SCIENTIFIC REPORTS
nature research

OPEN

csDMA: an improved bioinformatics tool for identifying DNA 6 mA modifications via Chou's 5-step rule

Received: 26 April 2019
Accepted: 24 August 2019
Published online: 11 September 2019

Ze Liu[1,2], Wei Dong[1,2], Wei Jiang[1,2] & Zili He[1,2]

DNA N^6-methyldeoxyadenosine (6 mA) modifications were first found more than 60 years ago but were thought to be only widespread in prokaryotes and unicellular eukaryotes. With the development of high-throughput sequencing technology, 6 mA modifications were found in different multicellular eukaryotes by using experimental methods. However, the experimental methods were time-consuming and costly, which makes it is very necessary to develop computational methods instead. In this study, a machine learning-based prediction tool, named csDMA, was developed for predicting 6 mA modifications. Firstly, three feature encoding schemes, Motif, Kmer, and Binary, were used to generate the feature matrix. Secondly, different algorithms were selected into the prediction model and the ExtraTrees model received the best AUC of 0.878 by using 5-fold cross-validation on the training dataset. Besides, the ExtraTrees model also received the best AUC of 0.893 on the independent testing dataset. Finally, we compared our method with state-of-the-art predictors and the results shown that our model achieved better performance than existing tools.

DNA N^6-methyldeoxyadenosine (6 mA) modifications were first discovered in *Bacteria* in 1955[1]. However, it had not received much attention as 5-methylcytosine (5mC) did. One important reason is that 6 mA modifications were thought to be only widespread in prokaryotes and unicellular eukaryotes, but rarely in multicellular eukaryotes[2,3]. Researchers have proposed several experimental methods to identify 6 mA modifications in the past few decades. The first method, developed by Dunn *et al.* in 1955, is a combination of ultraviolet absorption spectra, electrophoretic mobility, and paper chromatographic movement, but this method is relatively insensitive and cannot be used to detect 6 mA modifications in animals[1]. Then a restriction enzyme method was used to discover 6 mA modifications in 1978. However, this method can only find modified adenosines that occurred in the restriction enzyme target motifs[4]. With the development of high-throughput sequencing technology, thousands of 6 mA modifications were found in different multicellular eukaryotes. In 2015, Fu *et al.* found 6 mA modifications in 84% genes of *Chlamydomonas* by using 6 mA immunoprecipitation sequencing (6mA-IP-Seq)[5]. In 2016, Koziol *et al.* used dot blots, HPLC, and methyl DNA immunoprecipitation followed by sequencing (MeDIP-seq) to detect 6 mA modifications in vertebrates including *Xenopus laevis*, *mouse* and *human*[6]. In 2017, Mondo *et al.* observed that up to 2.8% of all adenines were methylated in early-diverging *fungi* by using single-molecule real-time (SMRT) sequencing[7]. In 2018, Zhou *et al.* found that about 0.2% of adenines in the *rice* genome were 6 mA methylated by using mass spectrometry, immunoprecipitation, and SMRT, and Zhang *et al.* observed that the 6 mA distribution in the *rice* and *Arabidopsis* genome were very similar by using 6mA-IP-seq[8,9]. As the experimental methods are time-consuming and costly, researchers are trying to predict DNA 6 mA modifications by using computational methods. Two prediction tools are reported up to now, i.e., iDNA6mA-PseKNC[10] and i6mA-Pred[11]. iDNA6mA-PseKNC is the first prediction tool for predicting 6 mA modifications in the *Mus musculus* genome and i6mA-Pred is the first identification method in the *rice* genome.

Predicting DNA 6 mA modifications based on computational algorithms is still in the infancy. However, in the parallel study of prediction of post-translational modification (PTM) sites, there are many PTM-predicting papers published by the previous researchers[12–22]. Although there is some detailed difference for each of the individual PTMs, the basic core is about the same. Thus, the feature extraction and classification methods proposed in these studies provide a valuable basis for the prediction of DNA 6 mA modifications. In this research, we aim

[1]College of Water Resources and Architectural Engineering, Northwest A&F University, Yangling, 712100, Shaanxi, China. [2]Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas, Ministry of Education, Northwest A & F University, Yangling, 712100, Shaanxi, China. Correspondence and requests for materials should be addressed to W.D. (email: dongw@nwafu.edu.cn)

Species	Dataset	Sequence identity threshold			
		0.95	0.90	0.85	0.80
Mouse	Positive	1,931	1,924	1,914	1,892
	Negative	1,885	1,866	1,844	1,836
Rice	Positive	880	879	878	876
	Negative	880	880	880	880
cross-species	Positive	2,811	2,803	2,792	2,768
	Negative	2,767	2,746	2,724	2,716

Table 1. Reduce sequence redundancy in the different datasets by using the CD-HIT-EST software.

to develop a prediction tool that can be used to predict DNA 6 mA modifications across species. The benchmark datasets created in the iDNA6mA-PseKNC and i6mA-Pred predictors were used and different algorithms were implemented to generate the final optimized model. 5-fold cross-validation was performed and the prediction results demonstrated that our model achieved a better performance than existing 6 mA prediction tools.

