*Bioinformatics*, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

*Subject Section*

# 4mCPred: Machine Learning Methods for DNA $N^4$-methylcytosine sites Prediction

Wenying He[1], Cangzhi Jia[2,*], Quan Zou[1,*]

[1] School of Computer Science and Technology, Tianjin University, Tianjin, China; [2] Department of Mathematics, Dalian Maritime University, Dalian, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** $N^4$-methylcytosine (4mC), an important epigenetic modification formed by the action of specific methyltransferases, plays an essential role in DNA repair, expression and replication. The accurate identification of 4mC sites aids in-depth research to biological functions and mechanisms. Because, experimental identification of 4mC sites is time-consuming and costly, especially given the rapid accumulation of gene sequences. Supplementation with efficient computational methods is urgently needed.

**Results:** In this study, we developed a new tool, 4mCPred, for predicting 4mC sites in *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Escherichia coli*, *Geoalkalibacter subterraneus* and *Geobacter pickeringii*. 4mCPred consists of two independent models, 4mCPred_I and 4mCPred_II, for each species. The predictive results of independent and cross species tests demonstrated that the performance of 4mCPred_I is a useful tool. To identify position-specific trinucleotide propensity (PSTNP) and electron-ion interaction potential features, we used the *F-score* method to construct predictive models and to compare their PSTNP features. Compared with other existing predictors, 4mCPred achieved much higher accuracies in rigorous jackknife and independent tests. We also analysed the importance of different features in detail.

**Contact:** zouquan@nclab.net or cangzhijia@dlmu.edu.cn.

**Availability:** The web-server 4mCPred, is accessible at http://server.malab.cn/4mCPred/index.jsp.

## 1 Introduction

Methylation is catalyzed by enzymes that participate in heavy metal modification, regulation of chromatin organization and regulation of gene expression. DNA methylation is a very important epigenetic modification that involving the addition of a methyl group to DNA. The most widespread DNA methylation modifications are N6-methyladenine (6mA), 5-methylcytosine (5mC) and N4-methylcytosine (4mC). These modifications are catalyzed by specific DNA methyltransferases (DNMTs) that transfer a methyl group to specific exocyclic amino groups (Wang, et al., 2016; Zacharias, 1993). 6mA and 5mC are found in both prokaryotes and eukaryotes (Hattman, 2005; Hattman, et al., 1978), whereas 4mC is present only in prokaryotes (Lyko, 2018; Sanchez-Romero, et al., 2015). 5mC is the most widely distributed methylation modification in higher eukaryotes (Hattman, et al., 1978) and the most widely studied. 5mC is a very important transcriptional repressor in genomes and is associated with the regulation of developmental processes and transcriptional silencing of transposons (Bestor, 2000; Tajima and Suetake, 1998). In prokaryotes, 6mA plays important roles as a marker in DNA repair (Harrison and Karrer, 1985; Messer and Noyer-Weidner, 1988) and in chromosome replication (Lu, et al., 1994; Ogden, et al., 1988) and cell regulation (Campbell, et al., 1990; Collier, et al., 2007). DNA $N^4$-methylcytosine, similar to 5mC, can participate in restriction modification systems (Ehrlich, et al., 1987). Some cytosine-specific DNA methyltransferases also play important roles in DNA repair, expression and replication (Glickman and Radman, 1980; Lu, et al., 1983; Pukkila, et al., 1983).

Although extensive research has been conducted on 5mC and 6mA modifications, studies on 4mC have been relatively limited because effective experimental approaches are lacking. In-depth study of these modification sites would help elucidate biological functions and mechanisms and aid the treatment of some genomic diseases. Experimental identification of 4mC sites is time-consuming and costly, especially given the rapid accumulation of gene sequences. An urgent need therefor exists to supplement approaches with efficient computational methods. Although various computational methods have been proposed as excellent complements to methylation sites prediction (Feng, et al., 2018), theoretical calculations and methods for 4mC site prediction are insufficient. Chen et al. recently (Chen, et al., 2017) developed an efficient prediction tool, iDNA4mC, which achieved accuracies of 78.04% (*Caenorhabditis elegans*), 81.16% (*Drosophila melanogaster*), 76.05% (*Arabidopsis thaliana*), 79.82% (*Escherichia coli*)), 81.53% (*Geoalkalibacter subterraneus*) and 84.04% (*Geobacter pickeringii*). Although iDNA4mC yielded quite satisfactory results, further improvement is needed.

The purpose of this study was to establish a new predictor, 4mCPred, to further improve the performance of 4mC site prediction. To fully extract information from benchmark datasets, position-specific preferences of trinucleotides and electron-ion interaction pseudopotential (EIIP) values of trinucleotides were used to transform DNA sequences into numerical vectors. To obtain a better generalization prediction model, an optimal feature selection process was adopted to select the optimal feature subset.

In accordance with Chou's five-step rule (Chou, 2011; Liu, et al., 2014), the following five essential steps must be carried out to establish a powerful statistical predictor: (i) construction of a valid benchmark dataset; (ii) formulation of an effective mathematical expression to model samples of interest; (iii) development of a suitable machine learning algorithm; (iv) operation of the model using cross-validation tests and (v) creation of a user-friendly web server for the predictor. The workflow of constructing the 4mCPred model is shown in Figure 1.

## 2 Materials and methods

### 2.1 Datasets

In this study, we adopted datasets constructed by Chen et al.(Chen, et al., 2017), to specifically identify DNA 4mC sites (Additional file 1: Datasets 1). We used these datasets for three reasons. First, the datasets were based on the MethSMRT database (Ye, et al., 2017) and thus are reliable. Second, sequence-homology reduction had been performed with an 80% threshold to avoid potential redundancy. Finally, and most importantly, the use of these datasets allowed us to fairly compare our newly proposed method with existing ones. The final datasets contained six species: *C. elegans*, *D. melanogaster*, *A. thaliana*, *E. coli*, *Geoalkalibacter subterraneus* and *Geobacter pickeringii*. (A sample with incomplete information was removed from the *Geoalkalibacter subterraneus* dataset.) A statistical summary of the six species datasets is provided in Table 1.

