Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbsd20

Using Chou's 5-steps rule to identify N^6-methyladenine sites by ensemble learning combined with multiple feature extraction methods

Zhongwang Zhang & Lidong Wang

To cite this article: Zhongwang Zhang & Lidong Wang (2022) Using Chou's 5-steps rule to identify N^6-methyladenine sites by ensemble learning combined with multiple feature extraction methods, Journal of Biomolecular Structure and Dynamics, 40:2, 796-806, DOI: 10.1080/07391102.2020.1821778

To link to this article: https://doi.org/10.1080/07391102.2020.1821778

JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
2022, VOL. 40, NO. 2, 796–806
https://doi.org/10.1080/07391102.2020.1821778

Taylor & Francis
Taylor & Francis Group

Using Chou’s 5-steps rule to identify N^6-methyladenine sites by ensemble learning combined with multiple feature extraction methods

Zhongwang Zhang and Lidong Wang

College of Science, Dalian Maritime University, Dalian, P.R. China

Communicated by Ramaswamy H. Sarma

ABSTRACT

N^6-methyladenine (m6A), a type of modification mostly affecting the downstream biological functions and determining the levels of gene expression, is mediated by the methylation of adenine in nucleic acids. It is also a key factor for influencing biological processes and has attracted attention as a target for treating diseases. Here, an ensemble predictor named as TL-Methy, was developed to identify m6A sites across the genome. TL-Methy is a 2-level machine learning method developed by combining the support vector machine model and multiple features extraction methods, including nucleic acid composition, di-nucleotide composition, tri-nucleotide composition, position-specific trinucleotide propensity, Bi-profile Bayes, binary encoding, and accumulated nucleotide frequency. For *Homo sapiens*, TL-Methy method reached the accuracy of 91.68% on jackknife test and of 92.23% on 10-fold cross validation test; For *Mus musculus*, TL-Methy method achieved the accuracy of 93.66% on jackknife test and of 97.07% on 10-fold cross validation test; For *Saccharomyces cerevisiae*, TL-Methy method obtained the accuracy of 81.57% on jackknife test and of 82.54% on 10-fold cross validation test; For *rice genome*, TL-Methy method achieved the accuracy of 91.87% on jackknife test and of 93.04% on 10-fold cross validation test. The results via these two test approaches demonstrated the robustness and practicality of our TL-Methy model. The TL-Methy model may be as a potential method for m6A site identification.

Abbreviations: Acc: accuracy; ANF: accumulated nucleotide frequency; BE: binary encoding; BPB: Bi-profile Bayes; DA: discriminant analysis; DNC: Di-nucleotide composition; KNN: k-Nearest neighbors; LR: logistic regression; m6A: N^6-methyladenine; MCC: Matthew’s correlation coefficient; NAC: Nucleic acid composition; RBF: Radial basis function; PSTNP: Position-specific trinucleotide propensity; Sn: sensitivity; Sp: specificity; SVM: support vector machine; TNC: tri-nucleotide composition

ARTICLE HISTORY
Received 22 March 2020
Accepted 29 August 2020

KEYWORDS
N^6-methyladenine; support vector machine; 2-level method; feature extraction; jackknife test; 10-fold cross validation test

1. Introduction

N^6-methyladenine (m6A) is a common chemical modification of eukaryotic RNA, which widely exists in eukaryotic mRNA and long non-coding RNA (Pan, 2013). m6A formation occurs via a complex composed of a variety of proteins, including METTL3, METTL14, WTAP, VIRMA, RBM15, RBM15B, HAKAI, METTL16, and ZC3H13 (Duan et al., 2019; Huang et al., 2019; Knuckles et al., 2018; Patil et al., 2016; Pendleton et al., 2017; Ping et al., 2014; Selberg et al., 2019; Wen et al., 2018; Yue et al., 2018). Studies have demonstrated that m6A modification regulates translation, cleavage, trafficking, localization, and RNA stability at the post-transcriptional level (Lin et al., 2019; Robbins Manke et al., 2005; Wang et al., 2018; Xu et al., 2018; Yue et al., 2015). In recent years, m6A has become a research hotspot in the field of life science.

A series of experimental methods for detecting m6A sites have been proposed to explore the biological functions of m6A, including two-dimensional cellulose thin-layer chromatography (2 D-TLC) (Grosjean et al., 2004), high performance liquid chromatography (HPLC) (Nees et al., 2014), coupling of liquid chromatography to mass spectrometry (LC-MS) (Chan et al., 2010), methylated RNA immunoprecipitation sequencing (MeRIP-seq) (Schwartz et al., 2013), photo-cross linking-assisted m6A sequencing (PA-m6A-seq) (Chen et al., 2015a), and methylation individual-nucleotide resolution cross-linking and immunoprecipitation (miCLIP) (Ke et al., 2015). m6A site detection is time-consuming and expensive using the aforementioned experimental approaches, so lots of machine learning and deep learning models have also been developed to facilitate detection.

Chen et al. (2017b) built a predictor, ‘MethyRNA’, based on nucleotide chemical properties and nucleotide frequency using a SVM classifier, whose accuracies reached 90.38% and 88.39% on *Homo sapiens* and *Mus musculus* datasets, respectively. ‘iN6-Methyl (5-step)’ was designed by Nazari et al. (2019), which used a deep learning model via Chou’s 5-step framework and general PseKNC to recognize the m6A sites in *S. cerevisiae*, *H. sapiens*, and *M. musculus*, and the corresponding accuracies were 75.38%, 91.11%, and 89.51%,

CONTACT Lidong Wang ldwang@dlmu.edu.cn College of Science, Dalian Maritime University, Dalian 116026, P.R. China

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07391102.2020.1821778.

respectively. By integrating the deep features with traditional handcrafted features, Wei et al. (2019) trained a support vector machine classification model with the accuracy 80.50%. Wei et al. (2018) trained an ensemble learning model by combining three feature extraction methods, named 'M6APred-EL', whose accuracy of M6APred-EL was 80.83%.m6A exhibits similar functions in prokaryotes and eukaryotes, and it closely related to a series of bioprocesses, such as DNA transcription, replication, and repair (Heyn & Esteller, 2015). To understand the biological functions of m6A in DNA, Tahir et al. (2019a) used a convolutional neural network (CNN) equipped with Chou's 5-step rule to build an intelligent computational model, iDNA6mA (5-step rule), which achieved 86.64% accuracy on *rice genome* dataset. Basith et al. (2019) identified m6A sites via a two-layer ensemble model against the *rice genome*, which was named 'SDM6A' and had 88.70% accuracy. Pian et al. (2020) used the transition probability information between adjacent nucleotides for identifying m6A sites, and gained a correct prediction accuracy of 89.72%.

In this study, a 2-level predictor model 'TL-Methy', was developed for identifying m6A sites. Four sub-models, TL-Methy_H, TL-Methy_M, TL-Methy_S, and TL-Methy_R, were trained on four different species training data set, respectively. Firstly, seven independent SVM models were trained by incorporating seven different features, and whose parameters were optimized with the aid of a grid search method. Secondly, the decision values that the probability value of the sample belongs to the positive category was taken as the output of the first level. The output of the first level was used as feature vectors for the second level and input to the second level SVM classifier. The results of the jackknife test suggested that the designed model could better identify m6A sites in *H. sapiens*, *M. musculus*, *S. cerevisiae*, and *rice genome*; accuracies were 91.68%, 93.66%, 81.57%, and 91.87%, respectively. Additionally, the 10-fold cross validation results were reliable as the accuracy levels achieved 92.23%, 97.09%, 82.54%, and 93.04%, respectively. The model flow chart is portrayed in Figure 1.

2. Materials and methods

As demonstrated by a series of recent publications (Chou, 2020a, 2020b, 2020c; Du et al., 2019; Kabir et al., 2020), to develop a powerful predictor for a biological system, one needs to follow 'Chou's 5-steps rule' (Chou, 2011): (1) collect or construct a valid benchmark dataset; (2) extract and represent the main characteristics of samples with an effective formulation; (3) introduce or design a powerful classification algorithm for conducting the prediction; (4) objectively evaluate the model by performing cross-validation tests; (5) establish a web-server for the predictor. The model built according to Chou's 5-steps rule has the following notable merits (Yang et al., 2020): (1) crystal clear in logic development; (2) completely transparent in model; (3) easy to be compare with the reported results by other investigators; (4) with high potential in stimulating other sequence-analyzing methods; (5) easy to be used by the majority of experimental scientists.