As demonstrated by a series of recent publications[10,13–19] and summarized in two comprehensive review papers[23,24], to develop a really useful predictor for a biological system, one needs to follow Chou's 5-steps rule (more detailed description can be found in https://en.wikipedia.org/wiki/5-step_rules.) to go through the following five steps: (1) construct a gold standard dataset to train and test the model; (2) encode samples with effective formulations; (3) conduct the prediction model with a powerful classifier; (4) evaluate model performance by using cross-validation tests and standard measures; (5) establish a user-friendly web-server for the predictor that can be accessible to the public. Below, we are to address these points one by one, making them crystal clear in logic development and completely transparent in operation.

Method

Dataset generation. Feng *et al*. created a DNA 6 mA benchmark dataset of the *M. musculus* genome in 2018[10]. The benchmark dataset includes 1,934 positive samples and 1,934 negative samples. Chen *et al*. launched a 6 mA benchmark dataset of the *rice* genome in 2019[11]. The benchmark dataset consists of 880 positive samples and 880 negative samples. The above two benchmark datasets were used to create the cross-species dataset and the CD-HIT-EST software[25] with different threshold was used to reduce sequence redundancy in the original datasets (Table 1). Finally, the cross-species dataset consists of 2,768 positive samples and 2,716 negative samples with the most rigorous threshold at 0.80, and the length of each sample is 41nt. To build a cross-species 6 mA prediction model, the stratified selection method was used and we random selected 80% samples for model training and the left 20% for independent testing. Finally, the training dataset consists of 2,214 positive samples and 2,214 negative samples, while the independent testing dataset includes 554 positive samples and 502 negative samples.

Feature encoding scheme. To construct a DNA 6 mA predictor, one of the most important but also most difficult issue is how to encode feature vector for each sequence, yet still retains most of the key patterns. The pseudo amino acid composition (PseAAC) was proposed by Chou *et al*. and has been widely used in nearly all the areas of computational proteomics[26,27]. Based on the PseAAC, four powerful software, such as 'PseAAC'[28], 'PseAAC-Builder'[29], 'propy'[30], and 'PseAAC-General'[31], were established: the former three are for generating various modes of Chou's special PseAAC[32]; while the 4th one for those of Chou's general PseAAC[23]. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of Pseudo K-tuple Nucleotide Composition (PseKNC)[33] was developed for encoding features of DNA/RNA sequences[34–36] that have proved very useful as well. Particularly, recently a very powerful web-server called 'Pse-in-One'[37] and its updated version 'Pse-in-One2.0'[38] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

K-mer pattern. *K* monomeric units (*k*-mers), are simply patterns of *k* consecutive nucleic acids[37] and have a total of 4^k kinds of nucleotide patterns for DNA/RNA. Such as 1-mer has 4 and 2-mer has 16 kinds of nucleotide patterns. To calculate the frequencies of *k*-mer nucleotide patterns, the length range *L* of the scanning region must be determined at first, and then the absolute frequencies of the *k*-mer nucleotide patterns are calculated from the start position to the *L*-*k*-1 position. Finally, the relative frequencies of *k*-mer patterns are calculated for each region. In this study, we set *k* as 2, 3, 4, and extracted $4^2 + 4^3 + 4^4 = 336$ kinds of *k*-mer nucleotide patterns for feature encoding.

KSNPF frequency. The KSNPF frequencies are nucleotide pairs separated by *k* arbitrary nucleotides and have been successfully employed for the prediction of mucin-type O-glycosylation sites[39] and phosphotase-specific dephosphorylation sites[40]. The KSNPF can be calculated using the following equation:

$$f(n1Gap(k)n2) = \frac{\mathrm{S(n1Gap(k)n2)}}{\mathrm{L - k - 1}} \quad (1)$$

where *n*1 and *n*2 represent a pair of sequence elements. For nucleotide, *n* stands for any one of A, C, G, T/U. Thus, there are $4^2 = 16$ combinations in each pair. *Gap*(*k*) stands for *k* arbitrary elements at intervals and S(n1Gap(k)n2) indicates the number of occurrences of the element pair. In this study, *L* represents the length of the nucleotide sequence, and the *k* was set as 1, 2, 3, 4, and the dimension of the KSNPF can be calculated by $4^2 \times 4 = 64$.

