Molecular Genetics and Genomics (2019) 294:1173–1182
https://doi.org/10.1007/s00438-019-01570-y

ORIGINAL ARTICLE

iN6-methylat (5-step): identifying DNA N^6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule

Nguyen Quoc Khanh Le[1]

Received: 3 February 2019 / Accepted: 25 April 2019 / Published online: 4 May 2019

Abstract

DNA N^6-methyladenine is a non-canonical DNA modification that occurs in different eukaryotes at low levels and it has been identified as an extremely important function of life. Moreover, about 0.2% of adenines are marked by DNA N^6-methyladenine in the rice genome, higher than in most of the other species. Therefore, the identification of them has become a very important area of study, especially in biological research. Despite the few computational tools employed to address this problem, there still requires a lot of efforts to improve their performance results. In this study, we treat DNA sequences by the continuous bags of nucleobases, including sub-word information of its biological words, which then serve as features to be fed into a support vector machine algorithm to identify them. Our model which uses this hybrid approach could identify DNA N^6-methyladenine sites with achieved a jackknife test sensitivity of 86.48%, specificity of 89.09%, accuracy of 87.78%, and MCC of 0.756. Compared to the state-of-the-art predictor as well as the other methods, our proposed model is able to yield superior performance in all the metrics. Moreover, this study provides a basis for further research that can enrich a field of applying natural language-processing techniques in biological sequences.

Keywords Skip gram · Continuous bag of words · DNA N^6-methyladenine · Support vector machine · FastText · DNA replication

Introduction

DNA N^6-methyladenine is a non-canonical DNA modification that occurs in different eukaryotes at low levels. It has been identified as an extremely important function of life. Recent studies have shown that the modification of DNA N^6-methyladenine sites is closely linked to a number of biological processes such as replication, transcription, and repair of DNA. The non-uniform distribution of N^6-methyladenine sites across the genome means that it is essential to characterize its position in the genome to better understand its biological functions. A study from (Zhou et al. 2018) has shown that about 0.2% of adenines are marked by DNA N^6-methyladenine in the rice genome, higher than in mammals and *D. melanogaster* (Zhang et al. 2015; Wu et al. 2016). It is desirable that there is a need to make an effort on this specific problem.

In general, DNA N^6-methyladenine is just one type of post-translational modifications (PTMs) and in the last few years, there are many PTM-predicting papers published by the previous investigators. According to these studies, there are many methods presented to predict PTM sites. For example, one of the most important approaches is to use PseAAC to generate features and perform supervised learning for classification. It has been widely used to identify human GPCRs N-linked glycosylation sites (Xie et al. 2013), S-nitrosylation sites (Xu et al. 2013a, b; Zhang et al. 2014), and phosphorylation sites (Qiu et al. 2016). Many studies also used an ensemble classifier to achieve a good performance in PTM sites predictions, such as (Jia et al. 2016; Qiu et al. 2016, 2018). Furthermore, a diversity of approaches was proposed in (Xu et al. 2014; Qiu et al. 2015; Chandra et al. 2018) to address this interesting problem.

In particular, due to the importance of DNA N^6-methyladenine sites, the identification of them has also become a very important study, especially in biological research.

Communicated by S. Hohmann.

✉ Nguyen Quoc Khanh Le
khanhle@ntu.edu.sg; khanhlee87@gmail.com

[1] Medical Humanities Research Cluster, School of Humanities, Nanyang Technological University, 48 Nanyang Ave, Singapore 639798, Singapore

It has attracted many researchers from different areas of biology, such as pure biology, system biology, and computational biology. For instance, an antibody against N^6-methyladenine sites can detect and enrich them in DNA (Fu et al. 2015; Greer Eric et al. 2015; Zhang et al. 2015). This antibody can be used to detect methylated DNA by dot blotting and to immunoprecipitate methylated DNA for sequencing. Moreover, DNA N^6-methyladenine sites can be detected via enzymes such as Dpn I and Dpn II (Lacks and Greenberg 1977). They can also be detected using single-molecule real-time (SMRT) sequencing (Flusberg et al. 2010; Fang et al. 2012). The latest techniques which were applied to identify N^6-methyladenine are genomic distributions and sequence motifs (Jones et al. 1998; Touzain et al. 2010; Smith and Meissner 2013). However, these experimental techniques are expensive and time-consuming. Therefore, there is a necessity to discover new computational strategies to supplant the experimental techniques in the investigation and characterization of N^6-methyladenine sites. Additionally, to date, as more DNA sequences have been discovered with the advancement of next-generation sequencing technologies, the number of DNA sequence entries is now a thousand times more than the number from about 25 years ago. Facing the surge of new DNA sequences discovered in the post-genomic era, there has been a desire to create automated computational prediction approaches for quick and accurate identification of N^6-methyladenine sites.