**Table 1. Statistical summary of six species datasets**

| Species | Number of Positive Samples | Number of Negative Samples |
|---|---|---|
| *C.elegan* | 1554 | 1554 |
| *D.melanogaster* | 1769 | 1769 |
| *A.thaliana* | 1978 | 1978 |
| *E.coli* | 388 | 388 |
| *G.subterraneus* | 905 | 906 |
| *G.pickeringii* | 569 | 569 |

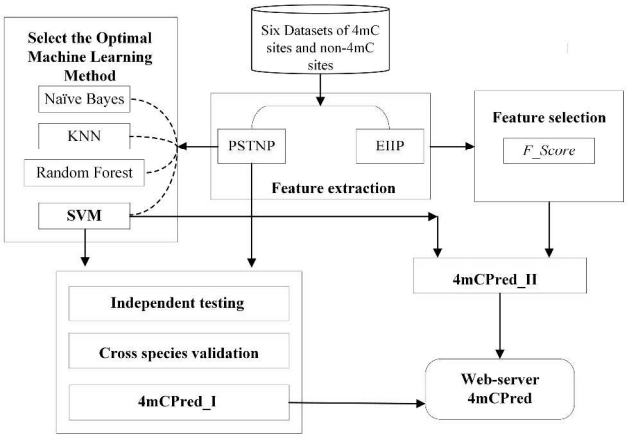



**Figure 1. The workflow of 4mCPred.**

### 2.2 Feature vector construction

In this study, we used a strategy based on two approaches, namely, position-specific trinucleotide sequence propensity (PSTNP) and EIIPs, to encode sample sequences.

#### 2.2.1 PSTNP

PSTNP, which describes differences in trinucleotides at each position between 4mC and non-4mC sequences, has been used successfully in many fields (Chen, et al., 2015; Chou, 2011; He and Jia, 2017; He, et al., 2018; Lin, et al., 2014; Pei Li, 2015). PSTNP reflects deformations in general pseudo amino acid composition (PseAAC), a concept first proposed in 2001 (Chou, 2001). Some recent studies have extended PseAAC to encompass pseudo dipeptide or pseudo *K*-spaced dipeptide composition (Chen, et al., 2016; Tang, et al., 2016; Yang, et al., 2016). Encouraged by successes based on the introduction of the PseAAC concept into computational proteomics, its application has been expanded to computational DNA and RNA fields (Guo, et al., 2014; Li, et al., 2015; Lin, et al., 2014; Lin, et al., 2017; Zhang, et al., 2016). Powerful web-servers named Pse-in-One (Liu, et al., 2015) and Pse-in-One 2.0 (Liu, et al., 2017) were recently established in succession to generate any general pseudo-amino acid composition or feature vectors of protein and DNA/RNA sequences in accordance with researches' needs. We believe, PSTNSP can complement existing nucleotide composition-based methods.

A PSTNSP profile $Z=(z_{i,j})_{64*39}$ can be constructed using the formula

$$z_{i,j} = F^{+}(\text{trinucleotide}_i \mid j) - F^{-}(\text{trinucleotide}_i \mid j) \quad (1)$$
$$(i = 1,2,...,64;\ j = 1,2,...39)$$

where $F^{+}(\text{trinucleotide}_i \mid j)$ and $F^{-}(\text{trinucleotide}_i \mid j)$ represent the frequency of the *i*-th trinucleotide at the *j*-th position in positive and negative training samples, respectively. In the present study, $\text{trinucleotide}_1$ is AAA, $\text{trinucleotide}_2$ is AAC, trinucleotide3 is AAG, …, $\text{trinucleotide}_{64}$ is TTT. The profile Z can then be applied to construct an (*L*-2)-dimensional feature vector $D=[\phi_1,\phi_2,...,\phi_u,...,\phi_{L-2}]^T$ for a sample DNA sequence of length *L* (i.e., *L*=41) as follow.

$$\phi_u = \begin{cases} z_{1,u}, & \text{when } N_u N_{u+1} N_{u+2} = \text{trinucleotide}_1 \\ z_{2,u}, & \text{when } N_u N_{u+1} N_{u+2} = \text{trinucleotide}_2 \\ z_{3,u}, & \text{when } N_u N_{u+1} N_{u+2} = \text{trinucleotide}_3 \\ \quad\vdots & \quad\vdots \quad\quad\vdots \\ z_{64,u}, & \text{when } N_u N_{u+1} N_{u+2} = \text{trinucleotide}_{64} \end{cases} \quad (1 \le u \le 39) \qquad (2)$$

### 2.2.2 EIIP values of trinucleotides

Nair and Sreenadhan (Nair and Sreenadhan, 2006) proposed the use of EIIPs to express the distribution of electron-ion energies along a DNA sequence. This approach has proven to be an effective feature encoding method. Because of its simple form and high computational efficiency, EIIP encoding method has been widely used to forecast items such as enhancers (He and Jia, 2017), hot spots in protein (Khan, 2011; Sahu and Panda, 2011) and nucleosome (Jia, et al., 2018).

**Table 2. EIIP values of nucleotides**

| Nucleotide | EIIP |
|---|---|
| A | 0.1260 |
| C | 0.1340 |
| G | 0.0806 |
| T | 0.1335 |

In this study, we used the average EIIP values of trinucleotides in a sequence to construct feature vectors. Nucleotide EIIP values of each nucleotide are given in Table2. The composition of each sample can be represented as a 64-dimensional feature vector (He and Jia, 2017) as follows:

$$D = [\text{EIIP}_{AAA} \cdot f_{AAA}, \text{EIIP}_{AAC} \cdot f_{AAC}, \cdots, \text{EIIP}_{TTT} \cdot f_{TTT}] \qquad (3)$$

In equation (3), the subscripts represent different trinucleotides. $\text{EIIP}_{xyz}$=$\text{EIIP}_x$+$\text{EIIP}_y$+$\text{EIIP}_z$ is the EIIP value of trinucleotide xyz ,where $x, y, z \in \{A, C, G, T\}$ , $f_{xyz}$ is the normalized frequency of trinucleotide xyz.