Nucleic shift density. Nucleic shift density encoding can be used to calculate the density of any nucleotide at the current position in its prefix string and has been used to encode nucleotide sequences in the iDNA-6mA-PseKNC predictor[10]. A nucleic shift density feature at any position can be defined as follows:

$$d_i = \frac{1}{N_i}\sum_{j=1}^{i} F(n_j), \quad F(n_j) = \begin{cases} 1 \ if \ n_j = q \\ 0 \ other \ case \end{cases} \tag{2}$$

where q represents of any nucleotide at current position i, N_i is the length of the ith prefix string in the sequence. For example, the DNA sequence "CAGCTG". The Nucleic shift density of 'C' at the position 1, 2, 3, 4, 5 or 6 is $1/1 = 1$, $1/2 = 0.5$, $1/3 \approx 0.33$, $2/4 = 0.5$, $2/5 = 0.4$ or $2/6 \approx 0.33$, respectively. In this study, the length of each sample is 41nt. Thus, 41 Nucleic shift density features were generated for each sample.

Binary code. Binary encoding scheme is used to predict 6 mA modifications in the iDNA6mA-PseKNC predictor[10]. For the nucleotide in position i, the Binary features can be defined as following:

$$\begin{cases} x_i = \begin{cases} 1 \ if \ n_i \in \{A, G\} \\ 0 \ if \ n_i \in \{C, T\} \end{cases} \\ y_i = \begin{cases} 1 \ if \ n_i \in \{A, T\} \\ 0 \ if \ n_i \in \{C, G\} \end{cases} \\ z_i = \begin{cases} 1 \ if \ n_i \in \{A, C\} \\ 0 \ if \ n_i \in \{G, T\} \end{cases} \end{cases} \tag{3}$$

In this research, the Binary encoding scheme generates a vector with $3 \times 41 = 123$ elements by characterizing each nucleotide, "A", "C", "G", or "T", with (1, 1, 1), (0, 0, 1), (1, 0, 0), or (0, 1, 0), respectively.

Motif score matrix. The MEME Suite (http://meme-suite.org/) consists of several motif-based sequence analysis tools. In this study, the MEME tool with differential enrichment mode was used and the maximum number of motifs was set to 10. The most enriched motifs were selected based on E-value and the motif matrixes were used for generating motif scores of each sample.

Performance evaluation. Five different classifiers, Random Forest, GradientBoosting, AdaBoost, ExtraTrees and SVM, were implemented by using Python. For Random Forest, GradientBoosting, AdaBoost, ExtraTrees Classifiers, 1,000 trees were selected for each of them. For SVM, grid research was used to search the best combination of C and *gamma* parameters. 5-fold cross-validation was used to evaluate the performance of our model. In a different fold of cross-validation, each subset was iteratively selected as a testing set, while the left 4 subsets were used to train the model. The mean results of the five experiments were finally used as the performance estimates of the algorithms.

Based on the Chou's symbols introduced for studying signal peptides[41,42], Four standard measures were derived and have been adopted by several recent publications[43–45]. The measures can be defined as follows:

$$\begin{cases} Sn = 1 - \dfrac{N_-^+}{N^+} \\ Sp = 1 - \dfrac{N_+^-}{N^-} \\ ACC = 1 - \left(\dfrac{N_-^+ + N_+^-}{N^+ + N^-}\right) \\ MCC = \dfrac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \end{cases} \tag{4}$$

where N^+ and N^- refer to the number of positive samples or negative samples, respectively. N_-^+ stands for the number of positive samples that were predicted to be negatives, N_+^- refers to the number of negative samples that were predicted to be positives. However, these measures are valid only for single-label learning issues. For the multi-label learning problems, whose appearances are more common in system biology[46], system medicine[47] and biomedicine[16], a completely different set of standard measures is needed[48]. Besides, the receiver operating characteristic curve (ROC) combined with the area under the ROC curve (AUC), the Precision-Recall curve combined with the average precision (AP), and the F1 score[49] were also used to evaluate the performance of different classifiers.

Using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein as shown in the systems of enzyme fast reaction[50], graphical rules in molecular biology[51], and low-frequency internal motion in biomacromolecules (such as protein and DNA)[52]. Particularly, what happened is that this kind of insightful implication has also been demonstrated in[53] and many follow-up publications[54–56]. The framework of csDMA is shown in Fig. 1.

