Bioinformatics, 33(22), 2017, 3518–3523
doi: 10.1093/bioinformatics/btx479
Advance Access Publication Date: 26 July 2017
Original Paper

OXFORD

Sequence analysis

iDNA4mC: identifying DNA N^4-methylcytosine sites based on nucleotide chemical properties

Wei Chen[1,*], Hui Yang[2], Pengmian Feng[3], Hui Ding[2] and Hao Lin[2,*]

[1]Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China, [2]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 and [3]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

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on May 4, 2017; revised on July 22, 2017; editorial decision on July 24, 2017; accepted on July 25, 2017

Abstract

Motivation: DNA N^4-methylcytosine (4mC) is an epigenetic modification. The knowledge about the distribution of 4mC is helpful for understanding its biological functions. Although experimental methods have been proposed to detect 4mC sites, they are expensive for performing genome-wide detections. Thus, it is necessary to develop computational methods for predicting 4mC sites.
Results: In this work, we developed **iDNA4mC**, the first webserver to identify 4mC sites, in which DNA sequences are encoded with both nucleotide chemical properties and nucleotide frequency. The predictive results of the rigorous jackknife test and cross species test demonstrated that the performance of **iDNA4mC** is quite promising and holds high potential to become a useful tool for identifying 4mC sites.
Availability and implementation: The user-friendly web-server, **iDNA4mC**, is freely accessible at http://lin.uestc.edu.cn/server/iDNA4mC.
Contact: chenweiimu@gmail.com or hlin@uestc.edu.cn

1 Introduction

5-Methylcytosine (5mC), N^6-methyladenine (6mA) and N^4-methylcytosine (4mC) are the three common DNA methylations (Davis *et al.*, 2013; Korlach and Turner, 2012; Roberts *et al.*, 2015) that have been detected in both prokaryotic and eukaryotic genomes (Blow *et al.*, 2016; Fu *et al.*, 2015; Greer *et al.*, 2015; Heyn and Esteller, 2015). These epigenetic modifications not only expand the genomic diversity, but also play profound roles in many biological processes. For example, 5mC often called the fifth base of DNA, is associated with differentiation, genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, aging and gene expression (Bergman *et al.*, 2003; Scarano *et al.*, 2005; Tao *et al.*, 2011). The biological consequences of 6mA include gene regulation, development, DNA replication, repair and expression (Fu *et al.*, 2015; Greer *et al.*, 2015; Zhang *et al.*, 2015).

Although 5mC and 6mA modifications are extensively studied, researches on 4mC have lagged behind due to the lack of effective detection methods. 4mC is catalyzed by the N-4 cytosine-specific DNA methyltransferase (DNMT) that specifically methylate the amino group at the C-4 position of cytosine in DNA (Timinskas *et al.*, 1995) as shown in Figure 1. Despite 4mC was found in 1983, our current knowledge about its biological functions is very limited. Similar to 5mC and 6mA, 4mC is also a member of the restriction modification (RM) systems and can protect the host DNA against degradation by restriction enzymes (Schweizer, 2008). In prokaryotes, 4mC can correct DNA replication errors and control DNA replication and cell cycle (Cheng, 1995; Messer and Noyer-Weidner, 1988; Modrich, 1991). However, other biological functions of 4mC are not fully elucidated. In order to reveal its new functions, it is necessary to develop various methods to identify 4mC sites.

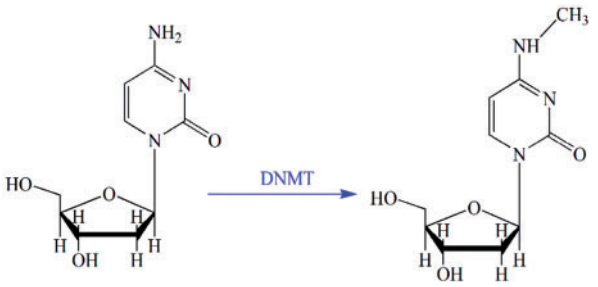

Fig. 1. An illustration to show the N^4-methylcytosine (4mC). The formation of 4mC is catalyzed by the N-4 cytosine-specific DNA methyltransferase (DNMT). In this modification, a methyl group (–CH3) is added to the nitrogen at the 4th position of the cytosine base