To date, in bioinformatics, to identify DNA N^6-methyladenosine sites, there are few studies that has been conducted such as iRNA-Methyl (Chen et al. 2015), pRNAm-PC (Liu et al. 2016), iMethyl-STTNC (Akbar and Hayat 2018), iRNA(m6A)-PseDNC (Chen et al. 2018), and iDNA6mA-PseKNC (Feng et al. 2019). Their performance results were high; however, the DNA N^6-methyladenosine sites in rice genome had various special functions and we can reach an improve accuracy in a more specific problem. To identify DNA N^6-methyladenosine sites in the rice genome, one computational method was recently adopted (Chen et al. 2019). In their work, they used nucleotide chemical properties and nucleotide frequency to identify DNA N^6-methyladenosine sites and reached 83.13% for cross-validation result. However, there is a need for improving their performance result, and in this study, we intend to address this issue with an innovative approach. Our idea is to transform the DNA sequences into a continuous bag of nucleobases using FastText model (Bojanowski et al. 2017) and then proceed to classify them with the use of effective machine-learning algorithms. FastText has been also applied in representing biological sequences such as protein sequences (Asgari et al. 2019) and enhancer sequences (Le et al. 2019), and here we extend this approach in a specific problem with a more in-depth analysis.

Based on this idea, we documented several key contributions of our study to the field of biology: (1) a new computational model for DNA N^6-methyladenosine sites in which it exhibited significant improvements beyond that of previous predictor, (2) a new framework constructed from continuous bag of nucleobases and supervised learning for identification of the DNA sequences with high performance, (3) a study that would provide much information to biologists and researchers as they are better equipped with the understanding of the DNA N^6-methyladenosine sites, and (4) a basis for further study to apply "continuous bag of nucleobases" feature in biological sequences.

As shown in a series of recent publications (Cai et al. 2018; Cheng et al. 2018; Song et al. 2018; Feng et al. 2019; Jia et al. 2019; Khan et al. 2019; Le et al. 2019), to develop a really useful sequence-based statistical predictor for a biological or biomedical system, one should observe the guidelines of 5-step rule (Chou 2011) so as to make the following five steps very clear: (1) how to construct or select a valid benchmark dataset to train and test the predictor; (2) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) how to properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (5) how to establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by-one.

Materials and methods

This study aims to use "continuous bags of nucleobases" representation and Support Vector Machine (SVM) (Chang and Lin 2011) to identify N^6-methyladenine sites in the rice genome. The study consists of two primary processes: using FastText (Bojanowski et al. 2017) to train vectors model and LibSVM to train supervised learning classification (as shown in Fig. 1). Using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein, as indicated by many previous studies on a series of important biological topics, (see, e.g., (Chou et al. 1979; Chou and Forsén 1980; Zhou and Deng 1984; Chou 1989, 1990; Althaus et al. 1993a, b; Kuo-Chen 2010; Zhou 2011), particularly in enzyme kinetics and protein folding rates (Chou and Forsén 1980; Chou 1990) as well as low-frequency internal motion (Chou et al. 1989). We explained the detailed description of our method as follows:

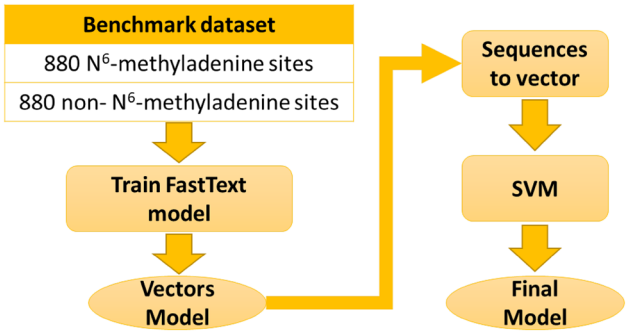

Fig. 1 Flowchart of this study. First, benchmark dataset was collected to train a FastText model and then we use this generated model to generate vectors from the input DNA sequences. In the end, LibSVM package was used to perform supervised learning on the generated vectors

Benchmark dataset

To develop a powerful supervised learning classification, there is a need to collect a high quality and objective benchmark dataset. Most machine-learning methods require achieving a high performance; this step is of the utmost importance. To overcome this issue, we re-used the benchmark dataset from the previous study (Chen et al. 2019). In this dataset, they collected N^6-methyladenine sites from NCBI Gene Expression Omnibus and the single-molecule real-time sequencing (Zhou et al. 2018). Later, they extracted DNA sequences into fragments of 41 bp with the N^6-methyladenine sites at the center. They also used CD-HIT (Fu et al. 2012) to remove the pairwise sequences which had the similarity of more than 60%. This study indeed set a little bit higher level than the common bioinformatics problems (usually at 30–40% similarity). However, it could be explained because of the sequence length used in this study. With fragments of 41 bp, we could reduce a lot of data if using the low similarity level. The similarity level of 60% not only ensures the difference in data but we did not also lose too much data. Finally, the benchmark dataset includes 880 positive samples (sequences with N^6-methyladenine sites) and 880 negative samples (sequences without N^6-methyladenine sites) which is available at http://lin-group.cn/server/i6mAPred/data.