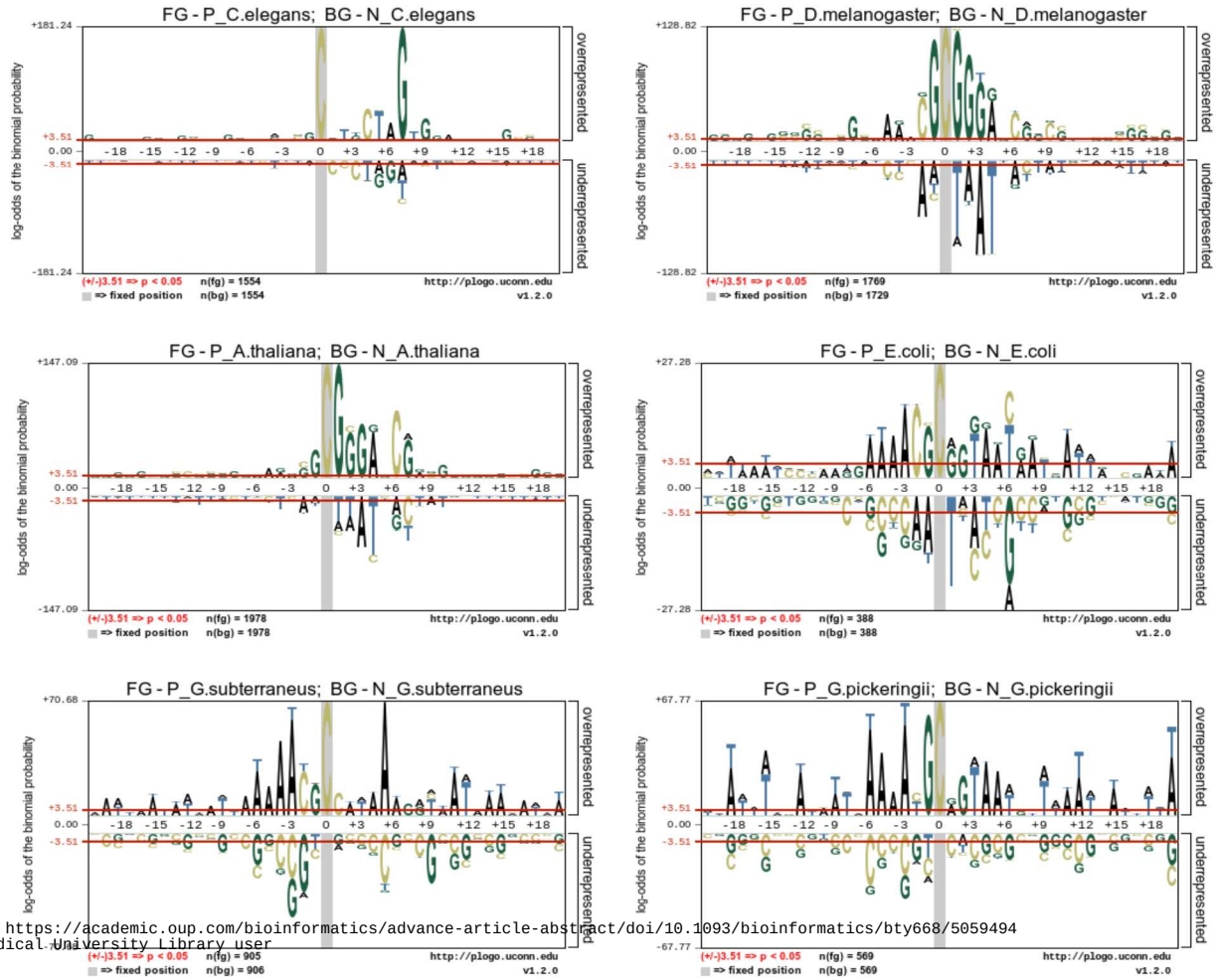
## 2.3 Feature selection

The modeling of high-dimensional data is often computationally expensive, and, because of the scarcity of meaningful data, obtaining a good generalization predictive model is difficult. Feature selection identifies the most significant subset of features of a target problem by eliminating redundant and irrelevant features (Senawi, et al., 2017; Wei and Billings, 2007). This search method helps enhance model classification performance, improves interpretability and reduces bias. To identify significant features, we applied the *F*-score method (Chen, et al., 2016) to compute the statistical value of each dimension. The *F-score* value of the *i*-th feature is defined as follows:

$$F - score(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n^+ - 1}\sum_{k=1}^{n^+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n^- - 1}\sum_{k=1}^{n^-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \qquad (4)$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$ and $\bar{x}_i^{(-)}$ are the average values of the *i*-th feature in whole, positive and negative datasets, respectively. $n^+$ and $n^-$ are respectively the number of positive and negative training samples, respectively, while $x_{k,i}^{(+)}$ and $x_{k,i}^{(-)}$ represent the *i*-th feature of *k*-th positive and negative samples, respectively. A high *F*-score implies that a feature has superior discrimination ability.

**Figure 2. Sequence logos between 4mC-modified and non-modified sites of *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Escherichia coli*, *Geoalkalibacter subterraneus* and *Geobacter pickeringii*. The n(fg) and n(bg) values at the bottom left of each pLogo plot indicate the number of sequences with the full 41 nucleotides used to generate the image. The red horizontal lines on each pLogo plot correspond to a significance threshold value of 3.51 ($p < 0.05$)**



### 2.4 Construction of a predictive model

Different machine learning methods have different varying abilities to learn categories. Considering the diversities of different machine learning method, we conduct contrastive experiment to choose the optimal machine learning method based on feature PSTNP. We trained widely used classifiers include Naïve Bayes (Rish, 2001), KNN (Cover and Hart, 1967), RF(Ho, 1995) and SVM (Cao, et al., 2014) , respectively.

Because the SVM received optimal predictive performance (See the Additional file 2: Table S1), and so other machine learning methods was not considered in the following steps. Since deep belief network (Cao, et al., 2016), recurrent neural network (Cao, et al., 2017) and two-layer neural network (Cao, et al., 2017) have successful applied in many areas, deep learning is also an important machine learning technique.

The support vector machine (SVM) (Hearst, et al., 1998) approach is an advanced machine learning method that is effective for pattern recognition and data analysis. Many studies have also demonstrated that the SVM algorithm is a good foreground application for bioinformatics (Jia and He, 2016; Jia, et al., 2017; Jia, et al., 2013; Wei, et al., 2015). In our study, the LIBSVM package (Chang and Lin, 2011) was used to train the SVM and build a model that could discriminate between 4mC and non-4mC sites. The radial basis function (RBF) $K(S_i, S_j) = \exp(-\gamma \|S_i - S_j\|^2)$ was selected as the kernel function, and the penalty parameter $C$ and kernel parameter γ were optimized by the grid search method. For different input features, C and γ were optimized by grid search method on 15-fold cross-validation test. The search spaces for C and γ are $[2^{-2}, 2^5]$ and $[2^{-5}, 2^2]$. The final value selected for each model also supplied in Additional file 2: Table S2.