Towards this goal, single-molecule real-time (SMRT) sequencing technology has been developed to detect 4mC sites (Flusberg *et al.*, 2010). However, SMRT sequencing technology is cost-ineffective for the analysis of thousands of genomes that already exist in the public domain (Yu *et al.*, 2015). Recently, Yu *et al.* developed the 4mC-TAB-seq method that is able to accurately identify 4mC sites without the interference of 5mC (Yu *et al.*, 2015). These experimental methods indeed promote the studies on 4mC. However, biological experiments are still laborious and costly in performing genome-wide detections. Therefore, it is necessary to develop computational methods for identifying 4mC sites.

The high-resolution experimental data made it feasible to develop computational methods. Hence, in the present study, we proposed a support vector machine based-predictor to identify the 4mC sites, in which the DNA samples are formulated by using the nucleotide chemical property and nucleotide frequency. In the following sections, we will first describe how the predictor was established. The performance assessment and the establishment of the web-server of the predictor will be subsequently described.

2 Materials and methods

2.1 Benchmark datasets

The positive samples (4mC site containing sequences) for *Caenorhabditis elegans* (*C.elegans*), *Drosophila melanogaster* (*D.melanogaster*), *Arabidopsis thaliana* (*A.thaliana*), *Escherichia coli* (*E.coli*), *Geoalkalibacter subterraneus* (*G.subterraneus*) *and Geobacter pickeringii* (*G.pickeringi*) genomes were obtained from the MethSMRT database (Ye *et al.*, 2017). Preliminary tests indicated that the best predictive results were achieved when the sequence length is 41 bp. Thus, the sequences of the positive samples are all 41 bp. In order to construct a high quality benchmark dataset, the following two procedures were performed. Firstly, as illustrated in the Methylome Analysis Technical Note, the Modification QV (modQV) score indicates that the IPD ratio is significantly different from the expected background. Since a modQV score of 30 is the default threshold for calling a position as modified, the sequences with the modQV no less than 30 are left for the subsequent analysis. Secondly, as elaborated in previous study (Chou, 2015), a dataset containing many redundant samples with high similarity has the low statistical representativeness. A computational model, if trained and tested by such a biased benchmark dataset, might yield overestimated accuracy. To get rid of redundancy and minimize the bias, the CD-HIT software (Fu *et al.*, 2012) with the cutoff threshold set at 80% was used to remove those sequences with high sequence similarity. After the above two procedures, we obtained 1, 554, 1, 769, 1, 978, 388, 906 and 569 positive samples in *C.elegans, D.melanogaster, A.thaliana, E.coli, G.subterraneus* and *G.pickeringii* genomes, respectively.

Table 1. Cluster of nucleotides based on chemical properties

Chemical property	Class	Nucleotides
Ring structure	Purine	A, G
	Pyrimidine	C, T
Hydrogen bond	Strong	C, G
	Weak	A, T
Functional group	Amino	A, C
	Keto	G, T

The negative samples (non-4mC site containing sequences) for the six species mentioned above were obtained by choosing the 41-bp long sequences satisfying the requirement that the cytosine in the center was not detected by the SMRT sequencing technology. By doing so, we could obtain a large number of negative samples in each species. Therefore, the number of negative samples will be dramatically larger than those of positive samples. To balance out the numbers between positive and negative samples in model training, the same number of negative samples was randomly picked out from the *C.elegans, D.melanogaster, A.thaliana, E.coli, G.subterraneus* and *G.pickeringii* genomes, respectively. Finally, we obtained six benchmark datasets as formulated by

$$S_k = S_k^+ \cup S_k^- \tag{1}$$

the subsets S_k^+ ($k = 1, 2, 3, \ldots, 6$) contains 1, 554, 1, 769, 1, 978, 388, 906 and 569 true 4mC site containing sequences, while the subsets S_k^- ($k = 1, 2, 3, \ldots, 6$) contains 1, 554, 1, 769, 1, 978, 388, 906 and 569 non-4mC site containing sequences in *C. elegans, D. melanogaster, A.thaliana, E. coli, G. subterraneus* and *G. pickeringii* genomes, respectively.

