Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

iDNA6mA-PseKNC: Identifying DNA N^6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC

Pengmian Feng[a], Hui Yang[b], Hui Ding[b], Hao Lin[b,d,**], Wei Chen[c,d,*], Kuo-Chen Chou[b,d,***]

[a] Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, School of Public Health, North China University of Science and Technology, Tangshan 063000, China
[b] Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China
[c] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, Tangshan 063000, China
[d] Gordon Life Science Institute, Boston, MA 02478, USA

ARTICLE INFO

Keywords:
PTMs
N^6-methyladenine
Nucleotide physicochemical properties
General PseKNC
Lingering density
Intuitive metrics

ABSTRACT

N^6-methyladenine (6mA) is one kind of post-replication modification (PTM or PTRM) occurring in a wide range of DNA sequences. Accurate identification of its sites will be very helpful for revealing the biological functions of 6mA, but it is time-consuming and expensive to determine them by experiments alone. Unfortunately, so far, no bioinformatics tool is available to do so. To fill in such an empty area, we have proposed a novel predictor called iDNA6mA-PseKNC that is established by incorporating nucleotide physicochemical properties into Pseudo K-tuple Nucleotide Composition (PseKNC). It has been observed via rigorous cross-validations that the predictor's sensitivity (Sn), specificity (Sp), accuracy (Acc), and stability (MCC) are 93%, 100%, 96%, and 0.93, respectively. For the convenience of most experimental scientists, a user-friendly web server for iDNA6mA-PseKNC has been established at http://lin-group.cn/server/iDNA6mA-PseKNC, by which users can easily obtain their desired results without the need to go through the complicated mathematical equations involved.

1. Introduction

As a dynamic DNA epigenetic modification, N^6-methyladenine (6mA) has been found in the following three kingdoms of life [1]: bacteria, archaea, and eukaryotes. DNA-adenine methyltransferase catalyzes the adenine methylation by adding a methyl group to the sixth position of the purine ring of the adenine [2,3], whereas its reversible modification (demethylation) is catalyzed by demethylase enzymes [4]. The first DNA 6mA demethylase was found in *Drosophila* and is belonging to the TET protein family. Recently, the AlkB family members ALKBH1 and NMAD-1 were observed to demethylate 6mA in DNA of mammals and *C. elegans*, respectively [1].

Being one kind of post-replication modification (PTM or PTRM), 6mA has participated in a broad spectrum of biological processes. In prokaryotes, 6mA has been found to be associated with a wide range of biological processes such as DNA replication [5], repair [6], transcription [7], and cellular defense [8–10]. Unlike the better-characterized RNA m6A, our knowledge about the potential roles of 6mA in eukaryotes is very limited; in other words, it is still in the infancy stage for eukaryotes [11]. Accordingly, identifying the genomic locations of 6mA will be very useful for the in-depth understanding of its biological functions.

To this end, a series of experimental techniques have been proposed to detect 6mA in both prokaryotes and eukaryotes such as ultra-high performance liquid chromatography coupled with mass spectrometry (UHPLC-ms/ms) [12], capillary electrophoresis and laser-induced fluorescence (CE-LIF) [13], methylated DNA immunoprecipitation sequencing (MeDIP-seq) [14], and single-molecule real-time sequencing (SMRT-seq) [15]. Although it is time-consuming and expensive to use experimental methods alone in performing genome-wide detection for 6mA sites, these techniques did play very important roles and provide key clues in stimulating the development of this important area. By using the sensitive detection techniques, 6mA sites have also been detected [16] in mouse and human cells.

* Correspondence to: W. Chen Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, Tangshan 063000, China.
** Correspondence to: H. Lao Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China.
*** Correspondence to: K.C. Chou Gordon Life Science Institute, Boston, MA 02478, USA.
E-mail addresses: hlin@uestc.edu.cn (H. Lin), wchen@gordonlifescience.org (W. Chen), kcchou@gordonlifescience.org (K.-C. Chou).

https://doi.org/10.1016/j.ygeno.2018.01.005
Received 27 October 2017; Received in revised form 24 December 2017; Accepted 7 January 2018

Recently, by integrating the publicly available SMRT sequencing datasets, the first database in this area, called "MethSMRT", was developed [17]. It hosts DNA methylomes and provides invaluable data for developing computational methods to predict the genomic localization of 6mA sites.

During the last few years, many powerful web-server predictors have been developed to identify various types of PTM sites in biological sequences (see e.g., [18–43] [16]). Unfortunately, none of them can be used to identify the 6mA sites in DNA. The present study was initiated in an attempt to fill such an empty area.

According to the Chou's 5-step rule [44], to develop a really useful sequence-based predictor for a biological system as done in a series of recent publications [36,37,45–50], one should make the following five procedures very clear: (i) bench mark dataset, (ii) sequence sample formulation, (iii) operation engine or algorithm, (iv) cross-validation, and (v) web server. Below, we are to address these procedures one by one.

2. Materials and methods

2.1. Benchmark dataset

In literature, the benchmark dataset usually consists of a training dataset and a testing dataset; the former is for the purpose of training a proposed model, while the latter is for the purpose of testing it, but as elucidated in Chou and Shen [51], it would suffice with one high quality benchmark dataset if the model is tested by the jackknife or subsampling (K-fold cross-validation) test [52] because the outcome thus obtained is actually a combination from many different independent dataset tests. Suppose the benchmark dataset in the current study is denoted by $\mathbb{S}$, which may be formulated as

$$\mathbb{S} = \mathbb{S}^{+} \bigcup \mathbb{S}^{-} \tag{1}$$

where $\mathbb{S}^{+}$ denotes positive subset, $\mathbb{S}^{-}$ as negative subset, and the symbol $\bigcup$ represents the union in the set theory.