Continuous bag of nucleobases representation

With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms [such as "optimization" algorithm (Zhang and Chou 1992), "covariance discriminant" or "CD" algorithm (Chou and Elrod 2002), "nearest neighbor" or "NN" algorithm (Hu et al. 2011), and "support vector machine" or "SVM" algorithm (Cai et al. 2006; Hu et al. 2011)] can only handle vectors as elaborated in a comprehensive review (Chou 2015). However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo-amino acid composition or PseAAC (Chou 2001b) was proposed. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics [e.g., a long list of references cited in (Chou 2017)]. Since it has been widely and increasingly used, recently three powerful open access softwares, called 'PseAAC-Builder' (Du et al. 2012), 'propy' (Cao et al. 2013), and 'PseAAC-General' (Du et al. 2014), were established: the former two are for generating various modes of Chou's special PseAAC; while the third one for those of Chou's general PseAAC, including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode [see Eqs. 9–10 of (Chou 2011)], "Gene Ontology" mode [see Eqs. 11–12 of (Chou 2011)], and "Sequential Evolution" or "PSSM" mode [see Eqs. 13–14 of (Chou 2011)]. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) (Chen et al. 2014) was developed for generating various feature vectors for DNA/RNA sequences that have proved very useful as well. Particularly, recently a very powerful web-server called 'Pse-in-One' (Liu et al. 2015) and its updated version 'Pse-in-One2.0' (Liu et al. 2017) have been established that can be used to generate any desired pseudo-component feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

Owing to generate a special feature for this dataset, we presented a different feature set called 'continuous bag of nucleobases'. The idea is to apply existing natural language-processing (NLP) models to biological sequences. It was premier introduced by (Asgari and Mofrad 2015) and successfully exploited in solving latter biological problems (Habibi et al. 2017; Vang and Xie 2017; Hamid and Friedberg 2018; Öztürk et al. 2018). However, with the use of Word2Vector to represent the biological sequences, these results had some certain drawbacks such as, not taking into account the internal structure of words and the out-of-vocabulary cases for unseen words. Therefore, one important issue needs to be addressed is that instead of using a distinct vector representation for the DNA word, there is a need to take into account the internal structure of each word. To perform this task, Facebook proposed FastText, which is an extension to Word2Vec that be able to treat each word as composed of character *n*-grams (Bojanowski et al. 2017). Therefore, the vector for a word is made up of the sum of

these character *n*-grams. FastText has been shown to be more accurate than Word2Vec vectors in many fields (Joulin et al. 2017; Yang et al. 2018). Inspired by the achievements of FastText, previous researchers have used it to represent biological sequences such as DNA sequence (Le et al. 2019) and protein sequence (Asgari et al. 2019).

According to these basic ideas, in this study, we divide the DNA samples into a sequence of words with one word being one nucleotide (A, C, G, or T), two continuous nucleotides, and so on. We then used FastText model to generate the 'continuous bag of nucleobases' representation, and each DNA word is thereby represented as a bag of character *n*-gram. By generating word embedding in this way, we ensure not to miss the sub-word information and it would be useful to perform classification. Figure 2 illustrates the idea of generating a "continuous bag of nucleobases" feature from DNA sequences. We represent the DNA word (ATGAC) by changing the *n*-gram values from 1 to 3 and the "continuous bag of nucleobases" are different between the three types. In this way, we are able to integrate the sub-word information for each DNA word according to each level of *n*-gram.

The objective of this step is to encode the nucleobases by asserting their distribution of vector space, thus supervised learning algorithms could adopt them. Supervised learning classification requires the same number of features for the input data. However, our DNA sequences are of different lengths, therefore, we fixed the dimension of the embedding vector to 100 to address this issue. This means that each sequence of DNA is represented as 100 real numerical values and it could be fed into any machine-learning classifier without preprocessing. Our rationale behind this idea is to emphasize that the biological words can be analogous to the motifs in DNA sequences. N^6-Methyladenine sites with similar motifs tend to have a higher degree of similarity. By taking this information into the dataset, we have more special features for a better prediction.

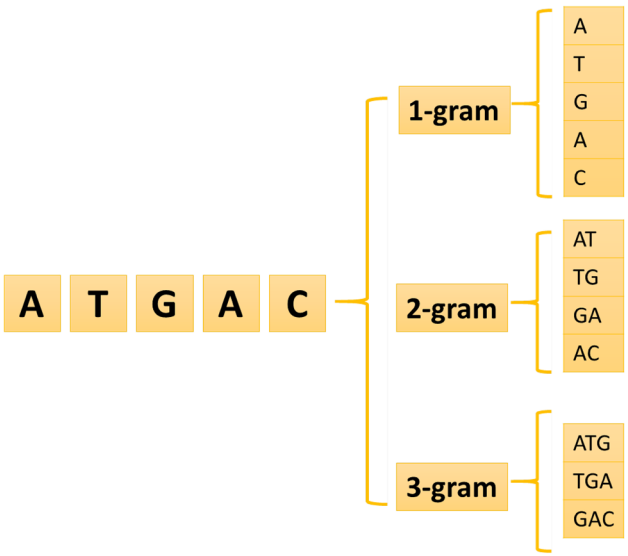

Fig. 2 Explanation of "continuous bag of nucleobases" for DNA N^6-methyladenine sequence. The word 'ATGAC' represents different results with different *n*-gram levels. Example: *n*-gram levels is from 1 to 3