### 2.5 Performance measurement

To evaluate the predictive ability of our model, we used the jackknife test, which generates unique results for a given benchmark dataset (Chou and Shen, 2010). To measure performance, we additionally used four metrics whose advantages have been analyzed in a series of studies (Liu, et al., 2017; Qiu, et al., 2016; Qiu, et al., 2015).

$$Sn = 1 - \frac{N_-^+}{N^+}$$
$$Sp = 1 - \frac{N_+^-}{N^-} \quad (5)$$
$$Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}$$
$$MCC = \frac{1 - (\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-})}{\sqrt{(1 + \frac{N_+^- - N_-^+}{N^+})(1 + \frac{N_-^+ - N_+^-}{N^-})}}$$

In these expressions, $N^+$ and $N^-$ are the total number of 4mC and non-4mC samples for each species, respectively, while $N_-^+$ and $N_+^-$ are respectively the number of 4mC samples incorrectly predicted as non-4mC samples for each species, and vice versa.

## 3 Results and discussion

### 3.1 Analysis of sequence-level elements of conserved motif specificity

To uncover differences in nucleotide distributions around 4mC and non-4mC sites of the six species, we submitted benchmark training sets of these species to the pLogo web server (O'Shea, et al., 2013) (http://plogo.uconn.edu/). Sequence logos representing significantly overrepresented and underrepresented nucleotides ($p$ = 0.05) at each position in the sequences of the six species' benchmark datasets are displayed in Figure 2. As shown in this figure, the nucleotide distributions of each species clearly differed from one another. Comparison of 1,554 4mC and 1,554 non-4mC sample sequences of *C. elegans*, for example, revealed that guanine (G) at position +7, cytosine (C) at position +4 and thymine (T) at position +5 were significantly enriched. In *D. melanogaster*, G and adenine (A) nucleotides displayed significant enrichment ($p < 0.05$) in the region from positions +1 to +4, while G and C were overrepresented at positions –1 and –2, respectively. The positions of nucleotide enrichment in *A. thaliana* were similar to those in *D. melanogaster*, except that G and C were enriched at positions –2, –1, +7 and +8. In *E. coli*, A was significantly enriched ($p < 0.05$) in both upstream (positions –6 to –3) and downstream (+4, +5, +8 and +11) regions, with G, C and T also enriched in different positions. In *Geoalkalibacter subterraneus* and *Geobacter pickeringii*, A was significantly enriched at most upstream and downstream positions. In addition, T was enriched at some positions, such as –15, +3 and +9, of *Geobacter pickeringii*. The above results indicate that the position of a nucleotide in a sequence is a key predictor for discriminating between 4mC and non-4mC sites.

Because motif visualizations are statistical representations of the overall dataset, internal correlations are lost for each DNA sequence. Moreover, these motifs are not based on the whole species. If only the observed motif is used as the basis of judgment, many false positive results will be obtained. The elucidation of motif-related information for specific positions is therefore needed. The use of a machine learning-based method, an approach that has proven effective in many areas, is also required.

### 3.2 PSTNP

The PSTNP approach, which has been shown to be an effective feature-extraction method, was adopted to encode the enhancer samples of the benchmark dataset (He and Jia, 2017). For the PSTNP encoding, each 41-bp long sample sequence in the benchmark was coded as a 39-dimensional numerical vector. The jackknife test was applied to produce unique results for each species and examine the performance of the training model. To further assess the performance of the PSTNP approach, ROC curves of the six species were constructed based on the results of the jackknife test (Figure 3). According to our results, our PSTNP feature-extraction method is an effective approach for improving prediction performance.

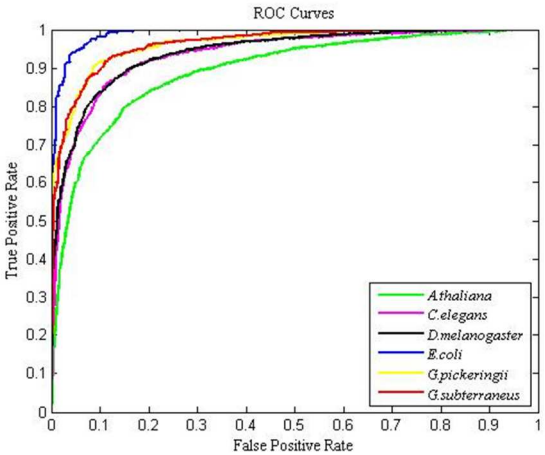



**Figure 3. ROC curves based on the jackknife test to assess the predictive performance of the PSTNP encoding scheme. AUC values of identified 4mC sites in**

***Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli*, *Geoalkalibacter subterraneus* and *Geobacter pickeringii* are 0.893, 0.936, 0.941, 0.963 and 0.963, respectively**

