6mAPred

calculated, p6mA has a larger auROC (0.8871) than MM-6mAPred (0.824).

Independent validation on *A. thaliana* dataset

We then performed independent validation on a dataset from another species. As a vital model flowering plant, *A. thaliana* is a good species to test our predictor, with its N^6-methyladenine modification landscape previously reported in 2015 [33]. The modification data of *A. thaliana* were downloaded from MethSMRT and a dataset for independent validation was constructed. We obtained 1055 non-redundant positive samples and 1055 non-redundant negative samples from the reference genome TAIR10. The dataset construction method was similar to the benchmark dataset.

We compared p6mA with iDNA6mA-Pred, iDNA6mA-PseKNC and MM-6mAPred by the *A. thaliana* dataset. As shown in Fig. 4a, iDNA6mA-PseKNC obtained the highest Sn (84.36%), but performed less well in the other three indicators, especially Sp (5.88%). The p6mA achieved better overall performance in comparison with the other predictors: the highest Sp

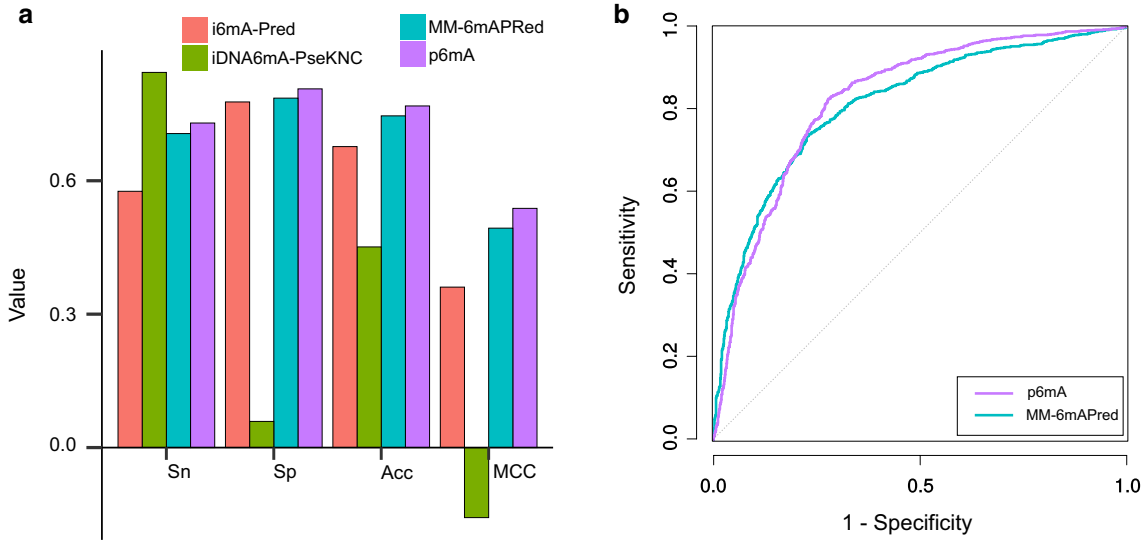

Fig. 4 Results of independent validation by *A. thaliana* dataset. **a** The performance of the 4 predictors. **b** ROC curves of p6mA and MM-6mAPred

(80.66%), Acc (76.82%), and MCC (0.5379). We also plotted MM-6mAPred's ROC curve and compared it with ROC curve of p6mA (Fig. 4b), p6mA has a higher auROC (0.8246) than that of MM-6mAPred (0.8141). Overall, we demonstrated the robustness of p6mA and its superiority over other existing methods by independent validation. The details of the performances can be found in Additional file 1: Table S2.

Software package introduction

To facilitate the application of our predictor, we implemented p6mA in R language, with the code stored in GitHub (https://github.com/Konglab404/p6mA). The feature extractor methods were also implemented, users just need to provide the input data in fasta format files. Each sequence of the input file should be a 41-bp-length sequence and the center position (e.g., the 21th nucleotide) is the A (adenine) for predict, like:

>human_seq_1101_hg19

ATAGTGTAGTGAGCGTACGTAACGTGAAGT GAGTGAGTAGC.

