Chemometrics and Intelligent Laboratory Systems 206 (2020) 104151

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A deep learning-based computational approach for discrimination of DNA N^6-methyladenosine sites by fusing heterogeneous features

Muhammad Tahir [a,b], Maqsood Hayat [a,*], Imran Ullah [d], Kil To Chong [b,c,**]

[a] Department of Computer Science, Abdul Wali Khan University, Mardan, 23200, KP, Pakistan
[b] Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, 54896, South Korea
[c] Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju, 54896, South Korea
[d] Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, 45320, Pakistan

ARTICLE INFO

Keywords:
N^6-methyladenine
DNA
Convolution neural network
One-hot encoding
Sequence analysis

ABSTRACT

N^6-methyladenine is post-replication modifications, which take place in the extensive range of DNA sequences and involved with a large number of different bioprocesses such as DNA repair, replication, cellular defense, and transcription in prokaryotes. Recently, various computational models were established to predict N^6-methyladenine sites within DNAs. However, one of the main issues in the precise prediction of N^6-methyladenine is the extraction of those features, which clearly define the characteristics of N^6-methyladenine sites. In this method, input sequences of DNA are expressed by one-hot representation in order to allow progressive convolution layers. To exhibit the hidden information from the recognized sequences, the convolution neural network (CNN) model is applied to automatically learn the abstract features. Then, we apply the tri-nucleotide Composition (TNC) feature extraction technique and concatenate with CNN features. Our proposed model achieved 98.05% accuracy for the S_1 benchmark dataset and 89.22% accuracy for the S_2 benchmark dataset. The classification rates demonstrated that the developed approach performed better compared to existing approaches in terms of all the evaluation measures. It is expected that the developed intelligent approach might be played a leading and progressive role for academia as well as industrial research in the area of genomics prediction. The code cv is attached here.

Author contribution

Muhammad TahirMaqsood HayatImran ullahKil To Chong

1. Introduction

The dynamic DNA epigenetic modification of N^6-methyladenine (6mA) has been reported in bacteria, archaea, and eukaryotes [1]. The addition of a methyl group to the sixth position of the purine ring of the adenine is catalyzed by DNA-adenine methyltransferase; however, the demethylation is catalyzed by demethylase [2,3]. The DNA 6mA demethylase was initially reported in Drosophila which belongs to the TET protein family. Currently, the members of the AlkB family namely: NMAD-1 and ALKBH1 were found in the demethylation of 6mA in the DNA of *C. elegans* and mammals, respectively [4]. The 6mA sites have been linked with many bioprocess i.e. DNA replication [5], repair [6], cellular defense [7–9], and transcription [10] in prokaryotes. The comprehension of 6mA roles in eukaryotes is still vague. Therefore, an in-depth study is required to explore the genomic location of 6 mA which will provide an understanding of the role of 6 mA in eukaryotes [3, 11–13].

During the previous decades, a series of laboratory approaches i.e., capillary electrophoresis through laser-induced fluorescence [14], ultra-performance liquid chromatography with mass spectrometry [15], real-time sequencing of single-molecule and methylated [16] was applied. DNA immune-precipitation sequencing [13] was developed to detect 6 mA in both eukaryotes and prokaryotes. Although, these laboratory techniques assist in the genome-wide detection of 6 mA sites yet they are reported very costly and laborious. With the fast and exponential growth of sequenced genomes, therefore, it is urgently required to develop accurate, automated, and low-cost methods to detect 6mA sites. In this regard, Chen et al. introduced a sequential-based approach known as iRNA-Methyl to predict m^6A site [17]. In this model, three RNA physiochemical properties of pseudo dinucleotide composition (PseDNC)

* Corresponding author. Department of Computer Science, Abdul Wali Khan University Mardan, 23200, KP, Pakistan.
** Corresponding author. Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, 54896, South Korea.
E-mail address: m.hayat@awkum.edu.pk (M. Hayat).

https://doi.org/10.1016/j.chemolab.2020.104151
Received 9 May 2020; Received in revised form 31 August 2020; Accepted 1 September 2020
Available online 4 September 2020
0169-7439/