2.1. Benchmark dataset

In this study, the benchmark datasets for m6A editing sites were derived from the work of *Chen et al.* (Chen et al., 2015b, 2019, 2017b). We first identified m6A editing sites in *H. sapiens* and *M. musculus* RNA datasets (Chen et al., 2017b). Therefore, to increase the generalizability of the model on DNA data, a *rice genome* dataset was used (Chen et al., 2019). The datasets for *H. sapiens*, *M. musculus*, and *rice genome* include 1130, 725, and 880 samples, respectively, and each with a sequence length of 41-bp. Therefore, to increase the generalizability of the model for different sequence lengths, the *S. cerevisiae* dataset was introduced (Chen et al., 2015b), with 51-bp sample lengths and 1307 sample numbers. The four datasets were defined using the following formula:

$$\mathcal{S}_k = \mathcal{S}_k^+ \cup \mathcal{S}_k^-, k = \begin{cases} 1 & \text{for } H.sapiens \\ 2 & \text{for } M.musculus \\ 3 & \text{for } rice\ genome \\ 4 & \text{for } S.cerevisiae \end{cases} \tag{1}$$

where the $\mathcal{S}_k (k = 1,2,3,4)$ denote four benchmark datasets, $\mathcal{S}_k^+$ and $\mathcal{S}_k^-$ denote positive and negative subset, respectively, and $\cup$ denotes the 'union' of the two sets (Chen et al., 2017a, 2017c). The specific information of positive and negative samples for each species is listed in Table 1.

2.2. Feature encoding scheme

With the advent of the post-genomic era, using a discrete model or a vector to express biological sequence information or key patterns is one of the most important and greatest challenges in computational biology. This is because existing optimization algorithms, such as optimization algorithm, covariance discriminant algorithm, nearest neighbor algorithm, support vector machine algorithm, and other machine learning algorithms can only handle numerical vectors, but cannot process biological sequences directly (Cai et al., 2006; Chou & Cai, 2003; Hu et al., 2011; Zhang & Chou, 1992). To make full use of the nucleotide sequence pattern information, *Chou et al.* proposed a pseudo-amino acid component (PseAAC) feature extraction method (Chou, 2011), which has been widely used in all branches of computational proteomics (Chou, 2020; Dehzangi et al., 2015; Meher et al., 2017). What's more, numeric vector can be extracted through the four powerful open access software: 'PseAAC' (Shen & Chou, 2008), 'PseAAC-Builder' (Du et al., 2012), 'propy' (Cao et al., 2013), and 'PseAAC-General' (Shen & Chou, 2008). The first three tools can help us to generate various modes of Chou's special PseAAC (Chou, 2009). 'PseAACGeneral' (Chou, 2011) covered all the special modes of feature vectors for proteins and the higher level feature vectors such as 'Functional Domain' mode (Cai et al., 2001, 2003), 'Gene Ontology' mode (Cai et al., 2004, 2005), and 'Sequential Evolution' or 'PSSM'

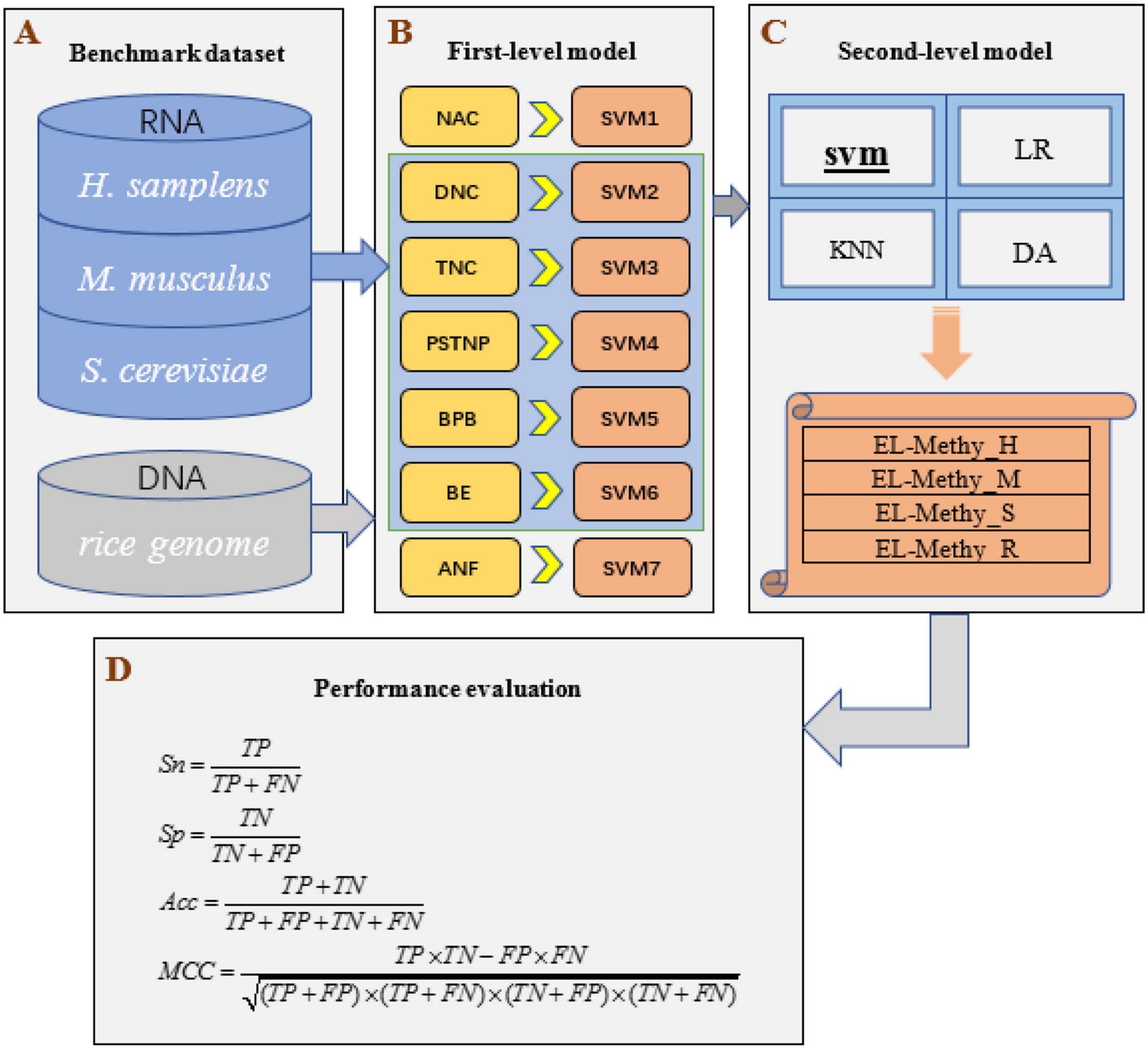

Figure 1. The flowchart and ensemble architecture of the TL-Methy. (a) Data collection and processing for *Homo sapiens, Mus musculus, Saccharomyces cerevisiae,* and *rice genome*; (b) The first-layer model framework of ensemble learning. Seven featured extraction methods were used to generate feature subsets to train seven single SVMs. Ensemble learning trains the second-level SVM by establishing a new feature space with seven single SVMs output decision values; (c) Selection of the second-level classifier. The decision values obtained from the first-level were input to each classier. Classifier selection for the second-level according to the results of jackknife test; (d) The measurements used to evaluate various models in terms of Sn, Sp, Acc, and MCC.

mode (Cai et al., 2010; Call et al., 2010). The concept of Pseudo K-tuple Nucleotide Composition (PseKNC) was also successful applied in the protein/peptide sequences (Chen et al., 2014, 2015c; Liu et al., 2018; Tahir et al., 2019b). In recent years, some useful software were developed, which can generate any desired protein/peptide and DNA/RNA feature, such as Pse-in-One (Liu et al., 2015), Pse-in-One 2.0 (Liu et al., 2017).

2.3. First layer feature information extraction

Each sequence sample was described in the following way:

$$S = N_1 N_2 N_3 \cdots N_L \tag{2}$$

where N_i denotes the i-th nucleotide in the sequence, $N_i \in \{A, C, G, U(T)\}$, and L denotes the length of the sequence.