The positive subset contains only, i.e., 6mA site-containing sequences, which were taken from the genome of *Mus musculus* in the MethSMRT database [17]. All these sequence samples are 41-bp long with the 6mA site located at the center. In order to construct a high quality benchmark dataset, the following two procedures were performed. Firstly, according to the Methylome Analysis Technical Note [17], those sequence samples with modQV > 30 were left out. Secondly, to reduce homology bias, those samples with pairwise sequence identity > 80% were removed by using the CD-HIT software [53]. Finally, we obtained 1934 positive samples for the positive subset.

The negative subset contains negative samples only, i.e., non-6mA site-containing sequences. They were obtained by choosing the 41-bp long sequences with "A" at their center but not being detected by the SMRT sequencing technology as of 6mA. By doing so, we could obtain a huge number of negative samples, from which we randomly picked 1934 samples to form the negative subset for the purpose of using a balance benchmark dataset to train the model [27,54–56].

The final benchmark dataset thus obtained is given in Supporting Information S1.

2.2. Sequence sample formulation

For a DNA sample with 41 bp, its most straightforward expression is

$$\mathbf{D} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\cdots\mathrm{N}_i\cdots\mathrm{N}_{41} \tag{2}$$

where

$$\mathrm{N}_i \in \{\mathrm{A\ (adenine)},\ \ \mathrm{C\ (cytosine)},\ \ \mathrm{G\ (guanine)},\ \ \mathrm{T\ (thymine)}\} \tag{3}$$

denotes the nucleotide at the *i*-th sequence position, and ∈ is the a symbol in the set theory meaning "member of".

Since all the existing machine-learning algorithms, such as Component-Coupled algorithm [57], SVM (Support Vector Machine) [58,59], LogitBoost [60], KNN (K-Nearest Neighbor) [61], PCA (Principal Component Analysis) [62], and RF (Random Forest) [28,40], can only handle vectors [24], we have to convert the sequential expression of Eq. (2) into a vector, but a vector defined in a discrete model might completely lose all the sequence-order information. To deal with this problem, the PseAAC (Pseudo Amino Acid Composition) was introduced [63,64]. Ever since the concept of pseudo amino acid composition or Chou's PseAAC [65–67] was proposed, it has been swiftly penetrated into many biomedicine and drug development areas [68,69] and nearly all the areas of computational proteomics (see e.g., [45,47,49,70–84] and a long list of references cited in two review papers [85,86]). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, its idea has been extended to deal with DNA/RNA sequences [26,36,46,48,87–90] in computational genomics via PseKNC (Pseudo K-tuple Nucleotide Composition) [91,92]. Recently, a very powerful web server called "Pse-in-One" [93] and its updated version "Pse-in-One 2.0" [94] were established, by which users can generate any pseudo components for both protein/peptide and DNA/RNA sequences according to their need or definition.

According to a recent review paper [92], the general form of PseKNC for **D** of Eq. (2) can be formulated as

$$\mathbf{D} = [\phi_1\ \ \phi_2\ \ \cdots\ \ \phi_u\ \ \cdots\ \ \phi_\Gamma]^{\mathbf{T}} \tag{4}$$

where the components $\phi_u(u = 1, 2, \ldots)$ and Γ is an integer; their values will depend on how to extract the desired features from the DNA sample; **T** is the transposing operator to a matrix or vector.

As shown in Eq. (3), DNA consists of four types of nucleotides. They can be classified into three different categories (Table 1): (1) from the angle of ring number, A and G have two rings, whereas C and T only one; (2) from the chemical functionality, A and C belong to amino group, while G and T to keto group; (3) from the angle of hydrogen bonding, C and G can be bonded to each other with three hydrogen bonds, but A and T with only two. All these properties would have different impacts to DNA's low-frequency internal motion [95,96] and its biological function [97–99] as well.

To incorporate these local features into Eq. (4), the following equation [100,101] is used to denote the *i*-th nucleotide in a DNA sequence

$$\mathrm{N}_i = (x_i, y_i, z_i) \tag{5}$$

where x_i, y_i, and z_i refer to the attributes of (1) ring structure, (2) functional group, and (3) hydrogen bonding, respectively (Table 1). Accordingly, the nucleotide A can be formulated as (1, 1, 1), C as (0, 1, 0), G as (1, 0, 0), and T as (0, 0, 1); or generally we have

$$x_i = \begin{cases} 1, & \text{if } \mathrm{N}_i \in \{\mathrm{A}, \mathrm{G}\} \\ 0, & \text{if } \mathrm{N}_i \in \{\mathrm{C}, \mathrm{T}\} \end{cases}; \quad y_i = \begin{cases} 1, & \text{if } \mathrm{N}_i \in \{\mathrm{A}, \mathrm{C}\} \\ 0, & \text{if } \mathrm{N}_i \in \{\mathrm{G}, \mathrm{T}\} \end{cases};$$

$$z_i = \begin{cases} 1, & \text{if } \mathrm{N}_i \in \{\mathrm{A}, \mathrm{T}\} \\ 0, & \text{if } \mathrm{N}_i \in \{\mathrm{C}, \mathrm{G}\} \end{cases} \tag{6}$$

To incorporate the sequence-coupled features into Eq. (4), we adopt the lingering density as defined below.

Table 1
Classification of nucleotides in DNA[a].