Here, we provided a summary of the FastText model for completeness. As originally developed by (Mikolov et al. 2013), given a sequence of nucleobases $n_1, n_2 \ldots n_T$, the vector model aims to maximize the average log likelihood:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(n_{t+j}|n_j), \tag{1}$$

where T is the total number of nucleobases in the whole N^6-methyladenine sites dataset, c is the context window size (i.e., the number of nucleobases on the left and right side of the target nucleobase), and $p(n_t+j|n_t)$ is defined as:

$$p\left(n_{\mathrm{O}}|n_{\mathrm{I}}\right)=\frac{\exp\left(v'_{n_{\mathrm{o}}},v_{n_{\mathrm{I}}}\right)}{\sum_{n=1}^{N}\exp\left(v'_{n_{\mathrm{o}}},v_{n_{\mathrm{I}}}\right)}, \tag{2}$$

where v_n and v'_n are two space vector representations of the nucleobase n. The subscripts O and I correspond to the output (context nucleobases) nucleobase and input (target) nucleobase, respectively. N is the total number of single nucleobases in the DNA vocabulary. In the typical NLP text corpus with a large vocabulary, the calculation of the log likelihood gradient becomes impractical. An approximation of the log likelihood is obtained by replacing every $\log p(n_{\mathrm{O}}|n_{\mathrm{I}})$ with:

$$\log\sigma\left(v'_{n_{\mathrm{o}}},v_{n_{\mathrm{I}}}\right)+\sum_{i=1}^{k}E_{n_i\sim P_n(w)}\left[\log\sigma\left(v'_{n_{\mathrm{o}}},v_{n_{\mathrm{I}}}\right)\right], \tag{3}$$

where

$$\sigma(x)=\frac{1}{1+\exp(-x)}, \tag{4}$$

and k are negative samples. This was motivated by the idea that a good model should be able to differentiate real data from negative data.

FastText implementation

To generate the "continuous bag of nucleobases" representation, we exploited FastText package (Bojanowski et al. 2017). In FastText, users are able to select a variety of hyperparameters for training, such as the type of model [continuous bag of words (CBOW) or skip gram], sampling

method, or loss function. Compared to Word2Vector model, there are some advantages of FastText as described in the previous article (Le et al. 2019). Each DNA word is represented as a continuous bag of nucleobases apart from the word itself. Therefore, the meaning of shorter words can be preserved, which can appear as *n*-grams of other words. This also enables the meanings for suffixes/prefixes to be captured naturally. This is why we can achieve outstanding performance in DNA sequence representations and classifications, notably in rare words, with the idea of using FastText in this study using information from character level (Bojanowski et al. 2017). NLP algorithms can be efficiently applied by considering DNA sequences as natural language phrases in a text corpus. In particular, DNA sequences are considered as natural language sentences, and nucleobases are considered as words. In this study, we trained FastText model with a vector space of 100 dimensions and a context window of size 8.

Support vector machine implementation

Support vector machine (SVM) is a common machine-learning method for classification, regression, and other learning tasks. SVM has been successfully applied to a lot of supervised learning problem in a variety of fields, such as computer vision, natural language processing, neuroimaging, and even bioinformatics. Since 2000, Chang and Lin developed a package namely LibSVM (Chang and Lin 2011) to help users easily apply SVM. The idea inside LibSVM is the implementation of sequential minimal optimization algorithm for SVM kernels. With that idea, LibSVM received a lot of positive responses and it has been applied more and more to a range of supervised learning classifications in bioinformatics. In this study, we also implemented the SVM algorithm using the LibSVM package. Moreover, due to many evidential improvements with radial basis function (RBF) kernel in this field (Le and Ou 2016a, b), we selected this kernel as our learning function. There are a lot of hyperparameters need to be tuned up in LibSVM, but the most important two are cost and gamma. Therefore, we also did a grid search to estimate the accuracy of each parameter combination to find the optimal cost and gamma in SVM [log2c ranged from − 5 to 15 (step = 2), log2 g ranged from 3 to − 15 (step = − 2)].

Assessment of predictive ability

Jackknife test is an approximately unbiased estimator of the generalization performance and it has been used in many bioinformatics applications to examine the quality of various predictors [see, e.g., (Chou 2001b; Chou and Elrod 2002; Chen et al. 2015)]. Therefore, we also used the jackknife test to examine the quality of our model. It also helps us easily compare our performance results with the previous work (Chen et al. 2019) because this one used the jackknife test too. In the jackknife test, if there are a total of N samples in the dataset, then the predictor is trained on $N-1$ training examples and tested on the remaining 1 data point. Then, the process is repeated for N times and the predicted label of each sample is predicted.

Although the traditional metrics copied from math books were often used in literature to measure the prediction quality of a prediction method, they are no longer good because of lacking intuitiveness and not easy-to-understand for most biologists. Particularly the MCC (the Matthews correlation coefficient), which is a very important metrics used for reflecting the stability of a prediction method. Fortunately, based on the Chou's symbols introduced for studying protein signal peptides (Chou 2001a, c), a set of four intuitive metrics were derived as given in Eqs. 5–8. Ever since then, the new set of intuitive metrics have been concurred and admired by a series of recent publications (see, e.g., (Xu et al. 2014; Le et al. 2019; Rahman et al. 2019; Tahir et al. 2019). Therefore, in a bid to evaluate the performance of the methods employed, we also adopted Chou's criterion. Some standard metrics were used, such as sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC) using below given formulae (TP, FP, TN, FN are true-positive, false-positive, true-negative, and false-negative values, respectively):

$$\text{Sensitivity} = 1 - \frac{N_-^+}{N^+}, \quad 0 \le \text{Sen} \le 1, \tag{5}$$

$$\text{Specificity} = 1 - \frac{N_+^-}{N^-}, \quad 0 \le \text{Spec} \le 1, \tag{6}$$

$$\text{Accuracy} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}, \quad 0 \le \text{Acc} \le 1, \tag{7}$$

$$\text{MCC} = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}}, \quad -1 \le \text{MCC} \le 1. \tag{8}$$