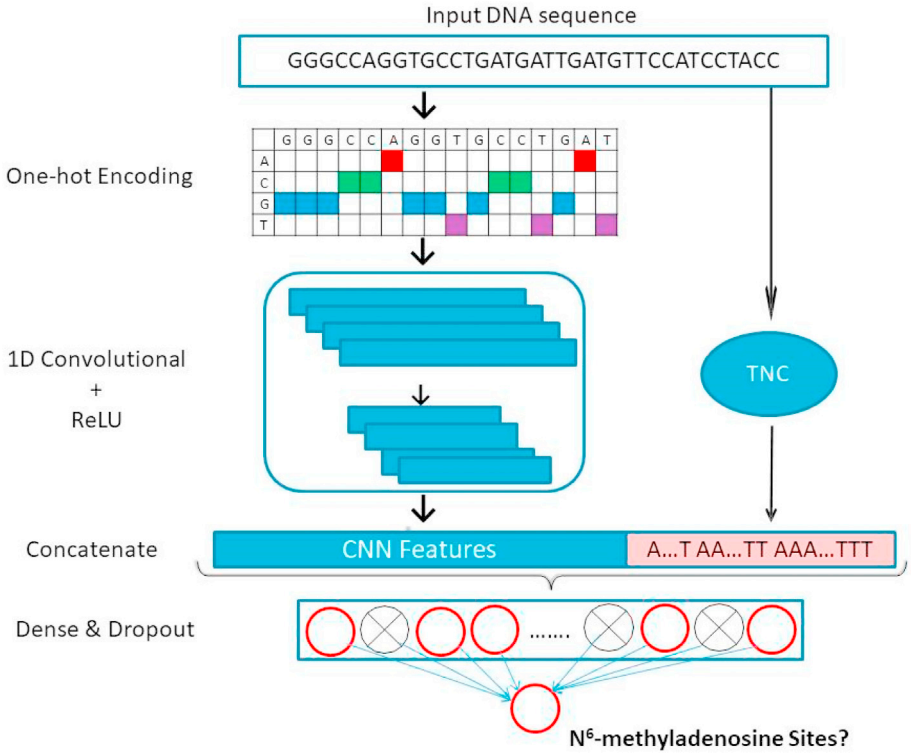

Fig. 1. The Architecture of the developed intelligent computational approach.

were used to formulate RNA samples [18,19]. Similarly, Zhou et al. introduced an intelligent sequence-based method named SRAMP, which incorporates three feature encoding methods: nucleotide pair spectrum encoding, k-nearest neighbor encoding, and positional binary encoding [20]. Similarly, Xiang et al. introduced an intelligent method namely: AthMethPre, which utilized the features of the position-independent k-mer nucleotide spectrum and positional flanking nucleotide sequence [21]. In a sequel, Lv et al., proposed a computational model to predict 4mC, 5hmC, and 6mA sites. In this model, various encoding schemes namely: binary encoding scheme, K-tuple nucleotide component, nucleotide frequency, mono-nucleotide, and nucleotide chemical property were utilized to express DNA examples, where the random forest was used for classification algorithm [22]. Likewise, Tang et al. developed a computational model namely: DNA4mC-LIP, in this method, initially systematically calculated the state-of-the-arts computational model for identification of 4mC in multiple species on independent datasets and then combined with a preliminary optimal weight [23]. In addition, most recently a computational model was established by Feng et al. namely: iDNA6mA-PseKNC to identify 6mA sites in DNA sequences using a support vector machine (SVM) algorithm in combination with the PseKNC feature vector [24].

Similarly, Chen et al. have proposed a sequence-based predictor (i6mA-Pred) for discrimination of 6mA sites in the rice genome. They have used SVM classifier and two various feature extraction techniques i.e., nucleotide frequency and nucleotide chemical [25]. In addition, most recently, Tahir et al. proposed an intelligent convolution neural network-based predictor to improve the performance of the existing model i6mA-Pred for the identification of 6 mA sites in the rice genome [26].