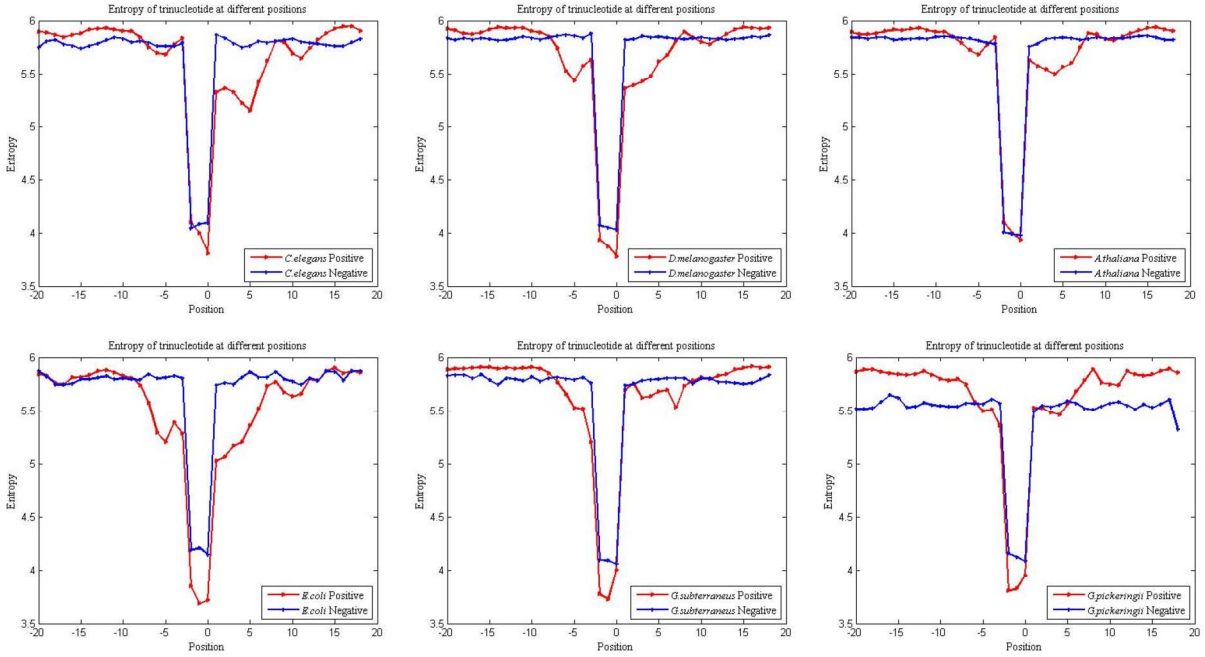



**Figure 4. The entropy of trinucleotides in 4mC and non-4mC samples**

### 3.2.1 Entropy analysis of specific trinucleotide positions

In this study, we conducted an entropy analysis (Fickett, 1996; Wu, et al., 2011) of specific trinucleotide positions in each species. The entropy of trinucleotides at each position can be calculated as:

$$E(i) = -\sum_{j=1}^{N} p(\text{trinucleotide}_j \mid i) \ln p(\text{trinucleotide}_j \mid i) \qquad (6)$$

where $p(\text{trinucleotide}_j \mid i)$ denotes the occurrence frequency of the $j$-th trinucleotide at the $i$-th position in the positive/negative training dataset and $N$ is the total number of trinucleotides (64). To avoid $p(\text{trinucleotide}_j \mid i)$ taking a value of zero, we then defined

$$\boldsymbol{p}'(\text{trinucleotide}_j \mid \boldsymbol{i}) = \frac{\boldsymbol{n}(\text{trinucleotide}_j \mid \boldsymbol{i}) + \boldsymbol{s}(\text{trinucleotide}_j \mid \boldsymbol{i})}{\boldsymbol{N}_i + \sum_j \boldsymbol{s}(\text{trinucleotide}_j \mid \boldsymbol{i})} \qquad (7)$$

as the new normalization of $p(\text{trinucleotide}_j \mid i)$ where $\boldsymbol{n}(\text{trinucleotide}_j \mid \boldsymbol{i})$ is the number of $\text{trinucleotide}_j$ occurrence at the $i$-th position and $\boldsymbol{N}_i$ denotes the total number of sample sequences in the corresponding dataset. $\boldsymbol{s}(\text{trinucleotide}_j \mid \boldsymbol{i})$ is the pseudo-count function, is given as:

$$\boldsymbol{s}(\text{trinucleotide}_j \mid \boldsymbol{i}) = \frac{\sqrt{\boldsymbol{N}_i}}{\boldsymbol{N}} \qquad (8)$$

For comparison, the entropy values of trinucleotides at different positions in 4mC samples and non-4mC samples of the six species are shown in Figure 4.

Entropy is obviously the measure of the conservatism of each position, with a low entropy value corresponding to higher conservation. As shown in Figure 4, 4mC and non-4mC samples had different entropy values at various positions around C sites.

To analyze differences in specific trinucleotide positions among species, we plotted all the entropy curves on one map. As shown in Additional file 2: Figure S1, PSTNPs of the six species were related but not identical, thus suggesting the possible complementary use of this method to study relationships between interacting species.

### 3.2.2 Independent testing of the PSTNP method

**Table 3. Results of independent tests of PSTNP and iDNA4mC proposed encoding schemes**

| Feature | Species | Sn (%) | Sp (%) | ACC (%) | MCC |
|---|---|---|---|---|---|
| PSTNP 4mCPred_I | *C.elegan* | **85.58** | **78.85** | **82.21** | **0.65** |
| | *D.melanogaster* | **83.9** | **81.36** | **82.63** | **0.65** |
| | *A.thaliana* | 76.52 | 76.52 | 76.52 | 0.53 |
| | *E.coli* | **84.62** | **80.77** | **82.69** | **0.65** |
| | *G.subterraneus* | **91.67** | **75.00** | **83.33** | **0.68** |
| | *G.pickeringii* | 86.84 | 68.42 | 77.63 | 0.56 |
| Chemiacal Properties (iDNA4mC) | *C.elegan* | 80.77 | 73.08 | 76.92 | 0.54 |
| | *D.melanogaster* | 74.58 | 77.97 | 76.27 | 0.53 |
| | *A.thaliana* | 80.30 | 77.27 | 78.79 | 0.58 |
| | *E.coli* | 96.15 | 69.23 | 82.69 | 0.68 |
| | *G.subterraneus* | 85.00 | 76.67 | 80.83 | 0.62 |
| | *G.pickeringii* | 81.58 | 78.95 | 80.26 | 0.61 |

When PSTNP is used to discriminate between 4mC and non-4mC sites, overestimation of the constructed training model due to information from

all of the training datasets is possible. To further evaluate the real performance of the model, we therefore performed independent testing. First, we randomly divided positive and negative training sets for each species' benchmark dataset into 15 subsets of approximately equal size. We then chose one of the 15 subsets to test a model for identifying 4mC sites in the six species that was trained using the remaining 14 subsets (Additional file 1: Datasets 2). We used the PSTNP encoding scheme and the iDNA4mC proposed encoding scheme to construct predictive models based on SVM using the optimal grid search parameters. The results of these tests are listed in Table 3. The PSTNP encoding scheme exhibited a good predictive performance, especially in *C. elegans*, *D. melanogaster*, *E. coli* and *Geoalkalibacter subterraneus*.