Category[b]	Attribute	Nucleotides	Code[c]
Ring structure (x_i)	Purine	A, G	1
	Pyrimidine	C, T	0
Functional group (y_i)	Amino	A, C	1
	Keto	G, T	0
Hydrogen bonding (z_i)	Stronger	C, G	1
	Weaker	A, T	0

[a] See the section of "Sequence Sample Formulation" for further explanation.
[b] See Eq. (5).
[c] See Eq. (6).

$$D_i = \frac{1}{\|L_i\|} \sum\nolimits_{j=1}^{\ell} f\,(\mathrm{N}_j) \tag{7}$$

where D_i is the density of the nucleotide N_i at the site i of a DNA sequence; $\|L_i\|$ is the length of the sliding substring concerned; ℓ denotes each of the site locations counted in the substring, and

$$f\,(\mathrm{N}_j) = \begin{cases} 1, & \text{if } \mathrm{N}_j\text{=the nucleotide concerned} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

For instance, suppose a DNA sequence “ATTGAC”. The lingering density of ‘A’ at the sequence position 1, 2, 3, 4, 5, or 6 is $1 = 1/1$, $0.5 = 1/2$, $0.33 \approx 1/3$, $0.25 = 1/4$, $0.40 = 2/5$, or $0.33 \approx 2/6$, respectively; that of “C” is $0 = 0/1$, $0 = 0/2$, $0 = 0/3$, $0 = 0/4$, $0 = 0/5$ or $0.16 = 1/6$, respectively; and so forth.

By combing Eq. (5) and Eq. (7), the i-th nucleotide of Eq. (2) can be uniquely defined by a set of four variables; i.e.,

$$\mathrm{N}_i = (x_i, y_i, z_i,\ \ D_i) \qquad i = 1,\ 2,\ \ldots, L \tag{9}$$

where L is the length of the DNA sequence concerned.

For example, the DNA sequence “ACGTA” can be expressed by the following five sets of digital numbers: (1, 1, 0, 1), (0, 1, 1, 0.5), (1, 0, 1, 0.33), (0, 0, 0, 0.25), (1, 1, 0, 0.40). Submitting these numbers into Eq. (4), we have

$$\begin{aligned} &\mathbf{D}(\text{ACGTA}) \\ &= [\,1\ \ 1\ \ 0\ \ 1\ \ 0\ \ 1\ \ 1\ \ 0.5\ \ 1\ \ 0\ \ 1\ \ 0.33\ \ 0\ \ 0\ \ 0\ \ 0.25\ \ 1\ \ 1\ \ 0\ \ 0.40\,]^{\mathbf{T}} \end{aligned} \tag{10}$$

meaning that the 5-nt nucleotide example can be defined by a 5 × 4 = 20-D (dimensional) PseKNC vector. Accordingly, for a 41-bp DNA sequence in the benchmark dataset $\mathbb{S}$ (cf. Supporting Information S1), Eq. (4) should become

$$\mathbf{D}(41\text{bp}) = [\phi_1\ \ \phi_2\ \ \cdots\ \ \phi_u\ \ \cdots\ \ \phi_{164}]^{\mathbf{T}} \tag{11}$$

2.3. Operation engine or algorithm

The prediction was operated by SVM (Support Vector Machine), which has been widely used in various areas of bioinformatics and computational biology (see e.g., [21,23,25,55,56,58,88,89,102–113]. Its basic idea has been elaborated in the aforementioned papers, and there is no need to repeat it here.

In the current study, the LibSVM package 3.18 was used to implement SVM, which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The SVM algorithm contains two uncertain quantities; one is the regularization parameter C, and the other is the kernel width parameter γ. They were optimized via an optimization procedure using the grid search approach as described by

$$\begin{cases} 2^{-5} \le C \le 2^{15} & \text{with step } \Delta C = 2 \\ 2^{-15} \le \gamma \le 2^{-5} & \text{with step } \Delta\gamma = 2^{-1} \end{cases} \tag{12}$$

where ΔC and $\Delta\gamma$ represent the step gaps for C and γ, respectively. Suppose the SVM output score for $\mathbf{D}$ (cf. Eqs. (2), (4), and (11)) is $\mathfrak{P}(\mathbf{D})$, it follows

$$\mathbf{D} \in \begin{cases} \text{6mA sample,} & \text{if } \mathfrak{P}(\mathbf{D}) > 0.5 \\ \text{non}-\text{6mA sample,} & \text{if } \mathfrak{P}(\mathbf{D}) \le 0.5 \end{cases} \tag{13}$$

For those readers who are interested in knowing more about SVM, see the papers [114,115] or a monograph [116] where a brief introduction or detailed description is given, respectively.

The predictor obtained via the aforementioned procedures is called iDNA6mA-PseKNC, where “i” stands for “identify”, “DNA6mA” for “N^6-methyladenine modification sites in DNA”, and “PseKNC” for “by incorporating nucleotide physicochemical properties into pseudo K-tuple nucleotide composition”.

2.4. Cross-validation

It is important to evaluate the quality of a new predictor or its performance. For this, we need to consider the following two problems. First, what metrics should be used to measure the predictor's quality? Secondly, what method should be adopted to calculate the metrics? Below, we are to address the two problems.

In literature, the following four metrics are often used to evaluate a predictor's quality [117]: (i) overall accuracy (Acc); (ii) stability (MCC); (iii) sensitivity (Sn); and (4) specificity (Sp). But, their formulations directly taken from math books are not intuitive and difficult to be understood by most biological scientists. Fortunately, using the symbols introduced by Chou [118] in studying signal peptides, the four metrics can be converted to a set of intuitive ones [18] as given below:

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \tag{14}$$

where N^+ represents the total number of positive samples investigated, while $N_-{}^+$ is the number of positive samples incorrectly predicted to be of negative one; N^- the total number of negative samples investigated, while $N_+{}^-$ the number of the negative samples incorrectly predicted to be of positive one.

With the metrics of Eq. (14), the meanings of Sn, Sp, Acc, and MCC have become crystal clear as confirmed in a series of follow-up studies for many different areas (see, e.g., [21,22,27–35,40,45,47,54–56,80,81,105,111,119–121]. However, it is instructive to point out that, with the sequence analysis studies going into a deeper level, increasing numbers of multi-label sequence samples have been emerging in system biology and medicine (see e.g., [122] [33,61,123–133]. To deal with this kind of multi-label systems, a much more sophisticated set of metrics is needed as elaborated in [134].