The relations between these symbols and the symbols in Eqs. (5–8) are given by:

$$\begin{cases} N_+^- = \text{FP} \\ N_-^+ = \text{FN} \\ N^+ = \text{TP} + N_-^+ \\ N^- = \text{TN} + N_+^- \end{cases} \tag{9}$$

We also used the receiver-operating characteristic (ROC) curves to further illustrate our model in various experiments. In addition, the area under the ROC curve (AUC) metric is

a scalar value that represents the overall performance of the model (Bradley 1997). The AUC score is always bounded between zero and one, and there is no realistic classification with an AUC less than 0.5. The AUC metric is, therefore, used to compare the efficiency of different models.

Results

Identifying N^6-methyladenine sites with different *n*-gram levels

In the first experiences, we would like to examine the optimal *n*-gram level by performing the experiments in which each DNA sequence was split to biological words of equal length from 1 to 10. ROC Cure and AUC (Bradley 1997) were used to evaluate the performance results via FastText implementation. Notice that in these experiments, we set the same FastText hyperparameters with different *n*-gram levels (from 1 to 10). This step is to ensure a fair comparison among the experiments. The result of identifying N^6-methyladenine sites is displayed in Fig. 3. As shown in Fig. 3, the increase of *n*-gram value will help to achieve a better performance result accordingly. If we only used the lower levels of *n*-gram (1 or 2), the AUC did not reach the highest point (only 0.61 and 0.71, respectively). However, when we used a bigger *n*-gram level, the performance results were indeed increased. It means that to maximize the performance of our model, we have to choose the big value for *n*-gram. The model only captures the special information in a high level of *n*-gram, increasing high level of *n*-gram will help to increase much in the results. We easily saw that the performance results with 8-g, 9-g, and 10-g (with AUC = 0.80) outperformed the other *n*-gram levels. Therefore, in this study, we chose one of these levels (*n*-gram = 8) to perform further experiments.

Comparative performance between SVM and FastText baseline model

A multinomial logistic regression (MLR) is provided by FastText, in which the vector of the sentence/document matches its features. Many text classification problems had been successfully addressed using this integrated model (Joulin et al. 2017). Instead of using FastText to classify N^6-methyladenine sites directly, we replaced the MLR by SVM to maximize the performance. Therefore, we would like to compare between FastText baseline classifier and SVM classifier. The purpose behind this comparison is to find out whether SVM could be used to replace the MLR algorithm in FastText. We used the same dataset with the optimal *n*-gram levels (8-g). We also performed a manual grid search to find the optimal parameters in FastText classifier. Subsequently, some of these parameters had been optimized are: epoch of 100, learning rate of 0.1, and loss function of softmax. With this step, we ensure a

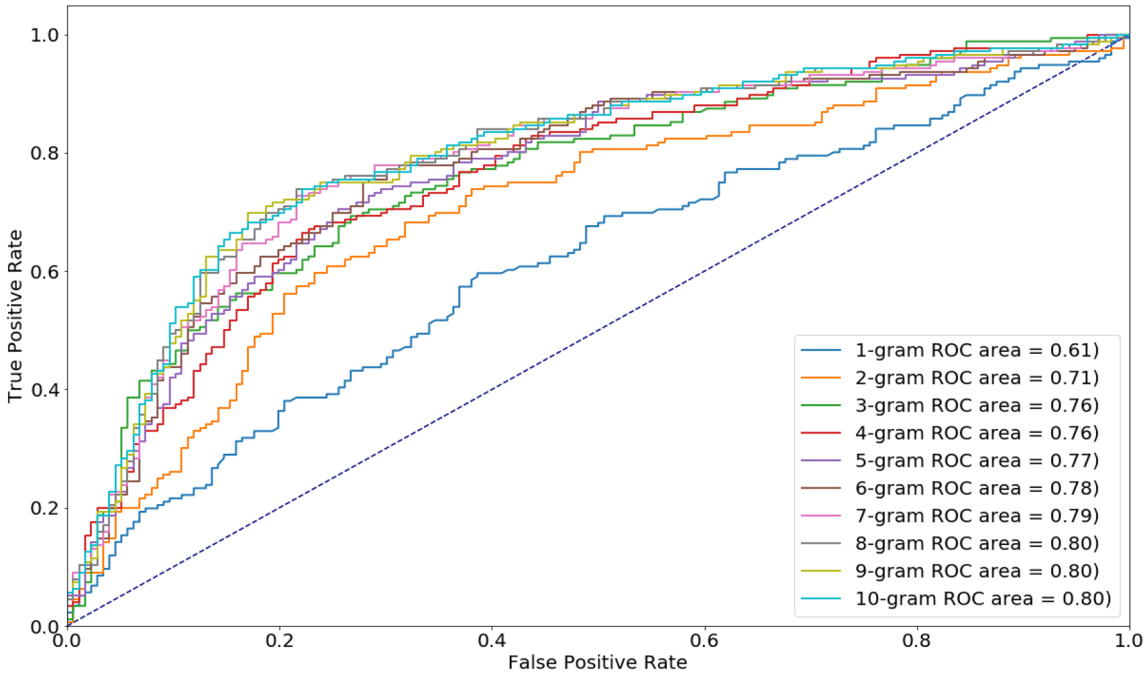

Fig. 3 Comparative performance results on identifying N^6-methyladenine sites with different *n*-gram levels. The result with 8-g, 9-g, and 10-g, (AUC = 0.80) outperforms other *n*-gram levels