The existing methods and models are mostly based on conventional machine learning approaches, and obtained considerable results but still have some space of improvement in terms of evaluation metrics. Some researchers [27–37] have also employed machine learning approaches for various sequencing datasets. The major issues reported in hand-design feature demonstration techniques are the reflection of inadequate information that does not clearly discern the motif of N^6-methyladenine and non- N^6-methyladenine sites. There needs a system that not only has the ability to learn automatically the pattern of N^6-methyladenine sites from sequences but also predicts the target class in high precision. The desired notion can be accomplished by using deep learning in order to learn the motif of the target class automatically from the provided data at multi-levels of abstraction. In speech recognition [38], cancer classification, information retrieval [39], natural language processing [24], image recognition [40], etc, the deep learning-based approach generated very successful and accurate outcomes.

In the present study, we propose an efficient and deep learning-based computational model to predict 6 mA sites. Firstly, we use the one-hot representation for the input sequence to allow the successive convolution layers. To discern the hidden information from the recognized sequences, the convolution neural network (CNN) model is utilized in order to automatically learn the abstract features. Then, we apply handcrafted features i.e., tri-nucleotide composition (TNC) feature extraction technique and concatenate with CNN features. As shown in the result and discussion, our proposed computational model yields remarkable performance than existing methods on both benchmark datasets, respectively.

2. Materials and methods

2.1. Datasets

Two benchmark datasets are selected and downloaded of N^6-methyladenosine sites from Feng's [24] and Chen's [25] works in order to compare the developed computational model with existing computational models. These datasets are mathematically expressed as:

$$\begin{aligned} S_1 &= S_1^+ \cup S_1^- \\ S_2 &= S_2^+ \cup S_2^- \end{aligned} \tag{1}$$

where S_1^+ represent the positive subset consists of 1934 number of sequences of N^6-methyladenosine sites while S_1^- representing the negative subset contains 1934 number sequences of non- N^6-methyladenosine

Table 1
Tuned Hyper-parameters of the CNN model.

Hyper-Parameter	Range
Convolution Layers	[1,1,2], [1–3]
Dropout	[0.25,0.3, 0.35, 0.4, 0.5,0.75]
Filter size	[3,5,7,9,11,13]
Filters	[4,8,16,32]

Table 2
The architecture of the proposed computational model.

Layers	Output Shape
Input-1	(41,4)
Conv1D (16,9,1)	(41,16)
Conv1D (8,7,1)	(41,8)
Dropout (0.35)	328
Input-2	(64)
Concatenate (Dropout output and Input-2)	388
Dense (1)	1

Table 3
Prediction results of the proposed computational model on S_1 and S_2 datasets.

Datasets	Accuracy	Sensitivity	Specificity	MCC	AUROC
S_1	98.05	95.88	100.00	0.961	0.982
S_2	89.22	88.03	90.52	0.784	0.939

sites. Similarly, S_2^+ represents the positive subset contains 880 numbers of sequences of 6 mA sites while S_2^- representing the negative subset consist of 880 non-6mA sites. The size of each sequence is 41-bp long residues with the N^6-methyladenosine site "A" located at the center.

2.2. Tri-nucleotide composition (TNC)

TNC is a feature-encoding scheme that represents genomics samples by forming a pair of three consecutive nucleotides and computes the occurrence of each pair. For example, in RNA sequence, the first pair is $N_1N_2N_3$, the second pair is $N_2N_3N_4$, and so forth, subsequently, $4 \times 4 \times 4 = 64$D features vector is formed [41,42]. TNC can be numerically defined as:

$$S = [f(AAA), f(AAC), f(AAG), f(AAT), ..., f(TTT)]^T \tag{2}$$

$$S = \left[f_1^{3-tuple} f_2^{3-tuple} f_3^{3-tuple} f_4^{3-tuple} \ldots f_{64}^{3-tuple}\right]^T \tag{3}$$

where $f_1^{3-tuple} = f(AAA)$ is the frequency of *AAA*, $f_2^{3-tuple} = f(AAC)$ is the *AAC* in DNA sequences; and so forth.