2.2 Nucleotide chemical property

The deoxyribonucleic acid is composed of four nucleic acids, namely Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The four nucleic acids have different chemical properties (Golam Bari *et al.*, 2014). In terms of ring structures, A and G are purines containing two rings, whereas C and T are pyrimidines containing one ring. When forming secondary structures, C and G form strong hydrogen bonds, whereas A and T form weak hydrogen bonds. In terms of chemical functionality, A and C can be classified into the amino group, while G and T can be classified into the keto group. Accordingly, the four nucleic acids can be classified into three different groups (Table 1).

In order to include these properties, three coordinates (x, y, z) were used to represent the chemical properties of the four nucleotides and the value of 0 and 1 was assigned to the coordinates. If x, y and z coordinates respectively stand for the ring structure, the hydrogen bond and the chemical functionality, each nucleotide can be encoded by (x_i, y_i, z_i), where

$$x_i = \begin{cases} 1 \text{ if } s_i \in \{A, G\} \\ 0 \text{ if } s_i \in \{C, T\} \end{cases}, y_i = \begin{cases} 1 \text{ if } s_i \in \{A, T\} \\ 0 \text{ if } s_i \in \{C, G\} \end{cases}, z_i = \begin{cases} 1 \text{ if } s_i \in \{A, C\} \\ 0 \text{ if } s_i \in \{G, T\} \end{cases} \tag{2}$$

Accordingly, A, C, G and T can be represented by the coordinates (1, 1, 1), (0, 0, 1), (1, 0, 0) and (0, 1, 0), respectively.

2.3 Nucleotide frequency

For the purpose of including nucleotide composition surrounding the 4mC sites as well, the density d_i of a nucleotide n_j at position i was defined as follows:

$$d_i = \frac{1}{|N_i|}\sum_{j=1}^{l} f(n_j), f(n_j) = \begin{cases} 1 \ if \ n_j = q \\ 0 \ other \ cases \end{cases} \quad (3)$$

where l is the sequence length, $|N_i|$ is the length of the ith prefix string $\{n_1, n_2, \ldots, n_i\}$ in the sequence, $q \in \{A, C, G, T\}$.

By integrating nucleotide chemical properties and nucleotide frequency, an l-bp long sequence will be encoded by a $(4 \times l)$-dimensional vector. For example, the DNA sequence 'CACGTC' will be represented as {{0, 0, 1, 1}, {1, 1, 1, 0.5},{0, 0, 1, 0.5}, {(1, 0, 0, 0.25)}, {0, 1, 0, 0.2}{0, 0, 1, 0.5}}.

2.4 Support vector machine

Support vector machine (SVM) is a powerful and popular method for pattern recognition and has been widely used in computational genomics (Chen *et al.*, 2015, 2016a,b, 2017; Lin *et al.*, 2014; Yang *et al.*, 2014). Its basic idea is to transform the input data into a high dimensional feature space and then determine the optimal separating hyperplane. In the current study, the implementation of SVM was carried out by using the LibSVM package 3.18, which is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The radial basis kernel function (RBF) was used to obtain the classification hyperplane. In the SVM operation engine, the grid search method was applied to optimize the regularization parameter C and kernel parameter γ using a grid search approach defined as

$$\begin{cases} 2^{-5} \le C \le 2^{15} & \text{with step of } 2 \\ 2^{-15} \le \gamma \le 2^{-5} & \text{with step of } 2^{-1} \end{cases} \quad (4)$$

The probability score obtained from SVM was used to make predictions. If the probability score is greater than 0.5, a cytosine will be predicted as a 4mC, otherwise, non-4mC.