The following three different cross-validation methods are often used to examine a predictor's performance [52]: (i) independent dataset test, (ii) subsampling (or K-fold cross-validation) test, and (iii) jackknife test. Of these three, however, the jackknife test is the least arbitrary and most objective [44]. Therefore, the jackknife test has been widely recognized and increasingly adopted by researchers to analyze the quality of various predictors (see e.g., [45,58,78,81,119,121,135–145]). In view of this, here we also used the jackknife test to examine the quality of the current prediction method. The jackknife test can exclude the “memory” effect since both the training dataset and testing dataset in a jackknife system are actually open, and each sample will be in turn moved between the two. Also, the arbitrariness problem intrinsic to the independent dataset and subsampling tests [44] no longer exists because the outcome derived via the jackknife test for a predictor is always the same on a given benchmark dataset.

3. Results and discussion

The jackknife test results for the iDNA6mA-PseKNC predictor on the benchmark dataset in Supporting Information S1 are given below.

$$\begin{cases} \mathrm{Sn} = 93.28\% \\ \mathrm{Sp} = 100.00\% \\ \mathrm{Acc} = 96.73\% \\ \mathrm{MCC} = 0.9300 \end{cases} \tag{15}$$

Since iDNA6mA-PseKNC is the first predictor ever developed for identifying N^6-methyladenosine sites in DNA, it is impossible to show its power via a comparison with its counterparts. However, the rates in

Table 2
A comparison of different classifiers in identifying 6mA site based on the same benchmark dataset via jackknife test.

Classifiers	Sn (%)	Sp (%)	Acc (%)	MCC
Naïve Bayes	93.54	93.79	93.67	0.87
BayesNet	93.54	98.34	96.04	0.92
J48	93.22	95.96	94.59	0.89
Random Forest	93.28	98.34	95.91	0.92
LogitBoost	93.00	96.76	94.88	0.90
SVM	93.28	100	96.73	0.93

Table 3
Predicted results by iDNA6mA-PseKNC on the samples collected from eight other genomes.

Genome	Number of samples	Number of corrected prediction	Success rate
Caenorhabditis elegans	121,192	110,146	90.86%
Arabidopsis thaliana	174,016	143,072	82.21%
Acidobacteria bacterium	12,546	11,562	92.16%
Alteromonadaceae bacterium	1637	1577	96.33%
E. coli	40,152	39,840	99.22%
Polycyclovorans algicola	6604	6081	92.08%
Ruminococcus flavefaciens	10,183	10,126	99.44%
Sphingomonas melonis	7479	7300	97.61%

Eq. (15) indicate that the prediction quality of iDNA6mA-PseKNC is indeed very high, with specificity reaching 100%, and overall accuracy and sensitivity $>$ 96% and 93%, respectively. Particularly, the predictor is also very stable as reflected by the fact of MCC $>$ 0.93 (cf. Eq. (14)).

3.1. Comparison of SVM with other classifiers

To demonstrate the right choice of using SVM for identifying 6mA site, we compared the predictive results by SVM against those by other classifiers, such as Naïve Bayes, BayesNet, J48, Random Forest, and LogitBoost that were implemented with their respective default parameters in WEKA [146]. Listed in Table 2 are the corresponding jackknife test results based on the same benchmark dataset. As we can see from the table, the SVM classifier achieves the highest rates in Acc and MCC, the two most important metrics [46,88] among the four in Eq. (14).

3.2. Validation on independent datasets

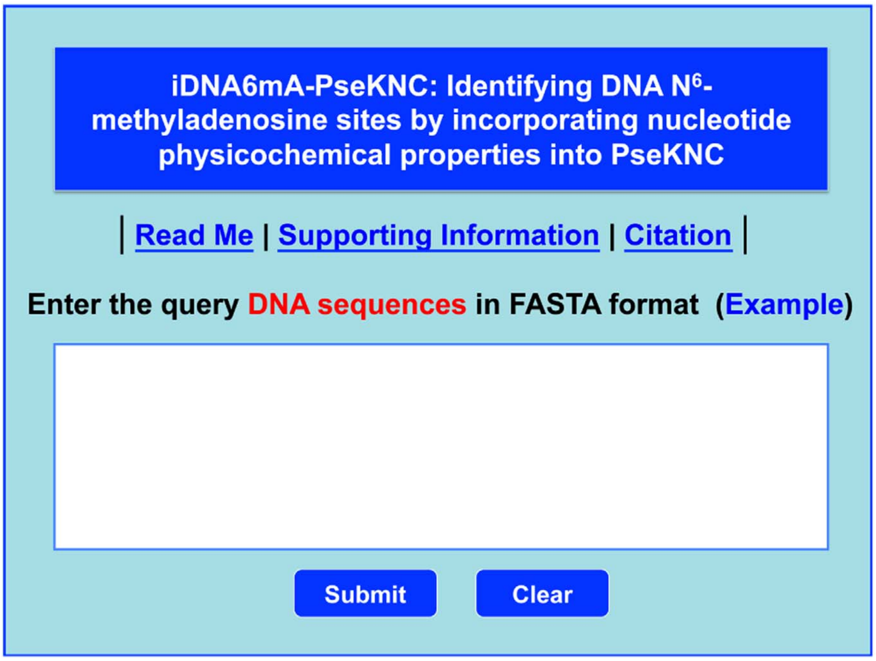
As mentioned above, the jackknife test is the most objective cross-validation approach [44,52,147,148] that has combined a series of different independent dataset tests, and hence for many cases there is no need to do independent dataset test again, but it would be instructive for practical applications [149] by performing the following independent dataset tests.