In this section, seven feature extraction techniques were employed to train the first layer SVM of the designed model, including nucleic acid composition (NAC), di-nucleotide composition (DNC), tri-nucleotide composition (TNC), position-specific trinucleotide propensity (PSTNP), Bi-profile Bayes (BPB), binary encoding (BE), and accumulated nucleotide frequency (ANF). The seven methods can convert sequence samples into numerical vectors for classifier model training.

2.3.1. Nucleic acid composition

The NAC is one of the most common encoding methods and appeared in surveying various biological sequences (Chen et al., 2015b, 2020; Liu et al., 2015, 2017; Sabooh et al., 2018). NAC contain the frequency of each nucleic acid type for each sample. Each sample is made up of four nucleotides, thus nucleic acid composition can be represented in the following manner:

Table 1. Benchmark datasets for different species.

Species	Attribute	Number of sample	Length of sequence
H. sapiens	positive	1130	*41-bp*
	negative	1130	
M. musculus	positive	725	
	negative	725	
rice genome	positive	880	
	negative	880	
S. cerevisiae	positive	1307	*51-bp*
	negative	1307	

$$f(t) = \frac{N(t)}{L}, \quad t \in \{A, C, G, U(T)\} \tag{3}$$

where *f*(*t*) represents the frequency of each nucleic acid type, *N*(*t*) is the number of nucleic acid types *t*, and *L* denotes the length of a nucleotide sequence. Finally, four dimensional feature vectors are extracted by using this method.

2.3.2. Di-nucleotide composition

The DNC contains occurrence information for all nucleotide pairs (Chen et al., 2020; Liu et al., 2015, 2017), which also has potential discriminant information for identifying m6A sites. Thus, DNC can be denoted as the following manner:

$$D(r, s) = \frac{N_{rs}}{L-1}, \quad r, s \in \{A, C, G, U(T)\} \tag{4}$$

where *D*(*r*, *s*) is the frequency of different di-nucleotide composition types, N_{rs} is the number of di-nucleotides represented by considering nucleic acid types *r* and *s*, and *L* is the length of a nucleotide sequence. One can extract $4 \times 4 = 16$ dimensional feature vector by employing DNC method.

2.3.3. Tri-nucleotide composition

The TNC considers the role of three nucleotide compositions (Chen et al., 2020; Liu et al., 2015). The formula for calculating it is similar to that of DNC, which is shown as follows.

$$D(x, y, z) = \frac{N_{xyz}}{L-2}, \quad x, y, z \in \{A, C, G, U(T)\} \tag{5}$$

where $D(x, y, z)$ is the frequency of different tri-nucleotide composition types, N_{xyz} is the number of corresponding tri-nucleotides *xyz*, and *L* is the length of a nucleotide sequence. One can obtain $4^3 = 64$ dimensional feature vector by using the feature extraction method TNC.

2.3.4. Position-specific trinucleotide propensity

The PSTNP reflects the overall trinucleotide content as well as the location information of the respective trinucleotide (Chen et al., 2020; He et al., 2019; Jia et al., 2018). PSTNP is employed to characterize the differences in trinucleotides at each location among m6A and non-m6A sequences. The specific encode flow is described as follows:

For a sequence with sample length *L*, a $(L-2)$ dimensional eigenvector is constructed:

$$D = [\phi_1, \phi_2, \ldots, \phi_u, \ldots, \phi_{L-2}]^T \tag{6}$$

where *T* denotes the transpose operator, and ϕ_u is defined by (7)

Table 2. Binary encoding.

Nucleotides	Binary encoding
A	(1,0,0,0)
C	(0,1,0,0)
G	(0,0,1,0)
U(T)	(0,0,0,1)

$$\phi_u = \begin{cases} z_{1,u}, & \text{when } N_uN_{u+1}N_{u+2} = AAA \\ z_{2,u}, & \text{when } N_uN_{u+1}N_{u+2} = AAC \\ z_{3,u}, & \text{when } N_uN_{u+1}N_{u+2} = AAG \\ \vdots & \qquad \vdots \\ z_{64,u}, & \text{when } N_uN_{u+1}N_{u+2} = UUU(TTT) \end{cases} \quad (1 \le u \le L-2) \tag{7}$$

$Z = (z_{i,j})_{64*(L-2)}$ can be represented using the following equation (8):

$$z_{i,j} = F^{+}(trinucleotide_i|j) - F^{-}(trinucleotide_i|j) \\ (i = 1, 2, \ldots, 64; j = 1, 2, \ldots, L-2) \tag{8}$$

where $F^{+}(trinucleotide_i|j)$ and $F^{-}(trinucleotide_i|j)$ are the frequency of the *i*-th trinucleotide at the *j*-th position in positive instances and negative instances, respectively. For samples, each sequence is converted into a $(L-2)$ dimensional numerical vector. For a sequence with length *L*, the position-specific trinucleotide propensity can be integrated as a matrix (9):

$$Z = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,L-2} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,L-2} \\ \vdots & \vdots & \vdots & \vdots \\ z_{64,1} & z_{64,2} & \cdots & z_{64,L-2} \end{bmatrix} \tag{9}$$

2.3.5. Bi-profile Bayes

The BPB also considers the sequence information of positive and negative samples (Shao et al., 2009), which has been successfully applied in many domains (Jia et al., 2017; Li et al., 2019a, 2019b). Here, each sample sequence of length *L* can be converted into a posterior probability eigenvector:

$$P = (x_1, \ldots, x_{L-1}, x_L, \ldots, x_{2*(L-1)})^T \tag{10}$$

where $x_i(i = 1, 2, \ldots, L-1)$ denotes the posterior probabilities of each nucleic acid of each location in positive sequence datasets and $x_i(i = L, \ldots, 2L-2)$ represents the posterior probabilities of each nucleic acid of each location in negative sequence datasets.

2.3.6. Binary encoding

In the binary encoding, each nucleotide is encoded by using a 4-dimensional binary vector (Chen et al., 2020). This encoding scheme is often used to convert nucleotide sequences with an equal length (Hayouka et al., 2017; Jing et al., 2019). The specific coding method is shown in Table 2.

2.3.7. Accumulated nucleotide frequency

The ANF encoding includes the nucleotide frequency information and the distribution of each nucleotide in the

Table 3. First-level classifier parameters.

Species	*H. sapiens*		*M. musculus*		*rice genome*		*S. cerevisiae*	
Feature	C	γ	C	γ	C	γ	C	γ
NAC	0.70711	0.17678	0.70711	0.25	22.6274	0.2500	16.0000	2.8284
DNC	32.0000	1.0000	22.6274	2.8284	2.82840	1.4142	22.6274	4.0000
TNC	16.0000	2.0000	8.0000	4.0000	2.00000	2.8284	22.6274	0.3536
PSTNP	1.0000	4.0000	22.6274	2.0000	11.3137	4.0000	11.3137	2.0000
BPB	0.2500	0.7071	0.2500	1.0000	8.0000	0.0442	22.6274	0.0884
BE	2.0000	0.0313	1.0000	0.0313	1.4142	0.1768	0.70711	0.0442
ANF	8.0000	0.125	11.3137	0.0884	1.4142	1.4142	22.6274	0.0625

sequence (Chen et al., 2020; Liu et al., 2019), and was formulated as the following equation (11):

$$d_i = \frac{1}{|s_i|}\sum_{i=1}^{L} f(s_i), f(q) = \begin{cases} 1 & if\ s_i = q \\ 0 & othercase \end{cases} \tag{11}$$

where L is the sequence length, $|s_i|$ is the length of the i-th prefix string $\{s_1, s_2, \ldots, s_i\}$ in the sequence, $q \in \{A, C, G, U(T)\}$.

2.4. Ensemble learning

Ensemble learning is the forefront of machine learning (Dietterich, 2002), which combines multiple machine learning machines to accomplish learning tasks. Ensemble learning combines the results of multiple classifiers through a certain rule, thereby obtaining better results than a single classifier (Liu et al., 2016a). In this study, seven SVM classifiers were trained using seven feature extraction techniques at the first level model. According to the prediction result of the first level SVM, the first column decision values were extracted and merged into a new feature space at the second level, which was used as the input features for the second level model. Thus, the input characteristic dimension of the second layer classifier was determined by the number of the first layer classifiers. Finally, the predicted results were obtained through the second layer classifier. Due to different feature selection techniques for different datasets, the final integration results may be different.