2.5 Performance evaluation

The performance of the proposed method was evaluated by using the following four metrics, namely sensitivity (*Sn*), specificity (*Sp*), accuracy (*Acc*) and the Mathew's correlation coefficient (*MCC*), which are expressed as

$$\begin{cases} Sn = \dfrac{TP}{TP + FN} \times 100\% \\ Sp = \dfrac{TN}{TN + FP} \times 100\% \\ Acc = \dfrac{TP + TN}{TP + FN + TN + FP} \times 100\% \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \end{cases} \quad (5)$$

where *TP*, *TN*, *FP* and *FN* represent true positive, true negative, false positive and false negative, respectively.

The ROC (receiver operating characteristic) curve was also used to evaluate the performance of the current method (Hanley and McNeil, 1982). Its vertical coordinate indicates the true positive rate (sensitivity) and the horizontal coordinate indicates the false positive rate (1-specificity). The area under the ROC curve (auROC) is an indicator of the performance quality of a binary classifier, i.e. the value 0.5 of auROC is equivalent to random prediction while 1 of auROC represents a perfect one.

Table 2. Jackknife test results of the method for identifying 4mC sites in different species

Species	Sn (%)	Sp (%)	Acc (%)	MCC
C.elegans	79.04	77.04	78.04	0.56
D.melanogaster	83.33	78.98	81.16	0.62
A.thaliana	76.55	75.54	76.05	0.52
E.coli	81.23	78.41	79.82	0.60
G.subterraneus	82.47	80.60	81.53	0.63
G.pickeringii	81.93	86.14	84.04	0.68

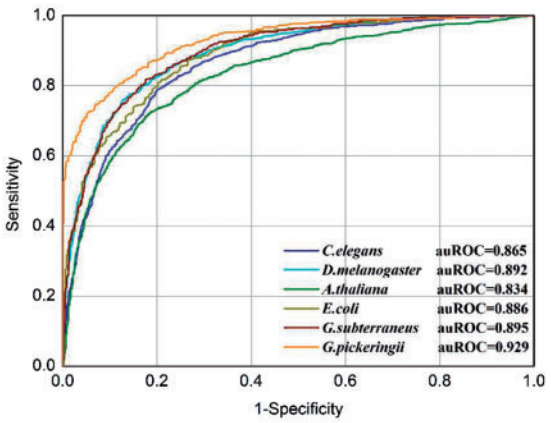

Fig. 2. A graphical illustration to show the performance of the model by means of the ROC curves obtained from the jackknife test. The vertical coordinate is the true positive rate (*Sn*) while the horizontal coordinate is the false positive rate (1-*Sp*)

3 Results and discussion

3.1 Identification of 4mC sites

The jackknife test (Chou, 2011) was used to examine the performance of the proposed method. In the jackknife test, each sample in the training dataset is in turn singled out as an independent test sample and all the properties are calculated without including the one being identified.

By encoding DNA sequence with nucleotide chemical property and nucleotide frequency, each 41-bp long sequence in the dataset was represented by a (4×41)=164-dimensional vector and used as the input data of SVM to perform predictions. The jackknife test results for identifying 4mC sites in each species were listed in Table 2. The accuracies of identifying 4mC sites in *C.elegans, D.melanogaster, A.thaliana, E.coli, G.subterraneus* and *G.pickeringii* are 78.04%, 81.16%, 76.05%, 79.82%, 81.53% and 84.04%, respectively. The corresponding *Sns*, *Sps* and *MCC* were also reported in Table 2.

To further demonstrate the performance of the proposed method, the ROC curves from the jackknife test were plotted as shown in Figure 2. The auROCs for identifying 4mC sites in *C. elegans, D. melanogaster, A. thaliana, E. coli, G. subterraneus* and *G. pickeringii* are respectively 0.865, 0.892, 0.834, 0.886, 0.895 and 0.929 (Fig. 2). All results demonstrate that our proposed models are powerful for identifying 4mC sites.