The output file of p6mA is a text file containing sequence names, scores, and predicted modification status. The detailed usage and installation guide and the example input/output files can also be found on the repository page.

Discussion

In this study, we built a 6mA predictor p6mA and showed that it is a more robust and competitive 6mA predictor than other existing ones, as determined using benchmark dataset and independent validation. We not only developed a convenient tool for predicting 6mA, but also indicated the portability of the position-based feature extraction method in biological subjects, especially in nucleotide modification prediction. Besides, recent research showed that different species may have different 6mA preference motifs, e.g., the AGAAT motif of *C. elegans* [34]. This phenomenon prompts us that it is necessary to build 6mA predictors by multiple species' data.

In the independent test by *A. thaliana* data, although p6mA obtained a higher auROC than that of MM-6mAPred, the ROC cures displayed that the specificity of p6mA needs to be improved. Due to the rarity of modified nucleotides in genome, some rare category exploration methods, e.g., RCLens [35], could also be adopted into uncommon nucleic acid modification prediction problems in the future.

As an epigenetic modification, methylation of adenine is a complex biological process that may be affected by other factors, such as chromatin topology and cytoplasmic physicochemical properties. Therefore, we will incorporate additional information to improve the performance of the discrimination ability in the future and different feature extraction methods may be used to construct a more powerful epigenetic modification predictor.

Conclusions

In summary, we developed a new bioinformatics tool, p6mA, for predicting 6mA-modified sites. We also implemented the predictor as an R software package for ease of use. p6mA was built by multiple species' data and it may help the investigation of 6mA modification pattern in different species' genomes.

Methods

Benchmark dataset construction

Fruit fly and worm 6mA-positive samples were obtained from the MethSMRT database [36] and human 6mA-positive samples were obtained from the HuaXia1 assembly [37]. To construct a high-quality dataset, 6mA sites with identification Qv scores, which represent the confidence level of a modification, of less than 30 (p-values < 0.001) were filtered. We extracted ±20 nt sequences from the 6mA sites for each sample to a final sequence length of 41 nt. To reduce sequence-homology bias, CD-HIT v4.6.8 [38] was utilized to generate non-redundant sequence sets with an identity threshold of 0.6.

The 6mA-negative samples from the above three species (i.e., fruit fly, worm, and human) were constructed by selecting non-6mA adenines randomly from the reference genomes (hg38, dm3, and ce10). To ensure negative sample quality, the non-6mA sites were not located in the ±500-bp flanking regions of positive 6mA sites. We also extracted ±20 nt sequences for each negative site as the negative samples, and sequence identity was also less than 0.6.

880 6mA-positive and 880 6mA-negative samples of rice were obtained from i6mA-Pred (http://lin-group.cn/server/i6mA-Pred) and retrieved by SMRT-seq [4]. All sequences were 41 nt long, with the 6mA site at the center position.

Finally, we constructed benchmark dataset from four species' data: i.e., rice, fruit fly, worm, and human. The final benchmark dataset contains 3040 positive samples and 3040 negative samples. Each DNA sequence in the study could be simplified as the formation:

$$\mathrm{Se} = N_1 N_2 N_3 \ldots N_{L-1} N_L, \tag{2}$$

where

$$N_i \in \left[\mathrm{A(adenine)}, \mathrm{C(cytosine)}, \mathrm{G(guanine)}, \mathrm{T(thymine)}\right] \tag{3}$$

represents the ith nucleotide in the sequence. Here, we used the following three sequence-based features: (1) electron–ion interaction pseudopotential (EIIP); (2) position-specific triple-nucleotide propensity (PSTNP), and (3) physicochemical properties. These feature extraction methods were implemented in our in-home R package *RTFE* (https://github.com/ritianjiang/RTFE), the details of which are introduced in the following sections. Briefly, we transformed each sample into a 172-dimensional numerical vector.

EIIP features

Electron–ion interaction pseudopotential, which reflects the electronic properties of nucleotides, was first used to predict the coding potential of genomic regions [39]. EIIP-based feature extraction methods were then widely applied in field prediction and classification, including the prediction of nucleosome positioning [40] and identification of E-gene signature [41].