2.3. The proposed method

In this study, deep learning i.e., convolution neural networks model has been utilized for discrimination of N^6-methyladenosine sites. CNN is generally used for image processing classification tasks and also useful in vision-related applications. Furthermore, they played quite a remarkable role as well in the sequential data processing. In this regard, our proposed CNN model captures a DNA sequence $R = \{R_1R_2R_3.....R_n\}$as an input, where n = 41 for S_1 and S_2 datasets, $R_i \in \{A, C, G, T\}$ and produces output. The DNA nucleotides are commonly required to be encoded numerically through the mostly used one-hot encoding technique before feeding to a CNN model and displayed as a vector with four channels i.e., A,G,C,T. The length of the vector depends on the value of *n*. In one-hot encoding approach, they are usually represented as A= (1, 0, 0, 0), G = (0, 0, 1, 0), C = (0, 1, 0, 0), and T = (0, 0, 0, 1). Fig. 1 demonstrates the architecture of the developed computational model.

Generally, in the deep learning network, a one-step process is known as a layer, which can be a pooling layer, a dropout layer, a loss layer, a convolution layer, a rectified linear unit (ReLU) layer, a normalization layer, a fully connected layer, etc. At the time of training different hyper-parameters are tuning i.e., the number of convolution layers, number of filters and its size, and finally dropout probability then convolution and dense layers. Table 1 presents an overview of these hyperparameters values. In addition, the suitable and precise parameters were selected on the base of outstanding outcome i.e., accuracy. Table 2 shows, the selected hyper-parameters for the prediction of RNA modification sites. Furthermore, ReLU is a nonlinear activation function that is applied in all convolution layers. However, the dropout regularization is applied to avoid the network from overfitting. Dropout(p) is a dropout operator with a probability of p and it is added to the dense layer which occasionally dropouts intermediate values from the previous layer by randomly assigning them to zero during training. The Concatenate function is used to combine a list of inputs. The Dense(x) layer is a fully-connected layer with x node and it is followed by a sigmoid activation function.

The developed CNN model can be mathematically defined as:

$$Conv(D)_{jf} = \mathrm{Re}LU\left(\sum_{s=0}^{S-1}\sum_{n=0}^{N-1} W_{sn}^{f} D_{j+s,n}\right) \tag{4}$$

$$f = w_{d+1} + \sum_{i=1}^{d} w_i x_i \tag{5}$$

$$f = w_{d+1} + \sum_{i=1}^{d} m_i w_i x_i \tag{6}$$

$$\mathrm{Re}LU(x) = \max(0, x) \tag{7}$$

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

Finally, the sigmoid function takes the prediction of whether an input DNA sequence is an N^6-methyladenosine site or not. Here for optimization and learning, Adam optimizer is applied where the learning rate is 0.0003, the batch size is 16, and the number of Epochs is 100. Execution of the developed model was carried out in Keras framework [43].

3. Performance evaluation

3.1. k-fold cross-validation (CV)

The literature demonstrated that computational models are widely evaluated by using various CV techniques like jackknife test, *k*-fold, and independent [30–33,37,44–50]. Researchers mostly utilized the concept of K-fold CV in their work due to its significance compared to the jackknife test. It required less computational time and generates unique results like the jackknife test. In this test, randomly divided the benchmark dataset into k subsets where one subset is used for testing and the rest k-1 subsets are provided for training data. The whole process is then reiterated k-times and finally, average scores of the k-folds are generated as an outcome of the computational model.

Here, the performance of the developed model is assessed by using *the 5*-fold CV technique by utilizing the prominent characteristics of it.