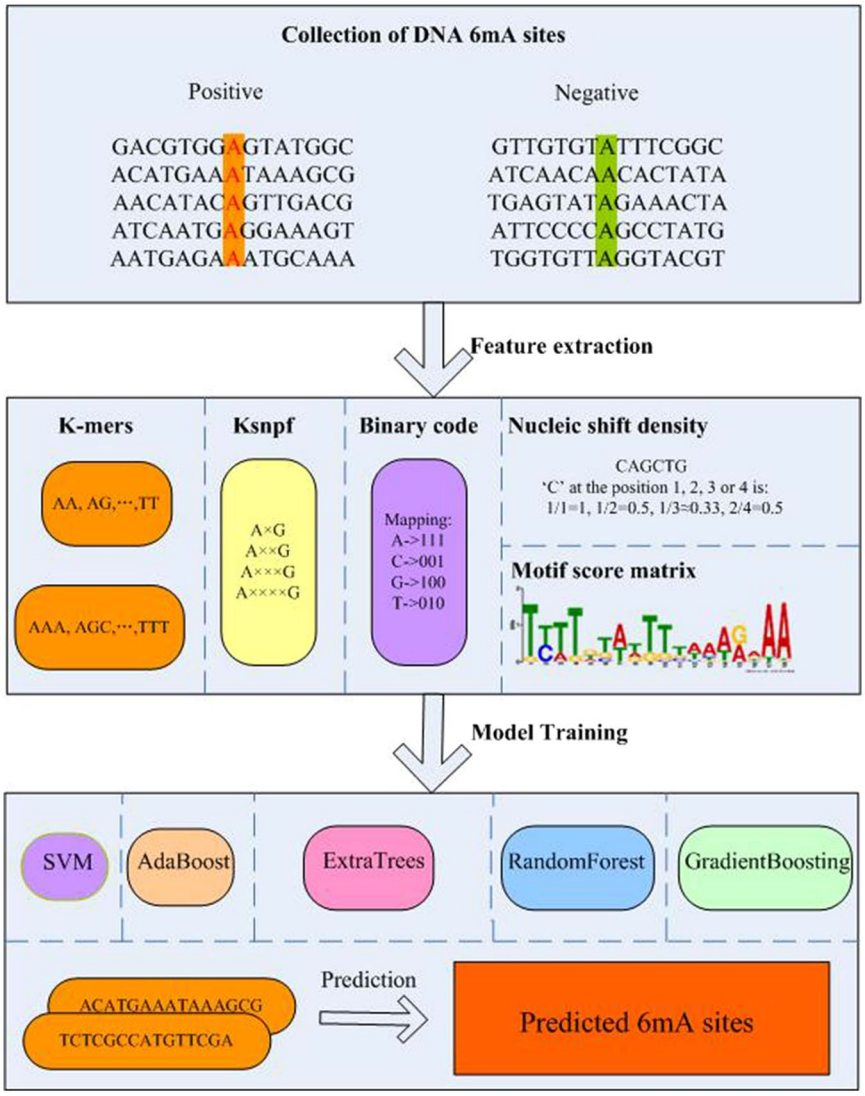

Figure 1. The framework of csDMA.

As pointed out by Chou *et al.*[57] and demonstrated in a series of recent publications[16–18], publicly accessible web-servers or online bioinformatics tools have significantly increased the impacts of bioinformatics on medical science[58], driving medicinal chemistry into an unprecedented revolution[59]. Accordingly, the datasets and online tool involved in this paper are all available at https://github.com/liuze-nwafu/csDMA.

Results

Differential enrichment motifs discovery. To find the enriched motifs in the flank of 6 mA sites, the MEME tool with differential enrichment mode was used and the maximum number of motifs was set to 10. We used the positive samples in the cross-species dataset as the input and treated the negative samples as the control sequences. The detailed information of the enriched motifs can be found in the supplementary materials. Consider the statistical significance of the motifs, the E-value lower than 0.05 was used to find the most statistically significant motifs and two motifs were selected. The first motif, NNNNNNNHHNHHNHWNTNTNWNNNWNYNNNNNNNNNNNNNN, with an E-value of 3.3e-18 was the most statistically significant. And the third motif ACCGATCSA, with an E-value of 2.9e-2, was also selected. The probability matrixes can also be downloaded from the MEME website which can be used to build motif score matrixes in the training process.

Model training with different feature subsets. To find the best combination of feature subsets, different feature subsets were selected into the Random Forest classifier and 5-fold cross-validation was used on the training dataset to evaluate the performance of our model. As shown in Fig. 2, the classifier received an AUC value of 0.866 only by using the Binary code features, which means that the Binary code features were the most significant features that can be used to distinguish positive samples from negative samples. Interestingly, this result was even slightly higher than using combined feature subsets, such as Motif and Binary, Ksnpf and Binary, which achieved an AUC value of 0.861 and 0.862, respectively. Besides, the model achieved the best AUC value of 0.871 when