The EIIP feature vector was constructed as follows:

$$D = [\mathrm{EIIP_{AAA}} \cdot f_{\mathrm{AAA}}, \mathrm{EIIP_{AAC}} \cdot f_{\mathrm{AAc}}, \cdots \mathrm{EIIP_{TTT}} \cdot f_{\mathrm{TTT}},], \tag{4}$$

where EIIP_{xyz} denotes the average EIIP value of three nucleotides (x, y, and z), f_{xyz} denotes the frequency of the 3-tuple nucleotides xyz in the sample sequence and x, y, $z \in$ (A, C, G, T). The EIIP values for the four nucleotides are:

$$\begin{cases} \mathrm{EIIP_A} = 0.1260 \\ \mathrm{EIIP_C} = 0.1340 \\ \mathrm{EIIP_G} = 0.0806 \\ \mathrm{EIIP_T} = 0.1335 \end{cases}. \tag{5}$$

Using this method, we generated 64 features.

PSTNP features

Position-specific triple-nucleotide propensity describes the differences in nucleotide composition at each position between the sequences with and without 6mA modification. As a statistics-based feature extraction method, PSTNP has been used to address multiple molecular biological problems, including DNA N^4-methylcytosine (4mC) site prediction [22], enhancer prediction [42], and σ70 promoter predictor [43].

Two subtypes of PSTNP were used in this study, i.e., single-stranded and double-stranded ($\mathrm{PSTNP_{SS}}$ and $\mathrm{PSTNP_{DS}}$, respectively). The $\mathrm{PSTNP_{SS}}$ features are based on the single-stranded characteristics of DNA and contain 64 (4^3) trinucleotides: AAA, AAC, AAG, …, TTT. Thus, for a sequence with a length of l-bp, the detailed information of the trinucleotide positions can be expressed by a $64 \times (l-2)$ matrix Z:

$$Z = \begin{bmatrix} Z_{1,1} & \cdots & Z_{1,l-2} \\ \vdots & \ddots & \vdots \\ Z_{64,1} & \cdots & Z_{64,l-2} \end{bmatrix}, \tag{6}$$

where the variable

$$Z_{i,j} = F^{+}\left(3\mathrm{mer}_i|j\right) - F^{-}\left(3\mathrm{mer}_i|j\right) \quad \left(i = 1, 2, \ldots, 64; j = 1, 2, \ldots l-2\right). \tag{7}$$

$F^{+}(3\text{mer}_i|j)$ and $F^{-}(3\text{mer}_i|j)$ denote the frequency of the ith trinucleotide (3mer_i) at the jth position in the positive and negative datasets, respectively. 3mer_1 is AAA, 3mer_2 is AAC,... 3mer_{64} is TTT in Eq. 6.

The sample in Eq. 1 can be expressed as the PSTNP_{SS} vector:

$$S = \left[\phi_1, \phi_2, \phi_3, \phi_4, \ldots, \phi_{l-2,}\right]^{\mathrm{T}}, \tag{8}$$

where T is the transpose operator and ϕ_v is defined as:

$$\phi_u = \begin{cases} Z_{1,u}, \text{ when } N_u N_{u+1} N_{u+2} = \text{AAA} \\ Z_{2,u}, \text{ when } N_u N_{u+1} N_{u+2} = \text{AAC} \\ Z_{3,u}, \text{ when } N_u N_{u+1} N_{u+2} = \text{AAG} \\ \vdots \\ Z_{64,u}, \text{ when } N_u N_{u+1} N_{u+2} = \text{TTT} \end{cases} (1 \le u \le l-2). \tag{9}$$

PSTNP_{DS} features characterize double-stranded position-specified information according to complementary pairing. We deemed A and T as identical, the same to C and G. Each sample could be converted into a sequence containing A and C only. For example, the DNA sequence "TCGAGT GAC" could be converted into "ACCACACAC". There are only eight (2^3) trinucleotides: AAA, AAC,..., CCC. Thus, for a sequence whose length is l-bp, detailed information on trinucleotide positions can be expressed by an $8 \times (l-2)$ matrix Z':

$$Z' = \begin{bmatrix} Z'_{1,1} & \cdots & Z'_{1,l-2} \\ \vdots & \ddots & \vdots \\ Z'_{8,1} & \cdots & Z'_{8,l-2} \end{bmatrix}, \tag{10}$$

where the variable

$$Z'_{i,j} = F^{+}\left(3\text{mer}_i|j\right) - F^{-}\left(3\text{mer}_i|j\right) \quad \left(i = 1, 2, \ldots, 8; j = 1, 2, \ldots l-2\right). \tag{11}$$