2.5. Performance evaluation

Sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthew's correlation coefficient (MCC) were widely adopted to measure the performance of the designed predictor (Liu et al., 2016b; Xu et al., 2013). They can be calculated using the following formulas:

$$Sn = \frac{TP}{TP + FN}, \quad 0 \le Sn \le 1 \tag{12}$$

$$Sp = \frac{TN}{TN + FP}, \quad 0 \le Sp \le 1 \tag{13}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \quad 0 \le Acc \le 1 \tag{14}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \\ -1 \le MCC \le 1 \tag{15}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. The sensitivity (Sn) and specificity (Sp) can characterize the prediction accuracy of the classifiers for positive and negative samples, respectively. Accuracy (Acc) means the prediction accuracy for both positive and negative samples. Matthew's correlation coefficient (MCC) mainly indicates the comprehensive predictive performance of the model (Wang & Yan, 2018). The MCC returns the values between -1 and $+1$, where $+1$ is a perfect prediction, 0 means no better than random prediction, and -1 indicates total disagreement between prediction and observation.

3. Results and discussions

Feature extraction and classifier selection are crucial for target identification and predictor model building within bioinformatics. This is a common method for considering the collective effects of feature extraction and classifiers. In this study, SVM was employed as the original operative machine (Chang & Lin, 2011). Three test strategies are often utilized to evaluate the performance of a predictor, including independence test, k-fold cross validation test, and jackknife test (Chou & Zhang, 1995). Among the three strategies, the independence test is the most objective method, the k-fold cross validation test is the most time-saving method, and the jackknife test is the most reliable method on account of one unique result being produced from multiple experiments (Cheng et al., 2018). The k-fold cross validation test is also equivalent to the jackknife test when k takes 1. In this study, we adopted 10-fold cross validation test and jackknife test to evaluate our predictor by comparing with other existing prediction methods. In the next stage, a series of discussions are offered around the use of Sn, Sp, Acc, and MCC for analyzing optimum predictors.

3.1. First-level classifier parameter

Ensemble learning is a combination of basic machine learning models. These models can be decision trees, neural networks, support vector machines, or other classical machine learning models. Compared with other models, SVM has a wider application in bioinformatics fields (Feng et al., 2019; Tang et al., 2018; Zhu et al., 2019). To simplify the experimental process and save experimental time, the first layer used SVM directly as the classifier, and for the second layer, the results of SVM were compared with other classifiers to optimize the model. For the first level, the radial basis function (RBF) was selected as the kernel function in the SVM model. The regularization parameter C (cost) and kernel

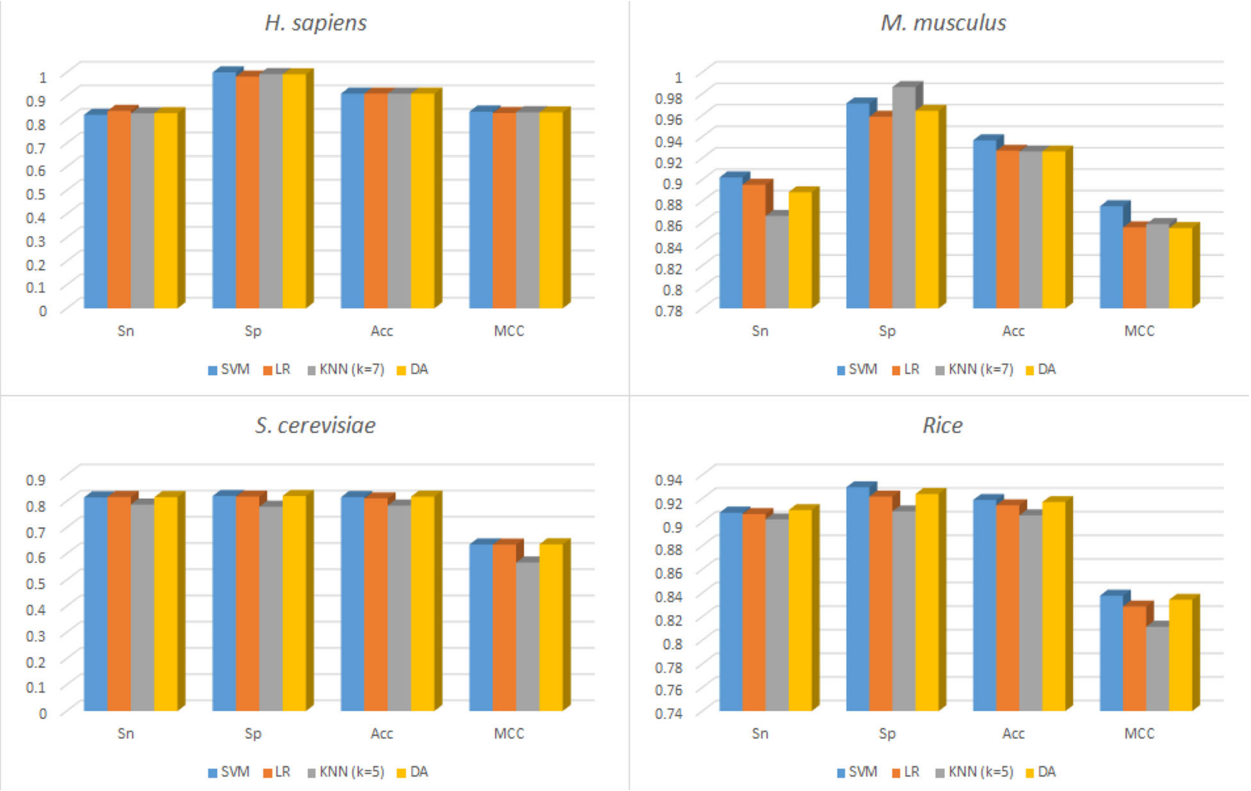

Figure 2. Classifier selection for second-level according to the results of jackknife test.

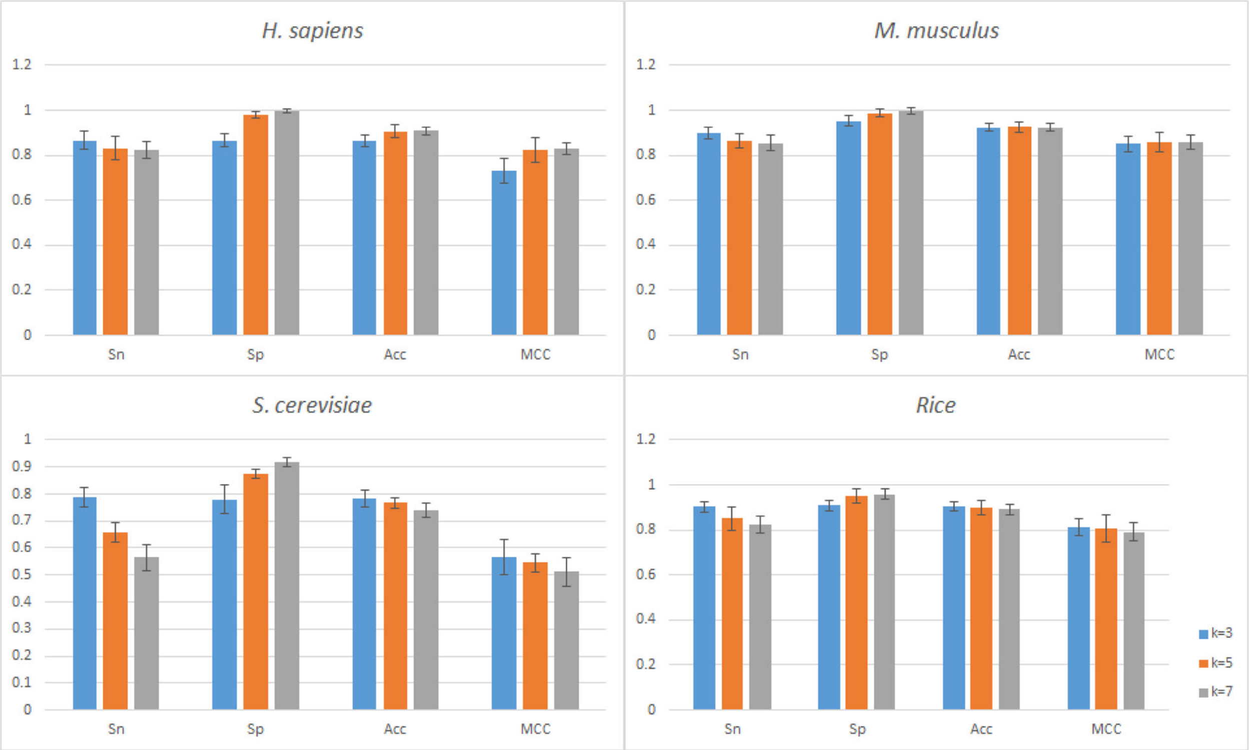

Figure 3. The performance of KNN with different *k* values.

parameter γ (gamma) were optimized by using the grid search method. The search spaces for C and γ were located in $[2^{-2}, 2^{5}]$ and $[2^{-5}, 2^{2}]$, respectively, and the search step size was set to 0.5. The values of C and γ of each group were optimized by means of SVM's classification accuracy in 15-fold cross validation, which were further used in the jackknife test. The specific parameters of each single classifier were shown in Table 3.