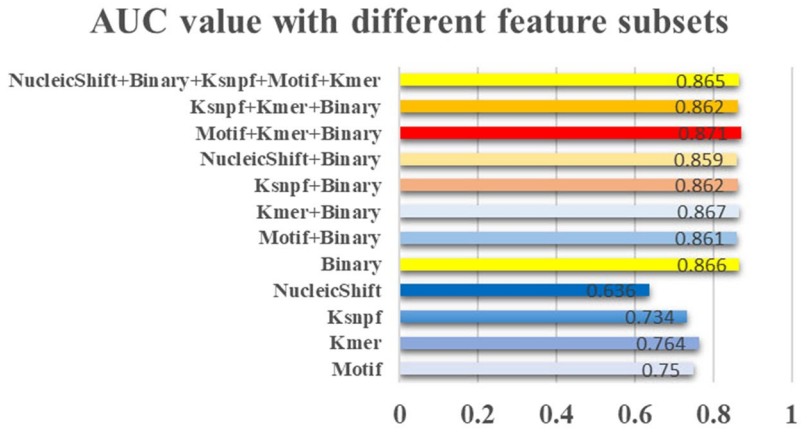

Figure 2. Model performance based on the different feature subsets. 1,000 decision trees were selected into the Random Forest classifier and 5-fold cross-validation was used to evaluate the performance of csDMA.

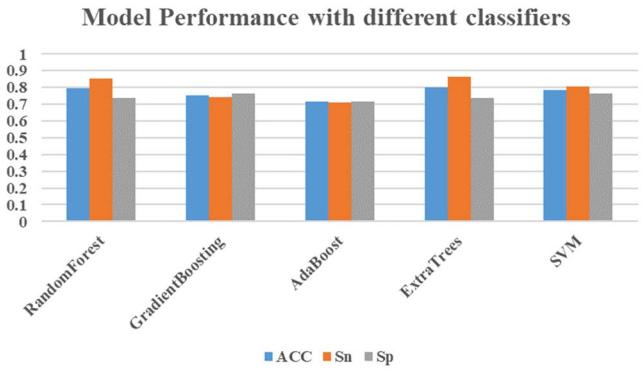

Figure 3. The model performance of different classifiers. The Motif, Kmer, and Binary feature subsets were selected into each classifier and the optimized parameters were used for model training. To evaluate the performance of each classifier, 5-fold cross-validation was used and Standard measures such as ACC, Sn and Sp were used to evaluate the performance of our model.

Algorithm	Sn	Sp	ACC	MCC	AUC	F1
RandomForest	0.853	0.735	0.794	0.593	0.871	0.806
GradientBoosting	0.743	0.762	0.752	0.506	0.818	0.750
AdaBoost	0.713	0.718	0.715	0.431	0.777	0.715
ExtraTrees	**0.864**	0.735	**0.799**	**0.603**	**0.878**	**0.811**
SVM	0.807	**0.764**	0.785	0.572	0.858	0.790

Table 2. Model performance of each algorithm on the training dataset. The highest value of each column is marked in bold.

three feature subsets Motif, Kmer, and Binary feature subsets were selected into the classifier. This result was even a little better than the model performance by using all feature subsets. Thus, we used the Motif, Kmer, and Binary encoding scheme to generate the optimized feature matrix.

Performance evaluation with different classifiers. Five different algorithms were implemented in this research. For the Random Forest, GradientBoosting, AdaBoost, ExtraTrees Classifiers, 1,000 trees were selected for each of them. For the SVM classifier, grid research was used to search the best combination of *C* and *gamma* parameters and the SVM classifier achieved the best performance with *C* of 0.98 and *gamma* of 0.01. To compare the performance of different classifiers, 5-fold cross-validation was used and each classifier was trained with the same fold. As shown in Fig. 3, the ExtraTrees classifier received the best ACC of 0.799 and Sn of 0.864, while the AdaBoost got the lowest ACC of 0.715, Sn of 0.713, Sp of 0.718. However, the ExtraTrees classifier performed not very well for predicting negative samples and received an Sp of 0.735, but it is only a little lower than those of other methods. A more detailed comparison of different classifiers is also shown in Table 2. What's more, the

Algorithm	Sn	Sp	ACC	MCC	AUC	F1
RandomForest	0.875	0.747	**0.814**	**0.630**	0.884	**0.832**
GradientBoosting	0.765	0.757	0.761	0.522	0.854	0.771
AdaBoost	0.776	0.719	0.749	0.496	0.814	0.764
ExtraTrees	**0.888**	0.729	0.813	0.628	**0.893**	**0.832**
SVM	0.843	**0.761**	0.804	0.607	0.875	0.819

Table 3. Model performance of the different algorithms on the independent testing dataset. The highest value of each column is marked in bold.