Following the same procedures as described in "Benchmark Dataset" section, we obtained eight sets of 6mA site-containing sequences from the genomes of (i) *Caenorhabditis elegans*, (ii) *Arabidopsis thaliana*, (iii) *Acidobacteria bacterium*, (iv) *Alteromonadaceae bacterium*, (v) *E. coli*, (vi) *Polycyclovorans algicola*, (vii) *Ruminococcus flavefaciens*, and (viii) *Sphingomonas melonis* genomes, respectively. All these sequences are also 41-bp long with the true 6mA site in the center, and their numbers are given in the 2nd column of Table 3. As we can see from the table, the success rates obtained by using the model trained by the benchmark dataset from *Mus musculus* to the genomes of other eight organisms are all very high, indicating that iDNA6mA-PseKNC is indeed quite promising and holds a high potential to become a useful tool in genome-wide analysis for identifying 6mA sites.

3.3. Web server

It has been clearly pointed out in [150] that user-friendly and publicly accessible web servers represent the future direction for developing practically more useful predictors. As shown by a series of recent publications [18–37,39–42,151,152], a new prediction method with the availability of a user-friendly web server would significantly enhance its impacts [24,86,148]. In view of this, the web server for the new predictor iDNA6mA-PseKNC has been established at http://lin-group.cn/server/iDNA6mA-PseKNC. Moreover, to maximize the convenience of most experimental scientists, a step-by-step guide of how to use the web server to get their desired results is given in given below.

(1) Click the link at http://lin-group.cn/server/iDNA6mA-PseKNC and you'll see the web server's top page as shown in Fig. 1.



iDNA6mA-PseKNC: Identifying DNA N^6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC

| Read Me | Supporting Information | Citation |

Enter the query DNA sequences in FASTA format (Example)

Fig. 1. A semi-screenshot for the top-page of the iDNA6mA-PseKNC web server.

(2) Either type or copy/paste the sequences of query DNA sequences into the input box at the center of Fig. 1. The input sequence should be in the FASTA format. For the examples of sequences in FASTA format, click the Example button right above the input box.

(3) Click on the Submit button to see the predicted result. For instance, if you use the two DNA sequences in the Example window as the input, after a few seconds, you will see the following on the screen of your computer. (i) Seq 1 contains 17 A (adenine) nucleotides, and only the ones in the positions 35, 41, 62, and 71 may be of 6mA modification. (ii) Seq 2 contains 14 A nucleotides, and only the ones in the positions 38, 62, and 71 may be of 6mA modification.

(4) Click the Supporting Information button to download the Supporting Information mentioned in this paper.

(5) Click on the Citation button to find the papers that have played the key roles in developing the current predictor of iDNA6mA-PseKNC.

4. Conclusions

The proposed predictor iDNA6mA-PseKNC is the first bioinformatics tool ever developed for identifying N^6-methyladenine (6mA) sites in DNA sequences. It not only achieves quite high success rates but is also with a web server, by which users can easily obtain their desired results without the need to go through the mathematical formulations. The reason of including the mathematical details in this paper is for its integrity, and for that they may be of use in stimulating the development of more powerful methods for predicting other PTM sites as well.

Although the model is trained by using the benchmark dataset derived from the genome of *M. musculus*, its success rates for identifying 6mA sites in many other species are also very high. It is anticipated that iDNA6mA-PseKNC will become a very useful high throughput tool for both basic research and drug development.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.01.005.

Acknowledgments

The authors wish to thank the two anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper. This work was supported by the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028), the Outstanding Youth Foundation of North China University of Science and Technology (No. JP201502), the Applied Basic Research Program of Sichuan Province (No. 2015JY0100), and the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015J144; ZYGX2015Z006; ZYGX2016J118; ZYGX2016J125; ZYGX2016J126), China Postdoctoral Science Foundation (No. 2015M582533).

References

[1] Z.K. O'Brown, E.L. Greer, N^6-Methyladenine: a conserved and dynamic DNA mark, Adv. Exp. Med. Biol. 945 (2016) 213–246.

[2] L.M. Iyer, S. Abhiman, L. Aravind, Natural history of eukaryotic DNA methylation systems, Prog. Mol. Biol. Transl. Sci. 101 (2011) 25–104.

[3] G.Z. Luo, C. He, DNA N(6)-methyladenine in metazoans: functional epigenetic mark or bystander? Nat. Struct. Mol. Biol. 24 (2017) 503–506.

[4] T.P. Wu, T. Wang, M.G. Seetin, Y. Lai, S. Zhu, K. Lin, Y. Liu, S.D. Byrum, S.G. Mackintosh, M. Zhong, A. Tackett, G. Wang, L.S. Hon, G. Fang, J.A. Swenberg, A.Z. Xiao, DNA methylation on N(6)-adenine in mammalian embryonic stem cells, Nature 532 (2016) 329–333.

[5] J.L. Campbell, N. Kleckner, *E. coli* oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork, Cell 62 (1990) 967–979.

[6] P.J. Pukkila, J. Peterson, G. Herman, P. Modrich, M. Meselson, Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in *Escherichia coli*, Genetics 104 (1983) 571–582.

[7] J.L. Robbins-Manke, Z.Z. Zdraveski, M. Marinus, J.M. Essigmann, Analysis of global gene expression and double-strand-break formation in DNA adenine methyltransferase- and mismatch repair-deficient Escherichia Coli, J. Bacteriol. 187 (2005) 7027–7037.

[8] S.E. Luria, M.L. Human, A nonhereditary, host-induced variation of bacterial viruses, J. Bacteriol. 64 (1952) 557–569.

[9] M. Meselson, R. Yuan, DNA restriction enzyme from *E. coli*, Nature 217 (1968) 1110–1114.

[10] S. Linn, W. Arber, Host specificity of DNA produced by *Escherichia coli*, X. In vitro restriction of phage fd replicative form, Proc. Natl. Acad. Sci. U. S. A. 59 (1968) 1300–1306.