### 3.2.3 Cross-species validation

Noteworthily, the number of experimentally verified 4mC sites strongly depends on species category, and the PSTNPs of various species are related. We were consequently motivated to explore whether knowledge transfer information can be applied to study the relationships of interacting species. For this purpose, we used the PSTNP method to train six species-specific predictors on species-specific 4mC data and then tested the resulting models against the 4mC data of the other species. These models were based on a transfer learning approach, which means that knowledge from the source domain (species-specific training data) was applied in the target domain (the other species). The transferred knowledge in this study was the PSTNP profile from the source domain. The predictive performances of the six predictors are shown as a heat map in Figure 5. In the heat map, each square represents the accuracy obtained for a given cross species, with a lighter color corresponding to a higher precision (see Additional file 2: Table S3 for full details). According to these results, models developed using different species had different predictive performances. An important point, however, is that models developed using *D. melanogaster*, *A. thaliana* and *E. coli* performed best when tested against those same three species, while *Geoalkalibacter subterraneus* and *Geobacter pickeringii* performed similarly with each other. We also used the iDNA4mC proposed encoding scheme to construct cross-species models based on SVM with optimal grid search parameters (see Additional file 2: Figure S2 and Table S4 for full details).

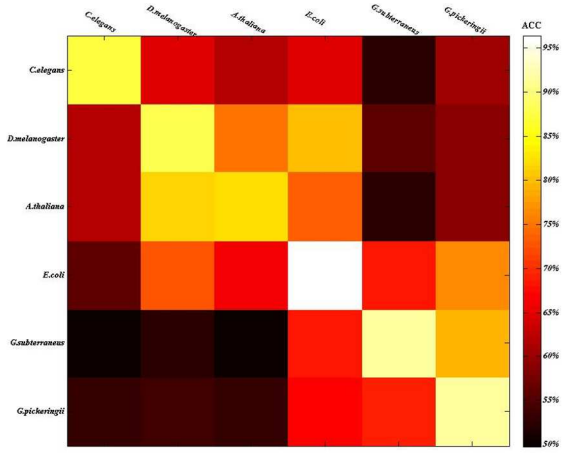


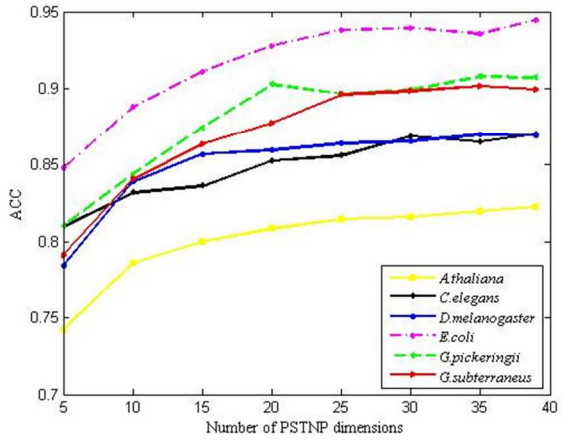

**Figure 5. Heap map showing cross-species prediction accuracies**

## 3.3 Improving predictive performance by incorporating PSTNP and EIIP

The construction of a comprehensive and proper feature space is an important step for the development of a powerful training model. To improve the prediction, we used the *F*-score metric to delete redundant features. A combination of hybrid features was generated from PSTNP and EIIP features on the basis of the *F*-score measurements.

First, we calculated 39 *F*-score values derived from PSTNP and sorted them in descending order. Second, we adopted an incremental feature selection approach to determine the optimal feature subsets. In this study, the feature subset initially comprised the top five optimal features as judged by the *F*-score, sorted in descending order. The five features with the next best *F*-scores were then added to the feature subset to generate a second feature subset. This process was repeated until all feature subsets were obtained. Third, we added EIIP features in the same way to the selected optimal subset of PSTNP features to find the best subset of all features. The predictive performances of PSTNP and hybrid features over different dimensions are shown in Figures 6A and 6B, respectively.

A.




B.

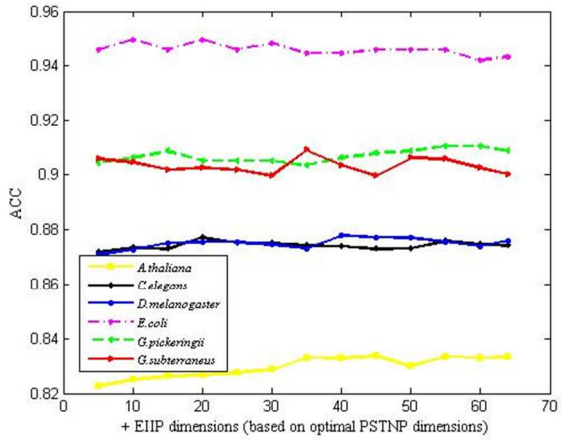



**Figure 6. Predictive performance of PSTNP (A) and hybrid (B) features over different dimensions**

In each step in this process, we chose the feature subset with the optimal ACC. As shown in Figure 6 (see Additional file 3 for full details), 39 features derived from PSTNP were retained in *A. thaliana*, *C. elegans*, *D. melanogaster* and *E. coli* compared with 35 features in *Geoalkalibacter subterraneus* and *Geobacter pickeringii*. When EIIP features were added, the predictive performance first increased and then decreased, and the best predictive performance was recorded for each

species. Detailed information can be found in Additional file 3. Heat maps of trinucleotide EIIP tendencies in positive samples are shown in Figure 7 for each species.

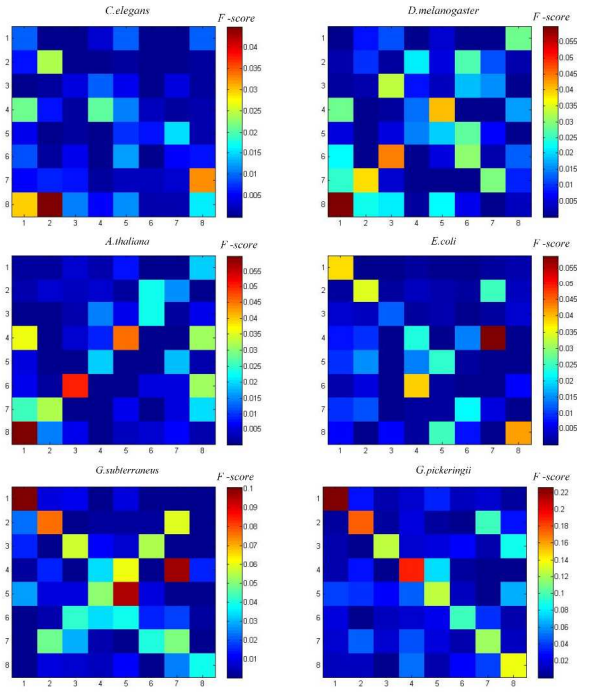



**Figure 7. A heat map for the *F-score* values of the 64 trinucleotides with EIIP values. Each square in a heat map corresponds to a unique trinucleotide, with trinucleotides arranged identically among maps (see Table S5 for full details). Deep red in the heat maps corresponds to the strongest recognition ability.**