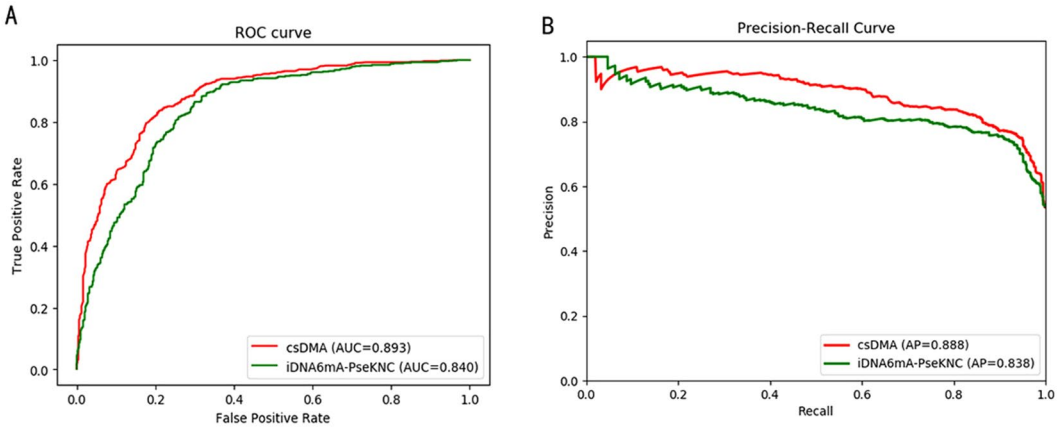

Figure 4. Performance comparison of csDMA and iDNA6mA-PseKNC. (**A**) The ROC curves of csDMA and iDNA6mA-PseKNC. (**B**) The Precision-Recall curves of csDMA and iDNA6mA-PseKNC.

ExtraTrees classifier also achieved the highest MCC of 0.603, AUC of 0.878 and F1 of 0.811. Thus, we used the ExtraTrees algorithm to train the optimized model.

The independent testing dataset was also used to further evaluate the performance of each classifier. Each classifier was trained on the whole training dataset and evaluated on the independent testing dataset. As shown in Table 3, the ExtraTrees classifier received the best Sn of 0.888, AUC of 0.893 and F1 of 0.832, while the SVM model got the highest Sp of 0.761. Interestingly, the performance of each classifier on the independent testing dataset was even a little higher than that on the training dataset, which suggests that the classifier will receive better performance with a larger training dataset.

Comparison with existing 6 mA predictors. The SVM-based tool iDNA6mA-PseKNC was also implemented in this research. Grid research was used to find the best *C* and *gamma*, and the iDNA6mA-PseKNC achieved the best performance with *C* of 0.336 and *gamma* of 0.02. The same fold used for training csDMA was also used for training iDNA6mA-PseKNC. The iDNA6mA-PseKNC predictor received Sn of 0.767, Sp of 0.769, ACC of 0.767, MCC of 0.536, and F1 of 0.767. Most of the measures were lower except Sp is higher than our model with the ExtraTrees classifier. To further compare the performance of the two algorithms. The ROC and Precision-Recall curves were also plotted in Fig. 4. Our model received an AUC of 0.893, while iDNA-6mA-PseKNC got an AUC of 0.840, which also demonstrates that our model achieved better performance than the iDNA6mA-PseKNC predictor.

To test the performance of our model across species, we compared the performance of csDMA and iDNA-6mA-PseKNC on the different datasets, i.e., Cross-species, *rice*, and *M. musculus* datasets. For each dataset, 5-fold cross-validation was performed and the previously optimized parameters were used. We used the same fold for training on different datasets. The five-round results of each measure were averaged and shown in Table 4. For the Cross-species dataset, iDNA6mA-PseKNC got an AUC of 0.844, while our model received a higher AUC of 0.879. For the *rice* dataset, iDNA6mA-PseKNC received an AUC of 0.896, while our model achieved a higher AUC of 0.923. For the *M. musculus* dataset, both models got the same AUC values, but our model also received higher MCC and F1 than those of iDNA6mA-PseKNC. All these results show that the proposed method is very accurate and can be used to predict 6 mA sites in different species.

Discussion

Unlike the prediction of m^6A modifications in mRNA, the identification of 6 mA modifications in DNA is still at the beginning. In this study, we developed an improved tool, called csDMA, for predicting 6 mA modifications in different species. Three feature encoding strategies were used to generate the feature matrix and different algorithms were selected into the model. For performance evaluation, 5-fold cross-validation and independent test were used and the ExtraTrees classifier received the best performance on the training and independent test

Algorithm	Species	Sn	Sp	ACC	MCC	AUC	F1
csDMA	Cross-species	0.863	0.735	0.799	0.603	0.879	0.811
	Rice	0.842	0.880	0.861	0.723	0.923	0.858
	M. musculus	0.932	1	0.966	0.935	0.974	0.965
iDNA6mA-PseKNC	Cross-species	0.762	0.769	0.765	0.531	0.844	0.764
	Rice	0.569	0.721	0.641	0.394	0.896	0.543
	M. musculus	0.869	1	0.935	0.877	0.974	0.930

Table 4. Model performance of each algorithm across species.

datasets. We also compared the performance of our tool with that of iDNA6mA-PseKNC. And the results showed that our model improved the recognition performance of DNA 6 mA modifications effectively.