[11] M.J. Koziol, C.R. Bradshaw, G.E. Allen, A.S. Costa, C. Frezza, Identification of methylated deoxyadenosines in genomic DNA by dA6m DNA immunoprecipitation, Bio. Protoc. 6 (2016).

[12] E.L. Greer, M.A. Blanco, L. Gu, E. Sendinc, J. Liu, D. Aristizabal-Corrales, C.H. Hsu, L. Aravind, C. He, Y. Shi, DNA methylation on N6-Adenine in *C. elegans*, Cell 161 (2015) 868–878.

[13] A.M. Krais, M.G. Cornelius, H.H. Schmeiser, Genomic N(6)-methyladenine determination by MEKC with LIF, Electrophoresis 31 (2010) 3548–3551.

[14] K.R. Pomraning, K.M. Smith, M. Freitag, Genome-wide high throughput analysis of DNA methylation in eukaryotes, Methods 47 (2009) 142–150.

[15] B.A. Flusberg, D.R. Webster, J.H. Lee, K.J. Travers, E.C. Olivares, T.A. Clark, J. Korlach, S.W. Turner, Direct detection of DNA methylation during single-molecule, real-time sequencing, Nat. Methods 7 (2010) 461–465.

[16] W. Chen, H. Yang, P. Feng, H. Ding, H. Lin, iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, Bioinformatics 33 (2017) 3518–3523.

[17] P. Ye, Y. Luan, K. Chen, Y. Liu, C. Xiao, Z. Xie, MethSMRT: an integrative database for DNA N^6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing, Nucleic Acids Res. 45 (2017) D85–D89.

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

[19] Y. Xu, X.J. Shao, L.Y. Wu, iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, PeerJ. 1 (2013) e171.

[20] W.R. Qiu, X. Xiao, W.Z. Lin, iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach, Biotechnol. Res. Int. 2014 (2014) 947416.

[21] Y. Xu, X. Wen, X.J. Shao, N.Y. Deng, iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition, Int. J. Mol. Sci. 15 (2014) 7594–7610.

[22] Y. Xu, X. Wen, L.S. Wen, L.Y. Wu, iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PLoS One 9 (2014) e105018.

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

[24] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.

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

[26] W. Chen, H. Tang, J. Ye, H. Lin, iRNA-PseU: Identifying RNA pseudouridine sites, Mol. Ther.–Nucleic Acids 5 (2016) e332.

[27] J. Jia, Z. Liu, X. Xiao, iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset, Anal. Biochem. 497 (2016) 48–56.

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

[29] J. Jia, Z. Liu, X. Xiao, B. Liu, iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC, Oncotarget 7 (2016) 34558–34570.

[30] J. Jia, L. Zhang, Z. Liu, pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC, Bioinformatics 32 (2016) 3133–3141.

[31] Z. Liu, X. Xiao, D.J. Yu, J. Jia, pRNAm-PC: predicting N-methyladenosine sites in RNA sequences via physical-chemical properties, Anal. Biochem. 497 (2016) 60–67.

[32] W.R. Qiu, B.Q. Sun, Z.C. Xu, iHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC, Oncotarget 7 (2016) 44310–44321.

[33] W.R. Qiu, B.Q. Sun, Z.C. Xu, iPTM-mLys: identifying multiple lysine PTM sites and their different types, Bioinformatics 32 (2016) 3116–3123.

[34] W.R. Qiu, X. Xiao, iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier, Oncotarget 7 (2016) 51270–51283.

[35] Y. Xu, Recent progress in predicting posttranslational modification sites in proteins, Curr. Top. Med. Chem. 16 (2016) 591–603.

[36] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Mol. Ther.–Nucleic Acids 7 (2017) 155–163.

[37] L.M. Liu, Y. Xu, iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC, Med. Chem. 13 (2017) 552–559.

[38] W.R. Qiu, S.Y. Jiang, B.Q. Sun, iRNA-2methyl: identify RNA 2′-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier, Med. Chem. 13 (2017) 734–743.

[39] W.R. Qiu, S.Y. Jiang, Z.C. Xu, iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide

composition, Oncotarget 8 (2017) 41178–41188.

[40] W.R. Qiu, B.Q. Sun, D. Xu, iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory, Mol. Inf. 36 (2017) (UNSP 1600010).

[41] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics (2017), http://dx.doi.org/10.1016/j.ygeno.2017.10.008.

[42] Y. Xu, C. Li, iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC, Med. Chem. 13 (2017) 544–551.

[43] X. Cheng, S.G. Zhao, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (2017) 58494–58503.

[44] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review), J. Theor. Biol. 273 (2011) 236–247.

[45] P.K. Meher, T.K. Sahu, V. Saini, A.R. Rao, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC, Sci. Rep. 7 (2017) 42362.

[46] B. Liu, S. Wang, R. Long, iRSpot-EL: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (2017) 35–41.

[47] H. Huo, T. Li, S. Wang, Y. Lv, Y. Zuo, L. Yang, Prediction of presynaptic and postsynaptic neurotoxins by combining various Chou's pseudo components, Sci. Rep. 7 (2017) 5827.

[48] B. Liu, F. Yang, K.C. Chou, 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function, Mol. Ther.–Nucleic Acids 7 (2017) 267–277.

[49] P. Tripathi, P.N. Pandey, A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition, J. Theor. Biol. 424 (2017) 49–54.

[50] W. Chen, P. Feng, H. Yang, H. Lin, iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, Oncotarget 8 (2017) 4208–4217.

[51] K.C. Chou, H.B. Shen, Review: recent progresses in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.

[52] C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

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

[54] J. Jia, Z. Liu, B. Liu, iPPBS-opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets, Molecules 21 (2016) 95.

[55] Z. Liu, X. Xiao, W.R. Qiu, iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77.

[56] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J. Biomol. Struct. Dyn. 33 (2015) 2221–2233.