In order to investigate the contributions of each feature for identifying 4mC sites, we established a series of predictors with a single

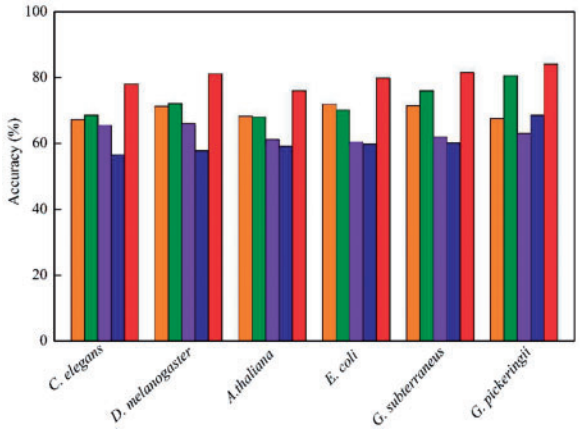

Fig. 3. The predictive accuracies obtained from the jackknife test for identifying 4mC sites in *C.elegans, D.melanogaster, A.thaliana, E.coli, G.subterraneus* and *G.pickeringii* genomes by using different kinds of parameters. Orange, green, purple, blue and red histograms stand for the accuracies obtained by the model trained using the ring structure, hydrogen bond, functional group, nucleotide frequency and the combination of all the four kinds of features, respectively

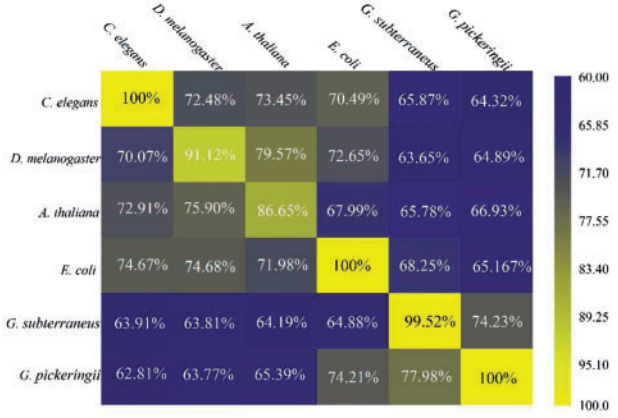

Fig. 4. The heat map showing the cross species prediction accuracies. Once a species-specific model was established on its own training dataset, it was tested on the data from the other six species

kind of feature (i.e. ring structure, hydrogen bond, chemical functionality or nucleotide frequency) and evaluated them on the benchmark dataset of each species. The predictive accuracies based on the jackknife test were shown in Figure 3.

It was found that among the four kinds of features, the predictors based on the ring structure or hydrogen bond yielded the highest accuracy for identifying 4mC sites in all the six species. However, their predictive accuracies are all lower than those obtained by using the combination of the four kinds of features. This result indicated that nucleotide chemical properties make the greatest contributions for 4mC site identification, whereas the nucleotide frequency only contributes slightly to 4mC site identification and plays complementary role in the prediction.

3.2 Cross species validation

To the best of our knowledge, the computational method for identifying 4mC sites has not been reported so far. Therefore, we could not provide the comparison analysis with previous results or confirm whether the performance of the current method is superior to other

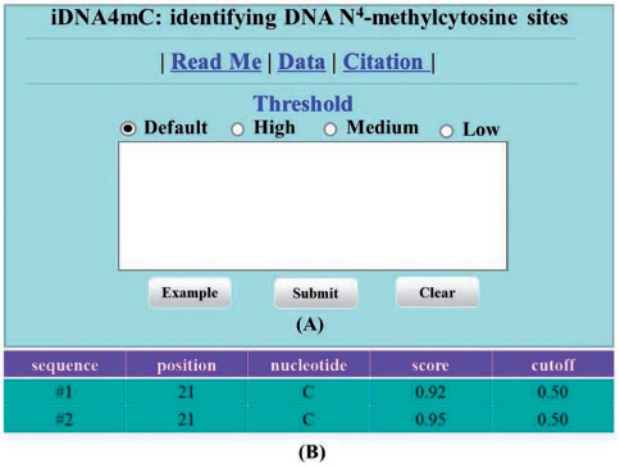

Fig. 5. (**A**) A semi-screenshot for the top-page of the **iDNA4mC** web-server. (**B**) The prediction result of the two example sequences. The information on modified position, modified nucleotide, prediction score, prediction cutoff are presented

methods. Since the training datasets of the proposed method were collected from different species, it is necessary to demonstrate whether a model trained with the data from one species could recognize the 4mC sites in other species. To demonstrate this point, we trained six species-specific models using the species-specific 4mC data and validated these models on the independent 4mC data of other species. The predictive accuracies thus obtained were shown in Figure 4.