### 3.4 Comparison with other methods

To further analyze the effectiveness of 4mCPred, we compared its two constituent models with iDNA4mC, a proposed model for predicting 4mC sites in *A. thaliana*, *C .elegans*, *D. melanogaster*, *E. coli*, *Geoalkalibacter subterraneus* and *Geobacter pickeringii*. As shown in Table 4, jackknife testing of these three predictors against the same benchmark dataset revealed that 4mCPred_II had the best performance. In addition, 4mCPred_I not only performed better than iDNA4mC in terms of prediction results, but also achieved an accurate recognition rate based on fewer features

## 4 Conclusions

In this study, we developed a new bioinformatics tool, 4mCPred, to recognize 4mC sites as a supplement to experimental approaches. Our 4mCPred tool, which extracts features based on PSTNP and EIIP values of trinucleotides and also exploits efficient feature selection, was shown to be robust and high performing according to jackknife testing. An entropy analysis indicated that trinucleotides surrounding 4mC sites had high PSTNPs. In addition, the PSTNPs of various species were related but not identical, thus suggesting their complimentary use to study relationships between interacting species. Physical chemical properties in the 4mC site environment were also analyzed. According to our jackknife evaluation, the 4mCPred_I model trained with PSTNP features had satisfactory results compared with an existing predictor, while the 4mCPred_II model trained with hybrid features achieved an even better performance. Remarkably, hybrid features conferred better performance on the latter model and may also have influenced the predictive ability of this model because of its higher dimensionality. We have provided both 4mCPred_I and 4mCPred_II, and users can apply either model. When 4mCPred_II is selected, we can get the best prediction performance; when 4mCPred_I is selected, we can better understand the performance of feature PSTNP within the result and carry out the interspecies detection. Since deep learning is an important supplementary to shallow sequence analysis, the future work may build a model combine with the two aspects. Our future work aims at extending this work to other bioinformatics sequence recognition.

**Table 4. Results of the comparison of iDNA4mC, 4mCPred_I and 4mCPred_II on the jackknife test**

| Method | Species (Dimension) | Sn(%) | Sp(%) | ACC(%) | MCC |
|---|---|---|---|---|---|
| iDNA4mC (Chemiacal Properties) | *C.elegan* (164) | 79.04 | 77.04 | 78.04 | 0.56 |
| | *D.melanogaster* (164) | 83.33 | 78.98 | 81.16 | 0.62 |
| | *A.thaliana* (164) | 76.55 | 75.54 | 76.05 | 0.52 |
| | *E.coli* (164) | 81.23 | 78.41 | 79.82 | 0.60 |
| | *G.subterraneus* (164) | 82.47 | 80.60 | 81.53 | 0.63 |
| | *G.pickeringii* (164) | 81.93 | 86.14 | 84.04 | 0.68 |
| 4mCPred_I (PSTNP) | *C.elegan* (39) | 87.13 | 86.87 | 87.00 | 0.74 |
| | *D.melanogaster (39)* | 86.94 | 86.94 | 86.94 | 0.74 |
| | *A.thaliana* (39) | 81.29 | 83.22 | 82.25 | 0.65 |
| | *E.coli* (39) | 95.62 | 93.30 | 94.46 | 0.89 |
| | *G.subterraneus* (39) | 89.94 | 89.96 | 89.95 | 0.80 |
| | *G.pickeringii* (39) | 89.98 | 91.39 | 90.69 | 0.81 |
| 4mCPred_II (F_PSTNP+ F_EIIP) | *C.elegan* (39+20) | 87.52 | 87.90 | 87.71 | 0.75 |
| | *D.melanogaster* (39+40) | 87.62 | 87.96 | 87.79 | 0.76 |
| | *A.thaliana* (39+45) | 82.96 | 83.77 | 83.37 | 0.67 |
| | *E.coli* (39+20) | 95.10 | 94.85 | 94.97 | 0.90 |
| | *G.subterraneus* (35+35) | 91.21 | 90.86 | 91.04 | 0.82 |
| | *G.pickeringii* (35+55) | 90.28 | 91.50 | 90.89 | 0.82 |

## Acknowledgments

The authors would like to thank the three anonymous reviewers for their constructive comments.

## Funding

The work was supported by the Natural Science Foundation of China (No. 61771331).

*Conflict of Interest:* none declared.

## References


Bestor, T.H. The DNA methyltransferases of mammals. *Hum Mol Genet* 2000;9(16):2395-2402.

Campbell*, et al.* E. coli oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. *Cell* 1990;62(5):967-979.

Cao, R.*, et al.* DeepQA: improving the estimation of single protein model quality with deep belief networks. *Bmc Bioinformatics* 2016;17(1):495.

Cao, R.Z.*, et al.* QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics* 2017;33(4):586-588.

Cao, R.Z.*, et al.* ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules* 2017;22(10).

Cao, R.Z.*, et al.* SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. *Bmc Bioinformatics* 2014;15.

Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2011;2(3):27.

Chen, W.*, et al.* iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 2016;7(13):16895-16909.

Chen, W., Lin, H. and Chou, K.-C. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Molecular BioSystems* 2015;11(10):2620-2634.

Chen, W.*, et al.* iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 2017;33(22):3518-3523.

Chen, X.X.*, et al.* Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. *BioMed research international* 2016;2016:1654623.

Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology* 2011;273(1):236-247.

Chou, K.-C. and Shen, H.-B. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Natural Science* 2010;2(10):1090.

Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 2001;43(3):246-255.

Collier, J., Mcadams, H.H. and Shapiro, L. A DNA methylation ratchet governs progression through a bacterial cell cycle. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104(43):17111-17116.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory* 1967;13(1):21-27.

Ehrlich, M.*, et al.* N4-methylcytosine as a minor base in bacterial DNA. *Journal of Bacteriology* 1987;169(3):939-943.

Feng, P.*, et al.* iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* 2018.