[57] G.M. Maggiora, Domain structural class prediction, Protein Eng. 11 (1998) 523–538.

[58] J. Chen, R. Long, X.L. Wang, B. Liu, dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation, Sci. Rep. 6 (2016) 32333.

[59] Q. Su, W. Lu, D. Du, F. Chen, B. Niu, Prediction of the aquatic toxicity of aromatic compounds to tetrahymena pyriformis through support vector regression, Oncotarget 8 (2017) 49359–49369.

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

[61] X. Xiao, P. Wang, W.Z. Lin, J.H. Jia, iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types, Anal. Biochem. 436 (2013) 168–177.

[62] Q.S. Du, S.Q. Wang, N.Z. Xie, Q.Y. Wang, R.B. Huang, 2L-PCA: a two-level principal component analyzer for quantitative drug design and its applications, Oncotarget 8 (2017) 70564–70578.

[63] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins Struct. Funct. Genet. 43 (2001) 246–255 (Erratum: ibid., 2001, Vol.44, 60).

[64] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.

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

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

[67] S.X. Lin, J. Lapointe, Theoretical and experimental biology in one —a symposium in honour of professor Kuo-Chen Chou's 50th anniversary and professor Richard Giegé's 40th anniversary of their scientific careers, J. Biomed. Sci. Eng. 6 (2013) 435–442.

[68] W.Z. Zhong, S.F. Zhou, Molecular science for drug development and biomedicine, Int. J. Mol. Sci. 15 (2014) 20072–20078.

[69] G.P. Zhou, W.Z. Zhong, Perspectives in medicinal chemistry, Curr. Top. Med. Chem. 16 (2016) 381–382.

[70] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo amino acid composition and support vector machine for prediction of enzyme subfamily classes, J. Theor. Biol. 248 (2007) 546–551.

[71] L. Nanni, A. Lumini, Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization, Amino Acids 34 (2008) 653–660.

[72] M. Esmaeili, H. Mohabatkar, S. Mohsenzadeh, Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses, J. Theor. Biol. 263 (2010) 203–209.

[73] S.S. Sahu, G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction, Comput. Biol. Chem. 34 (2010) 320–327.

[74] H. Mohabatkar, M. Mohammad Beigi, A. Esmaeili, Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo amino acid composition and support vector machine, J. Theor. Biol. 281 (2011) 18–23.

[75] M. Mohammad Beigi, M. Behjati, H. Mohabatkar, Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach, J. Struct. Funct. Genom. 12 (2011) 191–197.

[76] L. Nanni, A. Lumini, D. Gupta, A. Garg, Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (2012) 467–475.

[77] E. Pacharawongsakda, T. Theeramunkong, Predict subcellular locations of Singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of Chou's PseAAC, IEEE Trans. Nanobioscience 12 (2013) 311–320.

[78] S. Mondal, P.P. Pai, Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction, J. Theor. Biol. 356 (2014) 30–35.

[79] S. Ahmad, M. Kabir, M. Hayat, Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into Chou's general PseAAC, Comput. Methods Prog. Biomed. 122 (2015) 165–174.

[80] J. Jia, Z. Liu, X. Xiao, B. Liu, Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC), J. Biomol. Struct. Dyn. 34 (2016) 1946–1961.

[81] M. Khan, M. Hayat, S.A. Khan, N. Iqbal, Unb-DPC: identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC, J. Theor. Biol. 415 (2017) 13–19.

[82] Y.S. Jiao, P.F. Du, Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions, J. Theor. Biol. 416 (2017) 81–87.

[83] M. Rahimi, M.R. Bakhtiarizadeh, A. Mohammadi-Sangcheshmeh, OOgenesis_Pred: a sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition, J. Theor. Biol. 414 (2017) 128–136.

[84] B. Yu, L. Lou, S. Li, Y. Zhang, W. Qiu, X. Wu, M. Wang, B. Tian, Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising, J. Mol. Graph. Model. 76 (2017) 260–273.

[85] K.C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, Curr. Proteomics 6 (2009) 262–274.

[86] K.C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (2017) 2337–2358.

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

[88] W.R. Qiu, X. Xiao, iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. 15 (2014) 1746–1766.

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

[90] B. Liu, F. Yang, D.S. Huang, iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, Bioinformatics 34 (2018) 33–40.

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

[92] W. Chen, H. Lin, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. BioSyst. 11 (2015) 2620–2634.

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

[94] B. Liu, H. Wu, Pse-in-one 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nat. Sci. 9 (2017) 67–91.

[95] K.C. Chou, Low-frequency vibrations of DNA molecules, Biochem. J. 221 (1984) 27–31.

[96] G.M. Maggiora, B. Mao, Quasi-continuum models of twist-like and accordion-like low-frequency motions in DNA, Biophys. J. 56 (1989) 295–305.

[97] N.Y. Chen, S. Forsen, The biological functions of low-frequency phonons: 2. Cooperative effects, Chem. Scr. 18 (1981) 126–132.

[98] B. Mao, Collective motion in DNA and its role in drug intercalation, Biopolymers 27 (1988) 1795–1815.

[99] K.C. Chou, Review: low-frequency collective motion in biomacromolecules and its biological functions, Biophys. Chem. 30 (1988) 3–48.

[100] C.T. Zhang, Diagrammatization of codon usage in 339 HIV proteins and its biological implication, AIDS Res. Hum. Retrovir. 8 (1992) 1967–1976.

[101] C.T. Zhang, Analysis of codon usage in 1562 *E. Coli* protein coding sequences, J. Mol. Biol. 238 (1994) 1–8.

[102] P.M. Feng, W. Chen, H. Lin, iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, Anal. Biochem. 442 (2013) 118–125.

[103] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, Bioinformatics 30 (2014) 472–479.

[104] H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, H. Lin, W. Chen, iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels, Biomed. Res. Int. 2014 (2014) 286419.