3.2. Evaluation metrics

In machine learning, the predictive outcomes of the hypothesis are widely measured by using various statistical measurement metrics including Matthews's correction coefficient (MCC), Sensitivity (Sn), Accuracy (Acc), and Specificity (Sp) [51–56]. These metrics are mathematically expressed as:

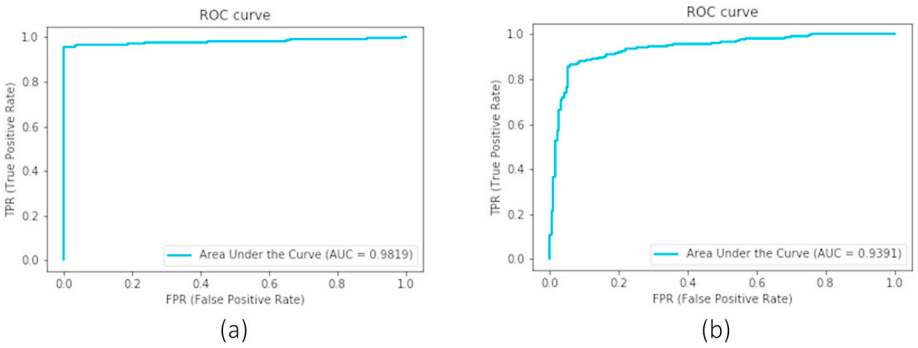

Fig. 2. The auROC curves of the developed computational model on S_1 and S_2 datasets.

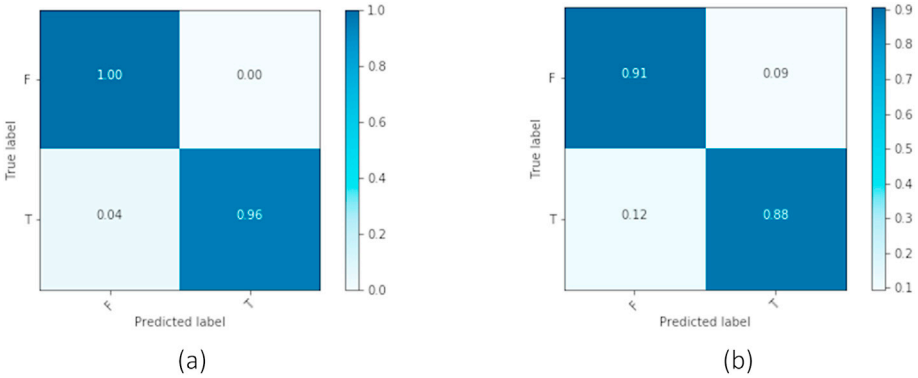

Fig. 3. The confusion matrix of the developed computational model on S_1 and S_2 datasets.

Table 4
Our proposed model comparison with existing models.

Datasets	Models	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC	auROC
S_1	Our proposed model	98.05	100.00	95.88	0.961	0.982
	iDNA6mA-PseKNC	96.73	100.00	93.28	0.930	Nil
	iDNA-MS	96.79	96.68	96.85	0.936	0.984
S_2	Our proposed model	89.22	90.52	88.03	0.784	0.939
	iDNA6mA (5-step rule)	86.64	86.59	86.70	0.73	0.931
	6 mA-Pred	83.13	83.30	82.95	0.66	0.886

$$\begin{cases} Acc = \dfrac{TN + TP}{TP + TN + FN + FP} \times 100\% \\ Sn = \dfrac{TP}{FN + TP} \times 100\% \\ Sp = \dfrac{TN}{FP + TN} \times 100\% \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \end{cases} \quad (9)$$

Here TN (true negative), TP (true positive), FN (false negative), and FP (false positive). However, Sn and Sp individually denote the capability of the predictor in the sense of true prediction of N^6-methyladenosine, and non- N^6-methyladenosine sites. (Acc) computes the overall performance of the model and considering a good measure in case of a balanced dataset. However, in case of an imbalanced dataset, mostly hypothesis bias toward the majority class, in that case, the minor class is totally ignored, consequently, the outcome of the hypothesis is unreliable. In such a scenario, MCC is the best measure to evaluate the performance of model in a real sense. Here the ratios of the negative and positive sequences are the same but MCC shows the success rates of the predictor on imbalance dataset.