Table 1 Comparison between our method and previous works in jackknife test

Method	Sn	Sp	Acc	MCC
MLR	67.95	65.91	66.93	0.339
PseDNC	63.52	65.57	64.55	0.29
i6mA-Pred	82.95	83.3	83.13	0.66
Our method	86.48 (+)	89.09 (+)	87.78 (+)	0.756 (+)

MLR multinomial logistic regression from FastText, PseDNC and i6mA-Pred: results from the previous paper

fair comparison between SVM and FastText because we also applied grid search in SVM. Table 1 shows the comparative performance results among different methods. As shown in Table 1, it is clear that the SVM classifier outperforms MLR on the same proposed dataset. Therefore, replacing MLR by an effective algorithm is a more significant solution than using the MLR baseline model.

Comparative performance between proposed method and the existing predictors

So far, our best model is the combination of the support vector machine classifier and the *n*-gram level of 8 for N^6-methyladenine sites prediction. In this section, we aim to compare the effectiveness of our proposed approach with the previous work (Chen et al. 2019) on the same dataset. The next comparison is with the PseDNC method that had been used in (Feng et al. 2019) to identify general N^6-methyladenine sites. As shown in Table 1, our model showed a jackknife cross-validation sensitivity of 86.48%, specificity of 89.09%, accuracy of 87.78%, and MCC of 0.756. The problem raised here is how to sustain the better performance of our algorithm, as compared to the other methods when it undergoes many cross-validation tests. To answer this question, we performed a paired *t* test to determine whether our SVM is significantly better (+) or worse (−), or even have no statistical difference compared to the other methods. The null hypothesis assumes that the true mean difference between the paired metrics is zero and the statistical significance is determined by p value = 0.05 (confidence level of 95%). It is clear that on average, our method outperforms previous predictors in almost measurement metrics. To detail, our model is able to improve the sensitivity of 3.53%, specificity of 5.79%, accuracy of 4.65%, and MCC of 0.096 compared to the state-of-the-art predictor (Chen et al. 2019). Therefore, we are able to present an innovative method for extracting features in DNA sequences in which outperforms other feature sets.

Discussion and conclusion

Improving the prediction of N^6-methyladenine sites is a crucial problem in bioinformatics. To resolve it, a number of scientists have made their efforts using various feature extraction methods and machine-learning algorithms. Based on the outstanding results of FastText in generating vectors and classifying natural language, applying it to DNA sequence classification is a vital task for biological researchers. In this study, we implemented FastText as a tool that is able to help us generate a "continuous bag of nucleobases" with sub-word information to identify DNA N^6-methyladenine sites in the rice genome. To the end, the SVM algorithm with hyperparameter optimization was employed to perform supervised classification on the generated vectors. With this method, we can interpret the DNA sequences as biological words and improve predictive performance. We evaluated the performance using jackknife cross-validation. On average, our method showed a jackknife cross-validation accuracy of 87.78% in identifying DNA N^6-methyladenine sites. The average sensitivity, specificity, and MCC reached 86.48%, 89.09%, and 0.756, respectively. Compared with the performance of the state-of-the-art predictor, this approach achieved an evident improvement in almost all the measurement metrics. Throughout this study, we have taken on this approach that uses a powerful model for the prediction of N^6-methyladenine sites and yielding high accuracy. The findings of this study could open a platform for further research that can interpret the biological words in DNA sequences. Moreover, scientists can use our approach to solve various computational biology problems in the future.

Furthermore, as pointed out in (Chou and Shen 2009) and demonstrated in a series of recent publications (see, e.g., (Feng et al. 2013; Lin et al. 2014; Qiu et al. 2014, 2016; Cheng et al. 2017; Le et al. 2017; Cheng et al. 2018; Le et al. 2018), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have significantly increased the impacts of bioinformatics on medical science (Chou 2015), driving medicinal chemistry into an unprecedented revolution (Chou 2017), we shall make efforts in our future work to provide a web-server for the prediction method presented in this paper.

Acknowledgements The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Funding The author received no funding for this work.