3.2 s-level classifier selection

At the first layer, seven SVM models were trained by combining with different feature extraction techniques, and their output decision values were used as inputs of the second layer model. At the second level, four classifiers were employed, which included support vector machine (SVM), k-nearest neighbors (KNN), logistic regression (LR), and discriminant analysis (DA). The comparison jackknife test results for each classifier are shown in Figure 2. When determining the number of neighbors of KNN, we selected the optimal number of neighbors from $\{3, 5, 7\}$ for different species data sets through rigorous 10-fold cross validation. It was eventually decided to select $k = 7$ in dealing with the *H. sapiens* and *M. musculus* datasets, and to select $k = 5$ in handling *S. cerevisiae* and *rice* datasets. The detailed results are illustrated in Figure 3, the error bars were also offered by calculating the standard deviation of the 10-fold cross validation results.

As depicted in Figure 2, the results of each classifier were relatively accurate, suggesting our multiple feature selection integration method to be more effective in predicting the m6A sites. Although the KNN achieved higher specificity, the accuracies of SVM were higher than those of other classifiers on four datasets. Meanwhile, the MCC value obtained by SVM was significantly higher than those of other classifiers in *rice genome* and *M. musculus* datasets. Finally, we chose SVM as the second-level classifier model.

Table 4. The 10-fold cross validation results for different species.

Method	Sn(%)	Sp(%)	Acc(%)	MCC
TL-Methy_H	83.72	98.94	92.23	0.837
TL-Methy_M	95.14	99.03	97.09	0.943
TL-Methy_R	92.39	93.75	93.04	0.862
TL-Methy_S	82.08	83.00	82.54	0.651

3.3. Rigorous 10-fold cross validation analysis

The rigorous 10-fold cross validation test was employed to further illustrate the reliability of the predicted performance of the model. In the first layer, the feature set was randomly divided into 10 parts, in which 9 parts were selected as the training set and the remaining part was used as the test set. The experiments were alternately carried out 10 times. The decision values of 10 test sets results were stored and the first column decision values were utilized to structure the input feature set of the second-level model. According to different training sets, different parameters were determined and the performance of the model was verified with the corresponding test sets. The results of the 10-fold cross validation are reported in Table 4. As can be seen, TL-Methy method achieved the accuracy of 92.23% for *H. sapiens*, of 97.09% for *M. musculus*, of 82.54% for *S. cerevisiae*, and of 93.03% for *rice genome*. The results showed that the 10-fold cross validation results were slightly better than the jackknife test results on four data sets. Therefore, our model was illustrated as a robustness and reliable method.

3.4. Performances comparison between first-level and second-level classifiers

Ensemble learning completes learning tasks by integrating multiple single classifiers. In this study, to verify that the two-layer model performs better than a single SVM, their performance on four sample datasets were evaluated. Results

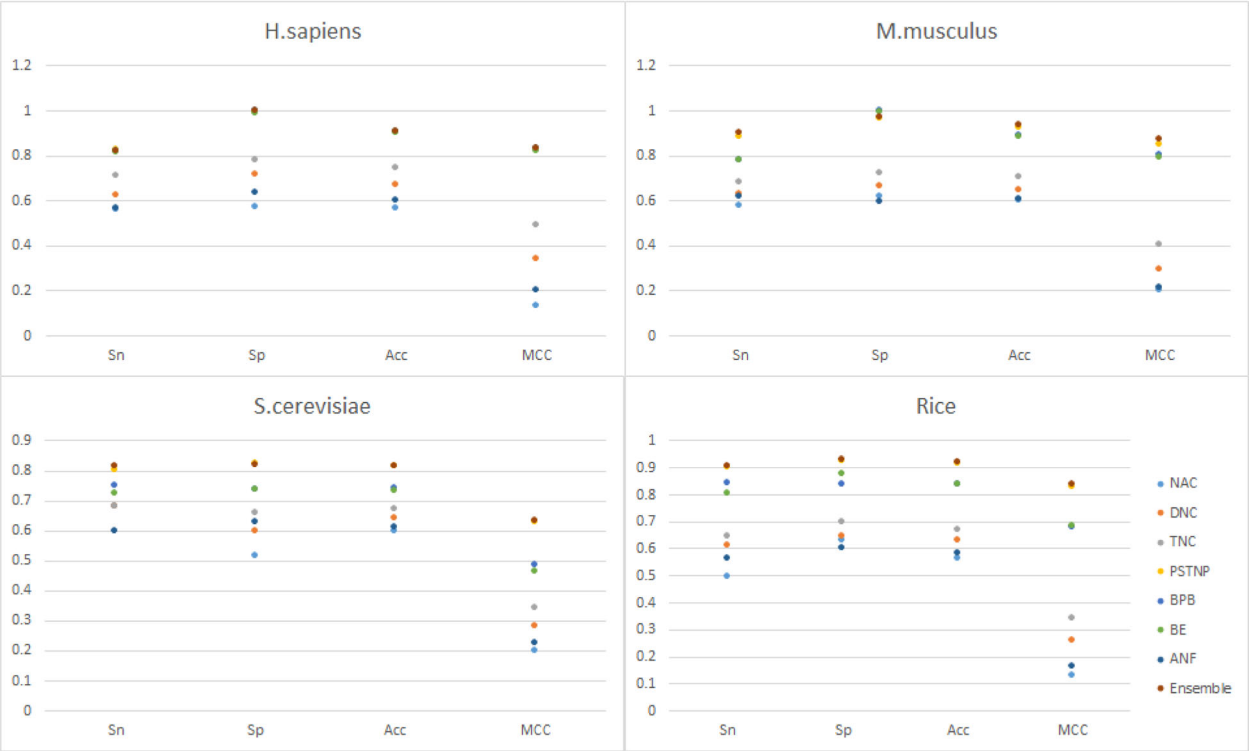

Figure 4. Performances comparison between two-layer ensemble learning algorithm and single SVM.

Table 5. Performances comparison on cross-species joint datasets with single-species datasets.

Method	Sn(%)	Sp(%)	Acc(%)	MCC
TL-Methy_H	82.14	100.00	91.68	0.835
TL-Methy_M	90.21	97.10	93.66	0.875
TL-Methy_H_M	80.59	100.00	90.30	0.822

Table 6. The jackknife test results of different methods.

Species	Model	Sn(%)	Sp(%)	Acc(%)	MCC
H. sapiens	**TL-Methy_H**	**82.14**	**100.00**	**91.68**	**0.835**
	MethyRNA	81.68	99.11	90.38	NAN
	M6AMRFS	82.04	100.00	91.02	0.834
M. musculus	**TL-Methy_M**	**90.21**	**97.10**	**93.66**	**0.875**
	MethyRNA	77.79	100.00	88.39	NAN
	M6AMRFS	78.90	99.59	89.24	0.803
rice genome	**TL-Methy_R**	**90.91**	**92.84**	**91.87**	**0.838**
	i6mA-Pred	82.95	83.30	83.13	0.660
	iDNA6mA(5-step rule)	86.70	86.59	86.64	0.730
	iN6-methylat(5-step)	86.48	89.09	87.78	0.756
	SDM6A	87.80	89.60	88.70	0.774
S. cerevisiae	**TL-Methy_S**	**81.48**	**82.02**	**81.75**	**0.635**
	RAM-ESVM	78.93	77.78	78.35	0.570
	M6AMRFS	75.21	73.30	74.25	0.486
	M6A-PXGB	76.38	75.98	77.13	0.535
	RAM-NPPS	78.42	80.87	79.65	0.590

*The jackknife test result of our model was marked in bold.

of jackknife test are shown in Figure 4. For the *H. sapiens* dataset, the average MCC increased from 0.664 to 0.834; For the *M. musculus* dataset, the average MCC increased from 0.629 to 0.875; For the *S. cerevisiae* dataset, the average MCC increased from 0.441 to 0.635; For the *rice genome* dataset, the average MCC increased from 0.440 to 0.837. In the two-layer learning model, the average MCC of the four species was 0.252 higher than that of the single SVM classifier, which illustrates two-layer learning model is superior to a single SVM classifier for identifying m6A sites. This also demonstrates that ensemble learning effectively enhances the prediction performance of the single classifier.