4. Results and discussion

Table 3 presents the experimental outcomes of our proposed model obtained on S_1 and S_2 benchmark datasets, while Fig. 2 displays the Receiver Operating Characteristics (ROC) curves, and Fig. 3 shows the confusion matrix of these two benchmark datasets. On the S_1 dataset, the proposed model achieved 98.05%, 95.88%, 100.00%, and 0.961 of Acc, sn, sp, and MCC, respectively, which is quite satisfactory. Similarly, on the S_2 dataset, the model yielded acc of 89.22% while other performance measuring parameters are 88.03% of sn, 90.52% of sp, and 0.784 of MCC. Furthermore, the ROC on S_1 and S_2 datasets are 0.982 and 0.939, respectively.

In Table 4, we can see that the prediction result of the developed computational approach outperforms than that of the state-of-the-art approaches on both S_1 and S_2 datasets. Comparison with other methods, the prediction outcome presents that our proposed computational method improved the performance of the S_1 benchmark dataset by 1.32% of accuracy, 0.031 of MCC, and 2.6% of sensitivity. Similarly, our proposed computational method improved the performance of the S_2 benchmark dataset by 3.93% of specificity, 2.58% of accuracy, 0.054 of MCC, and 1.33% of sensitivity. The success rate display that our proposed

computational model achieved remarkable outcomes than existing methods. The importance of our proposed model is ascribed with the utilization of two distinct feature representation methods one-hot encoding and TNC. One-hot encoding expresses genomic sequences by using CNN which reflects the clear pattern of the target class. The second method is discrete method, which computes the sequence order information. These methods compensate the weakness of each other as a result, CNN obtained quite promising results.

As reported by Ref. [57], publically-accessible web-server for the developed approach leads the future direction for various methods and predictors for bioinformatics and computational biology [12,28,58–62]. In fact, in the future, it has incredibly enhanced the power of bioinformatics and computational biology to push medical science into an unprecedented revolution; in this paper, we are going to endeavor to introduce a web-server for the proposed method.

5. Conclusions

Here, a novel intelligent deep learning-based approach is introduced for discrimination of N^6-methyladenosine sites. In this method, one-hot representation is applied for encoding the input sequence to allow successive convolution layers. To declare the hidden information from the recognized sequences, the CNNs model is used in order to automatically learn the abstract features. Then, we apply the tri-nucleotide composition (TNC) feature extraction technique and concatenate with CNN features. Furthermore, the predictive outcomes of the developed computational approach are better and outperform compared to the state-of-the-art approaches and give promising outcomes in terms of all evaluation metrics as well as show major performance enhancement on all two various benchmark datasets. In conclusion, the superiority of the proposed computational model determines that it has a great practical and supportive tool for drug discovery and scientific studies on N^6-methyladenosine sites.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Brain Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. NRF-2017M3C7A1044816).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2020.104151.