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Akbar S, Hayat M (2018) iMethyl-STTNC: identification of N^6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. J Theor Biol 455:205–211

Althaus IW, Chou JJ, Gonzales AJ, Deibel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F (1993a) Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J Biol Chem 268:6119–6124

Althaus IW, Gonzales AJ, Chou JJ, Romero DL, Deibel MR, Chou KC, Kezdy FJ, Resnick L, Busso ME, So AG (1993b) The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem 268:14875–14880

Asgari E, Mofrad MRK (2015) Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One 10:e0141287

Asgari E, McHardy AC, Mofrad MRK (2019) Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). Sci Rep 9:3577

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30:1145–1159

Cai Y-D, Feng K-Y, Lu W-C, Chou K-C (2006) Using LogitBoost classifier to predict protein structural classes. J Theor Biol 238:172–176

Cai L, Huang T, Su J, Zhang X, Chen W, Zhang F, He L, Chou K-C (2018) Implications of newly identified brain eQTL genes and their interactors in schizophrenia. Mol Ther Nucleic Acids 12:433–442

Cao D-S, Xu Q-S, Liang Y-Z (2013) propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29:960–962

Chandra A, Sharma A, Dehzangi A, Ranganathan S, Jokhan A, Chou K-C, Tsunoda T (2018) PhoglyStruct: prediction of phosphoglycerylated lysine residues using structural properties of amino acids. Sci Rep 8:17923

Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2:27

Chen W, Lei T-Y, Jin D-C, Lin H, Chou K-C (2014) PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal Biochem 456:53–60

Chen W, Feng P, Ding H, Lin H, Chou K-C (2015) iRNA-Methyl: identifying N^6-methyladenosine sites using pseudo nucleotide composition. Anal Biochem 490:26–33

Chen W, Ding H, Zhou X, Lin H, Chou K-C (2018) iRNA(m6A)-PseDNC: identifying N^6-methyladenosine sites using pseudo dinucleotide composition. Anal Biochem 561–562:59–65

Chen W, Lv H, Nie F, Lin H (2019) i6mA-Pred: Identifying DNA N^6-methyladenine sites in the rice genome. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz015

Cheng X, Xiao X, Chou K-C (2017) pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. Mol BioSyst 13:1722–1727

Cheng X, Xiao X, Chou K-C (2018) pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics 110:50–58

Chou KC (1989) Graphic rules in steady and non-steady state enzyme kinetics. J Biol Chem 264:12074–12079

Chou K-C (1990) Applications of graph theory to enzyme kinetics and protein folding kinetics: steady and non-steady-state systems. Biophys Chem 35:1–24

Chou K-C (2001a) Using subsite coupling to predict signal peptides. Protein Eng 14:75–79

Chou KC (2001b) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Struct Funct Bioinf 43:246–255

Chou KC (2001c) Prediction of protein signal sequences and their cleavage sites. Proteins: Struct Funct Bioinf 42:136–139

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

Chou K-C (2015) Impacts of bioinformatics to medicinal chemistry. Med Chem 11:218–234

Chou K-C (2017) An unprecedented revolution in medicinal chemistry driven by the Progress of Biological science. Curr Top Med Chem 17:2337–2358

Chou K-C, Elrod DW (2002) Bioinformatical analysis of G-protein-coupled receptors. J Proteome Res 1:429–433

Chou KC, Forsén S (1980) Graphical rules for enzyme-catalysed rate laws. Biochem J 187:829

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

Chou KC, Jiang SP, Liu WM, Fee CH (1979) Graph theory of enzyme kinetics: 1. Steady-state reaction system

Chou K-C, Maggiora GM, Mao B (1989) Quasi-continuum models of twist-like and accordion-like low-frequency motions in DNA. Biophys J 56:295–305

Du P, Wang X, Xu C, Gao Y (2012) PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal Biochem 425:117–119

Du P, Gu S, Jiao Y (2014) PseAAC-general: fast building various modes of general form of chou's pseudo-amino acid composition for large-scale protein datasets. Int J Mol Sci 15:3495

Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-Paz A, Chess A, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE (2012) Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nat Biotechnol 30:1232

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

Feng P, Yang H, Ding H, Lin H, Chen W, Chou K-C (2019) iDNA6 mA-PseKNC: identifying DNA N^6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics 111:96–102

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7:461

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152

Fu Y, Luo G-Z, Chen K, Deng X, Yu M, Han D, Hao Z, Liu J, Lu X, Doré Louis C, Weng X, Ji Q, Mets L, He C (2015) N^6-methyldeoxyadenosine marks active transcription start sites in chlamydomonas. Cell 161:879–892

Greer Eric L, Blanco Mario A, Gu L, Sendinc E, Liu J, Aristizábal-Corrales D, Hsu C-H, Aravind L, He C, Shi Y (2015) DNA methylation on N^6-adenine in *C. elegans*. Cell 161:868–878

Habibi M, Weber L, Neves M, Wiegandt DL, Leser U (2017) Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33:i37–i48

Hamid M-N, Friedberg I (2018) Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. Bioinformatics:bty937-bty937

Hu L, Huang T, Shi X, Lu W-C, Cai Y-D, Chou K-C (2011) Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. PLoS One 6:e14556

Jia J, Liu Z, Xiao X, Liu B, Chou K-C (2016) pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J Theor Biol 394:223–230

Jia J, Li X, Qiu W, Xiao X, Chou K-C (2019) iPPI-PseAAC(CGR): identify protein-protein interactions by incorporating chaos game representation into PseAAC. J Theor Biol 460:195–203

Jones PL, Jan Veenstra GC, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J, Wolffe AP (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet 19:187

Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp 427–431

Khan YD, Jamil M, Hussain W, Rasool N, Khan SA, Chou K-C (2019) pSSbond-PseAAC: prediction of disulfide bonding sites by integration of PseAAC and statistical moments. J Theor Biol 463:47–55