3.5. Performances comparison between cross-species joint datasets and single-species datasets

Although the designed model performed better for the four tested species, many instances in real life belong to unknown species. To demonstrate whether our method could recognize the m6A sites of unknown species, a predictive model, named TL-Methy_H_M, was designed. *H. sapiens* and *M. musculus* datasets were mixed to generate a cross-species joint dataset to train the model. The accuracy of the cross-species training model was 90.30% based on the jackknife test. Although the results of the cross-species dataset were slightly lower than those of the single-species dataset, the results were also considerable, which could indicate that it is an effective model for predicting unknown species. The specific performance comparison results are reported in Table 5.

3.6. Performances comparison with other methods

To evaluate the identification ability of the designed method, we compared the performances of the TL-Methy and the previously developed predictor in terms of Sn, Sp, Acc, and MCC. The obtained results for each method are listed in Table 6. The results showed that our method obtained more ideal results compared to the existing predictive system. Therefore, the designed method was favorable and can be considered as a potential tool for the prediction of m6A sites.

4. Conclusion

In this study, we designed a two-level predictor model, ‘TL-Methy’, for m6A site identification, in which the outputs of seven SVM models were taken as new inputs for the second-level SVM classifier. To verify the method validity, different classifiers and TL-Methy were compared, and the results demonstrated that TL-Methy were superior to those of other existing methods. Thus, the designed model could be a potential prediction of m6A sites. The related source and code are uploaded in https://github.com/LDWang-dlmu/N6-methyladenine. Accumulating researches on new findings or approaches, such as user-friendly and publicly accessible web-servers will provide an unprecedented revolution to the medicinal chemistry (Liu et al., 2018). In our future research, we will make efforts to demonstrate a web-server to display the findings that can be manipulated by users according to their needs.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (grant nos. 3132019323 and 3132019175).

ORCID

Lidong Wang http://orcid.org/0000-0003-1462-9443

References

Basith, S., Manavalan, B., Shin, T. H., & Lee, G. (2019). SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Molecular Therapy. Nucleic Acids*, *18*, 131–141. https://doi.org/10.1016/j.omtn.2019.08.011

Cai, Y. D., Feng, K. Y., Lu, W. C., & Chou, K. C. (2006). Using LogitBoost classifier to predict protein structural classes. *Journal of Theoretical Biology*, *238* (1), 172–176. https://doi.org/10.1016/j.jtbi.2005.05.034

Cai, Y. D., He, J. F., Li, X. L., Feng, K. Y., Lu, L., Feng, K. R., Kong, X. Y., & Lu, W. C. (2010). Predicting protein subcellular locations with feature selection and analysis. *Protein and Peptide Letters*, *17* (4), 464–472. https://doi.org/10.2174/092986610790963654

Cai, Y. D., Liu, X. J., & Chou, K. C. (2001). Artificial neural network model for predicting membrane protein types. *Journal of Biomolecular Structure & Dynamics*, *18* (4), 607–610. https://doi.org/10.1080/07391102.2001.10506692

Cai, Y. D., Ricardo, P. W., Jen, C. H., & Chou, K. C. (2004). Application of SVM to predict membrane protein types. *Journal of Theoretical Biology*, *226* (4), 373–376. https://doi.org/10.1016/j.jtbi.2003.08.015

Cai, Y. D., Zhou, G. P., & Chou, K. C. (2003). Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical Journal*, *84* (5), 3257–3263. https://doi.org/10.1016/S0006-3495(03)70050-2

Cai, Y. D., Zhou, G. P., & Chou, K. C. (2005). Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. *Journal of Theoretical Biology*, *234* (1), 145–149. https://doi.org/10.1016/j.jtbi.2004.11.017

Call, M. E., Wucherpfennig, K. W., & Chou, J. J. (2010). The structural basis for intramembrane assembly of an activating immunoreceptor complex. *Nature Immunology*, *11* (11), 1023–1029. https://doi.org/10.1038/ni.1943

Cao, D. S., Xu, Q. S., & Liang, Y. Z. (2013). propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics (Oxford, England))*, *29* (7), 960–962. https://doi.org/10.1093/bioinformatics/btt072

Chan, C. T., Dyavaiah, M., DeMott, M. S., Taghizadeh, K., Dedon, P. C., & Begley, T. J. (2010). A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. *PLoS Genetics*, *6* (12), e1001247.https://doi.org/10.1371/annotation/6549d0b1-efde-4aa4-9cda-1cef43f66b30.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27. https://doi.org/10.1145/1961189.1961199

Chen, K., Lu, Z. K., Wang, X., Fu, Y., Luo, G. Z., Liu, N., Han, D., Dominissini, D., Dai, Q., Pan, T., & He, C. (2015a). High-resolution N6-methyladenosine (m6A) map using photo-crosslinking-assisted m6A sequencing. *Angewandte Chemie (International ed. in English))*, *54* (5), 1587–1590. https://doi.org/10.1002/anie.201410647

Chen, W., Feng, P. M., Ding, H., Lin, H., & Chou, K. C. (2015b). iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. *Analytical Biochemistry*, *490*, 26–33. https://doi.org/10.1016/j.ab.2015.08.021

Chen, W., Feng, P. M., Yang, H., Ding, H., Lin, H., & Chou, K. C. (2017a). iRNA-AI: Identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget*, *8* (3), 4208–4217. https://doi.org/10.18632/oncotarget.13758

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. *Analytical Biochemistry*, *456*, 53–60. https://doi.org/10.1016/j.ab.2014.04.001

Chen, W., Lin, H., & Chou, K. C. (2015c). Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences. *Molecular Biosystems*, *11* (10), 2620–2634. https://doi.org/10.1039/C5MB00155B

Chen, W., Lv, H., Nie, F. L., & Lin, H. (2019). i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics (Oxford, England))*, *35* (16), 2796–2800. https://doi.org/10.1093/bioinformatics/btz015

Chen, W., Tang, H., & Lin, H. (2017b). MethyRNA: A web server for identification of N6-methyladenosine sites. *Journal of Biomolecular Structure & Dynamics*, *35* (3), 683–687. https://doi.org/10.1080/07391102.2016.1157761

Chen, W., Yang, H., Feng, P. M., Ding, H., & Lin, H. (2017c). iDNA4mC: Identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics (Oxford, England))*, *33* (22), 3518–3523. https://doi.org/10.1093/bioinformatics/btx479

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., Zhu, Y., Powell, D. R., Akutsu, T., Webb, G. I., Chou, K. C., Smith, A. I., Daly, R. J., Li, J., & Song, J. (2020). iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings in Bioinformatics*, *21*(3), 1047–1057. https://doi.org/10.1093/bib/bbz041

Cheng, X., Xiao, X., & Chou, K. C. (2018). pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. *Genomics*, *110* (4), 231–239. https://doi.org/10.1016/j.ygeno.2017.10.002

Chou, K. C. (2009). Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics*, *6* (4), 262–274. https://doi.org/10.2174/157016409789973707

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology*, *273* (1), 236–247. https://doi.org/10.1016/j.jtbi.2010.12.024

Chou, K. C. (2020). Progresses in predicting post-translational modification. *International Journal of Peptide Research and Therapeutics*, *26*(2), 873-888. https://doi.org/10.1007/s10989-019-09893-5

Chou, K. C. (2020a). Other Mountain Stones Can Attack Jade: The 5-Steps Rule. *Natural Science*, *12* (3), 59–64. https://doi.org/10.4236/ns.2020.123009