The i6mA-Pred predictor is another of the two existing tools for DNA 6 mA prediction. However, the research paper is still in the corrected proof phase and their method cannot be reached until our work finished. Fortunately, we acknowledge from their online abstract that the method received an ACC of 0.831 by using a jackknife test. As jackknife test will generate a fixed ACC on the same dataset and their dataset was also downloaded as the *rice* dataset in this study. Thus, we also evaluated the performance of our model on the rice dataset by using a jackknife test and our model received an ACC of 0.859, which is also higher than that of i6mA-Pred.

Although our model received a high performance on the *M. musculus* dataset, the performance on the *rice* and cross-species datasets were relatively low. In the future, more feature encoding schemes, such as genomic and structural features, will be used to improve the performance of csDMA. And also we will extend csDMA to other species, such as *human* and *Arabidopsis thaliana*.

References

1. Dunn, D. B. & Smith, J. D. Occurrence of a new base in the deoxyribonucleic acid of a strain of bacterium coli. *Nature.* **175**, 336–337 (1955).
2. Vanyushin, B. F., Belozersky, A. N., Kokurina, N. A. & Kadirova, D. X. 5-Methylcytosine and 6-Methylaminopurine in Bacterial DNA. *Nature.* **218**, 1066–1067 (1968).
3. Casadesus, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol and Molecular Biology Reviews.* **70**, 830 (2006).
4. Bird, A. Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *Journal of Molecular Biology.* **118**, 49–60 (1978).
5. Fu, Y. *et al*. N6-Methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell.* **161**, 879–892 (2015).
6. Koziol, M. J. *et al*. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nature Structural & Molecular Biology.* **23**, 24–30 (2016).
7. Mondo, S. *et al*. Widespread adenine N6-methylation of active genes in fungi. *Nature Genetics*. **49** (2017).
8. Zhou, C. *et al*. Identification and analysis of adenine N6-methylation sites in the rice genome. *Nature Plants.* **4**, 554–563 (2018).
9. Zhang, Q. *et al*. N(6)-Methyladenine DNA methylation in Japonica and Indica rice genomes and its association with gene expression, Plant Development, and Stress Responses. *Molecular Plant.* **11**, 1492–1508 (2018).
10. Feng, P. M. *et al*. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics*. **111**, 96–102 (2018).
11. Chen, W., Lv, H., Nie, F. & Lin, H. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics*. btz015 (2019).
12. Xu, Y. *et al*. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *Plos One*. **9**, e105018 (2014).
13. Chen, W., Feng, P., Ding, H., Lin, H. & Chou, K. C. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. *Analytical Biochemistry.* **490**, 26–33 (2015).
14. Chen, W., Tang, H., Ye, J., Lin, H. & Chou, K. C. iRNA-PseU: Identifying RNA pseudouridine sites. *Molecular Therapy-Nucleic Acids.* **5**, e332 (2016).
15. Jia, J., Zhang, L. X., Liu, Z., Xiao, X. & Chou, K. C. pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics.* **32**, 3133–3141 (2016).
16. Qiu, W. R., Sun, B. Q., Xiao, X., Xu, Z. C. & Chou, K. C. iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics.* **32**, 3116–3123 (2016).
17. Feng, P. *et al*. iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. *Molecular Therapy-Nucleic Acids.* **7**, 155–163 (2017).
18. Chen, W. *et al*. iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites. *Molecular Therapy-Nucleic Acid.* **11**, 468–474 (2018).
19. Qiu, W. R. *et al*. iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. *Genomics.* **110**, 239–246 (2018).
20. Li, F. *et al*. Positive-unlabelled learning of glycosylation sites in the human proteome. *BMC Bioinformatics.* **20**, 112 (2019).
21. Zhang, Y. *et al*. Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Briefings in Bioinformatics.* https://doi.org/10.1093/bib/bby079 (2018).
22. Chen, Z. *et al*. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Briefings in Bioinformatics.* https://doi.org/10.1093/bib/bby089 (2018).
23. Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology.* **273**, 236–247 (2011).
24. Chou, K. C. Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Current Medicinal Chemistry*, https://doi.org/10.2174/0929867326666190507082559 (2019).
25. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* **28**, 3150–3152 (2012).
26. Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins.* **43**, 246–255 (2001).
27. Chou, K. C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics.* **21**, 10–19 (2005).
28. Shen, H. B. & Chou, K. C. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Analytical Biochemistry.* **373**, 386–388 (2008).