References

[1] Z.K. O'Brown, E.L. Greer, N6-methyladenine: a Conserved and Dynamic DNA Mark, DNA Methyltransferases-Role and Function, Springer, 2016, pp. 213–246.
[2] L.M. Iyer, S. Abhiman, L. Aravind, Natural History of Eukaryotic DNA Methylation Systems, Progress in Molecular Biology and Translational Science, Elsevier, 2011, pp. 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.
[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, DNA methylation on N 6-adenine in mammalian embryonic stem cells, Nature 532 (2016) 329.
[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] S.E. Luria, M.L. Human, A nonhereditary, host-induced variation of bacterial viruses, J. Bacteriol. 64 (1952) 557.
[8] M. Meselson, R. Yuan, DNA restriction enzyme from E. coli, Nature 217 (1968) 1110.
[9] 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. Unit. States Am. 59 (1968) 1300–1306.
[10] 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.
[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-protocol (2016) 6.
[12] H. Yang, H. Lv, H. Ding, W. Chen, H. Lin, iRNA-2OM: a sequence-based predictor for identifying 2′-O-methylation sites in homo sapiens, J. Comput. Biol. 25 (2018) 1266–1277.
[13] K.R. Pomraning, K.M. Smith, M. Freitag, Genome-wide high throughput analysis of DNA methylation in eukaryotes, Methods 47 (2009) 142–150.
[14] A.M. Krais, M.G. Cornelius, H.H. Schmeiser, Genomic N6-methyladenine determination by MEKC with LIF, Electrophoresis 31 (2010) 3548–3551.
[15] E.L. Greer, M.A. Blanco, L. Gu, E. Sendinc, J. Liu, D. Aristizábal-Corrales, C.-H. Hsu, L. Aravind, C. He, Y. Shi, DNA methylation on N 6-adenine in C. elegans, Cell 161 (2015) 868–878.
[16] 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.
[17] W. Chen, P. Feng, H. Ding, H. Lin, K.-C. Chou, iRNA-Methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33.
[18] W. Chen, H. Tran, Z. Liang, H. Lin, L. Zhang, Identification and analysis of the N 6-methyladenosine in the Saccharomyces cerevisiae transcriptome, Sci. Rep. 5 (2015) 13859.
[19] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, K.-C. Chou, 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.
[20] Y. Zhou, P. Zeng, Y.-H. Li, Z. Zhang, Q. Cui, SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features, Nucleic Acids Res. 44 (2016) e91-e91.
[21] S. Xiang, Z. Yan, K. Liu, Y. Zhang, Z. Sun, AthMethPre: a web server for the prediction and query of mRNA m 6 A sites in Arabidopsis thaliana, Mol. Biosyst. 12 (2016) 3333–3337.
[22] H. Lv, F.-Y. Dao, D. Zhang, Z.-X. Guan, H. Yang, W. Su, M.-L. Liu, H. Ding, W. Chen, H. Lin, iDNA-MS: an Integrated Computational Tool for Detecting DNA Modification Sites in Multiple Genomes, Iscience, 2020, p. 100991.
[23] Q. Tang, J. Kang, J. Yuan, H. Tang, X. Li, H. Lin, J. Huang, W. Chen, DNA4mC-LIP: a linear integration method to identify N4-methylcytosine site in multiple species, Bioinformatics 36 (2020) 3327–3335.
[24] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, K.-C. Chou, iDNA6mA-PseKNC: Identifying DNA N6-Methyladenosine Sites by Incorporating Nucleotide Physicochemical Properties into PseKNC, Genomics, 2018.
[25] W. Chen, H. Lv, F. Nie, H. Lin, i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome, Bioinformatics 35 (2019) 2796–2800.
[26] M. Tahir, H. Tayara, K.T. Chong, iDNA6mA (5-step rule): identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule, Chemometr. Intell. Lab. Syst. 189 (2019) 96–101.
[27] W. Chen, H. Ding, X. Zhou, H. Lin, K.-C. Chou, iRNA (m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition, Anal. Biochem. 561 (2018) 59–65.
[28] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, K.-C. Chou, iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, Oncotarget 8 (2017) 4208.
[29] W. Chen, P.-M. Feng, H. Lin, K.-C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68-e68.
[30] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, K.-C. Chou, PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.
[31] W. Chen, H. Lin, K.-C. Chou, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. Biosyst. 11 (2015) 2620–2634.
[32] 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.
[33] 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.
[34] H. Lin, Q.-Z. Li, Eukaryotic and prokaryotic promoter prediction using hybrid approach, Theor. Biosci. 130 (2011) 91–100.
[35] H. Lin, Z.-Y. Liang, H. Tang, W. Chen, Identifying sigma70 promoters with novel pseudo nucleotide composition, IEEE ACM Trans. Comput. Biol. Bioinf (2017).
[36] M. Tahir, M. Hayat, iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC, Mol. Biosyst. 12 (2016) 2587–2593.
[37] 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 Progr. Biomed. 146 (2017) 69–75.
[38] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, Deep neural networks for acoustic