[105] W. Chen, P.M. Feng, H. Lin, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, Biotechnol. Res. Int. 2014 (2014) 623149.

[106] W. Chen, P.M. Feng, E.Z. Deng, H. Lin, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83.

[107] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, W. Chen, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.

[108] Y.N. Fan, X. Xiao, J.L. Min, iNR-Drug: predicting the interaction of drugs with nuclear receptors in cellular networking, Int. J. Mol. Sci. 15 (2014) 4915–4937.

[109] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, PLoS One 9 (2014) e106691.

[110] R. Xu, J. Zhou, B. Liu, Y.A. He, Q. Zou, Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach, J. Biomol. Struct. Dyn. 33 (2015) 1720–1730.

[111] B. Liu, L. Fang, S. Wang, X. Wang, Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, J. Theor. Biol. 385 (2015) 153–159.

[112] B. Liu, L. Fang, R. Long, X. Lan, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics 32 (2016) 362–369.

[113] W. Chen, H. Tang, H. Lin, MethyRNA: a web server for identification of N(6)-methyladenosine sites, J. Biomol. Struct. Dyn. 35 (2017) 683–687.

[114] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.

[115] Y.D. Cai, G.P. Zhou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophys. J. 84 (2003) 3257–3263.

[116] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Chapter 3, Cambridge University Press, 2000.

[117] J. Chen, H. Liu, J. Yang, Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, Amino Acids 33 (2007) 423–428.

[118] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.

[119] F. Ali, M. Hayat, Classification of membrane protein types using voting feature interval in combination with Chou's pseudo amino acid composition, J. Theor. Biol. 384 (2015) 78–83.

[120] W. Chen, P. Feng, H. Ding, H. Lin, Using deformation energy to analyze nucleosome positioning in genomes, Genomics 107 (2016) 69–75.

[121] Z. Ju, J.Z. Cao, H. Gu, Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC, J. Theor. Biol. 397 (2016) 145–150.

[122] X. Xiao, Z.C. Wu, A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites, PLoS One 6 (2011) e20592.

[123] X. Xiao, Z.C. Wu, iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites, J. Theor. Biol. 284 (2011) 42–51.

[124] K.C. Chou, Z.C. Wu, X. Xiao, iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Mol. BioSyst. 8 (2012) 629–641.

[125] W.Z. Lin, J.A. Fang, X. Xiao, iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, Mol. BioSyst. 9 (2013) 634–644.

[126] X. Cheng, X. Xiao, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, Gene 628 (2017) 315–321 (Erratum: ibid., 2018, Vol.644, 156).

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

[128] X. Cheng, X. Xiao, pLoc-mGneg: predict subcellular localization of gram-negative bacterial proteins by deep gene ontology learning via general PseAAC, Genomics (2017), http://dx.doi.org/10.1016/j.ygeno.2017.10.002.

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

[130] X. Cheng, S.G. Zhao, W.Z. Lin, X. Xiao, pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (2017) 3524–3531.

[131] X. Xiao, X. Cheng, S. Su, Q. Nao, pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of gram-positive bacterial proteins, Nat. Sci. 9 (2017) 331–349.

[132] X. Cheng, X. Xiao, pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information, Bioinformatics (2017), http://dx.doi.org/10.1093/bioinformatics/btx711.

[133] X. Cheng, S.G. Zhao, X. Xiao, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (2017) 341–346 (Corrigendum, ibid., 2017, Vol.33, 2610).

[134] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. BioSyst. 9 (2013) 1092–1100.

[135] G.P. Zhou, N. Assa-Munt, Some insights into protein structural class prediction, Proteins Struct. Funct. Genet. 44 (2001) 57–59.

[136] G.P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, Proteins Struct. Funct. Genet. 50 (2003) 44–48.

[137] Y.D. Cai, Prediction of membrane protein types by incorporating amphipathic effects, J. Chem. Inf. Model. 45 (2005) 407–413.

[138] M. Hayat, A. Khan, Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC, Protein Pept. Lett. 19 (2012) 411–421.

[139] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, J. Theor. Biol. 364 (2015) 284–294.

[140] Z.U. Khan, M. Hayat, M.A. Khan, Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model, J. Theor. Biol. 365 (2015) 197–203.

[141] R. Kumar, A. Srivastava, B. Kumari, M. Kumar, Prediction of beta-lactamase and its class by Chou's pseudo amino acid composition and support vector machine, J. Theor. Biol. 365 (2015) 96–103.

[142] K. Ahmad, M. Waris, M. Hayat, Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition, J. Membr. Biol. 249 (2016) 293–304.

[143] M. Kabir, M. Hayat, iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples, Mol. Gen. Genomics. 291 (2016) 285–296.

[144] M. Behbahani, H. Mohabatkar, M. Nosrati, Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition, J. Theor. Biol. 411 (2016) 1–5.

[145] M. Tahir, M. Hayat, M. Kabir, Sequence based predictor for discrimination of enhancer and their types by applying general form of Chou's trinucleotide composition, Comput. Methods Prog. Biomed. 146 (2017) 69–75.

[146] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics 20 (2004) 2479–2481.

[147] K.C. Chou, H.B. Shen, Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms, Nat. Protoc. 3 (2008) 153–162.

[148] H.B. Shen, Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms, Nat. Sci. 2 (2010) 1090–1103.

[149] D.W. Elrod, Protein subcellular location prediction, Protein Eng. 12 (1999) 107–118.

[150] H.B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2009) 63–92.

[151] W. Chen, H. Ding, P. Feng, H. Lin, iACP: a sequence-based tool for identifying anticancer peptides, Oncotarget 7 (2016) 16895–16909.

[152] B. Liu, H. Wu, D. Zhang, X. Wang, Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods, Oncotarget 8 (2017) 13338–13343.