Fickett, J.W. Quantitative discrimination of MEF2 sites. *Mol Cell Biol* 1996;16(1):437-441.

Glickman, B.W. and Radman, M. Escherichia coli mutator mutants deficient in methylation-instructed DNA mismatch correction. *Proc Natl Acad Sci U S A* 1980;77(2):1063-1067.

Guo, S.H.*, et al.* iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* 2014;30(11):1522-1529.

Harrison, G.S. and Karrer, K.M. DNA synthesis, methylation and degradation during conjugation in Tetrahymena thermophila. *Nucleic Acids Research* 1985;13(1):73.

Hattman, S. DNA-[adenine] methylation in lower eukaryotes. *Biochemistry (Mosc)* 2005;70(5):550-558.

Hattman, S.*, et al.* Comparative study of DNA methylation in three unicellular eucaryotes. *Journal of Bacteriology* 1978;135(3):1156-1157.

He, W. and Jia, C. EnhancerPred2.0: predicting enhancers and their strength based on position-specific trinucleotide propensity and electron-ion interaction potential feature selection. *Mol Biosyst* 2017;13(4):767-774.

He, W.*, et al.* 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Systems Biology* 2018;12(4):44.

Hearst, M.A.*, et al.* Support vector machines. *IEEE Intelligent Systems and their applications* 1998;13(4):18-28.

Ho, T.K. Random decision forests. In, *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*. IEEE; 1995. p. 278-282.

Jia, C. and He, W. EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features. *Sci Rep* 2016;6:38741.

Jia, C., Yang, Q. and Zou, Q. NucPosPred: Predicting species-specific genomic nucleosome positioning via four different modes of general PseKNC. *J Theor Biol* 2018;450:15-21.

Jia, C.Z., He, W.Y. and Yao, Y.H. OH-PRED: prediction of protein hydroxylation sites by incorporating adapted normal distribution bi-profile Bayes feature extraction and physicochemical properties of amino acids. *J Biomol Struct Dyn* 2017;35(4):829-835.

Jia, C.Z., Liu, T. and Wang, Z.P. O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. *Mol Biosyst* 2013;9(11):2909-2913.

Khan, A. G-protein-coupled receptor prediction using pseudo-amino-acid composition and multiscale energy representation of different physiochemical properties. *Analytical biochemistry* 2011;412(2):173-182.

Li, W.C.*, et al.* iORI-PseKNC: A predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. *Chemometr Intell Lab* 2015;141:100-106.

Lin, H.*, et al.* iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic acids research* 2014;42(21):12961-12972.

Lin, H.*, et al.* Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM transactions on computational biology and bioinformatics* 2017.

Liu, B.*, et al.* Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res* 2015;43(W1):W65-71.

Liu, B., Wu, H. and Chou, K.-C. 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* 2017;9(04):67.

Liu, B.*, et al.* Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* 2014;30(4):472-479.

Liu, L.M., Xu, Y. and Chou, K.C. 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* 2017;13(6):552-559.

Lu, A.L., Clark, S. and Modrich, P. Methyl-directed repair of DNA base-pair mismatches in vitro. *Proc Natl Acad Sci U S A* 1983;80(15):4639-4643.

Lu, M.*, et al.* SeqA: A negative modulator of replication initiation in E. coli. *Cell* 1994;77(3):413-426.

Lyko, F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat Rev Genet* 2018;19(2):81-92.

Messer, W. and Noyer-Weidner, M. Timing and targeting: the biological functions of Dam methylation in E. coli. *Cell* 1988;54(6):735.

Nair, A.S. and Sreenadhan, S.P. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation* 2006;1(6):197-202.

O'Shea, J.P.*, et al.* pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods* 2013;10(12):1211-1212.

Ogden, G.B., Pratt, M.J. and Schaechter, M. The replicative origin of the E. coli chromosome binds to cell membranes only when hemimethylated. *Cell* 1988;54(1):127-135.

Pei Li, M.G., Chunyu Wang, Xiaoyan Liu, Quan Zou. An overview of SNP interactions in genome-wide association studies. *Briefings in Functional Genomics* 2015;14(2):143-155.

Pukkila, P.J.*, et al.* Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in Escherichia coli. *Genetics* 1983;104(4):571-582.

Qiu, W.R.*, et al.* iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics* 2016;32(20):3116-3123.

Qiu, W.R.*, et al.* iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J Biomol Struct Dyn* 2015;33(8):1731-1742.

Rish, I. An empirical study of the naive Bayes classifier. In, *IJCAI 2001 workshop on empirical methods in artificial intelligence*. IBM New York; 2001. p. 41-46.

Sahu, S.S. and Panda, G. Efficient localization of hot spots in proteins using a novel S-transform based filtering approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2011;8(5):1235-1246.

Sanchez-Romero, M.A., Cota, I. and Casadesus, J. DNA methylation in bacteria: from the methyl group to the methylome. *Curr Opin Microbiol* 2015;25:9-16.

Senawi, A., Wei, H.-L. and Billings, S.A. A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. *Pattern Recognition* 2017;67:47-61.

Tajima, S. and Suetake, I. Regulation and function of DNA methylation in vertebrates. *Journal of Biochemistry* 1998;123(6):993.

Tang, H., Chen, W. and Lin, H. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. *Molecular bioSystems* 2016;12(4):1269-1275.

Wang, Y.*, et al.* N 6 -methyladenine DNA modification in the unicellular eukaryotic organism Tetrahymena thermophila. *European Journal of Protistology* 2016;58:94.

Wei, H.-L. and Billings, S.A. Feature subset selection and ranking for data dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2007;29(1).

Wei, L.*, et al.* Enhanced Protein Fold Prediction Method through a Novel Feature Extraction Technique. *IEEE Transactions on Nanobioscience* 2015;14(6):649-659.

Wu, Q., Wang, J. and Yan, H. An Improved Position Weight Matrix method based on an entropy measure for the recognition of prokaryotic promoters. *Int J Data Min Bioinform* 2011;5(1):22-37.

Yang, H.*, et al.* Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition. *BioMed research international* 2016;2016:5413903.

Ye, P., *et al.* MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res* 2017;45(D1):D85-D89.
Zacharias, W. Methylation of cytosine influences the DNA structure. *EXS* 1993;64:27.
Zhang, C.J., *et al.* iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* 2016;7(43):69783-69793.