F^{+} $(3\text{mer}_i|j)$ and F^{-} $(3\text{mer}_i|j)$ denote the frequency of the ith trinucleotide (3mer_i) at the jth position in the positive and negative datasets, respectively. 3mer_1 is AAA, 3mer_2 is AAC, ..., 3mer_8 is CCC in Eq. 10.

The sample in Eq. 1 can be expressed as the PSTNP_{DS} vector:

$$S' = \left[\phi'_1, \phi'_2, \phi'_3, \phi'_4, \ldots, \phi'_{l-2,}\right]^{\mathrm{T}}, \tag{12}$$

where S' is the converted sequence and T is the transpose operator. In this formula, ϕ'_v is defined as:

$$\phi'_u = \begin{cases} Z'_{1,u}, \text{when} N_u N_{u+1} N_{u+2} = \text{AAA} \\ Z'_{2,u}, \text{when} N_u N_{u+1} N_{u+2} = \text{AAC} \\ Z'_{3,u}, \text{when} N_u N_{u+1} N_{u+2} = \text{ACA} \\ \vdots \\ Z'_{64,u}, \text{when} N_u N_{u+1} N_{u+2} = \text{CCC} \end{cases} (1 \le u \le l-2). \tag{13}$$

Here, both PSTNP_{SS} and PSTNP_{DS} generated 39 features.

Physicochemical properties

The pseudo-amino acid composition (PseAAC) method has been successful used to address many computational proteomics problems [44–47] and hastened the application of the pseudo k-tuple nucleotide composition (PseKNC) method. In this study, we used a simplified Type-II PseKNC based on physicochemical properties, which can represent the long-range interaction between oligonucleotides. The physicochemical Type-II PseKNC feature was constructed as follows:

$$\text{Dp} = [d_1, d_2, d_4, d_4, \ldots d_{\Lambda}, d_{\Lambda+1}, \ldots d_{\lambda\Lambda}]^{\mathrm{T}}, \tag{14}$$

where d_i reflects the long-range sequence-order physicochemical effect of a DNA sequence whose length is L-bp and definition is:

$$\begin{cases} d_1 = \frac{1}{L-k-1} \sum_{i=1}^{L-k-1} J^1_{i,i+1} \\ d_2 = \frac{1}{L-k-1} \sum_{i=1}^{L-k-1} J^2_{i,i+1} \\ d_3 = \frac{1}{L-k-1} \sum_{i=1}^{L-k-1} J^3_{i,i+1} \\ \vdots \\ d_{\Lambda} = \frac{1}{L-k-1} \sum_{i=1}^{L-k-1} J^{\Lambda}_{i,i+1} \\ \vdots \\ d_{\lambda\Lambda-1} = \frac{1}{L-k-1} \sum_{i=1}^{L-k-\lambda} J^{\Lambda-1}_{i,i+\lambda} \\ d_{\lambda\Lambda} = \frac{1}{L-k-1} \sum_{i=1}^{L-k-\lambda} J^{\Lambda}_{i,i+\lambda} \end{cases} . \tag{15}$$

In Eq. 15, λ denotes the tiers or correlation ranks along a DNA sequence and should be set to a signless integer less than $L-k$. Λ is the number of physicochemical properties used in feature construction. $J^{\psi}_{i,i+m}$ denotes the correlation of the ψth physicochemical property between the ith dinucleotide ($N_i N_{i+1}$) and $(i+m)$th dinucleotide ($N_{i+m} N_{i+m+1}$). $J^{\psi}_{i,i+m}$ can be calculated by:

$$\begin{cases} J^{\psi}_{i,i+m} = H_{\psi}(N_i N_{i+1}) \cdot H_{\psi}(N_{i+m} N_{i+m+1}) \\ \psi = 1, 2, \ldots, \Lambda; m = 1, 2, \ldots, \lambda; i = 1, 2, \ldots, L-k-\lambda \end{cases}, \tag{16}$$

where $H_{\psi}(N_i N_{i+1})$ and $H_{\psi}(N_{i+m} N_{i+m+1})$ are the values of the ψth physicochemical property for dinucleotides $N_i N_{i+1}$ and $N_{i+m} N_{i+m+1}$, respectively. In this study, six double-stranded B-DNA physicochemical properties (e.g., rise, ring, shift, slide, tilt, and twist) from DiProGB (https://diprodb.leibniz-fli.de/ShowTable.php) were used.

Before substituting values into Eq. 16, the original property values were standardized by the formula:

$$H_{\psi}(N_i N_{i+1}) = \frac{H^0_{\psi}(N_i N_{i+1}) \cdot \left\langle H^0_{\psi}(N_i N_{i+1}) \right\rangle}{\text{SD}\left[H^0_{\psi}(N_i N_{i+1})\right]}, \tag{17}$$

where $H_{\psi}^{0}(N_iN_{i+1})$ is the original ψth physicochemical property value for N_iN_{i+1} and $\langle\bullet\rangle$ brackets are the average of quantity therein over the 16 different combinations of A, C, G, and T for N_iN_{i+1}. SD is the standard deviation of the corresponding 16 property values.

In this study, $\lambda=5$ and there were six ($\Lambda=6$) properties. This method generated 30 features.

Feature selection

Maximum relevance maximum distance (MRMD) [28] was used to select the features. The software package of MRMD was obtained from http://lab.malab.cn/soft/MRMD/index.html.

Gradient Boosting decision trees

The Gradient Boosting algorithm constructs a strong ensemble learner using multiple weak learners, such as decision trees, and has been applied in a series of biologically supervised classification projects, including prediction of gamma-aminobutyric acid type-A receptors and hot spots at protein–protein interfaces [48, 49]. The Extreme Gradient Boosting (XGBoost) algorithm proposed by Chen and Guestrin [29] is an efficient implementation of Gradient Boosting and has been used extensively by data scientists [50]. The R interface in xgboost v0.81.0.1 was used in this study.

Appropriate tuning of parameters can strengthen a predictor's discrimination ability. We performed parameter tuning by grid search, with three parameters thus optimized: i.e., maximum tree depth for base weak learners (max_depth, from 2 to 10, step by 1), learning rate (eta, from 0.1 to 0.9, step by 0.05), and gamma (gamma, from 0 to 0.2, step by 0.002). We herein used tenfold cross-validation to select the optimal parameters by accuracy.

Performance assessment

We used the jackknife test to evaluate the predictor's performance [51]. Four indices were adopted: i.e., sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews Correlation Coefficient (MCC). The indices were defined as:

$$\begin{cases} \mathrm{Sn} = 1 - \frac{N_-^+}{N^+} \\ \mathrm{Sp} = 1 - \frac{N_+^-}{N^-} \\ \mathrm{Acc} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \\ \mathrm{MCC} = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \end{cases}, \quad (18)$$

where N_-^+ is the number of positive samples incorrectly predicted to be negative, N^+ is the total number of positive samples, N_+^- is the number of negative samples incorrectly predicted to be positive, and N^- is the total number of negative samples. The four metrics above are valid only for single-label systems.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13072-020-00330-2.

Additional file 1: Tables S1, **S2.** Summary of performances of 4 predictors on benchmark dataset and *A. thaliana* dataset.

Abbreviations

XGBoost: Extreme Gradient Boosting; 6mA: N^6-Methyladenine; PSTNP: Position-specific triple-nucleotide propensity; EIIP: Electron–ion interaction pseudopotential; MRMD: Maximum relevance maximum distance; IFS: Incremental feature selection; ROC curve: Receiver operating characteristic curve; auROC: Area under ROC curve; Sn: Sensitivity; Sp: Specificity; Acc: Accuracy; MCC: Matthews Correlation Coefficient.

Acknowledgements

We thank Dr. Christine Watts and Mr. Zia Ur Rahman for help in proof-reading the manuscript.