We found that, except the models trained on the data from *G.subterraneus* and *G.pickeringii*, the models trained on the data from the other four species could identify the 4mC sites from the other species with the accuracies of more than 70%. Although the predictive accuracies of the *G.subterraneus* and *G.pickeringii*-specific models are slightly lower than those of the other four models for identifying 4mC sites in *C.elegans*, *D.melanogaster*, *A.thaliana* and *E.coli*, their accuracies are higher than those of other models for identifying 4mC sites in *G.subterraneus* and *G.pickeringii*.

3.3 Web-server

Based on the six benchmark datasets defined in Eq. 1, a predictor called **iDNA4mC** was established, where 'i' stands for 'identify' and '4mC' stands for 'N^4-methylcytosine'.For the conveniences of scientific community, a freely accessible online web-server was established for **iDNA4mC**, which is available at http://lin.uestc.edu.cn/server/iDNA4mC. The top page of **iDNA4mC** is shown in Figure 5A.

Inspired by a recent work (Zhao *et al.*, 2014), besides the default prediction threshold (the score of 0.5 obtained from SVM), three thresholds of high (0.97), medium (0.96) and low (0.90) stringency with the specificity values of 95%, 90% and 85% were also provided in the web-server. The step-by-step guide on the web-server is provided as follows:

First, either paste or type the query DNA sequences into the input box. The input sequence should be in the FASTA format that can be seen by clicking on the **Example** button.

Second, clicking the open circles in the **Threshold** module, the threshold concerned will be selected.

Third, click on the **Submit** button to see the predicted result. For example, if the two query sequences in the **Example** window were used as the input, and selecting the default threshold, after clicking the **Submit** button, users will obtain the results as shown in Figure 5B.

4 Conclusions

In this work, a predictor called **iDNA4mC** was proposed to identify 4mC sites in both prokaryotes and eukaryotes. In order to demonstrate its performance, the **iDNA4mC** was evaluated by using the rigorous jackknife test. The sensitivities, specificities, accuracies and Mathew's correlation coefficient derived from the jackknife test demonstrate that **iDNA4mC** is effective for identifying 4mC sites. The **iDNA4mC** was further evaluated by performing cross species validations. It is encouraging to see that the accuracies of cross species validations are also quite good. These results indicate that **iDNA4mC** is useful for identifying 4mC sites.

In **iDNA4mC**, not only the sequence-based information was considered by encoding DNA sequences using nucleotide frequency, but also three kinds of nucleotide chemical properties (ring structure, hydrogen bond and chemical functionality) were incorporated. Comparisons of the accuracies obtained from the models based on different features demonstrate that the ring structure and hydrogen bond have the largest contributions for 4mC site identification in both prokaryotes and eukaryotes. This could be interpreted as follows. DNMTs are composed of two domains, one large domain containing the binding sites for the cofactor S-adenosyl-L-methionine (SAM) and the flipped base, and one smaller domain that participates in DNA binding and recognition (Gong *et al.*, 1997; Jeltsch, 2001). In the DNA methylation reaction, the target nucleotide is flipped out of the DNA helix (Cheng and Blumenthal, 1996; Roberts and Cheng, 1998) and is bound into a hydrophobic binding pocket within the catalytic domain of the DNMT where the methyl group transfer takes place (Klimasauskas *et al.*, 1994). Hence, the nucleotide chemical properties especially ring structure and hydrogen bond might facilitate the structural change of DNA helix and the interactions between DNMTs and DNA sequence.

As an epigenetic modification, DNA methylation is a complicate progress. Besides nucleotide chemical property and nucleotide frequency, similar to 5mC, 4mC might be also affected by other factors, such as transcription factors and histone modifications. Therefore, we will incorporate such information to improve the performance of the predictor in the future work.

Acknowledgements

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

Funding

This work was supported by the Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244), 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), the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015Z006; ZYGX2016J118; ZYGX2016J125; ZYGX2016J223).