Chou, K. C. (2020b). Proposing 5-Steps Rule Is a Notable Milestone for Studying Molecular Biology. *Natural Science*, *12* (03), 74–79. https://doi.org/10.4236/ns.2020.123011

Chou, K. C. (2020c). The Development of Gordon Life Science Institute: Its Driving Force and Accomplishments. *Natural Science*, *12* (04), 202–217. https://doi.org/10.4236/ns.2020.124018

Chou, K. C., & Cai, Y. D. (2003). Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *Journal of Cellular Biochemistry*, *90* (6), 1250–1260. https://doi.org/10.1002/jcb.10719

Chou, K. C., & Zhang, C. T. (1995). Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology*, *30* (4), 275–349. https://doi.org/10.3109/10409239509083488

Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., & Sattar, A. (2015). Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. *Journal of Theoretical Biology*, *364*, 284–294. https://doi.org/10.1016/j.jtbi.2014.09.029

Dietterich, T. G. (2002). Ensemble learning. *The Handbook of Brain Theory and Neural Networks*, *2*, 110–125. https://doi.org/10.1007/978-1-4419-9326-7_1

Du, P. F., 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. *Analytical Biochemistry*, *425* (2), 117–119. https://doi.org/10.1016/j.ab.2012.03.015

Du, X. Q., Diao, Y. Y., Liu, H., & Li, S. (2019). MsDBP: Exploring DNA-Binding Proteins by Integrating Multiscale Sequence Information via Chou's Five-Step Rule. *Journal of Proteome Research*, *18* (8), 3119–3132. https://doi.org/10.1021/acs.jproteome.9b00226

Duan, H. C., Wang, Y., & Jia, G. F. (2019). Dynamic and reversible RNA N6-methyladenosine methylation. *Wiley Interdisciplinary Reviews. Rna*, *10* (1), e1507. https://doi.org/10.1002/wrna.1507

Feng, P. M., Yang, H., Ding, H., Lin, H., Chen, W., & Chou, K. C. (2019). iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics*, *111* (1), 96–102. https://doi.org/10.1016/j.ygeno.2018.01.005

Grosjean, H., Keith, G., & Droogmans, L. (2004). Detection and quantification of modified nucleotides in RNA using thin-layer chromatography. *In: RNA Interference, Editing, and Modification*. Springer. pp. 357–391, https://doi.org/10.1385/1-59259-775-0:357

Hayouka, Z., Bella, A., Stern, T., Ray, S., Jiang, H. B., Grovenor, C. R., & Ryadnov, M. G. (2017). Binary encoding of random peptide sequences for selective and differential antimicrobial mechanisms. *Angewandte Chemie (International ed. in English))*, *56* (28), 8099–8103. https://doi.org/10.1002/anie.201702313

He, W. Y., Jia, C. Z., & Zou, Q. (2019). 4mCPred: Machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics (Oxford, England))*, *35* (4), 593–601. https://doi.org/10.1093/bioinformatics/bty668

Heyn, H., & Esteller, M. (2015). An adenine code for DNA: A second life for N6-methyladenine. *Cell*, *161* (4), 710–713. https://doi.org/10.1016/j.cell.2015.04.021

Hu, L. L., Huang, T., Shi, X. H., 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*(1), e14556. https://doi.org/10.1371/journal.pone.0014556

Huang, J. B., Dong, X., Gong, Z., Qin, L. Y., Yang, S., Zhu, Y. L., Wang, X., Zhang, D. L., Zou, T. T., Yin, P., & Tang, C. (2019). Solution structure of the RNA recognition domain of METTL3-METTL14 N6-methyladenosine methyltransferase. *Protein & Cell*, *10* (4), 272–284. https://doi.org/10.1007/s13238-018-0518-7

Jia, C. Z., He, W. Y., & Yao, Y. H. (2017). OH-PRED: Prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids. *Journal of Biomolecular Structure & Dynamics*, *35* (4), 829–835. https://doi.org/10.1080/07391102.2016.1163294

Jia, C. Z., Yang, Q., & Zou, Q. (2018). NucPosPred: Predicting species-specific genomic nucleosome positioning via four different modes of general PseKNC. *Journal of Theoretical Biology*, *450*, 15–21. https://doi.org/10.1016/j.jtbi.2018.04.025

Jing, X. Y., Dong, Q. W., Hong, D., & Lu, R. Q. (2019). Amino acid encoding methods for protein sequences: A comprehensive review and assessment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, https://doi.org/10.1109/TCBB.2019.2911677.

Kabir, M., Ahmad, S., Iqbal, M., & Hayat, M. (2020). iNR-2L: A two-level sequence-based predictor developed via Chou's 5-steps rule and general PseAAC for identifying nuclear receptors and their families. *Genomics*, *112* (1), 276–285. https://doi.org/10.1016/j.ygeno.2019.02.006

Ke, S., Alemu, E. A., Mertens, C., Gantman, E. C., Fak, J. J., Mele, A., Haripal, B., Zucker-Scharff, I., Moore, M. J., Park, C. Y., Vågbø, C. B., Kusśnierczyk, A., Klungland, A., Darnell, J. E., & Darnell, R. B. (2015). A majority of m6A residues are in the last exons, allowing the potential for 3'UTR regulation. *Genes & Development*, *29* (19), 2037–2053. https://doi.org/10.1101/gad.269415.115

Knuckles, P., Lence, T., Haussmann, I. U., Jacob, D., Kreim, N., Carl, S. H., Masiello, I., Hares, T., Villaseñor, R., Hess, D., Andrade-Navarro, M. A., Biggiogera, M., Helm, M., Soller, M., Bühler, M., & Roignant, J. Y. (2018). Zc3h13/Flacc is required for adenosine methylation by bridging the mRNA-binding factor Rbm15/Spenito to the m6A machinery component Wtap/Fl(2)d. *Genes & Development*, *32* (5-6), 415–429. https://doi.org/10.1101/gad.309146.117

Li, T. Y., Song, R. Y., Yin, Q., Gao, M. Y., & Chen, Y. (2019a). Identification of S-nitrosylation sites based on multiple features combination. *Scientific Reports*, *9* (1), 3098. https://doi.org/10.1038/s41598-019-39743-9

Li, T. Y., Yin, Q., Song, R. Y., Gao, M. Y., & Chen, Y. (2019b). Multidimensional scaling method for prediction of lysine glycation sites. *Computing*, *101* (6), 705–724. https://doi.org/10.1007/s00607-019-00710-x

Lin, X. Y., Chai, G. S., Wu, Y. M., Li, J. X., Chen, F., Liu, J. Z., Luo, G. Z., Tauler, J., Du, J., Lin, S. B., He, C., & Wang, H. S. (2019). RNA m6A methylation regulates the epithelial mesenchymal transition of cancer cells and translation of Snail. *Nature Communications*, *10* (1), 2065. https://doi.org/10.1038/s41467-019-09865-9

Liu, B., Liu, F. L., Wang, X. L., Chen, J. J., Fang, L. Y., & 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* (W1), W65–W71. https://doi.org/10.1093/nar/gkv458

Liu, B., Wang, S. Y., Long, R., & Chou, K. C. (2016a). iRSpot-EL: Identify recombination spots with an ensemble learning approach. *Bioinformatics (Oxford, England))*, *33* (1), 35–41. https://doi.org/10.1093/bioinformatics/btw539

Liu, B., Wu, H., & Chou, K. C. (2017). 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*, *09*(04), 67–91. https://doi.org/10.4236/ns.2017.94007

Liu, B., Yang, F., Huang, D. S., & Chou, K. C. (2018). iPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics (Oxford, England))*, *34* (1), 33–40. https://doi.org/10.1093/bioinformatics/btx579

Liu, Z., Dong, W., Jiang, W., & He, Z. L. (2019). csDMA: An improved bioinformatics tool for identifying DNA 6mA modifications via Chou's 5-step rule. *Scientific Reports*, *9* (1), 1–9. https://doi.org/10.1038/s41598-019-49430-4

Liu, Z., Xiao, X., Yu, D. J., Jia, J. H., Qiu, W. R., & Chou, K. C. (2016b). pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. *Analytical Biochemistry*, *497*, 60–67. https://doi.org/10.1016/j.ab.2015.12.017

Meher, P. K., Sahu, T. K., Saini, V., & Rao, A. R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Scientific Reports*, *7* (1), 1–12. https://doi.org/10.1038/srep42362

Nazari, I., Tahir, M., Tayara, H., & Chong, K. T. (2019). iN6-Methyl (5-step): Identifying RNA N6-methyladenosine sites using deep learning mode via Chou's 5-step rules and Chou's general PseKNC. *Chemometrics and Intelligent Laboratory Systems*, *193*, 103811. https://doi.org/10.1016/j.chemolab.2019.103811.