Authors' contributions

QPK designed and supervised the project. HTW, FHX and GHL collected the data. HTW performed the data analyses. HTW implemented the R package. HTW, FHX, and QPK wrote the paper. GHL revised the paper. All authors read and approved the final manuscript.

Funding

The work was supported by Grants from National Key R&D Program of China (No. 2018YFC2000400), The Second Tibetan Plateau Scientific Expedition and Research (No. 2019QZKK0607), the National Natural Science Foundation of China (91749109, 81701394), and Key Research Program of Frontiers Science of the Chinese Academy of Sciences (QYZDB-SSW-SMC020).

Availability of data and materials

The datasets used and analyzed during the current study are available from MethSMRT database (https://sysbio.gzzoc.com/methsmrt) and i6mA-Pred (http://lin-group.cn/server/i6mA-Pred) (Readers can download the *.fasta sequence files of benchmark dataset and *A. thaliana* dataset from https://github.com/Konglab404/p6mA). p6mA was implemented by R language 3.6.1 and the code can also be found at https://github.com/Konglab404/p6mA.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] State Key Laboratory of Genetic Resources and Evolution/Key Laboratory of Healthy Aging Research of Yunnan Province, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China. [2] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China. [3] Kunming Key Laboratory of Healthy Aging Study, Kunming 650223, China. [4] KIZ/CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming 650223, China. [5] Kunming College of Life Science, University of Chinese Academy of Sciences, Beijing 100049, China.

Received: 14 October 2019 Accepted: 3 February 2020
Published online: 24 February 2020