Conflict of Interest: none declared.

References

Bergman, Y. *et al.* (2003) Epigenetic mechanisms that regulate antigen receptor gene expression. *Curr. Opin. Immunol.*, **15**, 176–181.

Blow, M.J. *et al.* (2016) The epigenomic landscape of prokaryotes. *PLoS Genet.*, **12**, e1005854.

Chen, W. *et al.* (2015) iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*, **490**, 26–33.

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

Chen, W. *et al.* (2016b) iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids Mol. Ther. Nucleic Acids*, **5**, e332.

Chen, W. *et al.* (2017) iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget*, **8**, 4208–4217.

Cheng, X. (1995) DNA modification by methyltransferases. *Curr. Opin. Struct. Biol.*, **5**, 4–10.

Cheng, X. and Blumenthal, R.M. (1996) Finding a basis for flipping bases. *Structure*, **4**, 639–645.

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

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

Davis, B.M. *et al.* (2013) Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr. Opin. Microbiol.*, **16**, 192–198.

Flusberg, B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.

Fu, L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Fu, Y. *et al.* (2015) N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell*, **161**, 879–892.

Golam Bari, A.T.M. *et al.* (2014) DNA encoding for splice site prediction in large DNA sequence. *MATCH Commun. Math. Comput. Chem.*, **71**, 241–258.

Gong, W. *et al.* (1997) Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. *Nucleic Acids Res.*, **25**, 2702–2715.

Greer, E.L. *et al.* (2015) DNA methylation on N6-adenine in *C. elegans*. *Cell*, **161**, 868–878.

Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Heyn, H. and Esteller, M. (2015) An Adenine code for DNA: a second life for N6-methyladenine. *Cell*, **161**, 710–713.

Jeltsch, A. (2001) The cytosine N4-methyltransferase M.PvuII also modifies adenine residues. *Biol. Chem.*, **382**, 707–710.

Klimasauskas, S. *et al.* (1994) HhaI methyltransferase flips its target base out of the DNA helix. *Cell*, **76**, 357–369.

Korlach, J. and Turner, S.W. (2012) Going beyond five bases in DNA sequencing. *Curr. Opin. Struct. Biol.*, **22**, 251–261.

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

Messer, W. and Noyer-Weidner, M. (1988) Timing and targeting: the biological functions of Dam methylation in *E. coli*. *Cell*, **54**, 735–737.

Modrich, P. (1991) Mechanisms and biological effects of mismatch repair. *Annu. Rev. Genet.*, **25**, 229–253.

Roberts, R.J. and Cheng, X. (1998) Base flipping. *Annu. Rev. Biochem.*, **67**, 181–198.

Roberts, R.J. *et al.* (2015) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **43**, D298–D299.

Scarano, M.I. *et al.* (2005) DNA methylation 40 years later: its role in human health and disease. *J. Cell. Physiol.*, **204**, 21–35.

Schweizer, H. (2008) Bacterial genetics: past achievements, present state of the field, and future challenges. *BioTechniques*, **44**, 633–634, 636-641.

Tao, Y. *et al.* (2011) Lsh, chromatin remodeling family member, modulates genome-wide cytosine methylation patterns at nonrepeat sequences. *Proc. Natl. Acad. Sci. USA*, **108**, 5626–5631.

Timinskas, A. *et al.* (1995) Sequence motifs characteristic for DNA [cytosine-N4] and DNA [adenine-N6] methyltransferases. Classification of all DNA methyltransferases. *Gene*, **157**, 3–11.

Yang, L. *et al.* (2014) Analysis and identification of toxin targets by topological properties in protein–protein interaction network. *J. Theor. Biol.*, **349**, 82–91.

Ye, P. *et al*. (2017) MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res*., **45**, D85–D89.

Yu, M. *et al*. (2015) Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite- sequencing. *Nucleic Acids Res*., **43**, e148.

Zhang, G. *et al*. (2015) N6-methyladenine DNA modification in Drosophila. *Cell*, **161**, 893–906.

Zhao, Q. *et al*. (2014) GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res*., **42**, W325–W330.