Kuo-Chen C (2010) Graphic rule for drug metabolism systems. Curr Drug Metab 11:369–378

Lacks S, Greenberg B (1977) Complementary specificity of restriction endonucleases of Diplococcus pneumoniae with respect to DNA methylation. J Mol Biol 114:153–168

Le NQK, Ou YY (2016a) Incorporating efficient radial basis function networks and significant amino acid pairs for predicting GTP binding sites in transport proteins. BMC Bioinf 17:183

Le NQK, Ou YY (2016b) Prediction of FAD binding sites in electron transport proteins according to efficient radial basis function networks and significant amino acid pairs. BMC Bioinf 17:298

Le NQK, Ho QT, Ou YY (2017) Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. J Comput Chem 38:2000–2006

Le NQK, Ho QT, Ou YY (2018) Classifying the molecular functions of Rab GTPases in membrane trafficking using deep convolutional neural networks. Anal Biochem 555:33–41

Le NQK, Yapp EKY, Ho QT, Nagasundaram N, Ou YY, Yeh HY (2019) iEnhancer-5Step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. Anal Biochem 571:53–61

Lin H, Deng E-Z, Ding H, Chen W, Chou K-C (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res 42:12961–12972

Liu F, Chen J, Fang L, Wang X, Liu B, Chou K-C (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res 43:W65–W71

Liu Z, Xiao X, Yu D-J, Jia J, Qiu W-R, Chou K-C (2016) pRNAm-PC: predicting N^6-methyladenosine sites in RNA sequences via physical–chemical properties. Anal Biochem 497:60–67

Liu B, Wu H, Chou K-C (2017) Pse-in-One 20: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nat Sci 9:67

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. ICLR Workshop

Öztürk H, Ozkirimli E, Özgür A (2018) A novel methodology on distributed representations of proteins using their interacting ligands. Bioinformatics 34:i295–i303

Qiu W-R, Xiao X, Chou K-C (2014) iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci 15:1746

Qiu W-R, Xiao X, Lin W-Z, Chou K-C (2015) iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. J Biomol Struct Dyn 33:1731–1742

Qiu W-R, Xiao X, Xu Z-C, Chou K-C (2016) iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget 7:51270

Qiu W-R, Sun B-Q, Xiao X, Xu Z-C, Jia J-H, Chou K-C (2018) iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics 110:239–246

Rahman MS, Aktar U, Jani MR, Shatabda S (2019) iPro70-FMWin: identifying Sigma70 promoters using multiple windowing and minimal features. Mol Genet Genom 294:69–84

Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. Nat Rev Genet 14:204

Song J, Li F, Takemoto K, Haffari G, Akutsu T, Chou K-C, Webb GI (2018) PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. J Theor Biol 443:125–137

Tahir M, Hayat M, Khan SA (2019) iNuc-ext-PseTNC: an efficient ensemble model for identification of nucleosome positioning by extending the concept of Chou's PseAAC to pseudo-tri-nucleotide composition. Mol Genet Genomics 294:199–210

Touzain F, Petit M-A, Schbath S, Karoui ME (2010) DNA motifs that sculpt the bacterial chromosome. Nat Rev Microbiol 9:15

Vang YS, Xie X (2017) HLA class I binding prediction via convolutional neural networks. Bioinformatics 33:2658–2665

Wu TP, Wang T, Seetin MG, Lai Y, Zhu S, Lin K, Liu Y, Byrum SD, Mackintosh SG, Zhong M, Tackett A, Wang G, Hon LS, Fang G, Swenberg JA, Xiao AZ (2016) DNA methylation on N^6-adenine in mammalian embryonic stem cells. Nature 532:329

Xie H-L, Fu L, Nie X-D (2013) Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. Protein Eng Des Sel 26:735–742

Xu Y, Ding J, Wu L-Y, Chou K-C (2013a) iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One 8:e55844

Xu Y, Shao X-J, Wu L-Y, Deng N-Y, Chou K-C (2013b) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ 1:e171

Xu Y, Wen X, Wen L-S, Wu L-Y, Deng N-Y, Chou K-C (2014) iNitro-tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS One 9:e105018

Yang X, Macdonald C, Ounis I (2018) Using word embeddings in twitter election classification. Inf Retr J 21:183–207

Zhang C-T, Chou K-C (1992) An optimization approach to predicting protein structural class from amino acid composition. Protein Sci 1:401–408

Zhang J, Zhao X, Sun P, Ma Z (2014) PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. Int J Mol Sci 15:11204–11219

Zhang G, Huang H, Liu D, Cheng Y, Liu X, Zhang W, Yin R, Zhang D, Zhang P, Liu J, Li C, Liu B, Luo Y, Zhu Y, Zhang N, He S, He C, Wang H, Chen D (2015) N^6-methyladenine DNA modification in drosophila. Cell 161:893–906

Zhou G-P (2011) The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism. J Theor Biol 284:142–148

Zhou GP, Deng MH (1984) An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochemical Journal 222:169

Zhou C, Wang C, Liu H, Zhou Q, Liu Q, Guo Y, Peng T, Song J, Zhang J, Chen L, Zhao Y, Zeng Z, Zhou D-X (2018) Identification and analysis of adenine N^6-methylation sites in the rice genome. Nat Plants 4:554–563