References

1. Dunn DB, Smith JD. Occurrence of a new base in the deoxyribonucleic acid of a strain of Bacterium coli. Nature. 1955;175(4451):336–7.
2. Vanyushin BF, Tkacheva SG, Belozersky AN. Rare bases in animal DNA. Nature. 1970;225(5236):948–9.
3. Xiao CL, Zhu S, He M, Chen D, Zhang Q, Chen Y, et al. N(6)-methyladenine DNA modification in the human genome. Mol Cell. 2018;71(2):306–18.
4. Zhou C, Wang C, Liu H, Zhou Q, Liu Q, Guo Y, et al. Identification and analysis of adenine N(6)-methylation sites in the rice genome. Nat Plants. 2018;4(8):554–63.
5. Linn S, Arber W. Host specificity of DNA produced by *Escherichia coli*, X. In vitro restriction of phage fd replicative form. Proc Natl Acad Sci USA. 1968;59(4):1300–6.
6. Meselson M, Yuan R. DNA restriction enzyme from *E. coli*. Nature. 1968;217(5134):1110–4.
7. Campbell JL, Kleckner N. *E. coli* oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. Cell. 1990;62(5):967–79.
8. Pukkila PJ, Peterson J, Herman G, Modrich P, Meselson M. Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in *Escherichia coli*. Genetics. 1983;104(4):571–82.
9. Vasu K, Nagaraja V. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev. 2013;77(1):53–72.
10. Wion D, Casadesus J. N6-methyl-adenine: an epigenetic signal for DNA–protein interactions. Nat Rev Microbiol. 2006;4(3):183–92.
11. Robbins-Manke JL, Zdraveski ZZ, Marinus M, Essigmann JM. Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase- and mismatch repair-deficient *Escherichia coli*. J Bacteriol. 2005;187(20):7027–37.
12. Pomraning KR, Smith KM, Freitag M. Genome-wide high throughput analysis of DNA methylation in eukaryotes. Methods. 2009;47(3):142–50.
13. Krais AM, Cornelius MG, Schmeiser HH. Genomic N(6)-methyladenine determination by MEKC with LIF. Electrophoresis. 2010;31(21):3548–51.
14. 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.
15. de Araujo Oliveira JV, Costa F, Backofen R, Stadler PF, Machado Telles Walter ME, Hertel J. SnoReport 2.0: new features and a refined Support Vector Machine to improve snoRNA identification. BMC Bioinform. 2016;17(Suppl 18):464.
16. Gupta Y, Witte M, Moller S, Ludwig RJ, Restle T, Zillikens D, et al. ptRNApred: computational identification and classification of post-transcriptional RNA. Nucleic Acids Res. 2014;42(22):e167.
17. Hertel J, Hofacker IL, Stadler PF. SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics. 2008;24(2):158–64.
18. Song K. Recognition of prokaryotic promoters based on a novel variable-window Z-curve method. Nucleic Acids Res. 2012;40(3):963–71.
19. Umarov RK, Solovyev VV. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PLoS ONE. 2017;12(2):e0171410.
20. Wu Q, Wang J, Yan H. An improved position weight matrix method based on an entropy measure for the recognition of prokaryotic promoters. Int J Data Min Bioinform. 2011;5(1):22–37.
21. 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.
22. He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. Bioinformatics. 2019;35(4):593–601.
23. Liu Z, Xiao X, Qiu WR, Chou KC. iDNA-methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal Biochem. 2015;474:69–77.
24. Fu Y, Luo GZ, Chen K, Deng X, Yu M, Han D, et al. N6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. Cell. 2015;161(4):879–92.
25. Iyer LM, Abhiman S, Aravind L. Natural history of eukaryotic DNA methylation systems. Prog Mol Biol Transl Sci. 2011;101:25–104.
26. Feng P, Yang H, Ding H, Lin H, Chen W, Chou KC. iDNA6mA-PseKNC: identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics. 2019;111(1):96–102.
27. Chen W, Lv H, Nie F, Lin H. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz015.
28. Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. BMC Syst Biol. 2016;10(Suppl 4):114.
29. Chen T, Guestrin C, editors. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining KDD'16. New York: ACM; 2016.
30. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.
31. Fickett JW. Quantitative discrimination of MEF2 sites. Mol Cell Biol. 1996;16(1):437–41.
32. Pian C, Zhang G, Li F, Fan X. MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov Model. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz556.
33. Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33(6):623–30.
34. Greer EL, Blanco MA, Gu L, Sendinc E, Liu J, Aristizabal-Corrales D, et al. DNA methylation on N6-adenine in *C. elegans*. Cell. 2015;161(4):868–78.
35. Lin H, Gao S, Gotz D, Du F, He J, Cao N. RCLens: interactive rare category exploration and identification. IEEE Trans Vis Comput Graph. 2018;24(7):2223–37.
36. 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.
37. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun. 2016;7:12065.
38. 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.
39. Nair AS, Sreenadhan SP. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation. 2006;1(6):197–202.
40. Jia C, Yang Q, Zou Q. NucPosPred: predicting species-specific genomic nucleosome positioning via four different modes of general PseKNC. J Theor Biol. 2018;450:15–21.
41. Cai R, Qian D, Wang D, Zhu P. E-gene signature method with biological and physical characteristics—case in p53 gene family. Comput Eng Appl. 2017;53(13):155–9.
42. He WY, Jia CZ. EnhancerPred2.0: predicting enhancers and their strength based on position-specific trinucleotide propensity and electron–ion interaction potential feature selection. Mol Biosyst. 2017;13(4):767–74.
43. He W, Jia C, Duan Y, Zou Q. 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. BMC Syst Biol. 2018;12(Suppl 4):44.
44. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins. 2001;43(3):246–55.
45. Liao B, Jiang JB, Zeng QG, Zhu W. Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition. Protein Pept Lett. 2011;18(11):1086–92.
46. Xu Y, Ding J, Wu LY, Chou KC. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE. 2013;8(2):e55844.
47. Xu Y, Wen X, Shao XJ, Deng NY, Chou KC. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int J Mol Sci. 2014;15(5):7594–610.
48. Liao Z, Huang Y, Yue X, Lu H, Xuan P, Ju Y. In silico prediction of gamma-aminobutyric acid type-A receptors using novel machine-learning-based SVM and GBDT approaches. Biomed Res Int. 2016;2016:2375268.

49. Wang H, Liu C, Deng L. Enhanced prediction of hot spots at protein–protein interfaces using extreme gradient boosting. Sci Rep. 2018;8(1):14285.
50. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241–50.
51. Chou KC. Using subsite coupling to predict signal peptides. J Protein Eng. 2001;14(2):75–9.

Publisher's Note