Nees, G., Kaufmann, A., & Bauer, S. (2014). Detection of RNA Modifications by HPLC Analysis and Competitive ELISA. *In: Innate DNA and RNA Recognition*. Springer. pp. 3–14, https://doi.org/10.1007/978-1-4939-0882-0_1

Pan, T. (2013). N6-methyl-adenosine modification in messenger and long non-coding RNA. *Trends in Biochemical Sciences*, *38* (4), 204–209. https://doi.org/10.1016/j.tibs.2012.12.006

Patil, D. P., Chen, C. K., Pickering, B. F., Chow, A., Jackson, C., Guttman, M., & Jaffrey, S. R. (2016). m6A RNA methylation promotes XIST-mediated transcriptional repression. *Nature*, *537* (7620), 369–373. https://doi.org/10.1038/nature19342

Pendleton, K. E., Chen, B. B., Liu, K. Q., Hunter, O. V., Xie, Y., Tu, B. P., & Conrad, N. K. (2017). The U6 snRNA m6A methyltransferase METTL16 regulates SAM synthetase intron retention. *Cell*, *169* (5), 824–835. https://doi.org/10.1016/j.cell.2017.05.003

Pian, C., Zhang, G. L., Li, F., & Fan, X. D. (2020). MM-6mAPred: Identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics (Oxford, England))*, *36* (2), 388–392. https://doi.org/10.1093/bioinformatics/btz556

Ping, X. L., Sun, B. F., Wang, L., Xiao, W., Yang, X., Wang, W. J., Adhikari, S., Shi, Y., Lv, Y., Chen, Y. S., Zhao, X., Li, A., Yang, Y., Dahal, U., Lou, X. M., Liu, X., Huang, J., Yuan, W. P., Zhu, X. F., … Yang, Y. G. (2014). Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Research*, *24* (2), 177–189. https://doi.org/10.1038/cr.2014.3

Robbins Manke, J. L., Zdraveski, Z. Z., Marinus, M., & Essigmann, J. M. (2005). Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase-and mismatch repair-deficient Escherichia coli. *Journal of Bacteriology*, *187* (20), 7027–7037. https://doi.org/10.1128/JB.187.20.7027-7037.2005

Sabooh, M. F., Iqbal, N., Khan, M., Khan, M., & Maqbool, H. (2018). Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC. *Journal of Theoretical Biology*, *452*, 1–9. https://doi.org/10.1016/j.jtbi.2018.04.037

Schwartz, S., Agarwala, S. D., Mumbach, M. R., Jovanovic, M., Mertins, P., Shishkin, A., Tabach, Y., Mikkelsen, T. S., Satija, R., Ruvkun, G., Carr, S. A., Lander, E. S., Fink, G. R., & Regev, A. (2013). High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*, *155* (6), 1409–1421. https://doi.org/10.1016/j.cell.2013.10.047

Selberg, S., Blokhina, D., Aatonen, M., Koivisto, P., Siltanen, A., Mervaala, E., Kankuri, E., & Karelson, M. (2019). Discovery of Small Molecules that Activate RNA Methylation through Cooperative Binding to the METTL3-14-WTAP Complex Active Site. *Cell Reports*, *26* (13), 3762–3771. https://doi.org/10.1016/j.celrep.2019.02.100

Shao, J., Xu, D., Tsai, S. N., Wang, Y., & Ngai, S. M. (2009). Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PloS One*, *4* (3), e4920. https://doi.org/10.1371/journal.pone.0004920

Shen, H. B., & Chou, K. C. (2008). PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical Biochemistry*, *373* (2), 386–388. https://doi.org/10.1016/j.ab.2007.10.012

Tahir, M., Tayara, H., & Chong, K. T. (2019a). iDNA6mA (5-step rule): Identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. *Chemometrics and Intelligent Laboratory Systems*, *189*, 96–101. https://doi.org/10.1016/j.chemolab.2019.04.007

Tahir, M., Tayara, H., & Chong, K. T. (2019b). 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. https://doi.org/10.1016/j.jtbi.2018.12.034

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., & Lin, H. (2018). HBPred: A tool to identify growth hormone-binding proteins. *International Journal of Biological Sciences*, *14* (8), 957–964. https://doi.org/10.7150/ijbs.24174

Wang, C. X., Cui, G. S., Liu, X., Xu, K., Wang, M., Zhang, X. X., Jiang, L. Y., Li, A., Yang, Y., Lai, W. Y., Sun, B. F., Jiang, G. B., Wang, H. L., Tong, W. M., Li, W., Wang, X. J., Yang, Y. G., & Zhou, Q. (2018). METTL3-mediated m6A modification is required for cerebellar development. *PLoS Biology*, *16* (6), e2004880. https://doi.org/10.1371/journal.pbio.2004880

Wang, X. F., & Yan, R. X. (2018). RFAthM6A: A new tool for predicting m6A sites in Arabidopsis thaliana. *Plant Mol. Biol*, *96* (3), 327–337. https://doi.org/10.1007/s11103-018-0698-9

Wei, L. Y., Chen, H. R., & Su, R. (2018). M6APred-EL: A sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Molecular Therapy. Nucleic Acids*, *12*, 635–644. https://doi.org/10.1016/j.omtn.2018.07.004

Wei, L. Y., Su, R., Wang, B., Li, X. T., Zou, Q., & Gao, X. (2019). Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites. *Neurocomputing*, *324*, 3–9. https://doi.org/10.1016/j.neucom.2018.04.082

Wen, J., Lv, R., Ma, H., Shen, H., He, C., Wang, J., Jiao, F., Liu, H., Yang, P., Tan, L., Lan, F., Shi, Y. G., He, C., Shi, Y., & Diao, J. (2018). Zc3h13 regulates nuclear RNA m6A methylation and mouse embryonic stem cell self-renewal. *Molecular Cell*, *69* (6), 1028–1038. https://doi.org/10.1016/j.molcel.2018.02.015

Xu, Y., Ding, J., Wu, L. Y., & Chou, K. C. (2013). iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PloS One*, *8* (2), e55844. https://doi.org/10.1371/journal.pone.0055844

Xu, Y. G., Zhao, W. L., Olson, S. D., Prabhakara, K. S., & Zhou, X. B. (2018). Alternative splicing links histone modifications to stem cell fate decision. *Genome Biology*, *19* (1), 133. https://doi.org/10.1186/s13059-018-1512-3

Yang, L., Lv, Y. L., Wang, S. Y., Zhang, Q., Pan, Y., Su, D. Q., Lu, Q. Z., & Zuo, Y. C. (2020). Identifying FL11 subtype by characterizing tumor immune microenvironment in prostate adenocarcinoma via Chou's 5-steps rule. *Genomics*, *112* (2), 1500–1515. https://doi.org/10.1016/j.ygeno.2019.08.021

Yue, Y., Liu, J., Cui, X., Cao, J., Luo, G., Zhang, Z., Cheng, T., Gao, M., Shu, X., Ma, H., Wang, F., Wang, X., Shen, B., Wang, Y., Feng, X., He, C., & Liu, J. (2018). VIRMA mediates preferential m6A mRNA methylation in 3'UTR and near stop codon and associates with alternative polyadenylation. *Cell Discovery*, *4* (1), 10. https://doi.org/10.1038/s41421-018-0019-0

Yue, Y., Liu, J. Z., & He, C. (2015). RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes & Development*, *29* (13), 1343–1355. https://doi.org/10.1101/gad.262766.115

Zhang, C. T., & Chou, K. C. (1992). An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci*, *1* (3), 401–408. https://doi.org/10.1002/pro.5560010312

Zhu, X. J., Feng, C. Q., Lai, H. Y., Chen, W., & Hao, L. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowledge Based Systems*, *163*, 787–793. https://doi.org/10.1016/j.knosys.2018.10.007