modeling in speech recognition: the shared views of four research groups, IEEE Signal Process. Mag. 29 (2012) 82–97.
[39] W. Qu, D. Wang, S. Feng, Y. Zhang, G. Yu, A novel cross-modal hashing algorithm based on multimodal deep learning, Sci. China Inf. Sci. 60 (2017), 092104.
[40] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.
[41] M. Kabir, M. Iqbal, S. Ahmad, M. Hayat, iTIS-PseKNC, Identification of Translation Initiation Site in human genes using pseudo k-tuple nucleotides composition, Comput. Biol. Med. 66 (2015) 252–257.
[42] S.-H. Guo, E.-Z. Deng, L.-Q. Xu, H. Ding, H. Lin, W. Chen, K.-C. Chou, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.
[43] F. Chollet, Keras: deep learning library for theano and tensorflow, URL: https://keras.io/k, 2015, 7.
[44] B. Liu, H. Wu, D. Zhang, X. Wang, K.-C. Chou, Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods, Oncotarget 8 (2017) 13338.
[45] A. Awazu, Prediction of nucleosome positioning by the incorporation of frequencies and distributions of three different nucleotide segment lengths into a general pseudo k-tuple nucleotide composition, Bioinformatics 33 (2016) 42–48.
[46] Z.-C. Xu, P. Wang, W.-R. Qiu, X. Xiao, iSS-PC: identifying splicing sites via physical-chemical properties using deep sparse auto-encoder, Sci. Rep. 7 (2017) 8222.
[47] M. Tahir, M. Hayat, S.A. Khan, A two-layer computational model for discrimination of enhancer and their types using hybrid features pace of pseudo K-tuple nucleotide composition, Arabian J. Sci. Eng. (2017) 1–9.
[48] M. Tahir, M. Hayat, Machine learning based identification of protein–protein interactions using derived features of physiochemical properties and evolutionary profiles, Artif. Intell. Med. 78 (2017) 61–71.
[49] M. Tahir, M. Hayat, K.T. Chong, Prediction of N6-methyladenosine sites using convolution neural network model based on distributed feature representations, Neural Network. (2020).
[50] M. Tahir, M. Hayat, S. Gul, K.T. Chong, An intelligent computational model for prediction of promoters and their strength via natural language processing, Chemometr. Intell. Lab. Syst. (2020), 104034.
[51] F.-Y. Dao, H. Lv, F. Wang, C.-Q. Feng, H. Ding, W. Chen, H. Lin, Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique, Bioinformatics 35 (2019) 2075–2083.
[52] F. Li, C. Li, T.T. Marquez-Lago, A. Leier, T. Akutsu, A.W. Purcell, A. Ian Smith, T. Lithgow, R.J. Daly, J. Song, Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome, Bioinformatics 34 (2018) 4223–4231.
[53] J. Song, F. Li, A. Leier, T.T. Marquez-Lago, T. Akutsu, G. Haffari, K.-C. Chou, G.I. Webb, R.N. Pike, PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy, Bioinformatics 34 (2018) 684–687.
[54] J. Song, Y. Wang, F. Li, T. Akutsu, N.D. Rawlings, G.I. Webb, K.-C. Chou, iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Briefings Bioinf. 20 (2019) 638–658.
[55] M. Tahir, M. Hayat, S.A. Khan, A two-layer computational model for discrimination of enhancer and their types using hybrid features pace of pseudo k-tuple nucleotide composition, Arabian J. Sci. Eng. 43 (2018) 6719–6727.
[56] H. Tayara, M. Tahir, K.T. Chong, iSS-CNN: identifying splicing sites using convolution neural network, Chemometr. Intell. Lab. Syst. 188 (2019) 63–69.
[57] K.-C. Chou, H.-B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2009) 63.
[58] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, K.-C. Chou, iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, Genomics 111 (2019) 96–102.
[59] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, K.-C. Chou, 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.
[60] H. Yang, W.-R. Qiu, G. Liu, F.-B. Guo, W. Chen, K.-C. Chou, H. Lin, iRSpot-Pse6NC, Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC, Int. J. Biol. Sci. 14 (2018) 883.
[61] F. Li, Y. Wang, C. Li, T.T. Marquez-Lago, A. Leier, N.D. Rawlings, G. Haffari, J. Revote, T. Akutsu, K.-C. Chou, Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods, Briefings Bioinf. 10 (2018).
[62] Z.-D. Su, Y. Huang, Z.-Y. Zhang, Y.-W. Zhao, D. Wang, W. Chen, K.-C. Chou, H. Lin, iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC, Bioinformatics 34 (2018) 4196–4204.