29. Du, P., Wang, X., Xu, C. & Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo amino acid compositions. *Analytical Biochemistry.* **425**, 117–119 (2012).
30. Cao, D. S., Xu, Q. S. & Liang, Y. Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics.* **29**, 960–962 (2013).
31. Du, P., Gu, S. & Jiao, Y. PseAAC-General: Fast building various modes of general form of Chou's pseudo amino acid composition for large-scale protein datasets. *International Journal of Molecular Sciences.* **15**, 3495–3506 (2014).
32. Chou, K. C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics.* **6**, 262–274 (2009).
33. Chen, W., Lei, T. Y., Jin, D. C., Lin, H. & Chou, K. C. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. *Analytical Biochemistry.* **456**, 53–60 (2014).
34. Chen, W. & Lin, H. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Molecular BioSystems.* **11**, 2620–2634 (2015).
35. Liu, B., Yang, F., Huang, D. S. & Chou, K. C. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics.* **34**, 33–40 (2018).
36. Tahir, M., Tayara, H. & Chong, K. T. iRNA-PseKNC(2methyl): Identify RNA 2′-O-methylation sites by convolution neural network and Chou's pseudo components. *Journal of Theoretical Biology.* **465**, 1–6 (2019).
37. Liu, B. *et al*. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research.* **43**, W65–W71 (2015).
38. Liu, B. & Wu, H. Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Natural Science.* **9**, 67–91 (2017).
39. Chen, Y., Tang, Y., Sheng, Z. & Zhang, Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinformatics.* **9**, 101 (2008).
40. Wang, X., Yan, R. & Song, J. DephosSite: a machine learning approach for discovering phosphotase-specific dephosphorylation sites. *Scientific Reports.* **6**, 23510 (2016).
41. Chou, K. C. Using subsite coupling to predict signal peptides. *Protein Engineering.* **14**, 75–79 (2001).
42. Chou, K. C. Prediction of signal peptides using scaled window. *Peptides.* **22**, 1973–1979 (2001).
43. Liu, B., Wang, S., Long, R. & Chou, K. C. iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics.* **33**, 35–41 (2017).
44. Cheng, X., Lin, W. Z., Xiao, X. & Chou, K. C. pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. *Bioinformatics.* **35**, 398–406 (2019).
45. Song, J., Wang, Y. & Li, F. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Briefings in Bioinformatics.* **20**, 638–658 (2018).
46. Cheng, X., Zhao, S. G., Lin, W. Z., Xiao, X. & Chou, K. C. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. *Bioinformatics.* **33**, 3524–3531 (2017).
47. Cheng, X., Zhao, S. G., Xiao, X. & Chou, K. C. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics.* **33**, 341–346 (2017).
48. Chou, K. C. Some remarks on predicting multi-label attributes in molecular biosystems. *Molecular Biosystems.* **9**, 1092–1100 (2013).
49. Song, J. *et al*. Transcriptome-wide annotation of m5C RNA modifications using machine learning. *Frontiers in Plant Science.* **9**, 519 (2018).
50. Chou, K. C. & Forsén, S. Diffusion-controlled effects in reversible enzymatic fast reaction system: Critical spherical shell and proximity rate constants. *Biophysical Chemistry.* **12**, 255–263 (1980).
51. Carter, R. E. & Forsén, S. A new graphical method for deriving rate equations for complicated mechanisms. *Chemica Scripta.* **18**, 82–86 (1981).
52. Chou, K., Chen, N. & Forsén, S. The biological functions of low-frequency phonons: 2. Cooperative effects. *Chemica Scripta.* **18**, 126–132 (1981).
53. Jiang, S. P., Liu, W. M. & Fee, C. H. Graph theory of enzyme kinetics: 1. Steady-state reaction system. *Scientia Sinica.* **22**, 341–358 (1979).
54. Shen, H. B., Song, J. N. & Chou, K. C. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *Journal of Biomedical Science and Engineering.* **2**, 136–143 (2009).
55. Chou, K. C. Graphic rule for drug metabolism systems. *Current Drug Metabolism.* **11**, 369–378 (2010).
56. Zhou, G. P. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. *Journal of Theoretical Biology.* **284**, 142–148 (2011).
57. Chou, K. C. & Shen, H. B. Recent advances in developing web-servers for predicting protein attributes. *Natural Science.* **1**, 63–92 (2009).
58. Chou, K. C. Impacts of bioinformatics to medicinal chemistry. *Medicinal Chemistry.* **11**, 218–234 (2015).
59. Chou, K. C. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Current Topics in Medicinal Chemistry.* **17**, 2337–2358 (2017).

Acknowledgements

This work was supported by the Start-up foundation of Northwest A&F University (Z109021809), the National Natural Science Foundation of China (51809218), and the Postdoctoral Research Foundation of China (2018M643744).

Author Contributions

Z.L. participated in conceiving and performing the experiments. W.D. and W.J. participated in analyzing the data. All authors contributed to the writing of the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-49430-4.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.